Using Reference Structures to Evaluate Co-word Structures: First Explorations

R.K. Buter and E.C.M. Noyons

buter@cwts.leidenuniv.nl, noyons@cwts.leidenuniv.nl
CWTS (Centre for Science and Technology Studies), Leiden University,
Wassenaarseweg 52, 2300RB Leiden (the Netherlands)

Abstract
We investigate the overlap between a variant on author-co-citation structure and a co-word structure, both created using hierarchical clustering. The overlap is a simple cross-tabulation of both structures. The rows or columns in this cross-tabulation may be sparse, showing a strong dominance of one cluster in another. But it may also be less sparse, showing overlap with most clusters of the other structure. A visualization of the cross-tabulation also reveals these structures.

Introduction

Bibliometric structures come in at least two important flavours. First, there are those built with the references between papers, like for example co-citation structures (Small, 1973) and author co-citation structures (White & Griffith, 1981). Second, there are those built with terms (words or phrases) from papers, which are commonly called co-word structures[1] (Callon, Law & Rip, 1986). The structures based on co-citation can be viewed as a representation of the use of the intellectual *base* of a field (Persson, 1994) or the "contemporary state of knowledge" (Small, 1999). Structures based on terms are said to represent themes within a science field (Callon, Law & Rip, 1986) or the "cognitive structure" of that field (Noyons, 1999).

Such bibliometric structures can serve as the basis of bibliometric maps: spatial representations created by putting similar sub-structures[2] closer together than those less similar. Such maps can, due to their analogy to geographic maps (Small, 1999) be called "landscapes of science" (Noyons, 1999).

Bibliometric structures and maps are tools in the analysis of a science field. For example, in White & McCain (1998), maps based on author-co-citation are used to analyze the field of 'Information Science'. And in Noyons, Luwel & Moed (1999) the field of 'Micro-Electronics' is analyzed and evaluated using co-word maps decorated with bibliometric indicators.

For the purpose of the analysis and especially the evaluation of a particular science field it is important that the generated structures are valid and credible: if the user does not believe the structure makes sense or is useful, there is no point in using it as an analysis tool. This was clearly put forward in Healey, Rothman & Hoch (1986).

But it can be hard to create such a credible and useful bibliometrics structure. For co-word structures, the vocabulary problem described in Furnas (1987) makes it particularly hard to assemble a set of words or phrases that is descriptive to all potential users of a map. And even if such a set of descriptive terms has been assembled, there is still the problem of selecting terms that are discriminative enough to present only a specific research theme within a field.

Since there are multiple (bibliometric) structures within a science field, we want to investigate if and how we might use the overlap of co-word and reference-based structures to evaluate and perhaps improve the first. For this purpose we use a variant of author-co-citation and cross-tabulate its overlap with a co-word structure. We should however note that the research described in this paper is on-going: therefore, the results presented in this paper have an explorative rather than a conclusive character.

Convergence of references and content

References and content share (at the very least) a similar purpose: both are carefully chosen or formulated by an author in order to establish a *communication* with a reader and to *persuade* that

[1] Somehow, even if the structure uses phrases, keywords or other kinds of labels instead of words.
[2] Depending on the type of main structure, sub-structures can be groups (clusters) of, for example, authors or terms, but also individual authors or terms.

Proceedings of ISSI 2005, pages 260–264.
Edited by P. Ingwersen and B. Larsen.

reader to keep on reading (Latour & Bastide, 1986; Law, 1986; Bazerman, 1981). Many studies have been published investigating how this takes shape in scientific papers, looking at the similarities and interactions between content and references. However, due to the nature of this paper we will only mention two.

In Braam, Moed & Van Raan, (1991) co-citation structures are analyzed using co-word maps. They conclude that the two structures only partly converge (overlap): the co-word structure reflects the subject matter in the current (or front of the) research, whereas the co-citation structure reflects the subject matter of the intellectual base of that research. This does not imply the two cannot be used together in a meaningful way; on the contrary, the two represent different structures yet they *also* overlap.

Additionally, in Peters, Braam & Van Raan (1995) the "word-profile similarity" between cited and citing publications is investigated. Their main conclusion is that there appears a "cognitive resemblance" to exist between citing and cited publications: papers cite papers with related content.

Database

We use the database of the field of *Bibliometrics and Scientometrics* described in Noyons (2004). The database includes all ISI-covered[3] references, also those made to papers outside the field. The familiarity of the field gives us the advantage of being able to evaluate the resulting structures ourselves, without having to consult external field experts.

Table 1: The number of papers and references in the whole database and in the selection used as the basis for the structures.

Set	**Size**	**%**
papers	1936	100
papers in 2002, 2003	556	29
references	16660	100
references from 2002, 2003	5951	35

Structures

Data and method

The basis for our structures is the set of papers published in 2002 and 2003 and all their references. Table 1 shows the number of papers in this selection.

We create two structures[4]: a reference-based one and a co-word structure. The one based on references is similar to author-co-citation, but uses an *author-organization-combination* (AOC) as introduced in Calero (2005) instead of (only) an author. The co-word structure uses Noun Phrases (NPs) extracted from titles and abstracts.

The elements of both structures are hierarchically clustered using *complete linkage* and the `hclust` method available in the R statistical environment (R, 2004).

Table 2 shows the number of papers covered by each structure, as well as their overlap.

[3] By this we mean a reference that is available as source paper in the database collected by the Institute for Scientific Information (ISI).

[4] We should note that our investigation used also a journal-to-journal citation based structure, which we will not present in this paper due to space constraints.

Table 2: The number of papers published in 2002, 2003 covered by the AOC and co-word structure and their overlap.

Structure	Size	%
AOC	212	38
co-word	284	51
AOC and co-word	148	27

AOC co-citation structure

An AOC is included if at least 6 references were made to papers associated with it, excluding self-citations[5]. These AOCs are clustered and, after investigation of the clustering structure cut into 6 clusters. We give tentative descriptions of the clusters in Table 3. Below, we will use the mnemonic labels to refer to the clusters.

Table 3: Mnemonic labels for the AOC clusters, a short description of their contents and the number of associated papers.

Label	Description	Size
bibl	bibliometrics, general research topics	113
info	information retrieval, visualization and metrics	120
lsi	Latent Semantic Indexing related research	38
web	web-o-metrics, electronic publication	20
misc	diverse subjects	21
ont	use of knowledge ontologies, DSS, Expert systems	7

Co-word structure

The co-word structure is created in two steps. First, a selection of descriptive terms is created, resulting in a set of 253 NPs. After that, the NPs are clustered, resulting in 11 clusters. Table 4 shows the clusters, the NPs associated with most papers and the mnemonic label.

Table 4: Mnemonic labels for the co-word clusters, the NPs associated with the most papers and the total number of associated papers

Label	NPs associated with the most papers in the cluster	Size
cit	citation analysis, co-citation	64
com	scientific communication, citation index	5
eval	bibliometric indicator, scientific performance	13
imp	bibliometric analysis, impact factor	47
info	information science	17
main	co-occurrence matrix, scholarly communication, bibliometric method	104
pat1	patent analysis, innovative activity	7
pat2	patent citation, knowledge flow	20
retr	information retrieval, latent semantic analysis	41
study	bibliometric study, research policy, international visibility	25
visu	information visualization, intellectual structure	35

Distribution of clusters over the co-word map

Figure 1 shows the cross-tabulation of the overlap between AOC co-citation clusters and co-word clusters. It shows the number of papers any two clusters have in common. The reader should disregard the column and row labelled "Entropy" for the moment: we will come back to this at the end of this section.

[5] Self-citations are calculated on the level of individual authors.

	bibl	info	lsi	web	ont	misc	*Entropy*
imp	26	15		1			1,21
eval	7	7					0,96
study	10	9		1	2		1,6
cit	20	24	1	6	4	2	2,69
main	32	30	3	7	4	1	2,72
retr	2	9	17				1,21
info	6	11	2	3			1,74
visu	6	10	2	1			1,4
pat1	4				2		0,98
pat2	7	3	1		11		1,63
Com	3	3	1				1,35
Entropy	5,14	4,84	2,23	2,14	2,23	0,92	

Figure 1: The number of papers AOC clusters and co-word clusters have in common and the entropy of the distribution of this overlap.

Looking at the "retr" row we see that the cluster shows a relatively high overlap of 17 out of a maximum of 38 papers (more on that below) with the "lsi" AOC cluster. This could be explained by noting that the 'Latent Semantic Indexing' research of the "lsi" cluster is used in 'Information Retrieval' ("retr") to extract descriptive terms from texts, but is little used elsewhere.
On the other hand, the co-word clusters "cit" and "main" with *all* AOC clusters. We will call this a *heterogeneous* convergence or overlap. This heterogeneity is also present in the AOC clusters "bibl" and "info", even more so than in any of the co-word clusters.

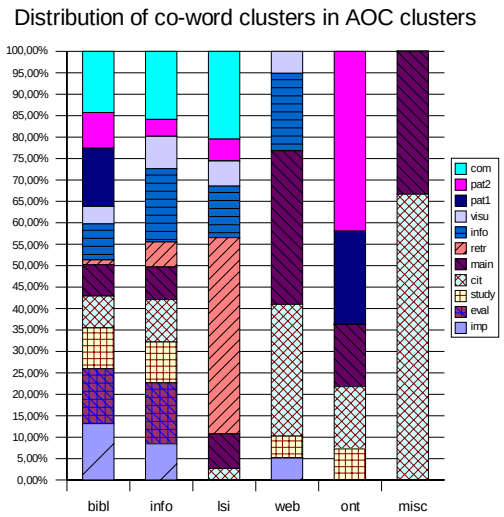

Figure 2a: A visualization of the distribution of co-word clusters in the AOC clusters

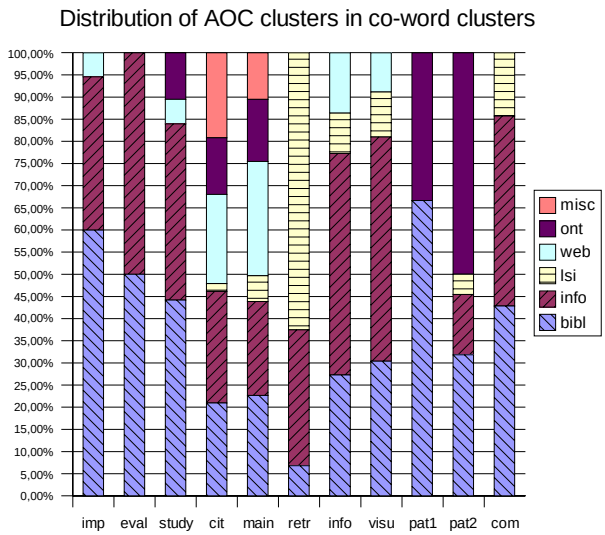

Figure 2b: A visualization of the distribution of AOC clusters in the co-word clusters

The cross-tabulation of Figure 1 is visualized in Figures 2a and 2b. These visualizations readily show the heterogeneity of the overlap between clusters. Again, note how overlap with the "lsi" AOC cluster dominates in the "retr" co-word cluster, visualized in the 6th bar in Figure 2b; and the "patchwork" of colours in the heterogeneous AOC clusters "bibl" and "info" and the heterogeneous co-word clusters "cit" and "main". But also note that the AOC clusters appear more "patched" than the co-word cluster does.

The visualizations of Figure 2a and 2b use normalized paper counts. To normalize a number of overlapping papers, we take the *minimum* of papers in either cluster, since this is the *maximum* number of papers the overlap can contain. For example, the co-word cluster "retr" contains 40 papers and the "lsi" cluster 38 (see Tables 3 and 4): so, their overlap of 17 papers is normalized relative to the 38 papers the individual clusters could have had in common.

To capture the heterogeneity of a particular cluster with a number, we calculate the *entropy*[6] of the distribution of clusters in a particular cluster: the higher this number, the more heterogeneous the overlap. The entropy in Figure 1 clearly correlates with our observations above: the "cit" and "main" clusters have the highest entropy of the co-word clusters and the "bibl" and "info" clusters have the highest among the AOC clusters. But the latter have much higher entropy, correlating with the difference in "patchwork" we also noted above.

Conclusions and further research

We have shown how a simple, straightforward way of cross-tabulating the overlap of different structures can reveal peculiarities in the overlap between these structures. But how should the particularities be interpreted? Are there specific NPs in the co-word structures responsible for the collation of many reference types in the "cit" and "main" clusters? And should the clustering of the AOCs be extended into more clusters, in order to distinguish other groups in the "bibl" and "info" clusters? Also, can we use the entropy measure to guide this process? If so, when can the entropy be regarded as too high? And finally, can these observations be generalized to cross-tabulations of other structures? Further investigation should give more definite answers to these questions.

References

Braam, R., Moed, H., & Van Raan, A.F.J. (1991). Mapping of Science by combined Co-Citation and Word Analysis I: Structural Aspects. *Journal of the American Society for Information Science*, 42(4), 233-251.

Bazerman, C. (1981). What written Knowledge does: Three Examples of Academic Discourse. *Philosophy of the Social Sciences*, 11(3), 361-387

Calero, C., Buter, R.K., Cabello Valdés, C., & Noyons, E.C.M. (2005). How to identify Research Groups using Publication Analysis: an Example in the Field of Nanotechnology. *to be published*.

Callon, M., Law, J., & Rip, A., (Eds.) (1986). *Mapping the Dynamics of Science and Technology, Sociology of Science in the real world*. The Macmillan Press Ltd., London.

Furnas, G.W., Landauer, T.K., Gomez, L.M. & Dumais, S.T. (1987). The Vocabulary Problem in Human System Communication: an Analysis and a Solution. *Communications of the ACM*, 30(11), 964-971

Healey, P., Rothman, H. & Hoch, P.K. (1986). An experiment in Science Mapping for Research Planning. *Research Policy*, 15, 233-251

Latour, B. & Bastide, F. (1986). Writing Science – Fact and Fiction. In Callon, Law & Rip (1986)

Law, J. (1986). The Heterogeneity of Texts. In Callon, Law & Rip (1986)

Noyons, E.C.M., Luwel, M. & Moed, H.F. (1999). Combining Mapping and Citation Analysis for Evaluative Bibliometric Purposes. *Journal of the American Society for Information Science*, 50, 115-131

Noyons, E.C.M. (1999). *Bibliometric Mapping as a Science Policy and Research Management Tool*. PhD Thesis. Centre for Science and Technology Studies, Leiden University. Leiden: DSWO Press.

Noyons, E.C.M. (2004). Science maps within a science policy context. In Moed, H., Glänzel, W., & Schmoch, U. (Eds.). *Handbook of Quantitative Science and Technology Research, The use of Publication and Patent Statistics in Studies of S&T Systems*, (pp. 237-255). Kluwer Academic Publishers, Dordrecht.

Persson, O. (1994). The Intellectual Base and Research Fronts of JASIS 1986-1990. *Journal of the American Society for Information Science*, 45, 31-38.

Peters, H., Braam, R., & Van Raan, A.F.J. (1995). Cognitive resemblance and citations relations in chemical engineering publications. *Journal of the American Society for Information Science*, 46(1), 9-21.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

Small, H. (1999). Visualizing Science by Citation Mapping. *Journal of the American Society for Information Science*, 50(9), 799-813

Small, H. (1973). Co-Citation in Scientific Literature: A New Measure of the Relation between Publications. *Journal of the American Society for Information Science,* 24, 265-269.

White, H. & Griffith, B.C. (1981). Author co-citation: A literature measure of Intellectual Structure. *Journal of the American Society for Information Science*, 32, 163-272

White, H. & McCain, K. (1998). Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995. *Journal of the American Society for Information Science*, 49(4), 327-355.

[6] The entropy is calculated by $-\sum p\log(p)$, using the normalized paper count as an estimate for p.