Information Visualization (2003) 0, 000–000

www.palgrave-journals.com/ivs

Co-Citation Count vs Correlation For Influence Network Visualization

Steven Noel[1]
Chee-Hung Henry Chu[2]
Vijay Raghavan[2]

[1]*Center for Secure Information Systems, George Mason University, Fairfax, VA, U.S.A.;* [2]*Center for Advanced Computer Studies, The University of Louisiana at Lafayette, Lafayette, LA, U.S.A.*

Correspondence:
Henry Chu, Center for Advanced Computer Studies, The University of Louisiana at Lafayette, PO Box 44330, Lafayette, LA 70504-4330, U.S.A. Tel: +1 337 482 6309; Fax: +1 337 482 5791
E-mail: cice@cacs.louisiana.edu

Abstract
Visualization of author or document influence networks as a two-dimensional image can provide key insights into the direct influence of authors or documents on each other in a document collection. The influence network is constructed based on the minimum spanning tree, in which the nodes are documents and an edge is the most direct influence between two documents. Influence network visualizations have typically relied on co-citation correlation as a measure of document similarity. That is, the similarity between two documents is computed by correlating the sets of citations to each of the two documents. In a different line of research, co-citation count (the number of times two documents are jointly cited) has been applied as a document similarity measure. In this work, we demonstrate the impact of each of these similarity measures on the document influence network. We provide examples, and analyze the significance of the choice of similarity measure. We show that correlation-based visualizations exhibit chaining effects (low average vertex degree), a manifestation of multiple minor variations in document similarities. These minor similarity variations are absent in count-based visualizations. The result is that count-based influence network visualizations are more consistent with the intuitive expectation of authoritative documents being hubs that directly influence large numbers of documents.
Information Visualization (2003) **00,** 000–000. doi:10.1057/palgrave.ivs.9500049

Keywords: Document collection visualization; co-citation analysis; influence networks; minimum spanning tree; graph layout

Introduction

Visualization of document-similarity structure contributes much to the understanding of relationships among documents in a collection. Such visual representations are ideal for rapid assimilation of large-scale structure, and are complementary to lower-level textual descriptions. A number of key applications can benefit from document-similarity visualizations, such as various forms of information retrieval, or bibliographic analyses of scientific disciplines.

A major development in information science was Garfield's introduction of indexes of literature citations.[1] As citations have relatively clear semantics regarding literature influences, citation-based analysis avoids many of the difficulties inherent in language-based analysis. Early forms of citation-based analysis used bibliographic coupling as a measure of similarity between pairs of documents. The dual form (co-citation) introduced by Small[2] has become much more popular. That is, the similarity between two documents is the number of documents that cite *them* in common (co-citation), rather than the number of documents that *they themselves* cite in common (bibliographic coupling).

Received: 14 March 2003
Revised: 5 September 2003
Accepted: 6 September 2003

Early efforts in visualizing document co-citation similarity employed cluster visualization. A more recently developed alternative is the influence network,[3,4] a citation-based visualization that depicts the influences that authors or documents have on each other. The influence network builds the (generalized) minimum spanning tree among a set of documents, with (inverse of) co-citation serving as inter-document distance. The network shows the minimal set of essential links among documents, which are interpreted as the influences within the collection. Branches in the influence network correspond to bifurcations of ideas in the evolution of science, with highly influential documents appearing near the center of the network, and the emerging research front appearing on the fringes.

Among the various forms of co-citation analysis in the literature, two main methods have been described for computing co-citation-based inter-document distances. The first method, *co-citation count*, uses co-citation counts as inter-document similarity. The second method, *co-citation correlation*, uses the correlation of citation patterns as the inter-document similarity. For both methods, inter-document distances are just the inverse of inter-document similarities.

Later, we show that co-citation counts and co-citation correlations are directly related, that is, correlations are counts that have been transformed to z scores (zero mean, unit variance). However, this seemingly minor difference has a large impact on the structure of the corresponding influence networks. In particular, transformation to z scores results in much finer granularity in similarities. That is, it is common for document pairs with the same number of co-citations to have slightly different means and variances in their patterns of citation, resulting in slightly different similarities.

But more importantly, for raw co-citation counts, highly cited documents are usually strongly similar to many other documents. Through conversion to z scores in computing co-citation correlations, this similarity advantage for highly cited documents is removed. The result is that influence networks for co-citation counts tend to have certain frequently cited documents as highly influential hubs of influence, a structural feature that is generally lacking in correlation-based networks.

The main premise of this paper is that similarities computed from raw co-citation counts (*vs* co-citation correlations) yield influence networks that better capture the semantics of document collection influences. To support our position, we offer analytical arguments as well as empirical examples. While both co-citation counts and correlations have appeared in the literature as similarities in various forms of co-citation analysis, a thorough description of their differences has largely been lacking.

In the next section, we formally describe co-citation count and correlation. Further section then reviews the computation and visualization of influence networks. In the penultimation section, we analyze the impact of the choice of similarity measure on the structure of the influence networks. We also show empirical results of influence networks generated using each of these measures. In the last section, we summarize our work and draw conclusions.

Co-citation count and co-citation correlation

Scholarly publications are similar to other hyperlink systems such as the World Wide Web, in that the documents are linked to each other. A journal article is linked to another through a citation, while web pages are hyperlinked to each other. These systems can in general be modeled as directed graphs. A graph edge from one document to another indicates a link from first to second. An adjacency matrix that corresponds to the citation graph can be formed so that the rows and columns correspond to citing and cited documents, respectively. Thus in adjacency matrix $\mathbf{A}$, element $a_{i,j} = 1$ indicates that document i cites document j, and $a_{i,j} = 0$ indicates a lack of citation. In the case of web page networks, the adjacency matrix entries indicate the presence or absence of links between two pages.

A co-citation[2] between two documents is the citing (or hypertext linking) of the documents by a third one, as shown in Figure 1. A measure of similarity between a pair of documents can be defined as the number of documents that co-cite the pair. This is known as co-citation count. Taken over all pairs of documents, the co-citation count similarity serves as a compact representation of the citation graph structure.

In terms of the adjacency matrix $\mathbf{A}$, co-citation count is a scalar quantity computed for pairs of matrix columns (cited documents). From columns j and k, the co-citation count $c_{j,k}$ is then

$$c_{j,k} = \sum_i a_{i,j} \cdot a_{i,k} = \mathbf{a}_j \cdot \mathbf{a}_k, \qquad (1)$$

which is the j,kth element of $\mathbf{A}^T\mathbf{A}$. Here $\mathbf{a}_j$ and $\mathbf{a}_k$ are the column vectors of $\mathbf{A}$, i indexes rows, $\mathbf{A}^T$ is the transpose of $\mathbf{A}$, and $\mathbf{x} \cdot \mathbf{y}$ is the vector dot (inner) product. Note that the product $a_{i,j}a_{i,k}$ represents a single co-citation occurrence, which the summation counts. The co-citation count $c_{j,j}$ of a document with itself is simply a citation count, that is, the number of times the document has been cited. To simplify the notation, we write the citation count of document j as c_j instead of $c_{j,j}$.

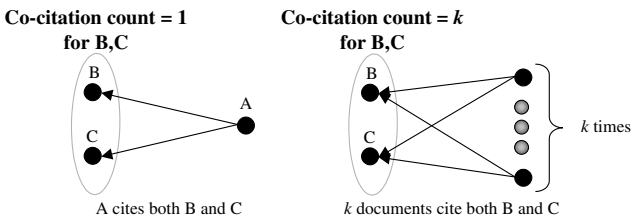


Figure 1 Co-citation count.

It is convenient to normalize the co-citation count $c_{j,k}$ through the linear transformation

$$\hat{c}_{j,k} = \frac{c_{j,k} - \min(c_{j,k})}{\max(c_{j,k}) - \min(c_{j,k})}, \tag{2}$$

yielding the normalized count $\hat{c}_{j,k} \in [0,1]$. Here $\min(\cdot)$ and $\max(\cdot)$ are the minimum and maximum functions, respectively. Standard clustering and minimum spanning tree algorithms assume dissimilarities (distances) rather than similarities.

One way to convert similarities to distances is through the nonlinear operation of multiplicative inversion. An inversion formula that avoids division by zero for normalized co-citation count $\hat{c}_{j,k} \in [0,1]$ is

$$d_{j,k} = \frac{1}{1 + \hat{c}_{j,k}}, \tag{3}$$

resulting in distance $d_{j,k}$ between documents j and k, normalized to $d_{j,k} \in [1/2, 1]$. Another way to convert normalized co-citation count $\hat{c}_{j,k}$ to distance $d_{j,k}$ is the linear transformation

$$d_{j,k} = 1 - \hat{c}_{j,k}. \tag{4}$$

In this case, $d_{j,k}$ is normalized to $d_{j,k} \in [0,1]$.

Another possible measure of co-citation similarity between two documents is the co-citation correlation. The most commonly applied is Pearson's product-moment correlation coefficient (or just Pearson's correlation). Pearson's correlation r for general variables x and y is defined as

$$\begin{aligned} r &= \frac{1}{n-1} \sum \left(\frac{x - \mu_x}{\sigma_x} \right) \left(\frac{y - \mu_y}{\sigma_y} \right) \\ &= \frac{\sum xy - (1/n)(\sum x)(\sum y)}{(n-1)\sigma_x \sigma_y}. \end{aligned} \tag{5}$$

Here μ_x and μ_y are means, σ_x and σ_y are standard deviations, and n is the number of observations.

Applying Equation (5) to the citation adjacency matrix $\mathbf{A}$, $x \equiv a_{i,j}$ and $y \equiv a_{i,k}$, so the correlation $r_{j,k}$ for columns (cited documents) j and k is

$$\begin{aligned} r_{j,k} &= \frac{\sum_i a_{i,j} a_{i,k} - (1/n)(\sum_i a_{i,j})(\sum_i a_{i,k})}{(n-1)\sigma_j \sigma_k} \\ &= \frac{c_{j,k} - (1/n)(\sum_i a_{i,j})(\sum_i a_{i,k})}{(n-1)\sigma_j \sigma_k}. \end{aligned} \tag{6}$$

Here $c_{j,k}$ is the co-citation count, and σ_j and σ_k are standard deviations for columns j and k. We see that Pearson's correlation for two columns of the adjacency matrix is simply the co-citation count with zero-mean unit-variance columns. Correlation is bipolar, that is, $r_{j,k} \in [-1, 1]$. As we seek a measure of similarity, we want the correlation absolute value. Correlation absolute value is then converted to a distance via either

$$d_{j,k} = \frac{1}{1 + |r_{j,k}|} \tag{7}$$

or

$$d_{j,k} = 1 - |r_{j,k}|. \tag{8}$$

An alternative is to invert the range of the correlation values, then scale and bias the range to [0,1]:

$$d_{j,k} = \tfrac{1}{2}(1 - r_{j,k}). \tag{9}$$

Let us examine an example computation of these co-citation (count and correlation)-based distances. We begin with a query to the Science Citation Index, keyword 'microtubules' for year 1999, and select the first 100 documents returned from the query. These 100 documents cite a total of 3070 unique documents. We retain only the more frequently cited documents, that is, those that have been cited six or more times. Figure 2 shows the resulting 100×45 citation matrix. In the figure, black indicates the presence of a citation and white indicates the lack of a citation.

Figure 3 shows co-citation counts and correlations for the microtubules data set. These are computed with Equations (4) and (8), respectively. The main diagonal of the correlations is unity; the main diagonal of the co-citation counts is the total number of times each document is cited. The correlation values are strongly biased toward positive values, although there are some negative values.

To better understand the differences in distance structure between co-citation count and co-citation correlation, we employ the dendrogram, a way of visualizing hierarchical clusters. The dendrogram visualization is a tree in which the leaves are individual documents, and non-leaves are the merging of two or more clusters at a given threshold distance value. The structure of hierarchical clusters yields additional insight into distance structure. For the dendrogram here, we use

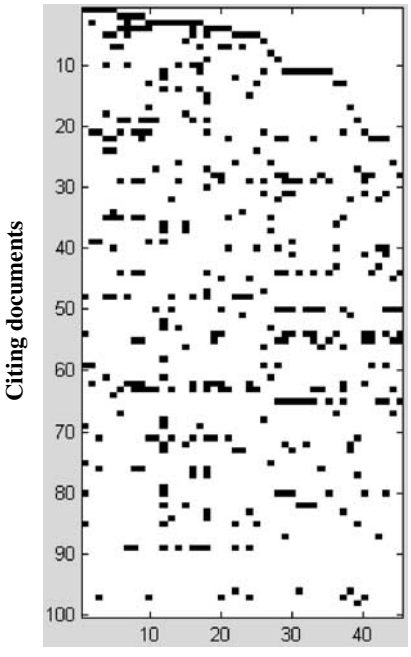
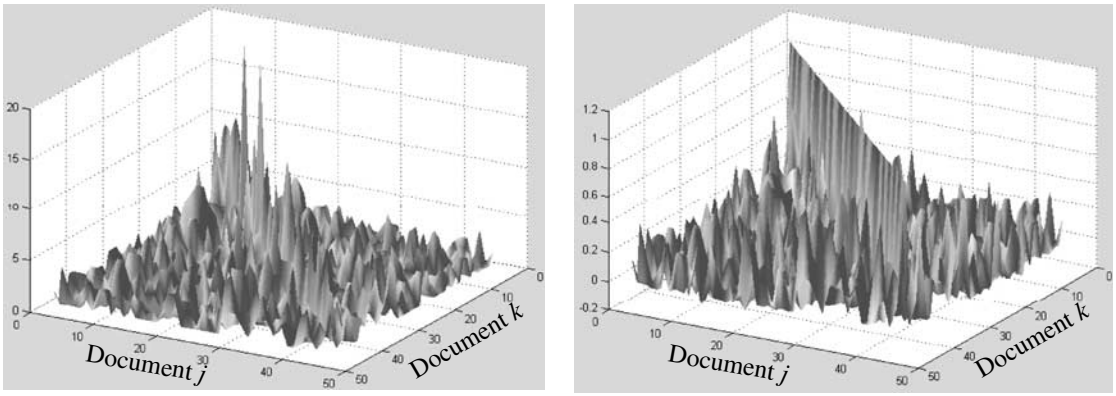
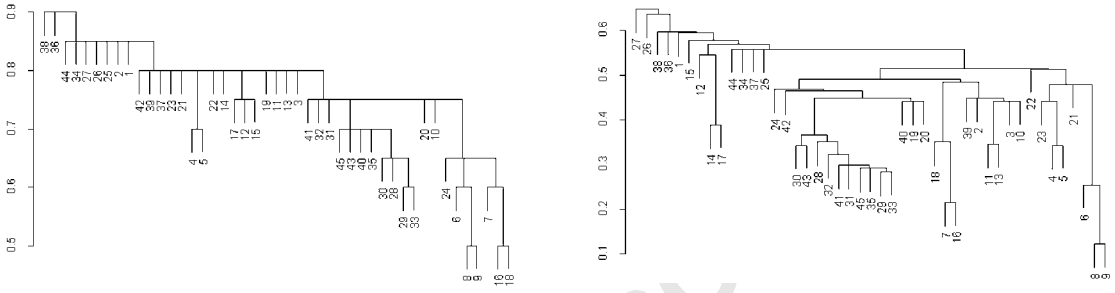


Figure 2 Citation matrix for SCI microtubules data set.



Co-citation count **Co-citation correlation**

Figure 3 Co-citation counts and correlations for SCI microtubules data set.



Co-citation count **Co-citation correlation**

Figure 4 Single-linkage dendrograms for co-citation count and correlation distances (microtubules data set).

the single-linkage clustering criterion,[5] which is closely related to the minimum spanning tree.

Figure 4 shows the dendrogram for the co-citation count and correlation data in Figure 3. There are general similarities in cluster membership between co-citation count and correlation. However, for count-based distances there is a tendency for multiple documents to merge into clusters at a common threshold value. This occurs because there are multiple document pairs with the same distance.

In contrast, for the correlation-based distances, the merges in Figure 3 are generally at all different threshold values. This is a consequence of the conversion to z scores (mean removal and variance normalization) of the citation matrix columns for correlations. As we show in the next section, these differences in distance structure greatly impact the structure of the minimum spanning tree, which in turn impacts the semantics of the document influence network.

Computation and visualization of influence networks

Chen[3] and Chen and Carr[4] have carried out significant work in applying Pathfinder network scaling to co-citation analysis. Pathfinder network scaling is a generalization of the minimum spanning tree, that is, the union of all possible minimum spanning trees. The interpretation is that Pathfinder network scaling provides a semantic network of intellectual influences among a collection of documents, as illustrated in Figure 5. The graph edges of the semantic network are considered to be essential links representing the most direct influences among documents. Documents near the center of the network are generally highly influential foundational works, while those near the perimeter represent the emerging research front.

Document distances based on co-citation (either count or correlation) can be modeled as fully connected, weighted, undirected graphs. We can define such a graph as $G = (V, E)$ with vertices V, edges E, and weight $w(u, v)$ of the edge $(u, v) \in E$, where $w(u, v)$ is the distance from document u to document v. A minimum spanning tree of G is then a set of edges $T \subseteq E$ that minimally connects all vertices V such that the sum of the edge weights

$$w(T) = \sum_{(u,v)\in T} w(u, v) \tag{10}$$

is minimized.

The most well-known algorithms for computing minimum spanning trees are Kruskal's algorithm and Prim's algorithm. Both can be easily implemented in $O(E \log V)$ time. In fact, Prim's algorithm can be made to run in

$O(E + V \log V)$ time. Thus for our completely connected distance graphs, in which $E = \Theta(V^2)$, Prim's algorithm becomes $O(V^2)$. For simplicity, we restrict subsequent discussion in this paper to (non-generalized) minimum spanning trees paper. However, our results can easily be extended to the generalized (Pathfinder network scaling) case.

It remains to put the minimum spanning trees in a form appropriate for visualization. Spatial coordinates must be induced for the minimum spanning tree vertices so that they can be placed on the visualization canvas. The only information at our disposal is the topology of the minimum spanning tree graph and the corresponding edge weights.

A classical method for visualizing distance data is multidimensional scaling. However, algorithms for multidimensional scaling tend to be slow, for example, $O(n^3)$ or $O(n^4)$. Also, most algorithms are non-iterative, that is, they must be restarted from the beginning each time, which can sometimes be a disadvantage. But more importantly, multidimensional scaling has been shown to perform poorly for the highly non-metric distances resulting from co-citations.[6]

As an alternative to multidimensional scaling, a frequently applied class of algorithms uses a model of forces and mechanical springs in placing nodes. Such spring models were developed initially for VLSI placement, and were popularized by the work of Eades.[7] But classical spring models attempt to minimize distance errors among all pairs of items, which perform poorly for significant deviations from Euclidean distances. We employ a type of spring model that abandons the idea of minimizing distance error over all pairs.[8,9] This heuristic has been shown to generate graph layouts that follow generally accepted aesthetic criteria, such as minimizing edge crossings, distributing vertices evenly, and reflecting underlying symmetries. Spring models generally have time complexity $\Theta(N^2)$, since each graph object interacts with every other one.

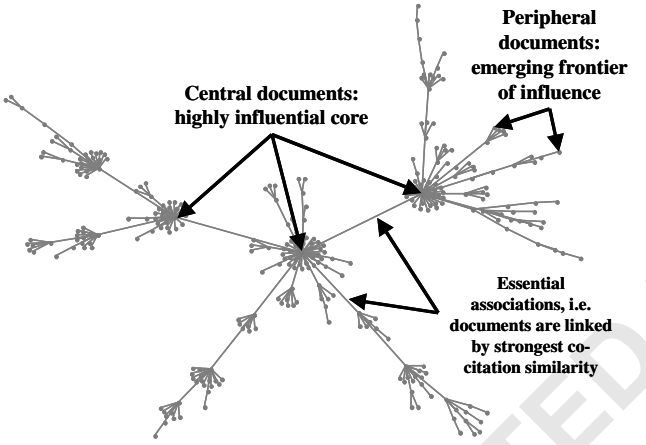


Figure 5 Minimum spanning tree visualization serves as network of document influences.

Interpretation and analysis of co-citation correlation

Correlations have been previously proposed[10,11] as measures of similarity between cited documents. The argument given in favor of correlations is to 'preserve patterns of co-citation over all [citing] documents.' In other words, correlation measures the similarity of cite/no-cite patterns for a pair of cited documents represented by their corresponding citation matrix columns. But in contrast to co-citation counts, there are some semantic as well as practical problems with co-citation correlations as document similarity measures.

For example, Figures 6–8 show results using the minimum spanning tree influence network described in the previous section. Here, we generate data sets by querying the Institute for Scientific Information's Science Citation Index (SCI) using the keyword 'Wavelets' for

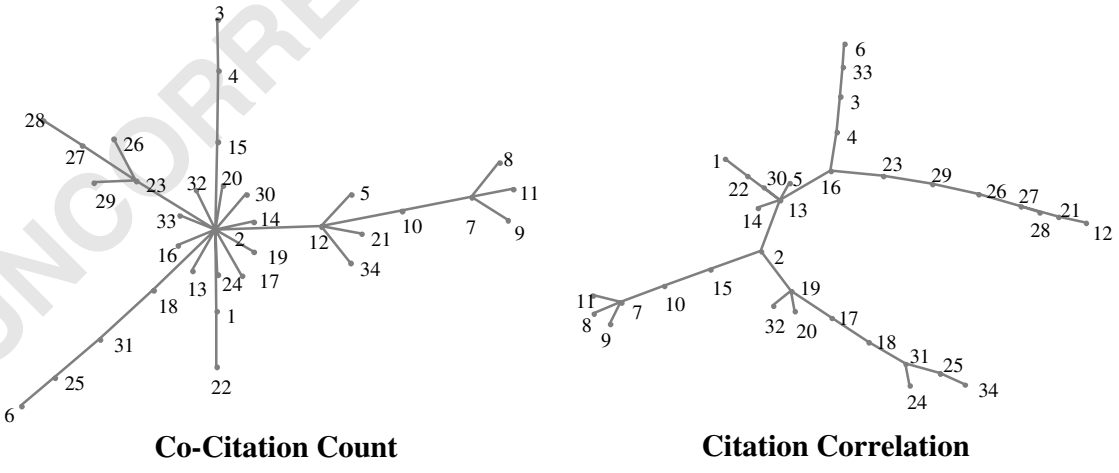


Figure 6 Minimum spanning tree placement for pairwise distances computed via co-citation count *vs* citation correlation, for data set 'Wavelets 1999 (1–100).'

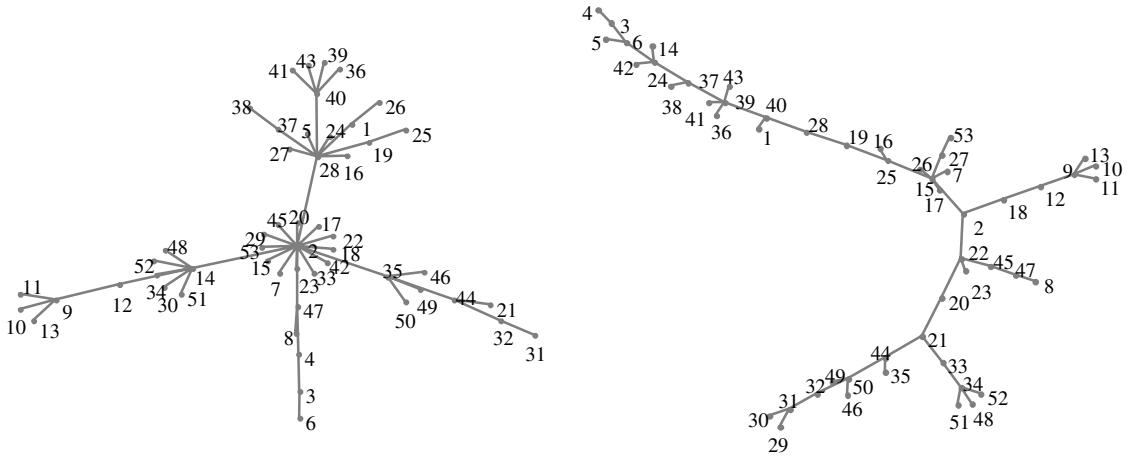


Figure 7 Minimum spanning tree placement for pairwise distances computed via co-citation count *vs* citation correlation, for data set 'Wavelets 1999 (1–150).'

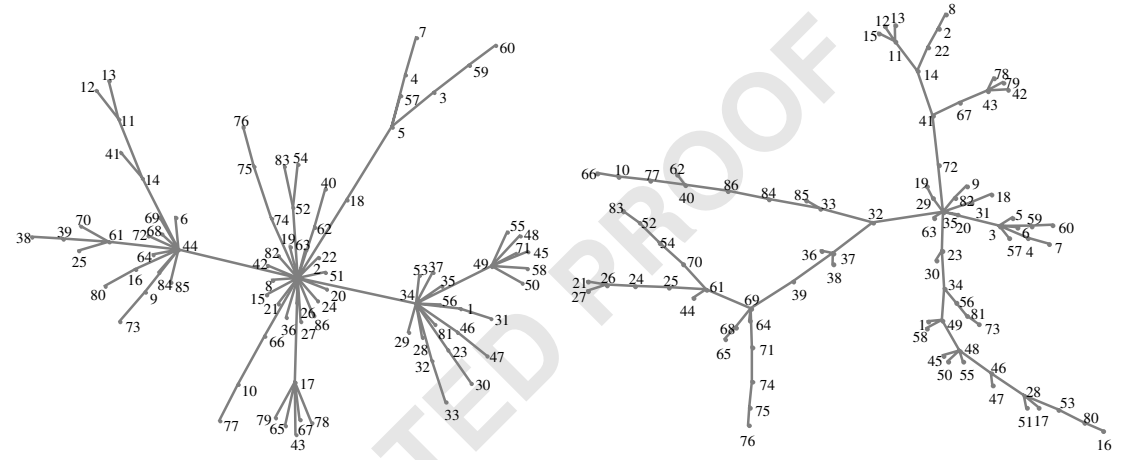


Figure 8 Minimum spanning tree placement for pairwise distances computed via co-citation count *vs* citation correlation, for data set 'Wavelets 1999 (1–200).'

documents from Year 1999. For the initial data set, we select the first 100 documents returned from the query as the citing documents. The set of unique documents cited by the initial set are subjected to a typical filtering operation in citation analysis: we retain only those cited documents that have been cited six or more times. The distances based on the co-citation counts and co-citation correlations of these cited documents are then used to build minimum spanning trees, resulting in the influence network visualizations in Figure 6. We repeat the process for larger sets by selecting the first 150 returned documents (Figure 7) and the first 200 returned documents (Figure 8).

As we described in the section on Co-citation count and co-citation correlation, document distances based on co-citation counts tend to have repeated values, while distances based on correlation do not. Thus in some sense, co-citation correlations have finer distance granularity. Because of this, in Figures 6–8, the influence networks for correlation-based distances tend to be more 'chained.' That is, the average degree of a minimum spanning tree vertex tends to be much lower for correlation-based distances. This causes edges that were incident on central vertices in count-based trees to be stretched out in chains over multiple vertices in the corresponding correlation-based trees. The remainder of this section analyzes this major structural difference and its semantic implications in more detail.

Correlation is one of several similarity measures that have been used to compare vectors in information

retrieval.[12] Other measures include: (i) the inner product, (ii) the cosine measure, which is the inner product normalized by the vector lengths, and (iii) the Dice measure, which is the inner product normalized by the sum of the vector elements. As mentioned in the Co-citation count and co-citation correlation section, correlation can be viewed as the inner product of two vectors that are normalized to have zero mean and unit variance. Measures that are not based on the inner product include the overlap measure and the χ^2 distance. But since the vectors in our context have only ones and zeros as elements, the overlap measure amounts to the inner product normalized by the shorter of the two vectors.

These similarity measures can be analyzed geometrically in terms of their angle, radial, and component-wise monotonicities, among other properties. In our context, the vector elements are zero or one so that the important property of a similarity measure is its angle monotonicity: whether the measure increases when the angle between the two vectors decrease. Co-citation count, as an inner product, has angle monotonicity. The correlation measure is not generally angle monotone, so that the correlation value of two vectors can decrease even if the angle between them is decreased.[12]

In[13] two requirements for an author co-citation similarity measure are described. First, the similarity between two authors should not decrease when additional authors are introduced into the group, if those new ones are not co-cited with the two authors being measured. Second, the order of the similarities between two pairs of authors should not switch when a group of new authors are included in the group, as long as those new authors are not co-cited with the two pairs of authors being measured. Unlike the χ^2 distance and the cosine measure, the correlation measure was shown to not satisfy either of these requirements.[13]

There is also a particular semantic problem with co-citation correlation in which the ones entries (corresponding to 'cite') of the adjacency matrix are treated the same as the 'no-cite' entries (zeros). In particular, if two columns (cited documents) are exactly the same, the correlation is unity, regardless of the actual number of ones *vs* zeros. For example, two columns with only a single one entry (in the same row) yield the same maximum correlation as two columns with all one entries. This disregards the actual number of co-citations. In short, correlations do not take into account the possibly different means and variances of the two columns, which in this case are important information.

An important point is that the ones in the citation matrix are not the same type of information as the zeros. Ones give an association between citing and cited papers. Zeros are not statements about a lack of association – they are more indicative of a lack of information about an association. So the presence of a one is much more important than the presence of a zero in terms of citation-based document similarity. Co-citation counts suit the task better since they measure only the simultaneous presence of ones.

From a practical standpoint, co-citation correlations are more computationally expensive than co-citation counts, since they involve removing means and normalizing variances. Count-based minimum spanning trees also make more efficient use of the placement area, 'filling' more of the plane in the sense of fractal dimension. Also, the vertices of minimum spanning trees resulting from co-citation correlations are usually more difficult to position for visualization. The finer distance granularity of co-citation correlations leads to trees with generally lower-degree vertices, resulting in larger numbers of local minima with respect to the graph layout algorithm. This problem tends to become more acute as the number of vertices increases.

From Equation (6), for citation adjacency matrix **A**, the co-citation correlation $r_{j,k}$ of columns (cited documents) j and k is

$$r_{j,k} = \frac{c_{j,k} - (1/n)(\sum_i a_{i,j})(\sum_i a_{i,k})}{(n-1)\sigma_j\sigma_k}, \tag{11}$$

where $c_{j,k}$ is the co-citation count, and σ_j and σ_k are standard deviations for columns j and k. Note that the summation terms in the numerator are the citation counts c_j and c_k. Since the term $a_{j,k}$ is either zero or one for all values of k, it can be shown that σ_j can be expressed in terms of c_j:

$$\sigma_j = \frac{\sqrt{c_j(n-c_j)}}{n}. \tag{12}$$

The correlation $r_{j,k}$ can therefore be written in terms of the co-citation and citation counts of the jth and the kth documents:

$$r_{j,k} = \frac{n^2 c_{j,k} - \frac{1}{n} c_j c_k}{(n-1)\sqrt{c_j c_k}\sqrt{(n-c_j)(n-c_k)}}. \tag{13}$$

We now use another example to illustrate the differences in influence network visualization that can result from using co-citation count *vs* co-citation correlation. Consider four documents A, B, C, and D with a citation adjacency matrix as shown in Figure 9, where the citing documents are sorted so that the shaded areas represent the one-entries, corresponding to 'cite' information. We can see from the citation adjacency matrix that A is cited by all documents that also cite B, C, and D so that we can consider A to be the authoritative document. We expect an influence map to show documents B, C, and D linked to A.

Next, we consider the co-citation count and the co-citation correlation to see which one results in a visualization that comes closer to our expectation. Let $n = 400$, $c_A = 200$, $c_B = c_C = c_D = 100$, $c_{BC} = 0$, $c_{BD} = 25$, and $c_{CD} = 75$. The normalized count and the correlation values are shown in Figure 10.

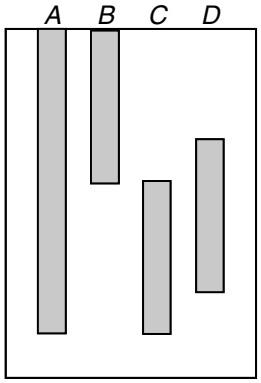


Figure 9 Citation adjacency matrix of the example case. The rows are sorted so that the shaded regions are the citation instances of documents A, B, C, and D.

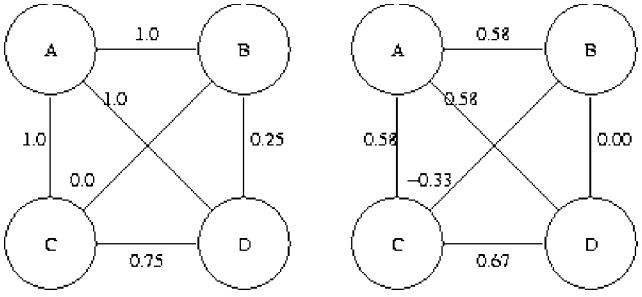


Figure 10 Co-citation normalized counts (left) and correlation values (right) for $n=400$, $c_A=200$, $c_B=c_C=c_D=100$, $c_{BC}=0$, $c_{BD}=25$, and $c_{CD}=75$.

The normalized counts can be converted to distances by Equation (3):

$$d = 1/(1+\hat{c}). \tag{14}$$

The distances and the resulting minimum spanning tree are shown in Figure 11. We can see that A is at the center and linked to documents B, C, and D.

The correlation values r can be converted to distances by Equation (8):

$$d^{(a)} = 1 - |r|. \tag{15}$$

The distances $d^{(a)}$ and the resulting minimum spanning tree are shown in Figure 12. A different, more semantically correct, conversion formula is by Equation (9):

$$d^{(s)} = (1-r)/2. \tag{16}$$

The distances $d^{(s)}$ and resulting minimum spanning tree using this conversion are shown in Figure 13. In both of the two minimum spanning trees, B and D are connected to A, while C is connected to D. This is similar to the chaining that we see in Figures 6–8. To see why chaining occurs in our example, note that from Figure 10, r_{CD} is greater than r_{AC}. The zero entries of C and D correlate well with each other and not with the one entries of A in the corresponding rows. This translates to

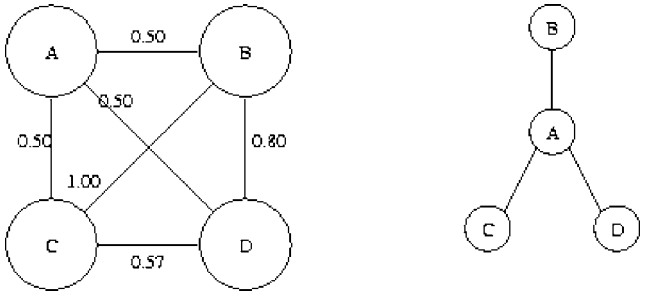


Figure 11 Distances converted from normalized counts (left) and the resulting minimum spanning tree (right).

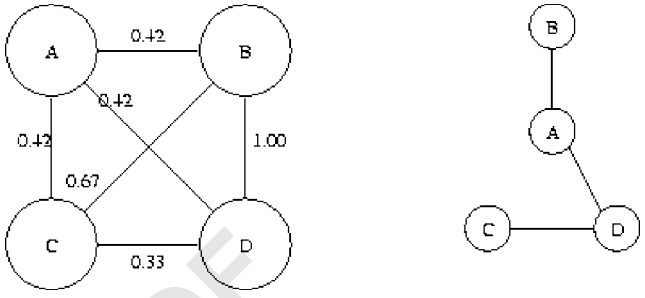


Figure 12 Distances converted from correlation values using $d=1-|r|$ (left) and the resulting minimum spanning tree (right).

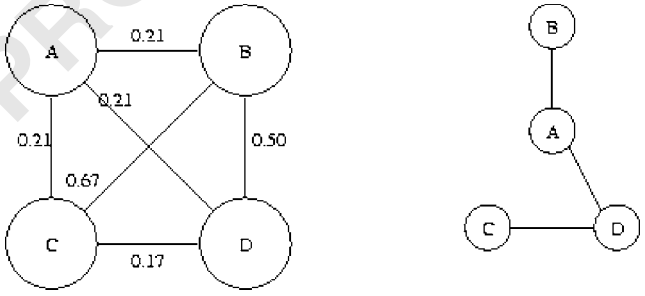


Figure 13 Distances converted from correlation values using $d=(1-r)/2$ (left) and the resulting minimum spanning tree (right).

a lower distance between C and D than that between C and A using either conversion formula. In the minimum spanning tree, D is connected to A, the authority. But since C is closer to D than it is to A, C is connected to D and not to A. We note that the two pairs (A, C) and (D, C) illustrate the angle non-monotonic property[12] of the correlation measure. The angle between A and C is smaller than that between D and C, and yet C is better correlated with D.

Although our numerical example shows only one case, the inequality $r_{CD}>r_{AC}$ holds for other cases. Let $c_A=\alpha n$, $c_C=c_D=\beta n$, and $c_{CD}=(1-\delta)\beta n$, where $\alpha>\beta$. This set of parameters corresponds to A as the authoritative document, while C and D overlap in being cited with each other. The factor $(1-\delta)$ is the fraction of overlap between

C and D in their citations. If δ is low, then C and D are co-cited most of the time. From Equation (13), we obtain the correlation values

$$r_{AC} = r_{AD} = \frac{n}{n-1}\sqrt{\frac{\beta}{\alpha}}\sqrt{\frac{1-\alpha}{1-\beta}},$$
$$r_{CD} = \frac{n}{n-1}\left(1 - \frac{\delta}{1-\beta}\right). \quad (17)$$

As long as $r_{CD} > r_{AC}$ holds, the distance between C and D is less than that between either C or D to A, and so A is not in the hub position of the visualization map. The inequality holds if

$$\delta < (1-\beta) - \sqrt{\frac{1-\alpha}{a}}\sqrt{(1-\beta)\beta} \quad (18)$$

If this upper bound to δ is small, the claim is that two closely related documents are linked to each other and not to an authoritative document in the influence map. One might argue that it is not unusual or overly counterintuitive. Nevertheless, Equation (18) shows when α is two to four times of β, regardless of the actual value of β, the bound on δ is between 0.3 and 0.5. This is not a very tight constraint so that even documents that are co-cited 50–70% of the times together can be better correlated with each other than with an authoritative document. In these cases, the influence network visualizations have the chaining effect that we see in Figures 6–8.

Early in this section, using example document collections, we show how co-citation count and co-citation correlation distances yield differing structures of document influence within the minimum spanning tree (Figures 6–8). We next do some formal analysis of some of the disadvantages of correlation as a co-citation distance measure. We now describe how count- and correlation-based co-citation distances impact the *semantics* (as opposed to structure) of influences in an example document collection.

In particular, we start with the minimum spanning trees for the document collection in Figure 8. For the tree computed from co-citation counts (left side of Figure 8), we select several documents that appear to be highly influential core documents within the collection. The selected documents are highlighted in Figure 14 (bibliographic details for these documents appear in Table 1). We then highlight the same selected documents in the tree computed from co-citation correlations (right side of Figure 8), yielding Figure 15.

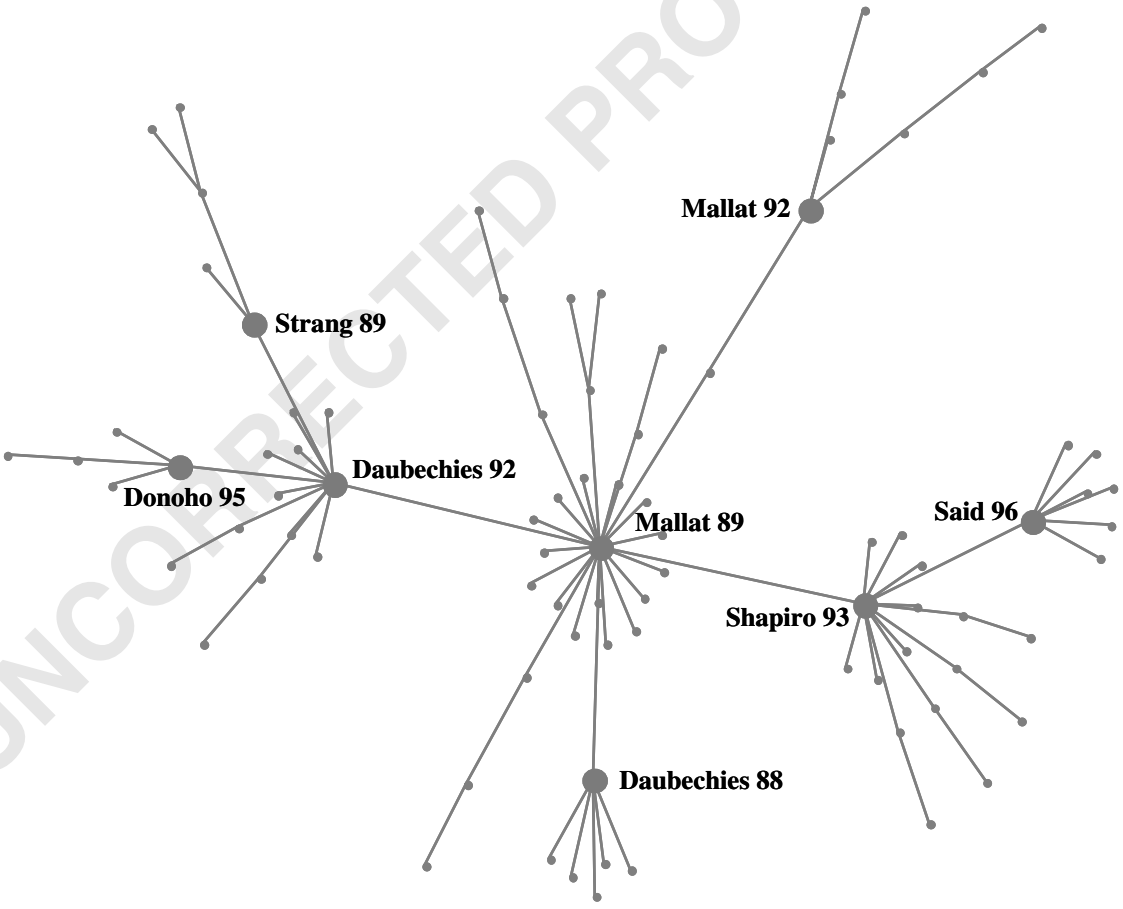


Figure 14 Using co-citation count, highly influential documents are prominent in minimum spanning tree.

The highlighted documents in Figure 14 are widely recognized as being highly influential within the wavelet community. For example, Ingrid Daubechies and Stephane Mallat have been called the 'mother and father of wavelets.' It is well known that Mallat's introduction of multiresolution analysis paved the way for wavelets in their modern form. (This was actually introduced in 1985. Mallat's own derivative works are apparently more highly cited, e.g. 'Mallat 89' in Figure 14). Daubechies later used Mallat's results to construct elegant orthonormal wavelet basis functions having compact support. The works of Mallat and Daubechies are the cornerstone of the highly efficient and widely used discrete wavelet transforms. It is safe to say that the remaining documents in the collection are essentially derived from the work of Daubechies and Mallat.

Table 1 Bibliographic information for highly influences documents in Wavelets 1999 (1–200) data set

Document	Details
Daubechies 88	DAUBECHIES I, 1988, COMMUN PUR APPL MATH, V41, P909
Daubechies 92	DAUBECHIES I, 1992 10 LECT WAVELETS
Donoho 95	DONOHO DL, 1995, J ROY STAT SOC B MET, V57, P301
Mallat 89	MALLAT SG, 1989 IEEE T PATTERN ANAL, V11, P674
Mallat 92	MALLAT S, 1992, IEEE T PATTERN ANAL, V14, P710
Said 96	SAID A, 1996, IEEE T CIRC SYST VID, V6, P243
Shapiro 93	SHAPIRO JM, 1993, IEEE T SIGNAL PROCES, V41, P3445
Strang 89	STRANG G, 1989, SIAM REV, V31, P614

The influence network of Figure 14 accurately captures the semantic relationships in the document collection. In particular, 'Mallat 89' is centrally located, and 'Daubechies 88' and 'Daubechies 92' are directly influenced by it. 'Shapiro 93' applies discrete two-dimensional wavelet transforms to image coding, and 'Said 96' refines Shapiro's method. 'Strang 89' was a popular and much more accessible description of wavelet fundamentals, and 'Donoho 95' applied wavelet transforms to statistical noise reduction.

Figure 15 uses co-citation correlations rather than co-citation counts. It is largely inaccurate as a semantic network of document influences. Indeed, it is essentially semantically opposite to the network in Figure 14. The

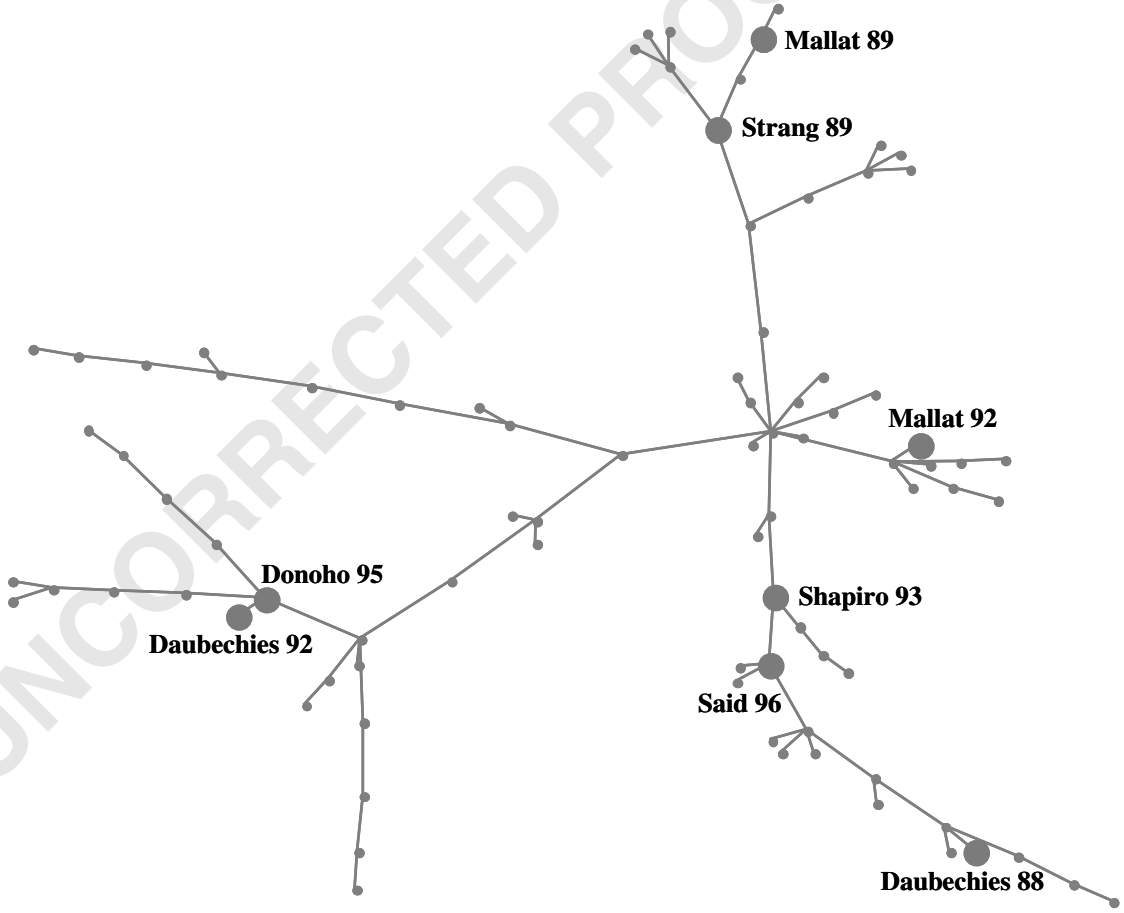


Figure 15 Using co-citation correlation, highly influential documents are no longer prominent.

Mallat and Daubechies publications now appear at the periphery rather than the center, and even the other frequently cited papers lack prominence within the network.

Summary and Conclusions

Influence network visualization provides much insight about a collection of documents based on the citation links among them. The appearance of such networks depends greatly on how two documents are judged to be similar. In this paper, the two candidates for computing such similarity are co-citation count and co-citation correlation.

We provide examples of influence network visualization for both similarity measures. We observe that the correlation-based visualizations tend to exhibit chaining effects. We analyze the co-citation correlation measure and illustrate the root cause of the chaining effects. We also examine the semantic implications for these two measures, that is, how well they capture the semantics of document collection influences.

In a document or author influence network visualization, one intuitively expects an authoritative document or author to be at the hub, with different groups of documents radiating outward from it. The count-based visualizations are much better matched to this expectation in comparison to correlation-based visualizations. A key observation is that while co-citation analysis in general offers a number of distinct advantages, one must be careful to match each aspect of the analysis to the underlying semantics of the problem domain.

References

1 Garfield E, Malin M, Small H. *Citation data as science indicators*. In: Elkana Y, Lederberg J, Merten R, Thackray A, Zuckerman H (Eds). Toward a Metric of Science: The Advent of Science Indicators. Wiley & Sons: New York, 1978; 179–207.
2 Small H. *Co-citation in the scientific literature: a new measure of the relationship between two documents*. Journal of the American Society of Information Science 1973; **24**: 265–269.
3 Chen C. *Visualising semantic spaces and author co-citation networks in digital libraries*. Information Processing & Management 1999; **35**: 401–420.
4 Chen C, Carr L. *Trailblazing the literature of hypertext: an author co-citation analysis (1989–1998)*. The 10th ACM Conference on Hypertext (Darmstadt, Germany, 1999); 51–60.
5 Gower J, Ross G. *Minimum spanning trees and single linkage cluster analysis*. Applied Statistics 1969; **18**: 54–64.
6 Noel S. *Data mining and visualization of reference associations: higher-order citation analysis*. Ph.D. dissertation, University of Louisiana, 2000.
7 Eades P. *A heuristic for graph drawing*. Congressus Numerantium 1984; **42**: 149–160.
8 Fruchterman T, Reingold E. *Graph drawing by force-directed placement*. Software – Practice and Experience 1991; **21**: 1129–1164.
9 Noel S, Chu C-H, Raghavan V. *Visualization of document co-citation counts*. The 6th International Conference on Information Visualization (London, U.K., 2002); 691–696.
10 McCain K. *Mapping authors in intellectual space: a technical overview*. Journal of the American Society for Information Science 1990; **41**: 433–443.
11 White H, McCain K. *Visualizing a discipline: an author co-citation analysis of information science, 1972–1995*. Journal of the American Society for Information Science 1998; **49**: 327–356.
12 Jones W, Furnas G. *Pictures of relevance: a geometric analysis of similarity measures*. Journal of the American Society for Information Science 1987; **38**: 420–442.
13 Ahlgren P, Jarneving B, Rousseau R. *Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient*. Journal of the American Society for Information Science and Technology 2003; **54**: 550–560.