WaveCluster: a wavelet-based clustering approach for spatial data in very large databases

Wavelets have found their way into numerous fields far removed from traditional signal analysis. WaveCluster was a new method for clustering large, d-dimensional datasets by decomposing the regions into grids and performing a wavelet transformation. The approximation from the transformation was used to effectively cluster and label items using known techniques. WaveCluster directly addressed some of the thorny issues associated with the growth of large spatial datasets by automating data mining in an efficient manner.

A number of methods for partitioning datasets were discussed, but a grid based formulation was chosen for its ability to be processed quickly. Only a single sweep of the data was needed to quantize the regions using various measures. The only issue with the grid system was how to automatically determine the scale and size of each cell. If the cells were too small then unnecessary clusters were created. If the cells were too large, then many clusters were joined together. A heuristic was developed to iteratively find an acceptably sized cell, but this had a high overhead and they did not define the interpretation of acceptable. In the end, they settled the matter by noting that domain specific knowledge was critical. Still, the grid method had significant advantages over k-means clustering (which required pre-knowledge of the number of clusters), hierarchical clusters (which were order sensitive) and density based algorithms (which was very slow).

The key in recognizing the usefulness of wavelets was seeing that the features of each cell could be transformed into a signal. Consequently, the feature space would have low frequency where there was clustering. By applying a low pass wavelet filter, this information could be retrieved. Using wavelets added other benefits, such as automatic unsupervised clustering, removal of noise and random data (by discarding the detail signals created from the decomposition) and the capability to use a multi-resolution process to iteratively find coarser clusters. A standard discrete wavelet transform was performed in their algorithm and by using the approximation signal, clusters emerged in their visualizations. To automate their discovery and labeling they needed to merely to find the connected components using a previously described algorithm. Then the original cells were labeled with their cluster, which was done through a lookup table (since the approximation data no longer had the original information). A point that was neglected was how this was done accurately though, since decomposition generally creates vectors of different sizes from the original data depending on their mode of padding. A direct correlation from cell to cell may not be entirely correct.

The wavelet transformation was much faster than previous methods and could be parallelized easily. Additionally, wavelet analysis could handle any shape, including concave objects. The notion of using wavelets in this domain was novel, but in hindsight must have been quite obvious to those familiar with the wavelet transformation. The authors also seemed to gloss over how well would the labeling process work for large, diverse datasets. The images shown were all well defined and could be visually confirmed for correctness. In real data, for example geographic and medical imagery, there might be significantly more complex clusters that even humans would have trouble recognizing. In these cases, determining the correct level of resolution and cell size might be very difficult, especially automatically.
Real Time Feature Extraction for the Analysis of Turbulent Flows

Turbulence simulations generate enormous datasets that take months or years to analyze. Moreover, despite the ability to simulate flows, how they form and evolve over time in three dimensions is still poorly understood. There is a strong motivation to gain insight into turbulence. The authors noted that if one percent of bursting activities could be suppressed in airflow over a jet, the industry would save 200 million dollars a year. Since a broad theoretical framework does not exist to describe flows, simulations resolve all time scales and lengths, even when dealing with continent sized weather simulations for example. One major hurdle is storing the data compactly so that it can be analyzed. Typically, researchers save only a small fraction of all time frames created, hoping that they can still capture the essence of the flow. The authors discussed an updated technique that performed analysis at the same time as the simulated data in order to save only the relevant events.

The key idea was to define a trigger that when matched, would cause all recent data around a certain point to be stored. Furthermore, feature extraction and visualization could then be done immediately. After an initial learning phase guided by researchers, subsequent simulations could be automatically classified, visualized and saved compactly. In doing so, the focus shifts first to the problem of clustering. Separating truly random events from larger objects that should be seen together could show underlying organization. To that extent, the eigenvalues of a velocity gradient grid could be used. To further highlight specific points of interest, volume filling was done which found connected components through a nearest neighbor search. To view the results, ray tracing was then performed on the groups of objects which illustrated the evolution of items within the flow over time.

There were further problems that needed to be handled in the data mining that would allow those studying the flows to be able to ask the right questions. The authors suggested that using training sets to help identify the location of special events was required. The training sets needed to handle clusters, patterns and then higher level features, such as patterns between clusters. Interestingly, in their wish list for a clusterer, the WaveCluster algorithm seemed well suited. The paper was surprisingly devoid of any indication at how successful the method was at representing only the important information of a flow. Moreover, whether they actually gained new insight into the evolution of flows was ignored. The technique is clearly one that is needed, but its effectiveness remains questionable after the paper. When running a simulation that will take months, discarding valuable data that cannot be regenerated is critical. Making sure that the right data is being saved should be backed up by more evidence it seems.