

Abstract

This dissertation introduces new methods for visualizing collections of linked documents, for enhancing the understanding of relationships among documents that are returned by information retrieval systems. The methodology employs a new class of inter-document distances that capture the information inherent in the link structure of a collection. In particular, the distances are computed through the process of association mining, which results in the identification of sets of items (called itemsets) that are jointly linked to sufficiently often. In the context that links are citations appearing in published literature, itemsets are interpreted as higher-order co-citations. The new distances retain a simple pairwise structure, and are consistent with important frequently occurring itemsets. This approach provides the advantage that standard tools of visualization, e.g. hierarchical clustering and the minimum spanning tree can still be applied, while the distance information upon which they are based is richer. This work also proposes a number of enhancements to the standard visualizations, which support information retrieval tasks. The approach is demonstrated with document sets extracted from the Science Citation Index citation database. More generally, this work is applicable to information spaces in which objects may be associated by reference, e.g. software engineering, communications networks, and the World Wide Web.