
# Co-Citation Analysis: The Methodology of SciVal Spotlight

June 2009

www.scival.com

ELSEVIER

# Overview

To date, research output has been evaluated based on the journal in which it is published. Each scientific journal is classified into a major field despite the fact that journals are progressively covering a wider array of disciplines and paradigms (sub-topics) that are not properly reflected in their field classification. The nature of this "square peg in a round hole" system leaves critical interdisciplinary research overlooked and unaccounted for – allowing only a simplistic and reductionist view of current institutional research initiatives. SciVal Spotlight was developed to overcome this knowledge gap by providing a broader and deeper view of research performance.

The purpose of this document is to describe how SciVal Spotlight creates paradigms (also known as clusters) to reveal interdisciplinary research trends within institutions. The method, generally called **co-citation analysis,** involves classifying the scientific literature into 'natural categories' (small groups of papers that correspond to specific problems that researchers are addressing).

When developing SciVal Spotlight, the methodological choices centered around three questions: What are the rules of thumb that indicate a good classification system? Why use co-citation analysis (instead of some of the other approaches)? What are the advantages to the way SciVal Spotlight does co-citation analysis (versus the way others do it)? This paper provides a summary of the answers to these questions.

## What makes a good classification system?

**The fundamental task at hand is to create a classification system that will group articles in a way that reflects both the content and granularity of how scientific activity is currently organized. The implicit assumption that leads us to seek a 'natural' classification system rather than an imposed (or disciplinary) classification system, is that researchers self-organize around problems or topics that have an associated literature. These problems (sometimes called paradigms or clusters) are considered the fundamental unit of analysis by many people that study the sociology of science.**

SciVal Spotlight uses three rules of thumb to identify a good classification system:

1. The size of a cluster should be between 4 and 100 documents
2. The size distribution should be log rank - log size linear (this will be explained later in this paper)
3. Some papers should be allowed to be in more than one cluster

**Size**: Academics trained in the sociology of science have made the reasonable observation that the number of people working on a problem is usually quite small (perhaps a dozen people). Considering that not everyone publishes every year, there emerged a rule of thumb, articulated by Henry Small (considered the 'father' of co-citation analysis) that any cluster with more than 100 papers per year was suspect (Small, 1973). He developed this rule of thumb by examining the coherence of the papers in larger clusters.

While there is general consensus that the maximum size is around 100, there is far less consensus about what the minimum should be. This issue mostly hinges on individual preference. Our experience is that a cluster with less than approximately 5 papers/year should be aggregated up into the most related cluster. Very small clusters can give a false signal about the strength of a university; clusters with only a few paper are so small that they can be considered inconsequential.

**Size distribution:** One of the standard procedures in statistical analysis is to create a graph that represents a well known phenomenon and then apply transforms that make the graph to appear as a straight line. Transforms can be applied to the data or to the axes of the graph. Most graphs can be transformed to appear as a straight line. Deviations from this often imply that something is wrong with the data.

The graph on the left of *Figure 1* is what one normally sees when one generates a graph of cluster size (number of articles in a group) on the y axis and rank on the x axis (the largest group is ranked #1). One starts with a very large cluster. Size drops rapidly and then tails out with many small clusters.

Hundreds of examples from many fields of science suggest that the best transform for a size versus rank distribution is to transform both axes to logarithms (log). Most size vs. rank distributions will appear as a straight line on such a graph, as shown on the right side of *Figure 1*.

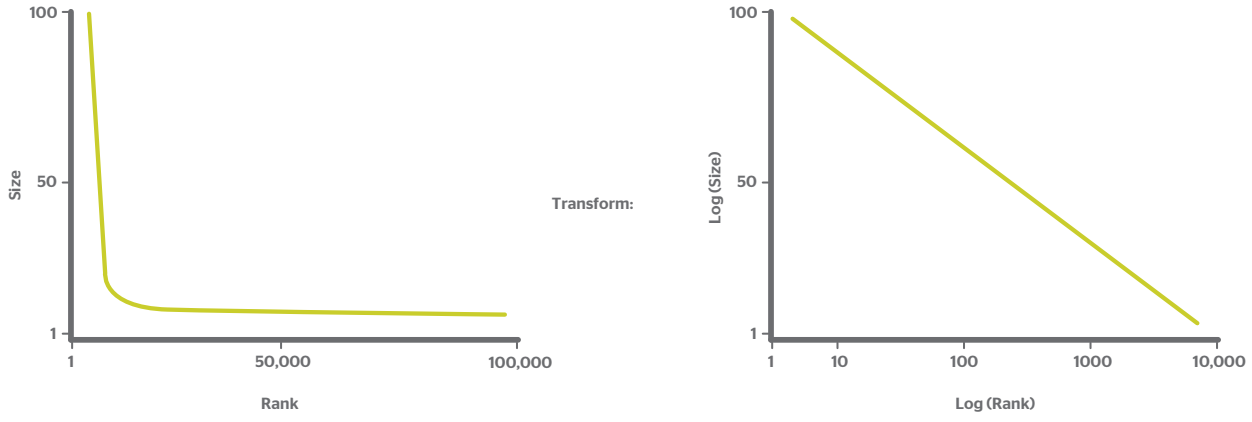
The bottom line for a size vs. rank distribution is this: a solution isn't 'good' unless the log size - log rank picture is close to linear. This appears to be true for a vast set of phenomena (such as the size of cities in the world or the number of times words occur in books). It's surprising how linear they all become when subjected to this log-log transform.

**Assigning a paper to more than one cluster:** Each cluster is a group of documents that are related to a focused problem in science. Some documents, however, link multiple clusters. To be accurate, a classification system needs to capture this fact, instead of assuming that a document is only reflecting the work in only one cluster. This common sense approach, assigning papers to more than one cluster where appropriate, is used in SciVal Spotlight, and enables us to calculate groups of clusters that are linked through common papers.

The ability to assign a paper to more than one cluster may sound obvious, but as the next section explains (the different ways that articles are clustered), this is generally ignored.

**Fig. 1**

The transform that makes a size distribution into a straight line.




'Taking the log' means that one morphs the picture (as if it is pushed down from the top and pushed in from the right) so that the 'distance' between the 1st observation and the 10th observation is the same as the distance between the 10th observation and the 100th observation. In this case we've labeled the 1, 10, 100, 1000, and 10,000th observations on the x axis. Notice that the distances between successive powers of 10 are the same.

**Summary: So far, there are three rules of thumb that guide SciVal Spotlight. First, make sure clusters are not too big or too small. Second, check if the distribution conforms to theory (gross deviations mean something may be wrong). And third, allow papers to be in multiple categories.**

## Why use co-citation analysis?

The two most common ways to cluster documents are to use **similarity** or **deconstruction and assignment.**

**Similarity:** The similarity approach calculates first-order or direct similarities between pairs of documents. If one is using common entries in bibliographies to compare two documents, this method is called bibliographic coupling.

The problem with bibliographic coupling is that it forces a paper to be in only one category. This is not a problem when doing a focused web search, and there is interest in only one category. But it can be an issue when reviewing a paper that intentionally references the literature from five categories. Review papers, or papers that show the link between multiple topics, are forced into one category by bibliographic coupling. Key information about the relationships between clusters are lost if papers are placed in only one cluster.

*Text* can also be used to determine the similarity between documents. Titles, abstracts, keywords, and even the full text of documents are often used to determine how similar two documents are. Using this approach, two documents are similar if they use the same words in similar proportions. For example, textual similarities are the basis of the approaches used by the 'related literature' features in PubMed and Google™ Scholar. The difficulties with the textual approach stem from the fact that language is naturally ambiguous, and also require extremely large amounts of computation time.

**Deconstruction and assignment:** The alternative approach to using direct similarity is to use a multi-step process in which:

1. elements associated with documents are identified
2. those elements are clustered using similarity,
3. the original documents are assigned to element clusters.

Co-citation analysis is an example of a deconstruction-assignment method that uses the references at the end of the document. First, references in a document are identified. Second, the relatedness between these references are calculated (how many times two references occurred in the same document). Third, the references are clustered using a transform of the co-occurrence matrix. And finally, the original documents are assigned to these reference clusters.

Co-word analysis is an example of a deconstruction method that focuses on the words in the document. The process used is the same for co-citation, but with words rather than references. The words or phrases that are important are identified and the relatedness between words are calculated (based on co-occurrence). Finally, the words are clustered and documents are assigned to these word clusters.

From one perspective, co-word and co-citation are similar. A citation is simply a very long string of linked words. A word (or phrase) is a very short string. Longer strings of words have more information embedded in them. Therefore, co-citation is going to capture more information because the document element (the reference) has more embedded information. However, there is additional information in words and phrases (as well as author names) that, if combined with references, should provide marginal improvements in accuracy.

Deconstruction and assignment approaches are not used if a quick answer is needed. They typically require more computation time than direct similarity methods. We believe, however, that they create a more accurate solution.

**Summary: A deconstruction-assignment approach must be used to assign a paper to more than one category. Because co-citation does this, it is far superior to bibliographic coupling. There may be better deconstruction-assignment approaches on the horizon (combining text, author names and references). This is an ongoing area of research.**

## What are the advantages of SciVal Spotlight's method of co-citation analysis?

This section details how SciVal Spotlight uses co-citation analysis, and the reasons behind selecting thresholds, clustering algorithms and assignment algorithms that are different than other researchers.

**Step 1: Selecting the corpus**

The tradition is to use all documents that are published in one year as the basis for an analysis. This allows one to see how clusters change year by year. It is possible, however, to select longer time periods (such as two or more years) that overlap. Few people have explored this possibility because of the increase in computational cost. A potential advantage to using a longer time window is increased stability – far fewer clusters will be born or die. However, an accompanying disadvantage would be a potential loss in early warning of trends due to smoothing of the data.

**Summary: SciVal Spotlight follows the tradition and uses one year of publications in order to pick up current trends.**

**Step 2: Selecting the references**
The number of references used in co-citation analysis has increased dramatically over the past 40 years. Years ago, a very high citation threshold was required to limit the reference sets to a size that could be managed using the computing resources of the time. Most studies done 10 or more years ago were limited to about 50,000 references. This created a major disciplinary bias – some disciplines were far more represented than others. To overcome this bias, researchers changed their methods to sample by discipline (about the top 1% of the most cited references by discipline). This increased the number of references to about 100,000 highly cited references.

Several years ago Klavans and Boyak published a study showing that disciplinary bias would decrease dramatically if the number of references was increased (Klavans & Boyack, 2006b). It was concluded that if one keeps about 1 million references or more, one can safely say that bias becomes unimportant, and that all disciplines are reasonably represented.

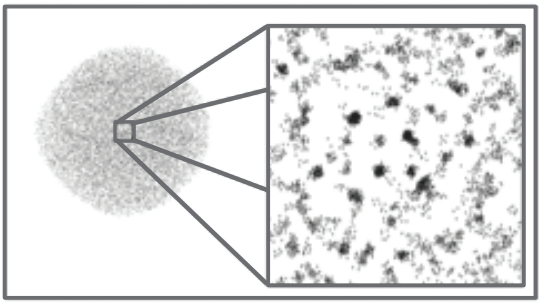
The thresholds used are even lower than those of a few years ago. SciVal Spotlight selects all reference papers that are cited 5 times or more. For recent papers (those published 3 years ago or less) we use an even lower threshold (2 to 4 citations, depending on age). The result of these low thresholds is that we select over 2 million of the most highly cited references to use in our co-citation model. All disciplines are well represented.

Increasing the number of references has a number of additional benefits. The more references that are used, the more accurately one can link current papers to clusters and tell that clusters over time are linked. If a current paper is only linked to a cluster by one reference, it is considered to be a very tenuous link – an ambiguous signal, so to speak. However, if a current paper is linked to a cluster by two or more references (the odds against matching two references are much higher than the odds of matching one), the assignment is unambiguous. This is one reason why a very large number of reference papers is preferred – most of our current paper assignments in SciVal Spotlight are unambiguous. Using lower numbers of reference papers increases the numbers of ambiguous paper assignment, resulting in a model with lower accuracy.

**Summary: The threshold SciVal Spotlight uses for selecting references has three benefits. It resolves the issue of disciplinary bias (all disciplines are well represented). It increases the number of current papers that can be unambiguously assigned to clusters. And it makes the linking of clusters over time more accurate (clusters are linked via the references they have in common).**

**Step 3: Calculating the relatedness between these references**
There are many different but commonly used ways to calculate paper-paper relatedness. Clearly, in order for a technique to provide valid results, it must be accurate. Over the years, a number of studies have been conducted to address the concern over accuracy by measureing and comparing the accuracy of different similarity measures and processes (Boyack, Klavans, & Börner, 2005; Klavans & Boyack, 2006a, 2006b). These studies indicated that the most accurate way to generate co-citation clusters was to use a modified cosine index based on co-citation counts for similarity, and to run the resulting matrix of cosine values through a visualization program which would assign each reference paper an (x,y) position on a 2-D plane. The visualization program is DrL (formerly known as VxOrd), a force-directed placement algorithm with edge cutting. An example layout is provided below.

**Fig. 2**
An example of the output from the visualization program. Each point is a reference paper. Points that are close to each other, and that are connected by an edge are highly related.



Interestingly, it was noted that the (x,y) positions and distances between papers on the "map" were a more accurate similarity metric than the original calculated similarities themselves. In effect, the visualization program acts as a filter that reduces the inherent noise in the system.

This measure has been further improved in the past 2 years. DrL uses a random walk routine; thus use of different starting conditions generates slightly different results. There is very little difference in overall accuracy between these different runs of the visualization algorithm. But there is a difference in the conclusion that references are 'close' or 'distant'. So we run the program 10 times and then take a consensus. This gives much more robust and accurate results than running the visualization algorithm only one time.

**Summary: The assumption is that better measures will result in better answers. The actual measures are more accurate than other measures proposed, as published in JASIST and Scientometrics. Additional details are available in the referenced papers.**

**Step 4: Clustering the references**

There are a large number of clustering algorithms that are in use today. They come in two standard varieties – supervised and unsupervised. Many commonly used algorithms are of the supervised type, meaning that they take a portion of data for which the groups and their characteristics are already known, "train" the algorithm using the known data, and then use that training to place new data in those existing groups. Supervised algorithms are not suitable for generating clusters representing a structure of science because (a) there are no "gold standards" (groups whose cluster assignments are already known) and (b) the face of science is always changing.

Unsupervised algorithms are those that do not use training data. Of these, there are algorithms (such as the common k-means approach) in which the user must specify the number of clusters a priori, and there are algorithms which allow the data to self-organize into emergent groupings. So-called agglomerative clustering techniques are of the latter type, and our approach is a variation of an agglomerative approach that is commonly called average-link clustering.

SciVal Spotlight's average-link clustering algorithm was tailored to work specifically with the (x,y) positional data output of the DrL layout algorithm. As shown in *Figure 2*, if the DrL output is plotted, the output looks like a set of clusters with white space between the clusters. The clusters are visually separated for the most part. Our average-link algorithm was designed to find these visual boundaries in the DrL output, and to assign references to the appropriate clusters.

There are two advantages to this approach. First, the number of clusters depends on the picture, not on an a priori decision. If one decides on the number of clusters a priori, the number will likely be too small or too large, and will result in clusters that should either be broken up or merged. Our clusters thus represent the emergent structure of science as defined by the reference sets, and not on any human decision.

Second, this method tends to result in clusters that range between 4 and 100 references. There are typically only a handful of clusters (out of tens of thousands) that have more than 100 references. SciVal Spotlight's clusters are thus within the size range that denotes a good classification system.

**Summary: We believe our cluster solutions are more accurate because (a) the underlying measures of relatedness are more accurate, (b) the two dimensional visualization is more accurate than the high dimensional data that creates the visualization, (c) the results can be easily validated and (d) the results provide more information about the actual shape of the cluster.**

**Step 5: Assigning current papers to these clusters**

Each current paper is assigned fractionally to clusters based on the references in the paper. In the 1980's, only 50-60% of the current papers could be assigned in this way because there were only 100,000 references to assign them to. This resulted in many ambiguous assignments (e.g. only one matching reference for a current paper).

Over 92% of the current papers *that have references*[1] are now assignable because of the significant increase in the number of references. The number of ambiguous assignments also drops dramatically with the increase in the number of references. For example, SciVal Spotlight does not need to assign a current paper to every ambiguous match. It drops out the ambiguous matches if there is an unambiguous match for that paper. This ensures that the fractional assignment of current papers better represents where the paper truly belongs. SciVal Spotlight also assigns the four previous years of current papers to the reference structure identified above. This allows one to identify publications trends associated with a specific set of highly cited references.

**Summary: Using the SciVal Spotlight model, more current papers are assigned and with higher confidence levels, because of the significant increase in the references that are included in the model.**

[1] Note that around 15% of current papers (types AR, CP, RE, SH, LE, NO) in Scopus do not have references (19.6% in 2003, 14.2% in 2007). We assume that these are citations that are imported from Medline; Medline does not have reference lists. None of these papers can be assigned in the current process. We will be recommending changes in the future to address this shortcoming.

## CONCLUSION

In designing the SciVal Spotlight methodology, Elsevier has striven to be true to two main principles: accuracy and transparency.

Some of the rules of thumb for identifying an accurate solution include cluster size, cluster distribution, and allowing papers to be assigned to multiple categories. Because a deconstruction-assignment approach must be used to assign a paper to more than one category, SciVal Spotlight uses co-citation analysis instead of bibliographic coupling or co-word analysis. The key differences in how SciVal Spotlight uses co-citation methodology include:

- Increasing the number of reference papers in the model
- Developing and validating a better measure of paper-paper relatedness
- Developing a more intuitive clustering algorithm for reference papers
- Increasing the number of assigned current papers to paradigms
- Improving the confidence level in assigning current papers to paradigms

These improvements to the methodology have been published in peer-reviewed literature, and rely on open source algorithms so that transparency is maintained. This is especially important because of this departure from the traditional way that co-citation analysis has been performed.

Future versions of SciVal Spotlight will continue this tradition of accuracy and transparency.

## REFERENCES

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. Scientometrics, 64(3), 351-374.

Klavans, R., & Boyack, K. W. (2006a). Identifying a better measure of relatedness for mapping science. Journal of the American Society for Information Science and Technology, 57(2), 251-263.

Klavans, R., & Boyack, K. W. (2006b). Quantitative evaluation of large maps of science. Scientometrics, 68(3), 475-499.

Small, Henry (1973), Co-citation in the scientific literature: A new measure of the relationship between two documents, Journal of the American Society for Information Science, 24(4), 265-269.


For more information about our SciVal suite of products, please contact your nearest Elsevier Regional Sales Office.

**North, Central and South America**

E-Customer Service Department
3251 Riverport Lane
Maryland Heights, MO 63043

**Email**: usinfo@funding.scival.com
**Tel**: 1 888 615 4500
(+1 314 523 4900, if calling from outside the USA and Canada)

**South America**

E-Customer Service
Rua Sete de Setembro, 111/16 Andar
Rio de Janeiro - RJ - 20050-006 - Brazil

**Tel**: +55 21 3970 9300
**Fax**: +55 21 2507 1991
**Email**: brinfo@spotlight.scival.com

**Europe, Middle East and Africa**

E-Customer Service
P.O. Box 211
1000 AE Amsterdam,
The Netherlands

**Tel**: +31 20 485 3767
**Fax**: +31 20 485 3432
**Email**: nlinfo@spotlight.scival.com

**Japan**

E-Customer Service
1-9-15 Higashi-Azabu, Minato-ku
Tokyo 106-0044, Japan

**Tel**: +81 3 5561 5034
**Fax**: +81 3 5561 5047
**Email**: jpinfo@spotlight.scival.com

**Asia and Australasia**

E-Customer Service
3 Killiney Road #08-01
Winsland House 1
Singapore 239519

**Tel**: +65 6 349 0222
**Fax**: +65 6 733 1050
**Email**: sginfo@spotlight.scival.com

www.scival.com



ELSEVIER