Visualizing Association Mining Results through Hierarchical Clusters

Steven Noel
Center for Secure Information Systems
George Mason University
snoel@gmu.edu

Vijay Raghavan and C.-H. Henry Chu
Center for Advanced Computer Studies
University of Louisiana at Lafayette
[raghavan,cice]@cacs.louisiana.edu

Abstract

We propose a new methodology for visualizing association mining results. Inter-item distances are computed from combinations of itemset supports. The new distances retain a simple pairwise structure, and are consistent with important frequently occurring itemsets. Thus standard tools of visualization, e.g. hierarchical clustering dendrograms can still be applied, while the distance information upon which they are based is richer. Our approach is applicable to general association mining applications, as well as applications involving information spaces modeled by directed graphs, e.g. the Web. In the context of collections of hypertext documents, the inter-document distances capture the information inherent in a collection's link structure, a form of link mining. We demonstrate our methodology with document sets extracted from the Science Citation Index, applying a metric that measures consistency between clusters and frequent itemsets.

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 [1] first perform a keyword search, and use results from prior analyses of 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 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, thus helping to alleviate cyberspace information overload.

Previous approaches in this area have typically analyzed Web hyperlinks directly to determine page relationships [2], 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 that exploit 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 [3]. 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 [4].

Similarity among objects by common reference has recently received some attention in the form of association mining [5]. 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 association mining 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. The approach we take here involves the discovery of frequently occurring itemsets of arbitrary cardinalities, and the assigning of importance to them according to their support frequencies.

In a collection of itemsets, pairs of itemsets can overlap, so that there is a combinatorial explosion in the numbers of sets the user has to potentially deal with. We 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 [6].

The next section introduces our higher-order generalization of co-citations, and describes how they can be included in inter-document distances. Section 3 then proposes the augmented dendrogram clustering visualization for information retrieval, and demonstrates the effects of our new distances on the dendrogram. In Section 4, we define a metric that measures consistency between clusters and frequently occurring itemsets, and apply the metric to a number of test cases with real-world data extracted from a literature citation database. Section 5 then summarizes our work and highlights its conclusions.

2. Distances from Itemset Supports

Hyperlink systems, e.g. the World Wide Web or science citations, can in general be modeled as directed graphs. A graph edge from one document to another indicates a link from first to second. It is convenient to apply a matrix formulation for the development of link-based document distances. In fact, for actual implementation this leads to the direct application of matrix data structures and operations usually found in programming languages.

In the matrix formulation, a binary adjacency matrix is formed that corresponds to the linkage graph. We take the convention that adjacency matrix rows are for citing documents and columns are for cited documents. Thus for adjacency matrix **A**, element $a_{i,j} = 1$ indicates that document *i* cites document *j*, and $a_{i,j} = 0$ is the lack of citation.

A co-citation between two documents is the citing (or hypertext linking) of the two documents by another one [3]. A measure of similarity between a pair of documents is the number of documents that co-cite the pair. This is known as citation count. Taken over all pairs of documents, the co-citation count similarity serves as a compact representation of citation graph structure.

A central component in classical citation analysis is clustering based on co-citations as a measure of similarity. In the case of co-citations, an association is made between two documents according to the number of times they are co-referenced, i.e. through hypertext links or literature citations. The purpose of clustering is to form larger sets of documents that are more strongly associated with one another than they are to documents outside the cluster.

Traditional citation analysis typically applies single-linkage clustering, because of its lower computational complexity [7]. But because of its very weak clustering criterion, single-linkage can have problems unless the data are inherently well clustered. Given the improved performance of modern computing machines, it becomes feasible to apply stronger clustering criteria in citation analysis.

For the stronger clustering criterion of complete linkage, *all* similarities for the three pairs need to exceed the threshold before the documents constitute a single cluster. But it is still possible to construct cases in which there is not even one document that cites all of the clustered documents simultaneously. The complete-linkage criterion is a necessary but not sufficient condition for the simultaneous citing of all documents in a cluster.

We propose a generalization of the co-citation similarity in which sets of cardinality above two are considered for co-citation. That is, we defin e a similarity among a set of cited documents that is based on the number of times all the members of the set are simultaneously cited. Because the similarity is among more than two documents, we consider it to be higher order than pairwise.

In our matrix formulation, itemset supports are computed for sets of columns (cited documents) of the adjacency matrix, just as they are computed for pairs of columns in computing co-citation counts. For itemset I of cardinality $|I|$, whose member documents correspond to columns $j_1, j_2, \ldots, j_{|I|}$, its scalar support $\zeta(I)$ is

$$\zeta(I) = \sum_i a_{i,j_1} a_{i,j_2} \cdots a_{i,j_{|I|}} = \sum_i \prod_{\alpha=1}^{|I|} a_{i,j_\alpha} ,$$

where i indexes rows (citing documents). Just as for pairwise co-citations, the term $a_{i,j_1} a_{i,j_2} \cdots a_{i,j_{|I|}}$ represents single co-citation occurrences, which are now generalized to higher orders. The summation then counts the individual higher-order co-citation occurrences.

A central problem in data mining is the discovery of frequent itemsets. In the context of hypertext systems, such frequent itemsets represent groups of highly similar documents based on higher-order co-citations. But managing and interacting with itemsets for information retrieval is problematic. Because of the combinatorially exploding numbers of itemsets and their overlap, user interaction becomes unwieldy.

Also, standard tools of analysis and visualization such as clustering assume an input matrix of pairwise distances. Mathematically, distances for all document pairs correspond to a fully connected distance graph. But the generalization to higher-order distances means that the distance graph edges are generalized to *hyperedges*, that is, edges that are incident upon more than two vertices. It is difficult to generalize clustering algorithms to such distance hypergraphs.

Our approach to this problem is to apply standard clustering algorithms, but with pairwise distances that include higher-order co-citation similarities. The new distances we propose are thus a hybrid between standard pairwise distances and higher-order distances. For information retrieval visualization, users need only deal with disjoint sets of items, rather than combinatorial explosions of non-disjoint itemsets. The approach is designed such that member documents of frequent itemsets are more likely to appear together in clusters.

We extend the standard model by computing document similarities from higher-order support features by summing supports over all itemsets that contain the document pair in question. More formally, the itemset support feature summation is

$$s_{j,k} = \sum_{\{I \mid j,k \in I\}} \zeta(I).$$

This yields the similarity $s_{j,k}$ between documents j and k, where $\zeta(I)$ is the support of itemset I.

We then introduce a nonlinear transformation $T[\zeta(I)]$ to be applied to the itemset supports $\zeta(I)$ before summation. The transformation T is super-linear (asymptotically increasing more quickly than linearly), so as to favor large itemset supports. The hybrid similarity $s_{j,k}$ then becomes

$$s_{j,k} = \sum_{\{I \mid j,k \in I\}} T[\zeta(I)]. \quad (1)$$

A straightforward approach for reducing computational complexity for hybrid pairwise/higher-order distances is to exclude itemsets whose supports fall below some threshold value, denoted *minsup*. The worst-case complexity of the problem of finding all frequent itemsets is exponential. But algorithms have been proposed that empirically scale linearly with respect to both the number of transactions and the transaction size [6]. We then simply exclude from the summation in (1) all itemsets with supports $\zeta(I)$ below *minsup* = m, i.e.

$$s_{j,k} = \sum_{\{I \mid j,k \in I, \zeta(I) \geq m\}} T[\zeta(I)]. \quad (2)$$

It is convenient to normalize the similarity $s_{j,k}$ through the linear transformation

$$\hat{s}_{j,k} = \frac{s_{j,k} - \min(s_{j,k})}{\max(s_{j,k}) - \min(s_{j,k})}, \quad (3)$$

yielding the normalized similarity $\hat{s}_{j,k} \in [0,1]$. Standard clustering algorithms assume *dissimilarities* rather than similarities. We convert the normalized similarity $\hat{s}_{j,k}$ to a dissimilarity (distance) through additive inversion, i.e.

$$d_{j,k} = 1 - \hat{s}_{j,k}. \quad (4)$$

This results in distance $d_{j,k}$ between documents j and k, normalized to $d_{j,k} \in [0,1]$.

We have derived a theoretical guarantee for the nonlinear transformation of itemset-support features in the similarity computation. In the discussion below, we provide a sketch of this theoretical result. It begins with a proof that the most frequent itemset can always be made a cluster, given a large enough degree of nonlinearity in the transformation. The proof relies on the fact that for a super-linear transformation $T = T_p = T(\zeta; p)$ of itemset supports ζ in (1) or (2), as the degree of nonlinearity p increases, $T_p(\zeta)$ with a larger ζ is asymptotically bounded from below by $T_p(\zeta)$ with a smaller ζ. Since the term with largest ζ asymptotically dominates the distance summation, the result is that documents in the most frequent itemset are asymptotically closer to one another than to any other documents, thus forming a cluster.

We then generalize the proof to cover the clustering of arbitrary itemsets in terms of their relative supports and document overlap. The result is that more frequent itemsets asymptotically form clusters at the expense of less frequent itemsets that overlap them. If there is no overlapping itemset with more support, then a given itemset will form a cluster for a sufficiently large value of the nonlinearity parameter p. For overlapping itemsets with equal support, the itemset members not in common asymptotically form clusters, and each of the members in common are asymptotically in one of those clusters, though there is no guarantee for exactly which of them.

This theoretical guarantee provides no upper bound on the necessary degree of nonlinearity of p to ensure itemset clustering for a given data set. But empirically, we have found that the transformation

$$T(\zeta) = \zeta^{p} \quad (5)$$

with $p = 4$ usually results in the most frequent itemsets appearing together in clusters.

In this section, we proposed a new class of co-citation based inter-document distances. These are a hybrid between pairwise distances and higher-order distances. The new hybrid distances retain a simple pairwise structure, but are better able to match higher cardinality itemsets than are standard pairwise distances. The next section applies these hybrid distances to hierarchical clustering visualizations, in support of information retrieval tasks.

3. Hierarchical Clusters with Higher-Order Co-Citations

Clustering plays a central role in information retrieval. In classical work, clustering based on co-citation similarity is known to correspond well to individual fields of knowledge [7]. For information retrieval, results from simple keyword queries can be clustered into more refined topics. Co-citation-based clustering provides a narrowing of search results, by allowing the user to focus on documents in pertinent clusters only. This helps alleviate the potentially tedious task of manually reviewing large lists of search results. Also, co-citation analysis can broaden search results by providing alternative documents linked by co-citation.

Three important heuristics for clustering are *single-linkage*, *average-linkage*, and *complete-linkage*. These heuristics are agglomerative, at each step merging clusters that have the closest distance between them. Arguments have been given for all three heuristics in terms of desirable clustering characteristics.

For our experiments with real-world document collections, clusters resulting from the 3 criteria are generally significantly different. This suggests that the documents are not inherently distributed as well-separated clusters. In particular, we have seen ample real-world examples of single-linkage "chaining." This is in direct contrast to the classical notion that typical document collections have well defined clusters, so that single-linkage is adequate [7]. However, at least one author has suggested that single-linkage clustering may be inadequate for citation analysis [8].

The *dendrogram* is a tree visualization of a hierarchical clustering. Leaves of the dendrogram tree are individual documents, at the lowest level of the hierarchy. Non-leaf nodes represent the merging of two or more clusters, at increasing levels of the hierarchy. A node is drawn as a horizontal line that spans over its children, with the line drawn at the vertical position corresponding to the merge threshold distance.

We now demonstrate our approach to itemset-based clustering, in which we compare frequent itemsets to graph-theoretic clustering. The demonstration employs data extracted from a literature citation database, the Institute for Scientific Information's Science Citation Index (SCI). We do the itemset/clustering comparison by a novel augmentation of the dendrogram with members of frequent itemsets, which allows an easy visual assessment of our proposed itemset-matching metric. For the example, we do an SCI query with keyword "wavelet*" for the year 1999. The first 100 documents returned by the query cite 1755 documents. We filter these cited documents by citation count, retaining only those cited three or more times, resulting in a set of 34 highly cited documents.

We then compute complete-linkage, average-linkage, and single-linkage clusters for the set of 34 highly cited documents. Here we first apply the standard pairwise method of computing co-citation based distances. The resulting augmented dendrogram is shown in Figure 1. The dendrogram is augmented by the addition of graphical symbols for members of frequent 4-itemsets, added at the corresponding tree leaves.

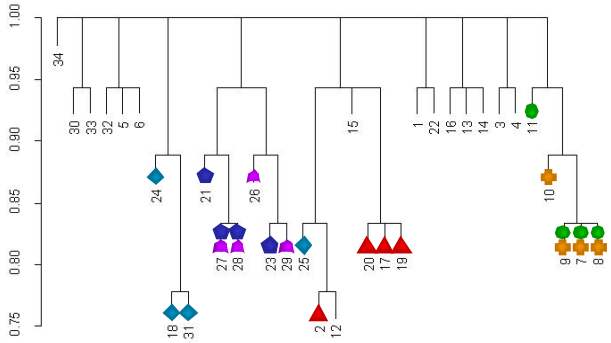
In this example, the most frequently occurring 4-itemset is {2, 17, 19, 20}▲. For complete linkage, documents 17, 19, and 20 of this itemset are a possible cluster. These documents apply wavelets to problems in the field of chemistry, and are well separated from the rest of the collection, both thematically and in terms of co-citations. But including document 2 (a foundational wavelet paper by wavelet pioneer Mallat) in this cluster would require the inclusion of documents 12, 15, and 25, which are not in the itemset. These three additional documents are another foundational wavelet paper by Mallat, and two foundational papers by fellow pioneer Daubechies.

For single linkage, there is even less cluster/itemset consistency. The itemset {2, 17, 19, 20}▲ is possible within a cluster only by including 8 other documents. We interpret this as being largely caused by single linkage chaining. In general, the application of clustering to mere pairwise co-citation similarities is insufficient for ensuring that itemsets of larger cardinality appear as clusters, even with complete-linkage.

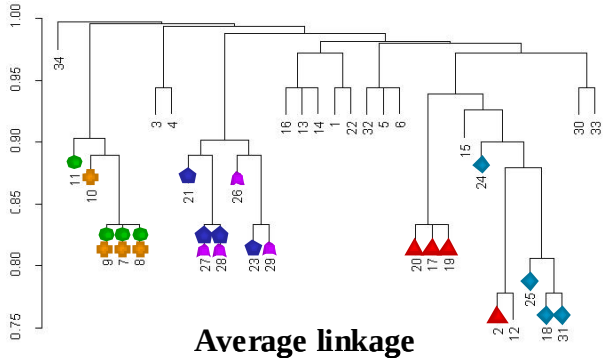
As a comparison with standard pairwise distances, Figure 2 shows complete-linkage clusters computed with our hybrid pairwise/higher-order distances. It considers three separate cases, each case being taken over multiple values of itemset cardinality χ. The three cases are $\chi = 2,3$; $\chi = 2,3,4$; $\chi = 3,4$. Here the itemset supports $\zeta(I)$ are nonlinearly transformed by $T[\zeta(I)] = [\zeta(I)]^4$, with distances computed via (1), (3), and (4).

Consistency between clusters and frequent itemsets is considerably improved with our hybrid distances. The most frequent itemset {2,17,19,20}▲ forms a cluster for

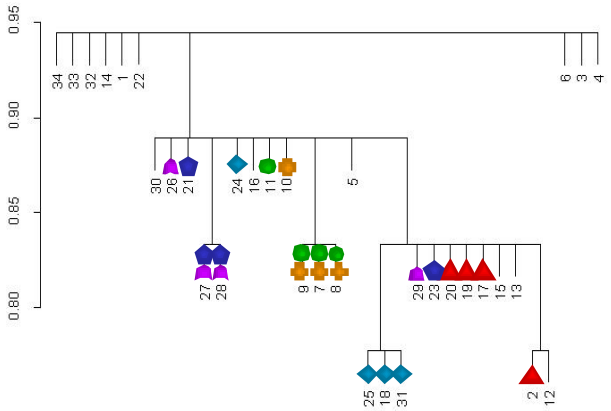
the two cases $\chi = 2,3,4$ and $\chi = 3,4$. However, the case $\chi = 2,3$ has inconsistency for the most frequent itemset. It is apparent that the lowest-order (pairwise) supports are the source of the disagreement. Lower order supports are generally larger than higher-order supports, and thus tend to dominate the summation in (1).



Complete linkage

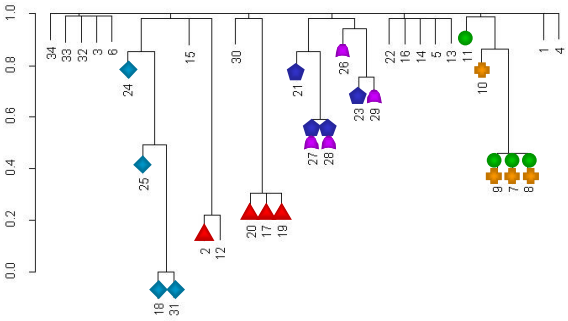


Average linkage

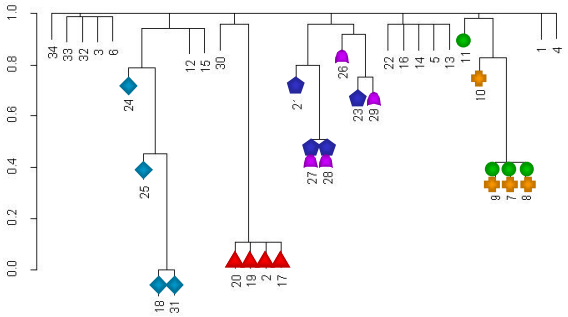


Single linkage

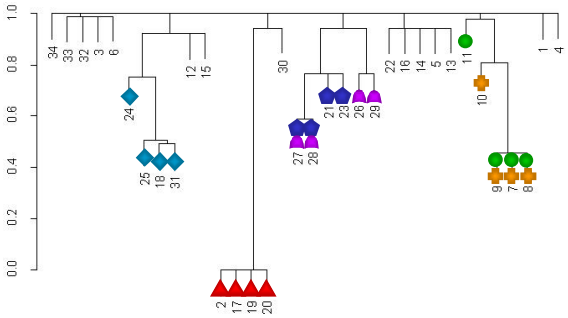
Figure 1. For standard co-citation document distances, there is considerable inconsistency between clusters and frequent itemsets.



(Cardinality 2)^4 + (Cardinality 3)^4



(Cardinality 2)^4 + (Cardinality 3)^4 + (Cardinality 4)^4



(Cardinality 3)^4 + (Cardinality 4)^4

Figure 2. Distances that include higher-order co-citations yield much improved consistency between clusters and frequent itemsets.

We suggest that clustering dendrograms such as Figure 2 can provide greatly enhanced understanding of document sets resulting from information retrieval keyword searches. Augmentation with symbols for frequent itemset members allows direct identification of these important itemset. The visualized clusters and itemsets are computed from distances based on citation (link) information, thus complementing text-based analysis. As retrieval output, the user would be presented with a modified form of the dendrogram shown here, which has been rotated clockwise by 90° and augmented with text corresponding to each document.

The augmented dendrogram we propose can help guide the navigation of retrieval results. For example, small representative samples can be examined from a number of separate clusters, allowing the user to more quickly identify documents of potential interest. Once a promising cluster is identified, the user can then focus solely on documents within it.

4. Experimental Results

This paper proposes new methods of visualizing hypertext document clusters for information retrieval. The methodology employs a new class of inter-document distances that is a hybrid between standard pairwise and higher-order distances. These higher-order distances are analogous to itemset supports in association mining.

In this section, we apply our proposed hybrid document distances to real-world hypertext. In particular, we apply them to data sets from the Institute for Scientific Information's Science Citation Index (SCI). Science citations are a classical form of hypertext, and are of significant general interest. The assumption that links imply some form of content influence also holds reasonably well for citations. This is in contrast to Web hypertext, in which links could be for more general purposes, such as navigation.

The SCI data sets we employ are described in Table 1. For each data set, the table gives the SCI query keyword and publication year(s), the number of citing documents resulting from the query, and the number of documents they cite after filtering by citation count. For the data sets 1, 5, and 9, results are included for both co-citations and bibliographic coupling, yielding data sets 2, 6, and 10 (respectively), for a total of 10 SCI data sets.

Our empirical tests apply a metric that compares clustering to frequent itemsets, determining whether given itemsets form clusters comprised only of the itemset members. In other words, we determine the minimal-cardinality cluster that contains all the members of a given itemset, and compare that cluster cardinality to the itemset cardinality. This is then averaged over a number of itemsets, to yield an overall itemset-matching metric for a clustering.

More formally, let $\pi = \{\pi_1, \pi_2, \ldots, \pi_{k_1}\}$ be a partition of items consistent with the hierarchical clustering merge tree. Furthermore, let $I = \{I_1, I_2, \ldots, I_{k_2}\}$ be a set of itemsets. Then for each itemset $I_i \in I$, there is some block of the partition $\pi_j \in \pi$ such that $|\pi_j|$ is minimized, subject to the constraint that $I_i \subseteq \pi_j$. We call this π_j the minimal cluster containing the itemset.

Table 1. Details for SCI data sets.

Data Sets	Query Keyword	Years	Citing Docs	Cited Docs
1, 2	adaptive optics	2000	89	60
3	collagen	1975	494	53
4	genetic algorithm* and neural network*	2000	136	57
5,6	quantum gravity AND string*	1999-2000	114	50
7	wavelet*	1999	100	34
8	wavelet*	1999	472	54
9,10	wavelet* AND brownian	1973-2000	99	59

Once a minimal (cardinality) cluster π_j is found for an itemset, a metric can be defined for measuring the extent to which the itemset is consistent with the cluster. This metric $M(\pi, I_i)$ is simply the portion of the cluster occupied by the itemset, or in terms of set cardinalities,

$$M(\pi, I_i) = \frac{|I_i|}{|\pi_j|}.$$

The metric $M(\pi, I)$ is defined for a set of itemsets I by averaging $M(\pi, I_i)$ over $I_i \in I$, that is,

$$M(\pi, I) = \frac{1}{|I|}\sum_{I_i \in I} M(\pi, I_i) = \frac{1}{|I|}\sum_{I_i \in I}\left(\frac{|I_i|}{|\pi_j|}\right). \qquad (6)$$

The itemset-matching metric $M(\pi, I)$ takes its maximum value of unity when $I_i = \pi_j$, indicating the best possible match between itemsets and clusters, and its minimum value $M(\pi, I) = |I_i|/n$, indicating the poorest possible match. Figure 3 illustrates the itemset-matching clustering metric $M(\pi, I)$.

Table 2 shows how hybrid pairwise/higher-order distances are computed for the experiments with SCI data sets. The table shows the itemset cardinalities $|I|$ that are applied in the distance formula (1), and the itemset support nonlinearity parameters p for itemset nonlinearity $T(\zeta) = \zeta^p$. For each hybrid distance formula, we compare metric values to those for standard pairwise distances, applying the same cardinalities in (6) as in (1).

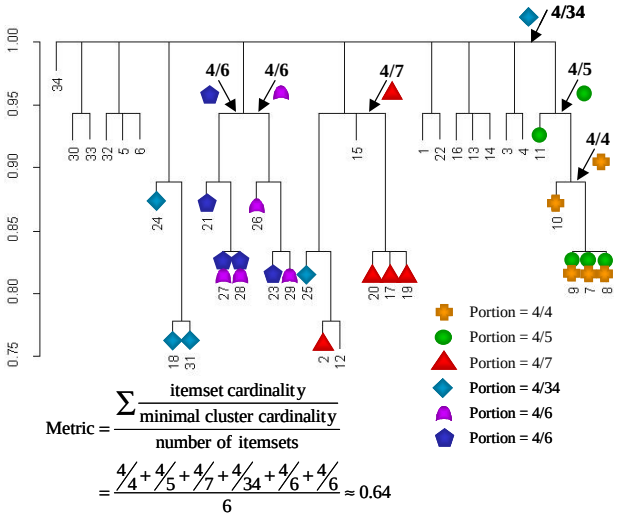


Figure 3. Itemset-matching clustering metric is the average portion occupied by an itemset within the minimal cluster containing it.

The comparisons are done for each combination of complete-linkage, average-linkage, and single-linkage clustering. We also compare for each combination of the most frequent itemset, the 5 most frequent itemsets, and the 10 most frequent itemsets in (6). For test cases in which nonlinearity parameter value $p = 4$ yields relatively low metric values, we include additional test cases with $p = 6$. Since there are 25 different combinations of itemset cardinalities and support nonlinearities, each with $3 \times 3 = 9$ different combinations of clustering criteria and sets of most frequent itemsets, there are $9 \times 25 = 225$ total test cases.

Table 2. Itemset cardinalities and support nonlinearities for hybrid pairwise/higher-order distances.

Data Sets	[Itemset Cardinality, Support Nonlinearity]
1	[3,4], [3,6], [4,4], [4,6]
2,6	[3,4], [4,4], [4,6]
3,5,7,8,9,10	[3,4], [4,4]
4	[3,4], [3,6], [4,4]

Table 3 compares clustering metric values for the test cases in Table 2. The table classifies cases as having metric values for hybrid distances equal to, greater than, or less than metric values for standard pairwise distances. Higher metric values correspond to clusters being more consistent with frequent itemsets. For most test cases (169 of 225 versus 27 of 225), the new hybrid pairwise/higher-order distances result in better consistency with frequent itemsets in comparison to standard pairwise distances. For a relatively small number of cases (29 of 225), the metric values are equal.

Table 3. Clustering metric comparisons for standard pairwise (P.W.) versus higher-order (H.O.) distances.

Data set	H.O.=P.W.	H.O.>P.W.	H.O.<P.W.	Cases
1	6	16	14	36
2	7	15	5	27
3	0	18	0	18
4	1	24	2	27
5	3	13	2	18
6	2	22	3	27
7	2	16	0	18
8	5	13	0	18
9	3	14	1	18
10	0	18	0	18
Totals	29	169	27	225

For the 20 combinations of 10 data sets and 2 itemset cardinalities, we find that itemset support nonlinearities with p=4 are usually sufficient (15 of 20) for a good match to frequent itemsets. Otherwise nonlinearities with p=6 in (5) are sufficient, in all but one of the 20 combinations. Here we consider a clustering metric value greater than about 0.7 to be a good match. This corresponds to a frequent itemset comprising on average about 70% of a cluster that contains all its members.

Table 4. Clustering metrics for hybrid distances with full computational complexity (*minsup* 0) versus hybrid distances with reduced complexity (*minsup* 2).

Data set	(*minsup* 2) = (*minsup* 0)	(*minsup* 2) > (*minsup* 0)	(*minsup* 2) < (*minsup* 0)	Cases
3	18	0	0	18
5	18	0	0	18
8	18	0	0	18
9	18	0	0	18
10	11	0	7	18
Totals	83	0	7	90

For data sets 3, 5, 8, 9, and 10, we compute itemset-matching clustering metrics resulting from the complexity reduction technique in (2), and compare metric values to the full-complexity method in (1). The clustering metric results of the 90 cases from the five extracted data sets are summarized in Table 4. The

results show that excluding itemset supports below *minsup* generally has little effect on clustering results, particular for smaller values of *minsup*. However, there is some degradation in metric values for higher levels of *minsup*. Here "degradation" means that metric values are smaller when some itemset supports are excluded, corresponding to a poorer clustering match to frequent itemsets.

We offer the following interpretation for the *minsup*-dependent degradation in clustering metric. Members of frequent itemsets are typically frequently cited documents overall. Such frequently cited documents are likely to appear in many itemsets, even less frequent itemsets. Thus there are likely to be many itemsets below *minsup* that contain these frequently cited documents. Excluding itemsets below *minsup* then removes the supports that these itemsets contribute to the summations in computing hybrid distances.

5. Summary and Conclusions

We have proposed a new methodology for visualizing association mining results. A central component in the methodology is a new class of inter-item distances computed directly from itemset supports. While the distances are computed from supports of itemsets of arbitrary cardinality, they still retain a simple pairwise structure. These hybrid pairwise/higher-cardinality distances thus allow the direct application of standard visualization tools such as clustering dendrograms. However, they require much less complex user interaction compared to that of working directly with all frequent itemsets. The hybrid distances are computationally feasible via fast algorithms for computing frequent itemsets. We provided a theoretical guarantee that under the new hybrid distances, consistency between clusters and frequent itemsets is attainable independent of the clustering criterion.

We proposed the application of the hierarchical clustering dendrogram for association mining visualization, which enables quick comprehension of complex distance relationships among items. We also introduced new augmentations of the dendrogram to support the visual mining process by adding item-descriptive text, and graphical symbols for members of frequent itemsets.

We tested the effects of our new hybrid distances on hierarchical clustering dendrograms, for large numbers of test cases with data extracted from the Science Citation Index. In particular, we compared dendrograms between a variety of hybrid distance formulas and standard pairwise distances, in terms of a metric that measures consistency between clusters and frequent itemsets. For the majority of the test cases, metric values were higher for our hybrid distances, indicating better consistency between clusters and frequent itemsets. We also conducted experiments showing that excluding itemset supports below some minimum value has relatively little effect on itemset/cluster consistency, so that it is reasonable to apply fast algorithms for computing the frequent itemsets needed for the hybrid distances.

This is the first known application of clustering for the visualizing association mining results. The pairwise structure of the new hybrid inter-item distances allows the visualization of frequent itemsets of arbitrary cardinalities. The nonlinear transformation of itemset supports helps ensure consistency between clusters and important frequent itemsets, making frequent itemsets less likely to be obscured during visualization. As a more basic contribution, this work represents a first step towards the unification of association mining and clustering visualization.

6. References

[1] M. Henzinger, "Link Analysis in Web Information Retrieval," *Bulletin of the Technical Committee on Data Engineering*, 23, special issue on Next-Generation Web Search, pp. 3-8, September 2000.

[2] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, January 1998, pp. 668-677.

[3] H. Small, "Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents," *Journal of the American Society of Information Science*, 24, pp. 265-269, 1973.

[4] J. Hartigan, *Clustering Algorithms*, John Wiley & Sons, New York, 1975.

[5] R. Agrawal, T. Imielinski, A. Swami, "Mining Association Rules between Sets of Items in Large Databases," in *Proceedings of the 1993 International Conference on the Management of Data*, Washington, DC, May 1993, pp. 207-216.

[6] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules," in *Proceedings of the 20th International Conference on Very Large Databases*, Santiago, Chile, September 1994, pp. 487-499.

[7] E. Garfield, *Citation Index: Its Theory and Application in Science, Technology, and Humanities*, John Wiley & Sons, New York, 1979.

[8] H. Small, "Macro-Level Changes in the Structure of Co-Citation Clusters: 1983-1989," *Scientometrics*, 26, pp. 5-20, 1993.