

Chapter 1

Introduction

In some respect, the World Wide Web is like a vast library without an index system. Search engines are thus critical in finding Web pages of interest. Traditionally, search engines rank their results according to how well pages match keywords in the user query. In contrast, more innovative search engines such as Google [Henz00] first perform a keyword search, and then analyze the structure of Web hyperlinks to generate page ranks, independent of user queries for the selected pages. However, the results for these link-based search engines are still displayed as ranked lists, just as for traditional search engines.

Simple linear lists cannot adequately capture many of the complex hyperlink relationships among Web pages. Techniques from the field of information visualization [Tuft91][Card99] can help in this regard, making complex relationships more readily understandable. Visualization augments serial language processing with eye/brain parallel processing. Thus, the goal of visualization techniques is to enable users to recognize patterns in Web link structure, in turn helping to alleviate cyberspace information overload.

Previous approaches in this area have typically analyzed Web hyperlinks directly to determine page relationships [Klei98], or have relied on measures of similarity that only consider joint referencing of pairs of pages. The approach proposed in this work relies instead on measures of similarity among sets of pages of arbitrary cardinality. In

particular, the similarity among a set of pages is based on the number of other pages that jointly link to them.

The proposed similarity measures are inspired by the concept of co-citations, introduced in classical information retrieval in the context of citations appearing in published literature [Whit89]. Co-citations reduce complex citation or hyperlink graphs to simple scalar similarities between documents or Web pages. Co-citation based similarities allow the direct application of standard tools developed in other areas of science, such as cluster analysis [Vena94] and the minimum spanning tree [Corm96].

Similarity among objects by common reference has recently received some attention in the form of association mining [Agra93], which is a sub-field of data mining. While they are not usually recognized as such, what are defined as itemsets in association mining can be interpreted as generalized co-citations. Similarities between pairs of documents in co-citation analysis can be generalized to reflect the impact of sets of documents of arbitrary, larger cardinality that are jointly cited. Thus, itemsets are interpreted as higher-order co-citations.

This work is the first known application of itemsets to the visualization of link structures. Important (frequently occurring) higher-order itemsets are often obscured by the mere pairwise treatment of traditional co-citation analysis [Smal73]. The approach I take here involves the discovery of frequently occurring itemsets of arbitrary cardinalities, and the assigning of importance to them according to their frequencies. The generalization of co-citations to itemsets also enables user-oriented clustering [Bhuy91a][Bhuy91b][Bhuy97], where the user is allowed to specify weight of importance to larger sets of documents, beyond just pairs.

Because a collection of itemsets is not a disjoint set, there is a combinatorial explosion in the numbers of sets the user has to potentially deal with. I propose a novel approach to the problem of presenting results of association mining to users, which involves embedding higher-order co-citations (itemset supports) into pairwise document similarities. This hybrid of pairwise and higher-order similarities greatly reduces the complexity of user interaction, while being significantly more consistent with higher-order co-citations than standard pairwise similarities. It also admits the application of fast algorithms developed for data mining, which are empirically known to scale linearly with problem size [Agra94].

Mathematically, pairwise similarities can be modeled as a fully connected graph, to which clustering or minimum spanning tree algorithms can be applied. In the case of higher-order similarities, this graph is generalized to a hypergraph, i.e. a graph whose edges span more than just pairs of vertices. My approach of embedding higher-order co-citations in pairwise similarities eliminates the difficult task of forming clusters or minimum spanning trees directly from a hypergraph. Instead, standard algorithms may be directly applied.

The importance of clustering in information retrieval is well known [Baez99]. Link analysis in general provides a broadening of search results, by identifying documents that are linked to the initial set of documents matching the query. Clustering, in addition, can provide a narrowing of search results, by allowing the user to focus on documents in pertinent clusters only, while excluding other documents. In other words, as a result of this work, link analysis can be applied for both broadening and narrowing of search results.

The application of the proposed higher-order similarities to clustering algorithms greatly increases the tendency for important frequently occurring itemsets to appear together in clusters. This tendency is measured by a new metric I introduce specifically for comparing clusters to frequently occurring itemsets.

Moreover, I offer a theoretical guarantee that there is always a sufficient degree of nonlinearity one can apply to itemset supports (frequencies of occurrence) such that more frequent itemsets get placed together in clusters at the expense of less frequent ones. This guarantee relies on asymptotic growth bounds for nonlinearly transformed itemset supports. More specifically, the nonlinearly transformed support of the most frequently occurring itemset asymptotically bounds from above the nonlinearly transformed supports of all other itemsets. This means that distances between documents in the most frequent itemset can all be made smaller than distances to any documents outside that itemset, thus guaranteeing that the most frequent itemset will form a cluster. This argument can be extended to cover all other itemsets, based on their relative supports and overlap of documents.

My method of embedding itemset supports in pairwise similarities is particularly successful when the more frequently occurring itemsets are comparatively sparse. I therefore investigate citation itemset support distributions. That is, I show the frequency of occurrence of co-citations of a given order (itemsets of a given cardinality), for various science citation data sets.

For reasons of computational feasibility with large document collections, citation analysis has traditionally used the single-linkage clustering criterion only [Garf79]. Given the computational power of modern machines, stronger clustering criteria such as

average or complete linkage becomes feasible. I show that in the context of citation databases, single-linkage clustering alone is insufficient for completely characterizing the cluster structure of typical document collections. In particular, clustering results are usually quite different for each of the clustering criteria.

Previous approaches to co-citation based clustering either exclude visualization altogether, or visualize a single clustering corresponding to *a priori* numbers of clusters or single threshold similarity [Garf79][Smal93]. Instead, I apply the dendrogram visualization [Vena94], which shows the hierarchy of clusters for all possible thresholds, with no *a priori* requirement for the desired number of clusters. This is the first time that the dendrogram has been proposed for the visualization of either hypertext systems or document citation databases.

I introduce the concept of an “augmented dendrogram” for the visualization of significant (document) item associations. The augmented dendrogram highlights items that are a part of the same itemset, via graphical glyphs. This extension of the standard dendrogram allows the simultaneous visualization of both hierarchical clusters and important higher-cardinality itemsets.

The feasibility of the augmented dendrogram depends on a sufficiently small number of highlighted itemsets having items in common. When an item appears in too many highlighted itemsets, the augmented dendrogram becomes unwieldy. At this point one must rely on non-augmented dendrograms computed from the new hybrid pairwise/higher-order distances.

The dendrogram augmentation also includes the addition of textual information for documents being clustered. The leaves of the dendrogram tree correspond to

individual documents. The augmented dendrogram adds document bibliographic details to each leaf, thus supporting information retrieval.

The minimum spanning tree has been shown to provide a network of literature influences among collections of documents [Chen99a][Chen99b]. In this dissertation, I apply my new higher-order document similarities to minimum spanning tree visualizations. In particular, I investigate the effects that higher-order distance functions have on the influences of documents that are members of frequently occurring itemsets.

I propose three new metrics for measuring the effects of distances on frequent itemset members within the minimum spanning tree influence network. The first metric measures the connectedness of itemset members in the network. This is for testing the hypothesis that hybrid pairwise/higher-order distances increase the connectedness of members of frequently occurring itemsets. The other two metrics measure, respectively, the direct and total influences of an itemset member. They help test the hypothesis that the new hybrid distances increase the influence that members of frequently occurring itemsets have within the network.

I also introduce a novel method for the landscape visualization of a minimum spanning tree's node density, based on the wavelet transform. This visualization is considered "2.5-dimensional," being a two-dimensional landscape surface embedded in three dimensions. The landscape surface offers depth cues to help users recognize node positions. It also helps to alleviate the disorientation that often occurs with three-dimensional visualizations, since humans are adept at navigating landscapes. For this visualization I apply a force-directed layout algorithm for positioning nodes of the minimum spanning tree [Fruc91].

I introduce a novel approach to clustering based on the landscape visualization of the minimum spanning tree. The visualization is modified to show clusters by retaining only the tree edges between documents of the cluster. For example, single-linkage clusters are visualized by removing minimum spanning tree edges larger than some threshold amount.

Unlike the single-linkage approach, which is applied to the original edge distances, I propose the application of the threshold to the edge distances induced by the force-directed layout algorithm. The result is a new type of clustering in which clusters are oriented to highly influential documents, and highly influential documents themselves are placed in separate clusters. This is in contrast to traditional clustering methods, in which highly influential documents are placed *together* in clusters by virtue of the relatively small distances between them.

Interestingly, such clusters correspond approximately to connected components of the wavelet density landscape after the application of a threshold. Changes to the threshold value result in a nesting of connected components, which corresponds to a clustering hierarchy. Overall, I interpret the new wavelet landscape visualization as a form of spatial, hierarchical, fuzzy clustering.

I introduce the novel “augmented minimum spanning tree” for visualizing significant document associations. This extension of the standard minimum spanning tree visualization highlights documents that are part of the same itemset, allowing them to be readily identified within the tree. Like the augmented dendrogram, the augmented minimum spanning tree includes text for document nodes, as an aid to information retrieval.

The proposed methods of data mining and visualization are evaluated using data sets extracted from the Institute for Scientific Information's (ISI) Science Citation Index (SCI). The SCI is a component of ISI's Web of Science [WOS00], which provides access to citation databases that cover over 8,000 international journals. The application of data mining and visualization to science citations is consistent with the interests of this work's sponsor, the U. S. Department of Energy's Office of Scientific and Technical Information (DoE OSTI) [OSTI00].

But more generally, my approach is applicable to any information space in which objects may be associated by reference, particularly spaces modeled by directed graphs. Examples abound in such areas as software engineering, market analysis, communications networks, and perhaps most notably the World Wide Web.

The next chapter reviews previous approaches, and provides the background and further motivation for this work. It first describes literature citation analysis in the area of bibliometrics. It then covers analyses of link structure for information retrieval and visualization, which often rely on results from classical citation analysis. Next it describes association mining, including fast algorithms for computing frequently occurring itemsets. It then shows how advances in information visualization can contribute to comprehension of some potentially complex relationships among linked objects.

Chapter 3 introduces itemset supports as indicators of higher-order co-citation similarities, and describes my proposal for embedding them into pairwise similarities for clustering visualizations. It begins with some foundational issues in co-citation analysis, including the conversion of similarities to dissimilarities (distances) to facilitate the

application of clustering algorithms. It then describes hierarchical topological clustering, in particular single-linkage, average-linkage, and complete-linkage clustering, and describes the dendrogram visualization of cluster hierarchies.

Chapter 3 also introduces a metric that compares a given clustering to a set of significant itemsets, e.g. ones that occur frequently. The metric helps guide the design of the new inter-document distances that include higher-order co-citation similarities, i.e. hybrid pairwise/higher-order distances. The metric is then applied in a number of computational experiments with literature citation data sets, to test my proposed approach to document clustering.

Chapter 4 investigates methods of reducing the computational complexity of inter-document distances. It first applies fast algorithms for computing more frequently occurring itemsets in hybrid distances. It then proposes a model for itemset support distributions, the rapid decay of the distributions for larger supports providing additional evidence that fast algorithms for computing frequent itemsets scale linearly with problem size. Chapter 4 also offers and some experimental evidence that simple schemes for weighting of transactions or documents in computing pairwise distances is insufficient for consistency between clusters and frequent itemsets.

Chapter 5 covers the application of higher-order co-citations to the minimum spanning tree visualization. It first introduces the minimum spanning tree problem and algorithms for solving it. Next it describes the force-directed algorithm for positioning nodes of the minimum spanning tree. It then proposes three separate itemset-based evaluations of the minimum spanning tree: a metric for average number of connected components on the tree formed by an itemset, a metric for average vertex degree of an

itemset member, and a metric for the numbers of tree descendants of itemset members. Finally, Chapter 5 describes the minimum spanning tree landscape visualization and its interpretation as a novel approach for clustering.

Chapter 6 summarizes this dissertation, and highlights its conclusions. It also suggests ideas for future work in this area, including higher-order similarities for user-oriented clustering, inferring association rules from hierarchical clusters, applying maximal frequent itemsets (frequent itemsets that are not subsets of other ones), and extending the minimum spanning tree visualization to three dimensions.