TECHNICAL RESEARCH REPORT

Using Categorical Information in Multidimensional Data Sets: Interactive Partition and Cluster Comparison

by Jinwook Seo and Ben Shneiderman

TR 2005-102

Using Categorical Information in Multidimensional Data Sets: Interactive Partition and Cluster Comparison

Jinwook Seo[12*] and Ben Shneiderman[123†]
Department of Computer Science[1],
Human-Computer Interaction Laboratory, Institute for Advanced Computer Studies[2], and
Institute for Systems Research[3]
University of Maryland, College Park, MD 20742 USA

ABSTRACT

Multidimensional data sets often include categorical information. When most columns have categorical information, clustering the data set by similarity of categorical values can reveal interesting patterns in the data set. However, when the data set includes only a small number (one or two) of categorical columns, the categorical information is probably more useful as a way to partition the data set. For example, researchers might be interested in gene expression data for healthy vs. diseased patients or stock performance for common, preferred, or convertible shares. For these cases, we present a novel way to utilize the categorical information together with clustering algorithms. Instead of incorporating categorical information into the clustering process, we can partition the data set according to categorical information. Clustering is then performed with each subset to generate two or more clustering results, each of which is homogeneous (i.e. only includes the same categorical value for the categorical column). By comparing the partitioned clustering results, users can get meaningful insights into the data set: users can identify an interesting group of items that are differentially/similarly expressed in two different homogeneous partitions. The partition can be done in two different directions: (1) by rows if categorical information is available for each column (e.g. some columns are from disease samples and other columns are from healthy samples) or (2) by a column if a column contains categorical information (e.g. a column represents a categorical attribute such as colors or sex). We designed and implemented an interface to facilitate this interactive partition-based clustering results comparison. Coordination between clustering results displays and comparison results overview enables users to identify interesting clusters, and a simple grid display clearly reveals correspondence between two clusters.

CR Categories: I.6.9.c Information visualization, H.5.2 User Interfaces, H.5.2.f Graphical user interfaces, I.5.2.b Feature evaluation and selection, H.2.8.d Data mining, H.1.2.a Human factors, H.2.8.c Data and knowledge visualization, H.2.8.h Interactive data exploration and discovery

Keywords: information visualization, exploratory data analysis, dynamic query, feature detection/selection, statistical graphics.

* jinwook@cs.umd.edu
† ben@cs.umd.edu

1 INTRODUCTION

Multidimensional and multivariate data sets can productively analyzed by cluster analysis to find related groups of items. In our work in microarray gene expression analysis we developed a rich environment for exploration and discovery [9, 10]. The data values were real valued and could be normalized to create multidimensional data sets. However, many of the biologists we worked with had data sets that included categorical information such as labels for healthy vs. diseased samples for which the goal was to compare gene expression levels to determine which genes might have higher or lower expression levels in the diseased samples. Other researchers were comparing male and female patients, and we found similar requests from stock market analysts, meteorologists, and others.

To accommodate these requests, we developed new features for the Hierarchical Clustering Explorer that enabled users to specify the partition of samples and then conduct comparisons among items. The partition was based on a value in the data set and then hierarchical clustering was applied to each partition. Users define clusters in each partition by moving a dynamic query slider called the minimum similarity bar (Figure 1). A typical user would create 10-15 clusters in each partition and then look for similarities and differences in items; a very tedious process. To accelerate this work, we took inspiration from the rank-by-feature framework that ranks 1D and 2D projections according to some criteria such as correlation coefficient, entropy, or outlierness. For cluster comparison, the goal was to rank all clusters in one partition with clusters in the second partition by a similarity measure. If the clustering result from the first partition, *CR1*, has *n* clusters, and the clustering result from the second partition, *CR2*, has *m* clusters, then the matrix would have m x n cells, that could be color coded to show similarity of clusters. This color coded matrix, which enables users to focus their investigation, is the heart of this contribution. In addition we provide coordination between the clustering to show where items from one cluster wind up in the second partition and a detailed scatterplot to rapidly identify similar items.

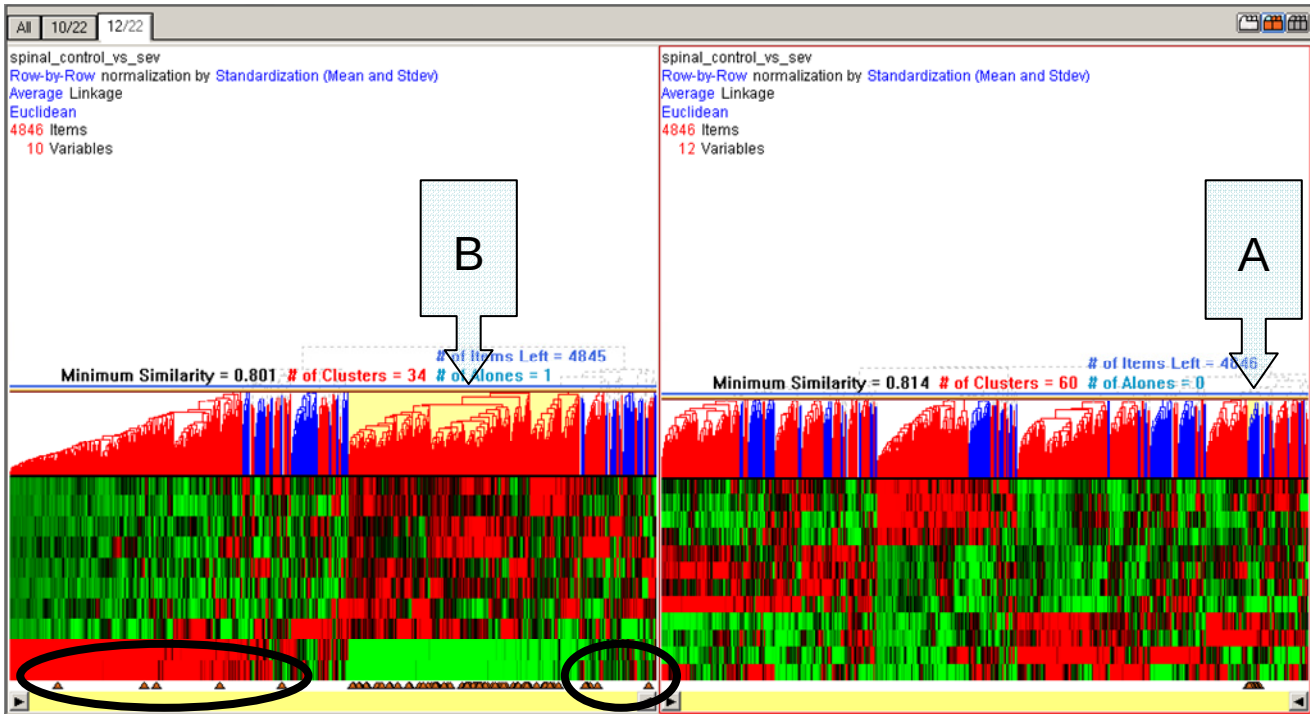

Figure 1: User creates two clustering windows, one for each partition of the data. Interaction for cluster comparisons. A click on a cluster on a dendrogram highlights items in the cluster on both dendrograms with orange triangles. In this figure, the user has clicked on a cluster on the right side of the rightmost dendrogram (A). The triangles on the left side show where the related items appear in the other clustering result. They are mostly within cluster (B) but five appear to the far left and four are to the right (black circles).

2 RELATED WORK

Most visualization tools use categorical variables to label displays distinctively to show the categorical information by using different sizes, colors, or shapes for different categorical values. Friendly [4] suggested several visualization techniques and graphical displays for categorical data, which include Fourfold display, Mosaic displays (similar to Treemap), and Association plots.

There has been much less work on clustering categorical data while there have been huge number of clustering algorithms for numerical data. Similarity between two items having categorical attributes should be calculated differently from that between items having only numerical attributes. Typically, co-occurrence measures and link analysis can be used to build a graph structure from a categorical data set, and then a graph partitioning algorithm or a traditional clustering algorithm generates clusters. [5, 6].

There is some relation to work that seeks to compare hierarchical structures, such as the Tree Juxtaposer [8] that highlights differing items and subtrees between two versions of a tree. The goal of showing relationships between two different hierarchies, such as a geographical hierarchy and a jobs hierarchy was supported by coordinated views in PairTrees [7]. Users could select a node in the geographic hierarchy, such as a state in the U.S., and that would produce highlights in the jobs hierarchy to identify which jobs were held by residents of the selected state. Similarly, if a job node were selected, that would produce a highlight in the geographic hierarchy to identify where those jobs were most frequent.

Adjacency matrix representations such as the Matrix Browser [14] show relationships between items, typically link relationships between nodes in a graph. These adjacency matrices are of order n x n for an n node graph. Adjacency matrices for bi-partite graphs with n nodes in one partition and m nodes in the second partition are close to what we are using in this work. Selections from two hierarchies were also shown in Matrix Zoom [1] which has similarities to our work. However, our emphasis is to enable users to compare clustering results to identify items that are noticeably different in performance across partitions.

Another source of related work is on reorderable or permutation matrices [12], which are often referred to as heatmaps in commercial systems such as Spotfire [13]. Our use of heatmaps is tied to the clusters in two dendrograms, and the similarity index we use represents features that are of interest to users seeking to identify items that are noticeably different in performance across partitions.

A further distinction in our work is the two levels of analysis. We start by trying to match clusters, and then drill down to identify the items that account for the similarity. The capacity to see the clusters and select individual items rapidly enables exploration of datasets with thousands of items. Also the capacity to see where items from a cluster in one partition fall in the other partition reveals differences across partitions. For example, by selecting a cluster with high gene expression levels for healthy patients, users can determine if some of these genes have lower expression levels in diseased patients.

3 PARTITIONING USING CATEGORICAL INFORMATION

Multidimensional data sets are represented in a table, where each row is a data item and each column is a dimension. be a multidimensional data set with n rows and m columns. Categorical information exists either in a row or in a column. First, a special row can have categorical information for each column. For example, a microarray data set usually includes more than two different phenotypes of samples (e.g. types of cancers and patient categories), and each sample is represented as a column and each gene is represented as a row in a data set. The phenotype information of each sample can be a category to partition columns as shown in Figure 2.

microarrays	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
sample type	A	A	A	A	A	B	B	B
$gene_1$								
$gene_2$								
……								
$gene_n$								

Figure 2: Partitioning of columns by the sample types A or B

The partitions of the original data sets into two or more smaller data sets each of which has all rows but has only a part of the columns with the same phenotype. Then each partition can be fed into a clustering algorithm to generate separate clustering results of rows. By comparing those clustering results, biologist might find an interesting group of genes that are similarly or differentially expressed in different groups of homogeneous samples.

Second, a column can have categorical information for data items. For example, a survey data set usually includes some categorical columns such as sex and education level. Each row represents a participant of the survey and each column represents information for participants. A categorical column such as sex shown in Figure 3 can be a category to partition the data set.

Survey	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
p_1								male
p_2								male
p_3								male
…..								
p_{n-1}								female
p_{n-2}								female

Figure 3: Partitioning of rows by attribute a_8

This kind of partition separates the original data set into smaller data sets each of which has all columns but has only rows that has the same categorical value for the categorical column. Then each partition is fed into a clustering algorithm to generate a clustering result of columns. Comparison of the clustering results can lead users to findings on the difference or similarity between male and female in terms of attributes relationships.

4 COMPARISON OF CLUSTERING RESULTS

Suppose two clustering results ($CR1 = \{CR1_i | i=0..n\}$ and $CR2 = \{CR2_j | j=1..m\}$) have been produced with two separate subsets of columns. Cluster comparisons depend on a definition of similarity. Correlation between average patterns of two clusters can be another possible similarity measure for two clusters. Set similarity measures can also be used to measure similarities between clusters. Most simple measure is $Sim1(CR1_i, CR2_j) = \frac{|CR1_i \cap CR2_j|}{|CR1_i \cup CR2_j|}$. This measure has also been used for tree node comparisons [8] and documents comparisons [2, 11]. While this measure is simple and (1-*Sim1*) is a metric distance measure, it penalizes pairs of a large and a small cluster. To compensate for the penalties, a heuristic similarity measure can be used:

$$Sim2(CR1_i, CR2_j) = \frac{\frac{|CR1_i \cap CR2_j|}{|CR1_i|} + \frac{|CR1_i \cap CR2_j|}{|CR2_j|}}{2} = |CR1_i \cap CR2_j| \cdot \frac{|CR1_i| + |CR2_j|}{2 \cdot |CR1_i| \cdot |CR2_j|}$$

This measure can be thought of as an arithmetic mean of the precision and recall values from information retrieval concepts:

$$\text{Precision} \quad P(i, j) = \frac{|CR1_i \cap CR2_j|}{|CR2_j|}$$

$$\text{Recall} \quad R(i, j) = \frac{|CR1_i \cap CR2_j|}{|CR1_i|}$$

The F-measure that is a harmonic mean of the precision (*P*) and recall (*R*) values can also be a possible similarity measure:

$$Sim3(CR1_i, CR2_j) = \frac{2\frac{|CR1_i \cap CR2_j|}{|CR1_i|}\frac{|CR1_i \cap CR2_j|}{|CR2_j|}}{\frac{|CR1_i \cap CR2_j|}{|CR1_i|} + \frac{|CR1_i \cap CR2_j|}{|CR2_j|}} = \cdot \frac{2|CR1_i \cap CR2_j|}{|CR1_i| + |CR2_j|}$$

In *Sim2* and *Sim3*, weighting by the size of each set keeps a small set from dominating the similarity value.

Graphical displays can provide useful overviews of the comparison results. We choose a grid display as shown in Figure 4 to show the overview. Each row represents a cluster from a clustering result, and each column represents a cluster from another clustering result, thus each grid cell represents a pair of clusters. Each cell is color-coded by a cluster similarity measure like an equation (*Sim1*, *Sim2*, or *Sim3*). Thus, similar cluster pairs can be preattentively identified in this display.

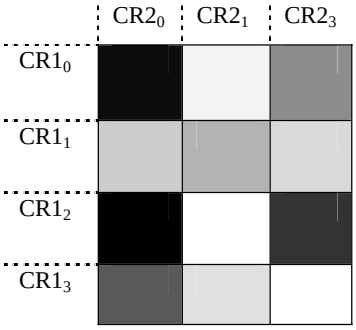

Figure 4: Overview of similarities between two clustering results where dark cells indicate high similarity

To show the similarity between a selected pair of clusters, we choose a revised scatterplot view as shown Figure 5. Each vertical or horizontal line represents an item in two clusters respectively. An intersection point has a blue square if the vertical item and the horizontal item are the same. The fraction of vertical or horizontal lines with a blue dot visualizes the similarity between two clusters. Linear alignment of blue dots on the scatterplot view tells us how similar the orders of items are in the two selected clusters. If blue dots are aligned along the diagonal line, the order of items in clusters is also similar to each other.

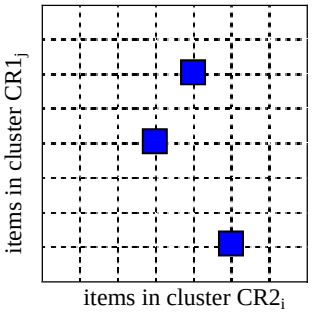

Figure 5: Scatterplot showing similarity between two clusters, $CR1_j$ with 9 items and $CR2_i$ with 8 items, where there are three items in common

5 EXAMPLES

We first explain our method with a simple data set ("Sleep in Mammals"). This data set has 62 mammals as rows and 7 variables (body weight, brain weight, nondreaming sleep hours, dreaming sleep hours, total sleep hours, maximum life span, gestation time, and overall danger index) as columns. Overall danger index is a categorical variable which has two categorical values, "high" and "low". Users can partition rows by this categorical variable and generate two partitions. By clustering each partition, users can generate two clustering results of columns.

In each dendrogram view (Figure 6), users generate two clusters using the minimum similarity bar. The overview of similarity measures (at the bottom left corner) for all four possible pairs of clusters. Two data black cells indicate that there are two perfectly matching pairs of clusters, but the revised scatterplot view (at the bottom right corner) shows that the arrangement of items in the two clusters are not the same. It turns out in the left dendrogram that the nondreaming sleep hours dominate total sleeping hours for the mammals with high overall danger index. This might mean that those mammals are too cautious to have a long dreaming sleep.

We applied our suggested graphical technique to a much larger biological data set on spinal cord injuries. A group of biologists studied molecular mechanisms of spinal cord degeneration and

Figure 6 An example of clustering results comparison with a small mammals sleep data set where there are 63 mammals and two categories by overall danger index. The overview of similarity measure (at the bottom left) shows two pairs of matching clusters by the two dark cells. The selected cluster pair turns out in the scatterplot view (at the bottom right) that they don't have the same structure since the blue dots are not aligned along the diagonal line.

repair [3]. They analyzed spinal cord above thoracic vertebrae T9 at various time points up to 28 days post injury. Mild, moderate and severe injury was examined. They were interested in finding group of genes that were similarly or differently expressed in two different groups of heterogeneous samples. The original data set has a special row containing a category of spinal cord samples: severity of injuries. We partitioned the original data set according to the categorical information to have two partitions; 10 control samples and 12 severe injury samples. Each partition was fed into a hierarchical clustering algorithm to generate two dendrograms in two separate tab windows in the hierarchical clustering explorer (HCE) [9]. Since the two dendrogram views are coordinated with each other and other views, users can click on a cluster in a dendrogram view and then the items in the cluster are highlighted with orange triangles in all other views including the other dendrogram view (Figure 1). Just by looking at where the orange triangles appear in the other dendrogram view, users can notice how items in a cluster are grouped in the other clustering result.

Cluster similarity measures and graphical displays facilitate this task by providing an overview of similarity measures for all possible pairs of clusters in the two clustering results. When users select the "Cluster Similarity" tool, a modeless dialog box pops up and users can drag and drop the target-shaped icon on dendrogram view tabs to choose two dendrograms to compare. The graphical overview of the comparison of two clustering results is shown at the bottom right corner of Figure 7. Each cell of the overview represents a pair of clusters. A mouseover event on the overview highlights the corresponding clusters in the selected dendrograms. The revised scatterplot view at the bottom right corner shows the overview of mapping of items between two clustering results.

6 CONCLUSION

Stimulated by user requests for capacity to compare partitions of data sets, we implemented an extension to the Hierarchical Clustering Explorer. Users studying different populations, such as healthy vs. diseased patients needed to identify the 5-50 genes with differing expression levels out of collections of 12,000 to 36,000 genes. We enabled users to partition the data and create clusters within each partition. Then they could look for similar or differing clusters, and drill down to find the specific genes that account for differences. The overview of cluster similarity provided by a heatmap display combined with rapid coordination among windows provides support for this challenging task. The current implementation can handle approximately 100 clusters each containing approximately 100 items. This is already very useful but scaling up is a natural next step. These concepts are difficult for some users to grasp, so effective training methods

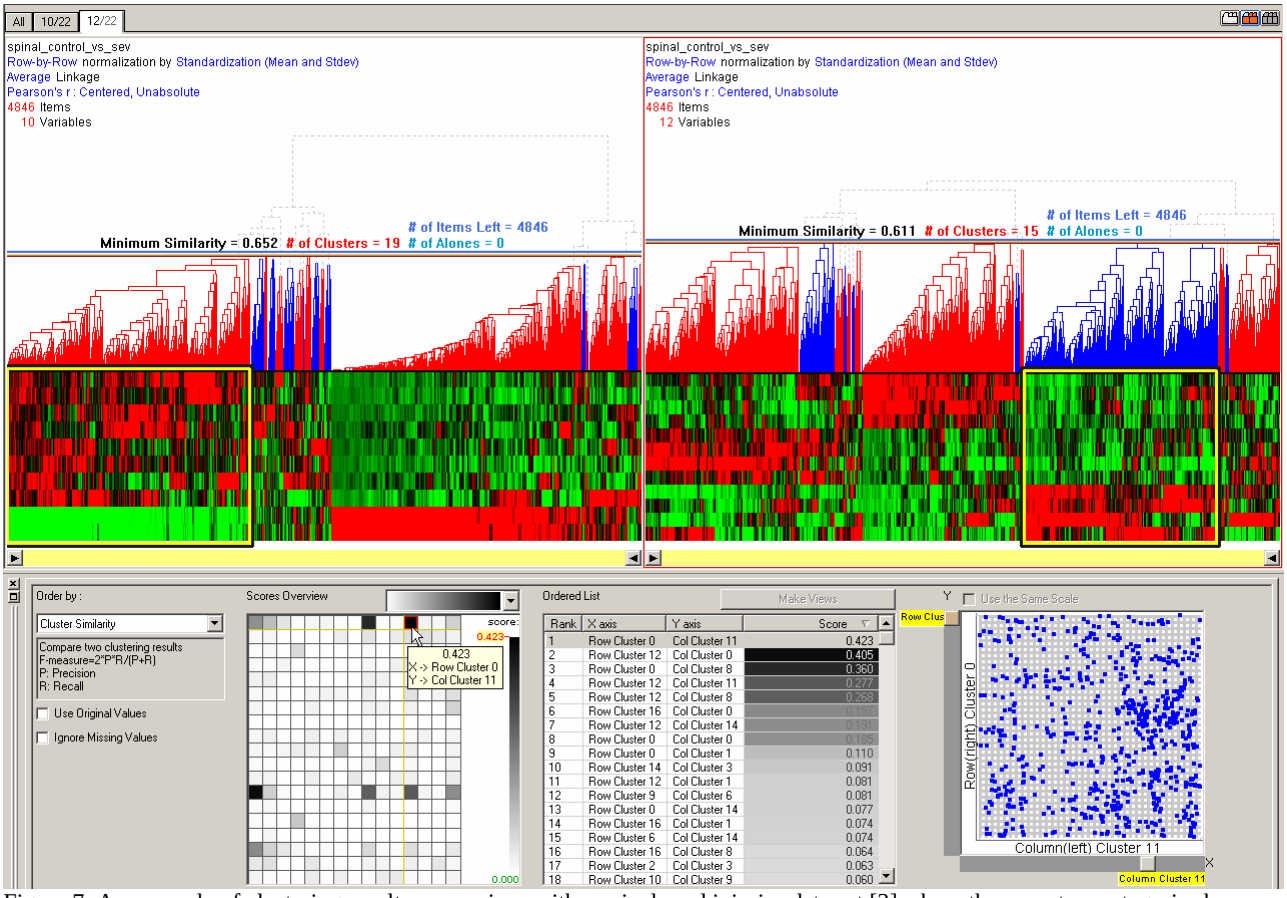

Figure 7: An example of clustering results comparison with a spinal cord injuries data set [3] where there are two categories by severity of injuries. The left dendrogram shows a clustering result with control samples, and the right dendrogram shows one with severe injuries samples. When users select a pair of clusters on the overview (the dark black square representing row cluster 0 on the left and column cluster 11 on the right), the selected clusters are highlighted with yellow rectangles in the dendrogram views, and the similarity between them is visualized in the scatterplot view on the lower right. *Sim3* was used as the similarity measure.

and understandable case studies would be helpful for new users.

Acknowledgement: This work was supported by N01 NS-1-2339 from the NIH.

REFERENCES

[1] J. Abello and F. van Ham, "Matrix Zoom: A Visual Interface to Semi-External Graphs," in *Proceedings of IEEE Symposium on Information Visualization*. Austin, TX, 2004, pp. 183-190.

[2] A. Z. Broder, "On the resemblance and containment of documents," in *Proceedings of the Compression and Complexity of Sequences*. Washington, DC: IEEE Computer Society, 1998, pp. 21-29.

[3] S. Di Giovanni, A. I. Faden, A. Yakovlev, J. S. Duke-Cohan, T. Finn, M. Thouin, S. Knoblach, A. De Biase, B. S. Bregman, and E. P. Hoffman, "Neuronal plasticity after spinal cord injury: identification of a gene cluster driving neurite outgrowth," *The FASEB Journal*, vol. 19, pp. 153-154, 2005.

[4] M. Friendly, *Visualizing Categorical Data*.: SAS Publishing, 2000.

[5] D. Gibson, J. M. Kleinberg, and P. Raghavan, "Clustering Categorical Data: An Approach Based on Dynamical Systems," *VLDB Journal*, vol. 8, pp. 222-236, 2000.

[6] S. Guha, R. Rastogi, and K. Shim, "ROCK: a robust clustering algorithm for categorical attributes," in *Proceedings of the 15th International Conference on Data Engineering*. Sydney, NSW, 1999, pp. 512-521.

[7] B. Kules, B. Shneiderman, and C. Plaisant, "Data Exploration with Paired Hierarchical Visualizations: Initial Designs of PairTrees," in *Proceedings of the Digital Government Research Conference*, 2003, pp. 255-260.

[8] T. Munzner, F. Guimbretiere, S. Tasiran, L. Zhang, and Y. Zhou, "TreeJuxtaposer: scalable tree comparison using Focus+Context with guaranteed visibility," *ACM Transactions on Graphics*, vol. 22, pp. 453-462, 2003.

[9] J. Seo and B. Shneiderman, "Interactively exploring hierarchical clustering results," *IEEE Computer*, vol. 35, pp. 80 - 86, 2002.

[10] J. Seo and B. Shneiderman, "A Rank-by-Feature framework for unsupervised multidimensional data exploration using low dimensional projections," in *Proceedings of IEEE Symposium on Information Visualization*. Austin, Texas, United States, 2004, pp. 65-72.

[11] N. Shivakumar and H. Garc´ia-Molina, "SCAM: A copy detection mechanism for digital documents," in *Proceedings of the 2nd Annual Conference on Theory and Practice of Digital Libraries*. Austin, TX, 1995, pp. 11-13.

[12] H. Siirtola, "Interactive cluster analysis," in *Proceedings of the Eighth International Conference on Information Visualization*. London, England, 2004, pp. 471-476.

[13] Spotfire, Spotfire DecisionSite, http://www.spotfire.com/

[14] J. Ziegler, C. Kunz, and V. Botsch, "Matrix browser: visualizing and exploring large networked information spaces," in *Proceedings of CHI '02 Extended Abstracts on Human Factors in Computing Systems*. Minneapolis, Minnesota, USA: ACM Press, 2002, pp. 602-603.