Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Interactive Visual Summarization of Multidimensional Data

Sarat M. Kocherlakota
Collaborative Environments
Renaissance Computing Institute (RENCI)
Chapel Hill, NC
smkocher@renci.org

Christopher G. Healey
Dept. of Computer Science
North Carolina State University
Raleigh, NC
healey@csc.ncsu.edu

***Abstract*—Visualization has become integral to the knowledge discovery process across various domains. However, challenges remain in the effective use of visualization techniques, especially when displaying, exploring and analyzing large, multidimensional datasets, such as weather and meteorological data. Direct visualizations of such datasets tend to produce images that are cluttered with excess detail and are ineffective at communicating information at higher levels of abstraction.**

To address this problem we provide a visual summarization framework to intuitively reduce the data to its important and relevant characteristics. Summarization is performed in three broad steps. First, high-relevance data elements and clusters of similar data attributes are identified to reduce a dataset's size and dimensionality. Next, patterns, relationships and outliers are extracted from the reduced data. Finally, the extracted summary characteristics are visualized to the user. Such visualizations reduce excess visual detail and are more suited to the rapid comprehension of complex data. Users can interactively guide the summarization process gaining insight into both how and why the summary results are produced.

Our framework improves the benefits of mathematical analysis and interactive visualization by combining the strengths of the computer and the user to generate high-quality summaries. Initial results from applying our framework to large weather datasets have been positive, suggesting that our approach could be beneficial for a wide range of domains and applications.

***Index Terms*—classification, decision-support, interaction, outlier detection, rule mining, summarization, visualization**

I. Introduction

Visualization has become an integral component of the knowledge discovery process across a wide variety of domains. Visual representations of large data collections allow users to rapidly analyze, explore and assimilate large amounts of information contained within these datasets. Effective visualizations help increase comprehension of such datasets [4], [6], [12], [21], which in turn allows users to make better informed decisions.

Visualization supports "sense-making" of the underlying data by focusing user attention on important characteristics such as patterns, trends, dependencies, clusters and outliers, among other aspects. However, existing visualization techniques are being continually challenged by the problem of effectively and meaningfully displaying larger and larger datasets [11], [21], [22]. Traditional approaches, including those based on glyphs, pixels, parallel coordinates, scatter-plot matrices and so on, often address this problem in a brute-force fashion; by increasing the visual detail displayed, for example, the number of graphical objects used to represent individual data elements, or the number of visual features used to represent values for each data attribute. Unfortunately, this approach works best for only a few million data elements and a handful of attributes at a time [22]. Real-world datasets frequently contain millions of elements, with each element encoding tens or hundreds of attributes. Attempting to visualize each individual element often produces cluttered, overloaded images where all elements are displayed simultaneously, or images that selectively restrict their contents to very small data subsets. In both cases, it can be difficult, even impossible for viewers to perform their analysis over the datasets as a whole.

Visualizing multidimensional data presents a related set of problems. For example, techniques that apply visual features like color and texture to represent individual data attributes are effective for low-dimensional datasets, but as the dimensionality increases, these methods can quickly run out of available features. Clearly, it is becoming increasingly difficult to produce visualizations that remain accurate and complete, while simultaneously ensuring that a user's cognitive abilities are not exceeded.

Problems of dataset size and dimensionality suggest that intelligent data compression or summarization could be invaluable to the visual exploration, analysis and comprehension of large, multidimensional datasets. Data summaries that highlight important aspects and characteristics of the underlying data could help users comprehend the data at higher levels of abstraction. This data pre-processing could also help transform large, multidimensional datasets into intermediate representations that can be more easily visualized using existing techniques. This could be particularly useful in the discovery of important behavioral characteristics and phenomena embedded within environmental data.

Although isolated methods of data summarization have been previously studied, we know of no system that intelligently combines different techniques to form an interactive summarization framework: a sequence of operations that allow users to observe, guide and explore within the summarization as it runs. This is the goal we are working towards through the research described in this paper.

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

A. Visual Summarization

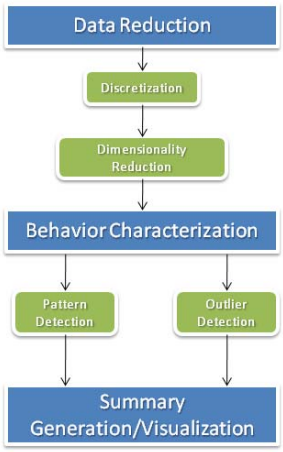

Fig. 1. Important phases of the summarization framework

Our visual summarization framework is designed to perform meaningful data processing of a multidimensional dataset. The framework reduces the underlying data to higher-level abstractions or "summaries" that contain important characteristics of the original data, and exclude extraneous detail. These characteristics include high-relevance elements and attributes, strong patterns and relationships, and outliers. The summaries are then visualized to the users.

In the past, mathematical analysis techniques such as association rule mining, clustering, multidimensional scaling, and outlier detection, among others, have been used to extract relevant data characteristics. These techniques are however more effective for analyzing low-dimensional data, as their complexity often makes them unsuited to larger, higher-dimensional datasets. These mathematical algorithms also operate in a "black box" fashion, where only final results produced are returned and intermediate analysis steps cannot be seen. These issues make such techniques unsuited for direct application for meaningful summarization.

A more structured approach chains together individual analysis algorithms into a series of supporting steps that systematically summarize a dataset. Including the ability to interact with the process and visualize intermediate execution steps helps to produce summaries that are relevant and meaningful to the user. This approach forms the basis of our visual summarization framework.

Consider a multidimensional dataset D containing m elements and n attributes $A = \{A_1, ..., A_n\}$, where each element $e_i \in D$ is a combination of n attribute values. An initial *data reduction* phase interactively discretizes attribute values into bins. The density or sparsity of each bin is used to select a subset of the data elements for further processing, reducing the size of the dataset. Inter-element similarities are then computed to cluster similar attributes, dividing the elements of interest into low-dimensional subsets. Next, a *behavior characterization* is performed on each attribute subset to extract patterns in the form of association rules. Outlier elements are also located during this phase. Both *data reduction* and *behavior characterization* include interactive visualizations to help users understand and guide the summarization process. Once elements of interest, attribute subsets, patterns and outliers have been extracted, they are combined into a summary form and then visualized to the user during a *summary generation and visualization* phase. Figure 1 shows the main steps of our summarization framework.

Throughout this paper, we will illustrate our approach using a weather dataset containing seven data attributes: *temperature*, *vapor pressure*, *cloud cover*, *frost days*, *precipitation*, *wet days* and *diurnal range*. This dataset was obtained from the Intergovernmental Panel on Climate Change (IPCC), and includes historical monthly averages for each attribute at $\frac{1}{2}^\circ$ latitude and longitude locations with positive elevation throughout the world for the period 1981-1990. In the analysis described in this paper, we focus on the month of April and for the area covering North America.

II. Related Work

A number of powerful mathematical techniques exist to identify potentially interesting properties of a dataset. Dimension ordering techniques are used to compute similar attribute subsets. For example, Value and Relation (VaR) [22] computes similarities between all pairs of attributes in a multidimensional dataset, then applies multidimensional scaling [17] to generate a 2D layout of the attributes that clusters similar attributes close to one another. Attribute subsets can be used as input to later algorithms, increasing their efficiency by reducing the size and dimensionality of the data they need to process.

Patterns and relationships defined as association rules can be constructed using decision tree-based techniques such as ID3 and C4.5 [20], or with counting methods such as Apriori [2] or frequent pattern growth [9]. An association rule takes the form $X \Rightarrow Y$ where X and Y denote one or more attribute-value pairs called itemsets. $X \Rightarrow Y$ implies that, if a data element contains the attribute values X, there is a strong likelihood it also contains the attribute values Y. Apriori techniques focus on counting the frequency of individual attribute values as well as combinations of values. These frequent itemsets are combined to identify strong association rules. Although techniques like inclusive and restrictive template matching [13] can be applied to speed up the rule mining process, association mining becomes more inefficient as the size and dimensionality of the underlying dataset increases.

Another useful dataset characteristic are clusters, collections of data elements where elements within a cluster are more similar to one another than to elements in different clusters. Clusters provide useful insight into data distributions, as well as relationships between the data elements. Techniques such as parallel coordinates, star coordinates [12], self-organizing maps [16] and multidimensional scaling [17] can visualize

multidimensional data in ways that attempt to highlight clusters and groups. Automatic cluster generation techniques include partitional and hierarchical approaches. Partitional clustering is generally more efficient, and includes the well-known k-means [19] algorithm. Although relatively simple and efficient, k-means is dependent on an initial assignment of cluster centers. Moreover, determining the number of clusters k that most accurately partitions the data can be time-consuming and may require a trial and error approach.

Outliers are data elements that differ from other elements by so much that they arouse suspicion of being generated by a separate mechanism or distribution function [10]. Outliers can be detected using distance-based [14] or densities of local neighborhoods [5] techniques. Distance-based algorithms choose a fixed distance d, then mark any element farther than d from its nearest neighbor as an outlier. Although these technique are efficient, they can fail to correctly locate outliers in a dataset with different densities of elements. The local neighborhoods density technique addresses this problem by calculating a local element density to define different distance thresholds d for outlier detection in different parts of the dataset. Both methods have difficulty locating outliers in high-dimensional space, however, since elements tend to be located far from every other element (i.e. a very sparse distribution), making it difficult to determine either locality or density [1]. Finding outliers often requires first constructing a low-dimension projection of relevant attributes, then locating outliers within this projection [1].

Techniques have also been proposed to perform summarization for visualization, for example, interactive data summarization (IDS) [18] and RuleViz [7], [8]. These systems combine automated analysis and interactive visualization for data pre-processing. IDS generates simple statistical measures using its combined approach. RuleViz focuses on characterizing datasets in more depth using association rules. However, RuleViz currently analyzes at most three attributes at a time, and does not offer any help to its users to select which three attributes to process. RuleViz is also restricted to display a single rule at a time. Finally, neither IDS nor RuleViz focus on finding other data characteristics such as high-relevance attribute subsets or outliers, and neither is designed to analyze and visualize high-dimensional data.

III. Data Reduction

The first step in our summarization framework applies data reduction operations, starting with discretization of attributes, then attribute partitioning. These operations are critical, since they compress both the size and the dimensionality of the data that later algorithms must process.

A. Attribute Value Discretization

First, each data attribute A_j is discretized into R_j equal-width ranges $r_{j,k}$ converting continuous attributes into discrete representations that are more efficient to manage. R_j is a user-defined value that can be unique for each A_j.

Next, the density of each value range $r_{j,k}$ is computed as the percentage of $e_i \in D$ whose attribute value $a_{i,j}$ for A_j falls in $r_{j,k}$'s interval. Densities are then used to categorize ranges as dense or sparse based on user-defined thresholds $\rho_{j,d}$ and $\rho_{j,s}$, respectively. A range $r_{j,k}$ is categorized as *dense* if $\rho(r_{j,k}) \geq \rho_{j,d}$ or *sparse* if $\rho(r_{j,k}) \leq \rho_{j,s}$.

This categorization helps identify elements and value ranges critical to frequent pattern detection as well as outliers, which are used to generate summaries. e_i belonging to dense ranges are more likely to lead to strong associations between attributes, while e_i belonging to sparse ranges are kept as potential outliers, since outliers are sometimes described as elements that have a very sparse presence within a dataset [1].

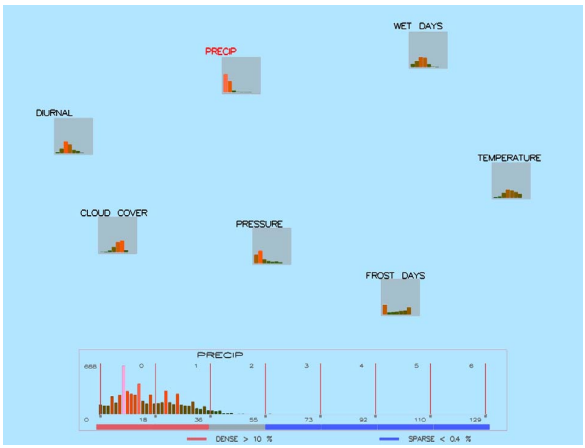

Fig. 2. The attribute layout for April for N. America along with the data distribution for *precipitation* showing ranges 0, and 1 categorized as dense and ranges 3 through 6 categorized as sparse.

B. Dimensionality Reduction

Once discretization is completed, the next step involves generation of low-dimensional subsets, in which attributes are partitioned into smaller clusters, where attributes within each cluster are strongly correlated.

To generate meaningful partitions, first correlations between all pairs of attributes A_j and A_l are calculated by comparing dense ranges from A_j to dense ranges in A_l. Given a pair of dense ranges $r_{j,k}$ and $r_{l,m}$, if a data element e_i's values fall within both ranges, then e_i is said to be common to the pair. The total number of such e_i is divided by the number of elements whose values fall in either range, to produce a correlation measure for the range pair. The highest correlation between dense range pairs for A_j and A_l is recorded as the correlation between the attributes in an $n \times n$ correlation matrix C.

Once pairwise correlations are computed, they are converted into a distance matrix C', where $C'_{i,j} = 1 - C_{i,j}$ (i.e. the higher the correlation between a pair of attributes, the closer they are to one another). Multidimensional scaling (MDS) [17] is then applied to C' to iteratively generate a 2D spatial layout that reflects the actual distances between attributes. The final spatial layout is meant to closely resemble the pairwise correlations between the attributes.

Data reduction operations are visualized using an attribute layout visualization shown in figure 2, which is also the initial visualization of our framework. It it consists of two main views: (1) the larger MDS generated view, and (2) the smaller data distribution view, displaying a single A_j at a time. In the *MDS view*, data attributes are represented by rectangular glyphs, which in turn contain small rectangular bars denoting both R_j (number of bars) and individual range densities (bar heights). The *data distribution view* displays the frequency distribution of values of A_j as well as range boundaries for each $r_{j,k}$ (A_j can be selected by clicking on the corresponding "attribute glyph" in the MDS view). Each data value is represented by a vertical bar, whose color and height corresponds to its frequency in D; short, dark green bars represent lower frequencies while tall, bright red and pink bars represent higher frequencies. Range categories are shown using horizontal bars placed directly below the corresponding ranges. Red bars denote dense ranges, blue denote sparse ranges, while grey bars represent neither categories. Current thresholds for density and sparsity are also displayed below the data distribution view.

Initially, the attribute layout is generated using default values. Users can choose to keep the initial configuration and the default parameters or modify each of these parameters uniquely for each A_j and generate a new updated layout. Controls are provided for this purpose via a dialog window. Choosing appropriate values for these parameters is critical to generating effective summaries. In figure 2 showing the weather data attributes, the values used were $R_j = 7$, $\rho_{j,d} = 10\%$), and $\rho_{j,s} = 0.4\%$) for all A_j.

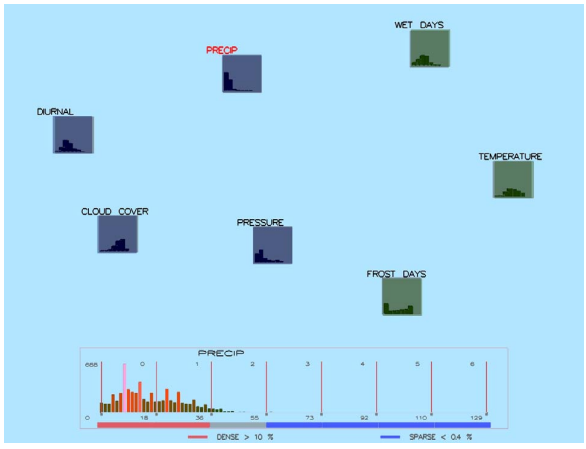

Fig. 3. Result of clustering using $k = 2$ clusters. *precip*, *cloud cover*, *vapor pressure* and *diurnal range* form blue attribute subset, while *temperature*, *wet days* and *frost days* form the green subset.

Following MDS, attribute subsets are computed using a k-means based clustering technique. The number of clusters k is defined by the user. In our clustering approach, we start by assigning k attributes randomly as cluster centers, and then assigning the remaining A_i to the nearest centers [15]. We then recompute cluster centers for each cluster and then reassign the remaining attributes to the new cluster centers. This step is repeated until the cluster centers stabilize.

Figure 3 shows the result of applying our k-means based clustering technique to the attribute layout generated in the previous step using $k = 2$. Cluster membership is shown using glyph colors. In figure 3, the blue cluster contains *precip*, *diurnal range*, *pressure* and *cloud cover*, while the remaining weather attributes belong to the green cluster. Users can interactively vary k to see how different numbers of clusters change the similar attribute subsets.

IV. Behavior Characterization

Once data reduction is complete, each similar attribute subset is examined for patterns, relationships, and outliers.

A. Pattern Detection

Pattern extraction is performed separately on each attribute subset generated in the previous step using a combination of association rule mining and interactive visualization techniques. As attribute subsets consist of correlated attributes, mining for association rules independently within each subset is much more efficient compared to searching D, but still allows us to correctly locate most or all of the strong association rules that exist in D.

We apply an Apriori-based technique [3] to generate association rules of the form $X \Rightarrow Y$, where X and Y are itemsets representing combinations of one or more attribute value ranges (items). In our approach, only dense attribute value ranges are used as individual items. As Apriori generates multiple rules for each frequent itemset, we only retain the rule with the highest confidence for each itemset. This prunes the number of rules that are generated and are to be visualized.

The association rules generated during this step are displayed using a radial layout technique shown in figure 4. For the weather dataset summarization, minimum support was set at 10% and minimum confidence at 85%. Attributes are represented by radial axes with attributes belonging to the same subset (i.e. cluster) placed alongside one another. Arcs represent individual association rules. Each arc connects individual items (i.e. attribute value ranges) that participate in the corresponding rule. Rectangles (denoting antecedents) or circles (denoting consequents) are displayed at points where the arcs intersect the axes of the attributes that feature in the corresponding rules. The attribute value ranges participating in the rule are also displayed in text form alongside these points.

For analyzing the weather dataset, we added *latitude* and *longitude* as additional attributes to each of the clusters generated during the attribute partitioning step. This allows us to correlate the observed weather variables with geographical locations. Both *latitude* and *longitude* value ranges were also discretized into 7 sub-ranges.

Figure 4 shows rules featuring various *latitude*, *longitude* values and *precip* value ranges, along with rules featuring *precip*, *pressure*, *cloud cover* and *diurnal range* values, among others. Also, various rules featuring *temperature* and *frost days* value ranges can be seen.

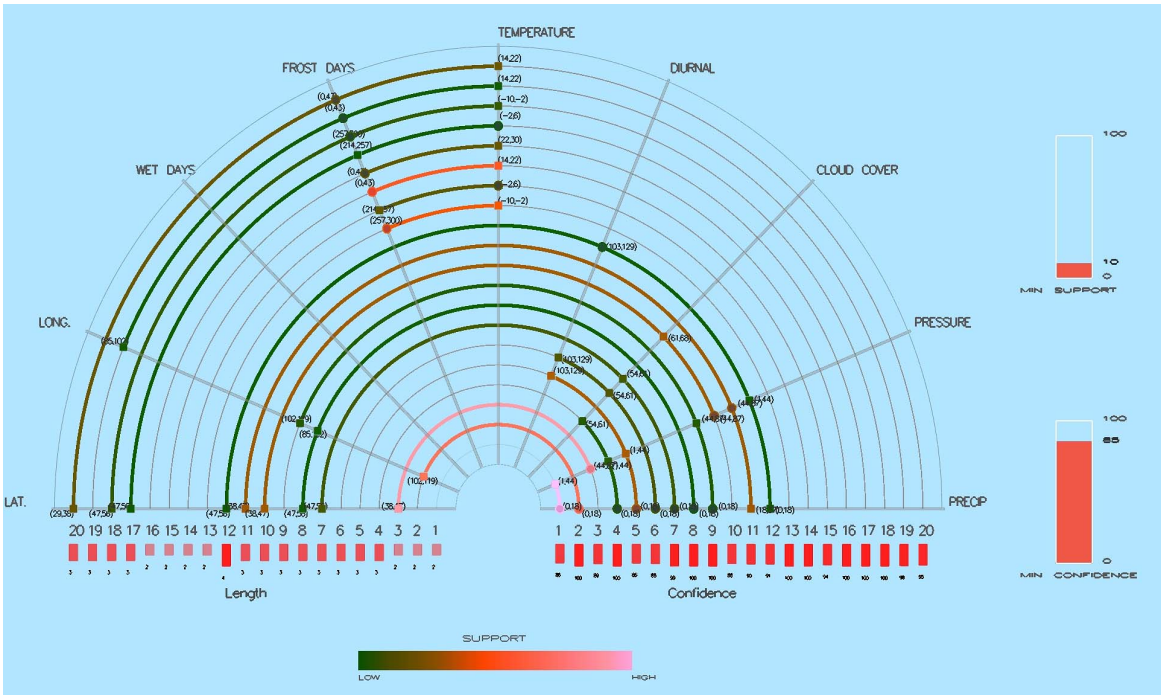

Fig. 4. The radial visualization showing the association rules extracted from the attribute subsets generated in the previous phase along with *latitude* and *longitude*. Bright red arcs denote high support rules while dark green arcs denote low support. Vertical bars below rule identifiers encode rule confidence (right area) and rule length (left area).

As the arcs have unique radii, they do not overlap thus allowing for clear visual differentiation of the individual rules. Also, using the radial layout, more attributes can be viewed in a smaller area. Arcs (i.e. rules) are also identified by unique numerical values displayed along both sides of the horizontal axis; the highest of these values also represents the total number of rules generated during the pattern mining step. Arc colors represent support of the rule, with dark green arcs signifying low support, and bright red and pink arcs signifying high support. Red, vertical bars displayed below the numerical labels denote the confidence (lower right) and the length (lower left) of the corresponding rules. Longer, deep red bars signify higher values.

Users can interact with the rule mining process by modifying minimum support and minimum confidence thresholds, and re-initiating the rule mining process using controls provided via a dialog window. Lower support and confidence thresholds could result in a large number of rules, while very high values could produce too few rules. We rely on the user's expertise to explore different support values and identify appropriate thresholds to generate rules that should be included within the data summaries. Users can also zoom in and out as well as rotate the visualization to gain more clarity. Figure 5 shows a magnified view of the radial visualization from figure 4.

B. Outlier Detection

A final characteristic that we identify are outliers. Our detection algorithm is focused on efficiency. In particular, we want to avoid computing pairwise distances between all pairs of elements, since this will negatively impact performance, particularly for datasets with large m and n. To achieve this goal, we modified the local neighborhoods technique [5] to use e_i from sparse attribute value ranges as candidate outliers.

The basic premise of the local neighbors approach is that, if an element e_i is much further from its k nearest neighbors than its neighbors are from their k neighbors (i.e. e_i's neighborhood is sparse relative to its neighbors neighborhoods), then e_i is likely to be an outlier. As computing the local neighborhoods to test every e_i requires significant computational effort, we avoid this issue by focusing on sparse ranges to identify potential outliers. This is because previous research suggests that outliers are elements that have a very sparse presence within a dataset [1]. We use a threshold on the sparseness of the local neighborhood of the potential outlier to filter the number of outliers detected. Varying this threshold to low or high changes the number of outliers detected; low threshold values result in a large number of outliers while high thresholds results in fewer outliers.

As with association rule mining, we perform outlier detection independently on similar attribute subsets. This further reduces the number of elements and attributes that need to be processed during the neighborhood searches and distance

calculations. It also avoids the problem of distances becoming less meaningful as dimensionality increases [1]. The similar attribute subsets act as lower-dimensional projections, compacting the elements into a smaller subspace and improving our ability to discriminate elements based on Euclidean distance measures.

Outliers are visualized using a parallel axes based display as shown in figure 6, in which the attributes are represented by vertical axes. Each outlier e_i is represented by a "polyline", with the individual line segments connecting the attributes values of the outlying element. To provide context, the frequency distribution of the attribute values is also displayed. Figure 6 displays the 3 outliers detected along with their corresponding attribute values from the attribute subsets of the weather dataset using an outlier threshold of 2.6.

V. Summary Generation and Visualization

Once important characteristics have been extracted, summaries generated from these characteristics are displayed to the user.

For summary generation purposes, we analyze the patterns generated from the previous phase to compute the number of frequent patterns that each attribute value range participates in. Attribute value ranges that participate in one or more frequent patterns are considered important. Additionally, we also maintain a count of the number of times each attribute value range occurs as an antecedent or as a consequent. This could help in identifying potentially dependent value ranges. We then compare pairs of important value ranges by computing the percentage of rules in which the range pairs participate in. This could inform the level of interaction between these value ranges. The co-participation measures are then transformed into a distance matrix (using a methodology similar to the one described in section 3). MDS is then used once again to generate a spatial configuration in which value ranges that co-occur more frequently in the extracted patterns cluster close to one another.

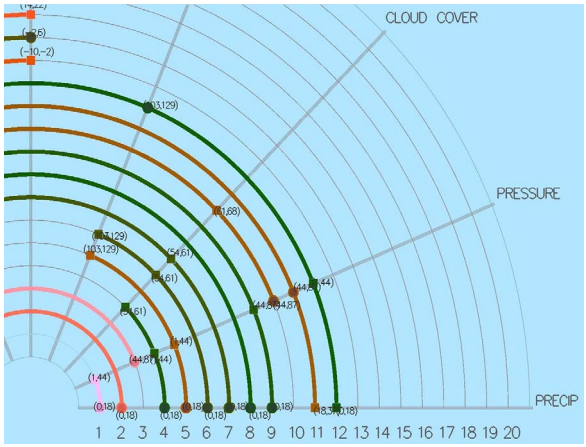

Fig. 5. A "zoomed-in" view of the radial visualization.

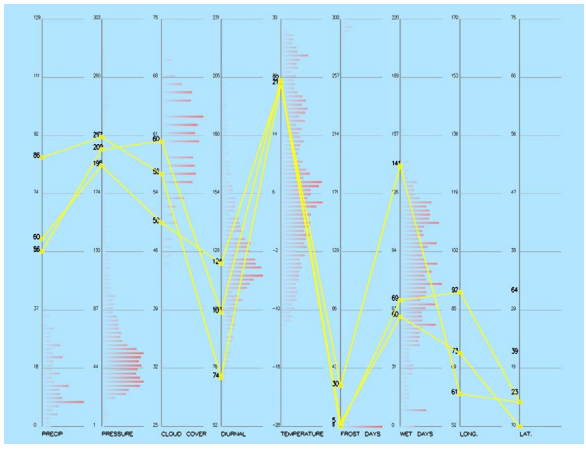

Fig. 6. Result of the outlier detection step displayed using a parallel axes view showing the 3 outliers (yellow poly-lines) found.

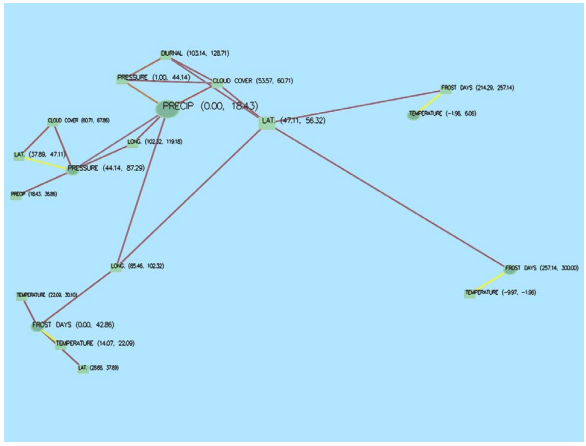

Fig. 7. The complete summmary visualization generated with nodes representing attribute value ranges. Larger nodes represent value ranges occurring more often in the extracted association rules. Lines connect value ranges occurring together in the association rules; bright yellow lines denote higher co-occurrence.

This configuration is then visualized to the user as shown in figure 7. Here, nodes represent important value ranges, with node size encoding the relative number of frequent patterns that feature the corresponding value range. Larger nodes indicates a higher participation in frequent patterns. Also, rectangular nodes represent majority antecedents while circles represent majority consequents. Text labels identify the attributes as well as the value ranges the nodes represent. Distances between nodes encode co-participation values; nodes closer to one another feature together more often than nodes away from one another. Lines connecting the nodes also reinforce this characteristic, with darker reds denoting low co-participation and bright yellows representing high co-participation. These characteristics can be seen more clearly in the detailed view seen in figure 8. As with the other visualizations users can

zoom in to areas of interest within the visualization.

Both figures 7 and 8 show the summaries of the weather dataset. For instance, we can see the relationship between *precip*$(0, 18.43)$, *pressure*$(1.00, 44.14)$, $(44.14, 87.29)$, *latitude*$(47.11, 56.32)$ and *longitude*$(85.46, 102.32)$, $(102.32, 119.18)$ value ranges. The close relationship between lower *temperature* ranges with higher *frost days* ranges and vice versa is also clear in figure 7, among others.

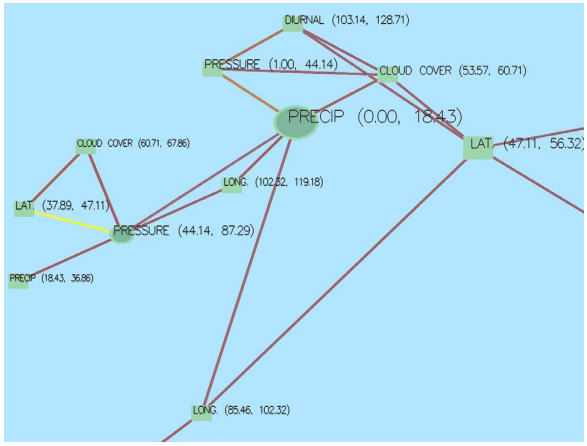

Fig. 8. A "zoomed-in" view of the summary visualization showing the relationships between *precip*, *cloud cover*, *vapor pressure*, *diurnal range* along with *latitude* and *longitude* value ranges.

VI. Conclusions

This paper presents a novel visual summarization approach designed to support the visualization and comprehension of large, multidimensional datasets such as weather and climate data. Our approach is aimed at generating relevant and concise summaries comprising of data characteristics such as important attribute value ranges, attribute subsets, patterns and outliers. Such summaries could then be visualized in place of the original dataset leading to more effective assimilation of the information contained within large, complex datasets. Our framework combines automated mathematical analysis with interactive visualization techniques to help users intuitively guide the summarization process and watch summary operations unfold. Users can also augment the summaries with their domain knowledge and expertise. Initial results of applying our framework to weather data have been promising leading us to believe that our framework could be helpful to climate and environmental researchers. The Renaissance Computing Institute (RENCI) is also currently collaborating with the State Climate Office of North Carolina, Raleigh, NC to explore further avenues for the application of our framework.

For future work, we are studying ways to enhance our framework's efficiency and effectiveness. We are currently working on incorporating more efficient analysis techniques, for e.g. hybrid clustering algorithms and FP-tree growth based association rule mining [9] among others. Another area of interest is in better integrating visual and non-visual techniques to help users guide the summarization process more intuitively. We are also interested in ways to include an user's domain knowledge within the summarization process, for e.g. using template-based mining techniques [13], to produce more relevant data summaries. We are also studying ways to both record the generated summary information as well as use this information. For instance, instead of sharing large, raw datasets, researchers could more efficiently and meaningfully share high-level summary information generated from the raw datasets. Finally, we would like to test and expand our framework's application by analyzing datasets across domains, for e.g. climate and socio-economic datasets among others, to further test and expand our framework.

Acknowledgment

The authors wish to thank Mark Brooks at the State Climate Office of North Carolina, Raleigh, NC for his valuable feedback on the summarization framework.

References

[1] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *SIGMOD Conference*, pages 37–46, 2001.

[2] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 1994.

[4] C. G. Beshers and S. K. Feiner. AutoVisual: Rule-based design of interactive multivariate visualizations. *IEEE Computer Graphics and Applications*, 13(4):41–49, July 1993.

[5] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *ACM SIGMOD Conference Proceedings 2000*, pages 93–104, 2000.

[6] U. M. Fayyad, G. P.-S., and P. Smyth. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press / The MIT Press, 1996.

[7] J. Han and N. Cercone. Ruleviz: A model for visualizing knowledge discovery process. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 244–253, 2000.

[8] J. Han, X. Hu, and N. Cercone. A visualization mode of interactive knowledge discovery systems and its implementations. *Information Visualization*, 2(2):105–125, June 2003.

[9] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of Data*, pages 1–12, New York, NY, 2000. ACM Press.

[10] D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

[11] C. G. Healey, R. S. Amant, and J. Chang. Assisted visualization of e-commerce auction agents. *Proceedings Graphics Interface 2001*, 2001.

[12] E. Kandogan. Visualizing multi-dimensional clusters, trends and outliers using star coordinations. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001.

[13] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. In N. R. Adam, B. K. Bhargava, and Y. Yesha, editors, *Third International Conference on Information and Knowledge Management (CIKM'94)*, pages 401–407. ACM Press, 1994.

[14] E. Knorr and R. Ng. Distance-based outliers in large data sets. In *VLDB Conference Proceedings*, 1998.

[15] S. Kocherlakota. *Interactive Visual Summarization for Visualizing Large Multidimensional datasets*. PhD thesis, Computer Science, North Carolina State University, Raleigh, North Carolina, 2006.

[16] T. Kohonen. *Self Organizing Maps*. Springer Verlag, Berlin, 1995.
[17] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978.
[18] N. Lesh and M. Mitzenmacher. Interactive data summarization: An example application. In *Proceedings of the working conference on Advanced visual interfaces*, pages 183 – 187, 2004.
[19] J. McQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
[20] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.
[21] J. Seo and B. Schneiderman. A rank-by-features framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proc. IEEE Conference on Information Visualization '04*, pages 65–72, 2004.
[22] J. Yang, A. Patro, S. Hang, N. Mehta, M. O. Ward, and E. A. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. In *Proceedings IEEE Symposium on Information Visualization 2004*, pages 73–80, 2001.