Visualizing Association Mining Results through Hierarchical Clusters

Steven Noel Vijay Raghavan and C.-H.Henry Chu

Presenter:Minawaer.Nulahemaiti

Presentation Outline

Motivation

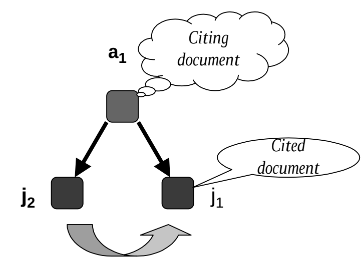
- *Web: a vast library without an index.*
- *Search engines rank their results in a keywords-based approach.*
- *Google is link-based search engines, results of it are still display as ranked lists*
- *Simple linear lists can't adequately capture many of the complex hyperlink relationships among web pages*
- *Information visualization can help making complex relationships more readily understandable.*
- *The visualization techniques is enable users to recognize patterns in web link structure,thus helping to alleviate cyberspace information overload.*

Distances from Itemset Supports

- *Definition*

Co-citation : A co-citation between two documents is the citing (or hypertext linking)of the two documents by another one.

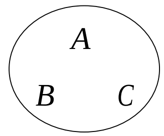
Co-citations reduce complex citation or hyperlink graphs to simple scalar similarities between documents or Web pages.

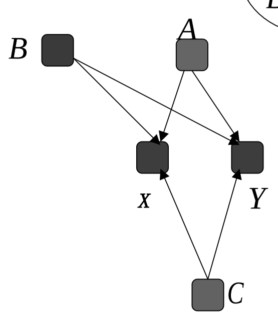


Co-citation(itemset support)

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

In the case of co-citations an associations is made between two documents according to the number of times they are co-referenced .





	X	Y
A	1	1
B	1	1
C	1	1

Association rule =3

Citation count(similarity)=2

Itemset support $\zeta(I)$

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

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

citing document(index row)

Single co-citations occurrences

Individual higher-order co-citation occurrences

The new distances we propose are thus a hybrid between standard pairwise distances and higher-order distances.

The itemset support feature summation is:

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

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

A nonlinear transformation $T[\zeta(I)]$ to be applied to the itemset supports $\zeta(I)$ before summation.

The transformation T is super-liner(asymptotically increasing more quickly than linearly),so as to favor large itemset supports

Itemset supports

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

A nonliner transformation

Reducing computational complexity for higher-order distances by exclude itemsets whose support < minsup.

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

normalized similarity

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

$$\hat{s}_{j,k} \in [0,1]$$

Standard clustering algorithm assume dissimilarities rather than similarities.

Dissimilarities (distance)

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

$$d_{j,k} \in [0,1]$$

Empirically,the transformation With p=4 usually results in the most frequent itemsets appearing together in clusters.

$$T(\zeta) = \zeta^{P} \qquad (5)$$

Frequently occurring

Itemset support

Hierarchical Clusters with Higher-Order Co-Citations

- Clustering is to form larger sets of documents that are more strongly associated with one another .
- Co-citation-based clustering provides a narrowing of search results.
- Reduce the manually reviewing large lists of search results.
- Co-citation analysis can broaden search results by providing alternative documents linked by co-citation.

Heuristics is a technique that gurantees an appropriate ,reasonable,acceptable solution but not the best one

Three important heuristics for clustering

- single-linkage----A method in linkage clustering where clusters are agglomerated based on their minimum distance set using the connectedness coefficient.
- average-linkage----based on a distance in between the minimum and maximum distance.
- complete-linkage---based on their maximum distance

These heuristics are agglomerative,at each step merging clusters that have the closest distance between them.

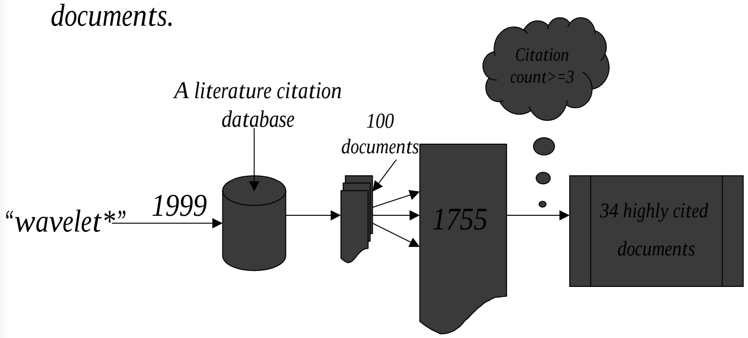
Dendrogram

1. *A tree visualization of a hierarchical clustering .*
2. *Leaves are individual documents*
3. *Non-leaf nodes represent the merging of two or more clusters .*
4. *A node is drawn as a horizontal line that spans over its children.*
5. *with the line drawn at the vertical position corresponding to the merge threshold distance.*

Demonstration

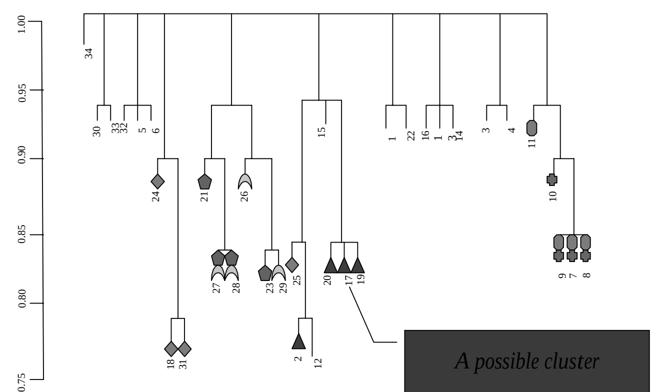
The demonstration employs data extracted from a literature citation database,the Institute for Scientific Information's Science Citation Index(SCI)

For the example---we do an SCI (Science Citation Index)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.*

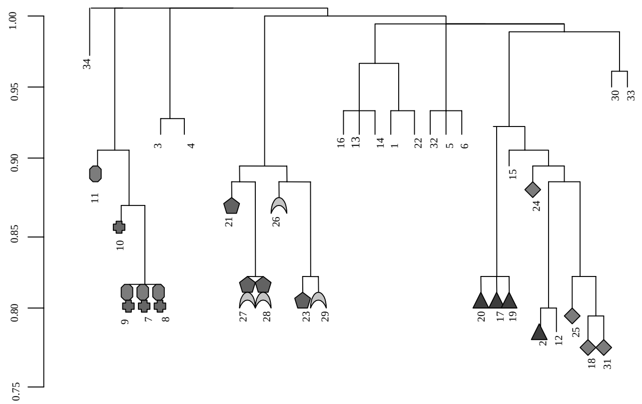


We then compute complete –linkage,average-linkage,and single-linkage clusters for the set of 34 highly cited documents.

In this example ,the most frequently occurring 4-itemset is {2,17,19,20} ▲



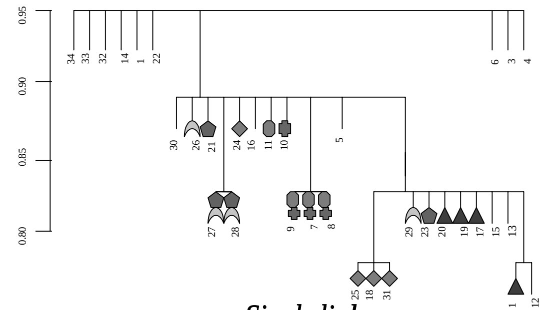
Complete linkage



Average linkage

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 .



Single linkage

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

As a comparison with standard pairwise distances,Figure 2 shows complete-linkage clusters computed with our hybrid pairwise/higher-order distance.

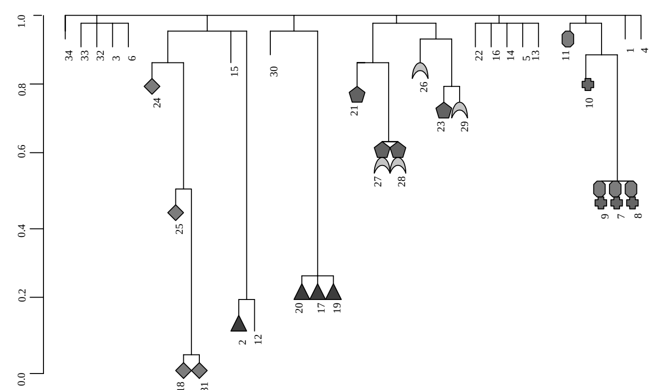
It considers three separate cases,each case being taken over multiple values of itemset cardinality X.

The three cases are x=2,3; x=2,3,4; x=3,4 Here the itemset supports

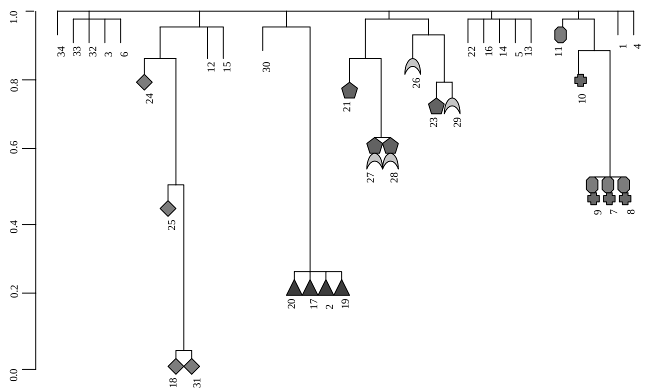
are nonlinearly transformed by $T[\zeta(I)]=[\zeta(I)]^4$, *with distances computed via (1),(2),(3) and(4).*

The most frequent itemset{2,17,19,20}form a cluster for the two cases x=2,3,4 and x=3,4.

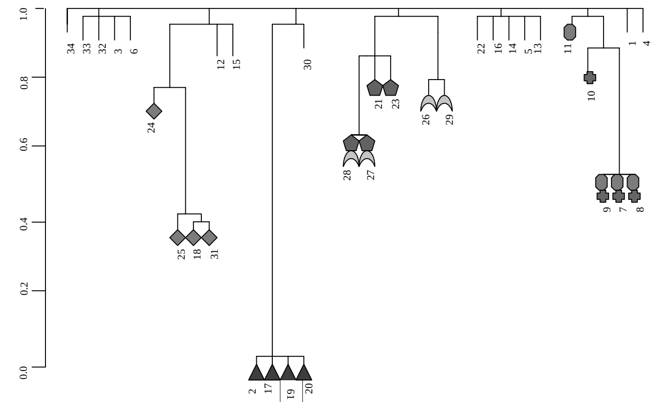
For the case=2,3 lower order supports are generally larger than high-order supports,and thus tend to dominate the summation(1).



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



(Cardinality 2)^4+(Cardinality 3)^4+(Cardinality4)^4



(Cardinality 3)^4+(Cardinality4)^4

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

Experimental Results

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

Our empirical tests apply a metric that compares clustering to frequent itemset,determining whether given itemsets form clusters comprised only of the itemset members.

$M(\pi, I_i)$ *is simply the portion of the cluster occupied by the itemset.*

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

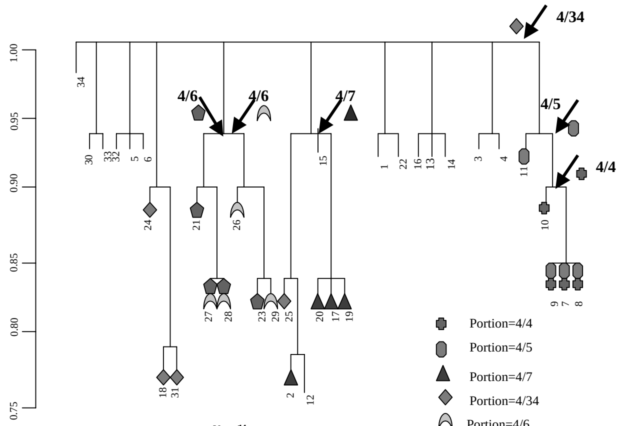
The metric $M(\pi, I)$ *is defined for a set of itemsets I by averaging* $M(\pi, I_i)$ *over* $I_i \in I$*, that is,*

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

A set of itemset (→ $|I_i|$)

A partition of items (→ $|\pi_j|$)

Figure 3 illusstrates the itemset-matching clustering metric $M(\pi, I)$.



$$\text{Metric} = \frac{\sum \frac{\text{itemset cardinality}}{\text{minimal cluster cardinality}}}{\text{number of itemsets}}$$

$$= \frac{4/4 + 4/5 + 4/7 + 4/34 + 4/6 + 4/6}{6} \approx 0.64$$

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

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]*

[3,4] ,[4,4] represent (Cardinality 3)^4+(Cardinality 4)^4

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

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*

- *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 member.*

  For the majority of the test case,metric values were higher for our hybrid distances,indicating better consistency clusters and frequent itemsets.

T he results show that excluding itemset supports below minsup generally has little effect on clustering results,particular for smaller values of minsup.

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

Date set	(minsup2)= (minsup 0)	(minsup2)> (minsup 0)	(minsup2)< (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

Summary and Conclusions

- *The hybrid distance are computationally feasible via fast algorithms for computing frequent itemsets.*
- *The hierarchical clustering dendrogram for association mining visualization enables quick comprehension of complex distance relationships among items.*
- *As a more basic contribution, this work represents a first step towards the unification of association mining and clustering visualization.*

Thank you!