

Chapter 6

Summary, Conclusions and Future Work

In this chapter, I summarize my work and describe the conclusions that can be drawn from it. This appears in the next section. In section 6.2, I discuss ideas for extending this work in the future.

6.1 Summary and Conclusions

In this dissertation, I have proposed new methods for visualizing collections of hypertext documents, leading to enhanced understanding of relationships among documents that are returned by information retrieval systems. A central component in the methodology is a new class of inter-document distances that includes information mined from a hypertext collection. These distances rely on a higher-order counterparts of the familiar co-citation similarity, in which co-citation is generalized from a relationship between a pair of documents to one between arbitrary numbers of documents. These document sets of larger cardinality are equivalent to itemsets in association mining.

The distances I propose are computed from higher-order similarities, but still retain a pairwise structure. These pairwise/higher-order hybrid distances allow the direct application of standard visualization tools such as clustering dendrograms and minimum spanning trees. However, they require much less complex user interaction compared to that of working directly with all frequent higher-order itemsets. The hybrid distances are

computationally feasible via previously proposed fast algorithms for computing frequent itemsets.

I provide a theoretical guarantee that consistency between clusters and frequent itemsets is attainable under the new hybrid distances. The guarantee is that there is always a sufficient degree of nonlinearity one can apply to itemset supports such that a more frequent itemsets forms a cluster at the expense of a less frequent itemset that overlaps it. While the guarantee does not include an upper bound on the necessary degree of nonlinearity, empirical results show that it is generally fairly low.

A similar guarantee can be supplied for the minimum spanning tree, in that there is always a sufficient nonlinearity degree such that a more frequent itemset will not be disconnected in the tree by a less frequent one. I show that for typical distributions of itemset supports, frequencies of occurrence of larger supports are very small. This contributes to the ability of the new hybrid distances to produce clusters or minimum spanning trees consistent with frequent itemsets.

I also propose the application of the hierarchical clustering dendrogram for information retrieval. The dendrogram enables quick comprehension of complex query-independent relationships among the documents, as opposed to the simple query-ranked lists usually employed for presenting search results. I introduce new augmentations of the dendrogram to support the information retrieval process, by adding document-descriptive text and glyphs for members of frequent itemsets. I also propose a metric for measuring the extent to which a clustering is consistent with frequent itemsets, which is based on the cardinality of the smallest cluster that contains all itemset members.

The minimum spanning tree has previously been proposed for visualizing co-citation relationships, the tree being interpreted as a network of document influences. In my work, I show that the new hybrid pairwise/higher-order distances tend to make the minimum spanning tree more consistent with frequent itemsets.

That is, when the hybrid distances are applied, frequent itemsets are more likely to be connected within the resulting trees, as measured by a metric that I propose based on numbers of itemset connected components. There is a slight tendency for the number of direct influences of frequent itemsets to increase, per a metric I propose based on itemset vertex degree. There is also a somewhat unpredictable tendency for the overall influence of a frequent itemset member to increase within the minimum spanning tree network, based on the number of its descendents in the tree.

This work represents the first known application of association mining in finding frequent itemsets for the purpose of visualizing hyperlink structures in information retrieval search results. The generalization of co-citation to higher orders helps prevent the obscuring of important frequent itemsets that often occurs with traditional co-citation based analysis, allowing the visualization of collections of frequent itemsets of arbitrary cardinalities. This work also represents a first step towards the unification of clustering and association mining.

3.2 Future Work

I now suggest ideas for extending this work. These ideas fall under the general headings of user-oriented clustering, the inference of association rules, the role of

maximal frequent itemsets, extensions of visualizations to higher dimensions, and the inference of citation semantics.

In *user-oriented clustering*, the user iteratively provides *a priori* domain knowledge to guide the clustering process. This is generally accomplished by the weighting of document pairs according to the importance of them being together in a cluster. The application of higher-order co-citation similarities allows sets of arbitrary cardinality to be weighted, providing a much richer form of cluster orientation.

For example, suppose documents A , B , and C should be clustered together, but B , C , and D should not. This orientation cannot be accomplished by mere pairwise weights. A high weight applied for the pair (B, C) on behalf of the triple (A, B, C) inadvertently increases the weight for (B, C, D) . But with higher-order co-citations, such weighting of the triples is trivial.

In association mining, the computation of frequent itemsets is often the first step in computing *association rules*, in which the presence of one itemset implies with some strength the existence of another, for non-overlapping itemsets. My results suggest that perhaps association rules can be inferred from the clustering dendrogram with the application of hybrid pairwise/higher-order distances.

An important type of frequent itemset in association mining is known as a *maximal* frequent itemset. This is a frequent itemset that is a subset of no other frequent itemset. In terms of the itemset lattice with partial order imposed by the subset relation, maximal frequent itemsets not “less than” any other itemset in the lattice. The role of maximal frequent itemsets in clustering with hybrid pairwise/higher-order distances remains to be explored.

In the landscape visualization I proposed for the minimum spanning tree, the tree is embedded in a vertex density estimate computed with the wavelet transform. The intersection of the surface with a threshold plane yields contour lines in the plane. The contours enclose areas containing clusters of minimum spanning tree vertices.

Now consider the extension of this to an additional spatial dimension. First position the minimum spanning tree vertices in a 3-dimensional volume rather than a plane. Then compute the wavelet-based vertex density in 3 dimensions. The application of a threshold volume to the 3-dimensional density results in contour surfaces, as opposed to the previous contour lines. These contour surfaces enclose volumes, each volume containing a cluster of vertices (documents). This is shown in Figure 6-1.

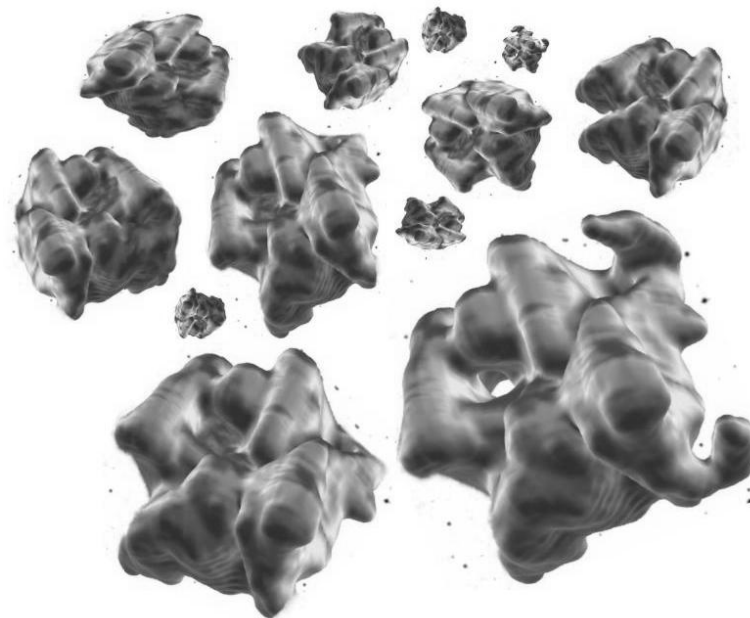


Figure 6-1: Extension of minimum spanning tree visualization to one higher spatial dimension.

Extending the minimum spanning tree landscape visualization should greatly improve the performance of algorithms for positioning tree vertices, since vertices would have more freedom in repositioning themselves during algorithm iterations. Perhaps the additional spatial dimension will also aid in the user's absorption of complex relationships in the minimum spanning tree visualization.

Currently the semantics of hypertext links such as science citations are lacking. The assumption is co-citations always imply some sort of similarity between documents, despite the common knowledge that the reasoning behind different citations or hyperlinks varies widely. Perhaps analyses of distributions of citations (links) within various documents, along with correlations between links and nearby text can help us make inferences about the meaning of the links.