Jointly published by Akadémiai Kiadó, Budapest
and Springer, Dordrecht

Scientometrics, Vol. 69, No. 2 (2006) 227–258

Mapping the structure of science through usage

JOHAN BOLLEN, HERBERT VAN DE SOMPEL

Research Library, Los Alamos National Laboratory, Los Alamos, NM (USA)

Science has traditionally been mapped on the basis of authorship and citation data. Due to publication and citation delays such data represents the structure of science as it existed in the past. We propose to map science by proxy of journal relationships derived from usage data to determine research trends as they presently occur. This mapping is performed by applying a principal components analysis superimposed with a k-means cluster analysis on networks of journal relationships derived from a large set of article usage data collected for the Los Alamos National Laboratory research community. Results indicate that meaningful maps of the interests of a local scientific community can be derived from usage data. Subject groupings in the mappings corresponds to Thomson ISI subject categories. A comparison to maps resulting from the analysis of 2003 Thomson ISI *Journal Citation Report* data reveals interesting differences between the features of local usage and global citation data.

Introduction

Science is an essential component of our globalized, technological civilization. From a social and political perspective it is therefore vital to build an understanding of its properties as a dynamic, social process. The scientific literature is rife with structural analysis and visualizations of the structure of science (CHEN & PAUL, 2001; NAGPAUL, 2002; BOYACK, 2004). The data sets used in such mappings rely mostly on citation

Received December 8, 2005

Address for correspondence:
JOHAN BOLLEN
Research Library, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA
E-mail: jbollen@lanl.gov

0138–9130/US $ 20.00

data (LEYDESDORFF, 2004a, 2004b), co-citation data (SMALL, 1973; MCCAIN, 1991; BRAAM et al., 1991a, 1991b) and co-authorship data (WAGNER & LEYDESDORFF, 2003; HE & SPINK, 2002; LIU et al., 2004; LIU et al., 2005; NEWMAN, 2001a, 2001b) which serve as proxies to the underlying social phenomena (EVERETT & PECOTICH, 1991). However, this focus on citation and authorship data has a number of limitations related to the nature of the publishing process:

Publication and citation delay. Most peer-reviewed articles are published and cited well after they are written and submitted (LUWEL & MOED, 1998; RINIA et al., 2001). Citation and authorship data will therefore reflect past scholarly trends. Although ADAMS (2005) shows how future citation rates can be predicted from early citation data, the problem remains that in fast changing domains, such as genomics and biochemistry, publication and citation delays will hamper our ability to study scholarly trends as they occur.

Citation bias. Citation and co-authorship are public phenomena and therefore subject to strong social desirability biases (NEDERHOF, 1985; KING & BRUNER, 2000), for example the perceived need to cite popular articles and co-author with prestigious team leaders. Such biases may obfuscate important, but implicit trends in the scientific community.

Process bias. Science is an iterative process. A publication is but the end-result or by-product of this process. Well before publication takes place researchers analyze the relevant literature, perform the actual research, distribute preliminary results to colleagues, etc. Examining the structure of science solely on the basis of publication data will provide only a partial picture of this process.

Granularity. Each phase of the scientific process produces valuable scholarly results, e.g. raw data files, analysis software, technical reports and literature reviews. By studying the structure of science on the basis of publication data only, we ignore the many other products of scientific activity.

There exist, however, a number of interesting precursors and proxies to publication data. It is commonly known that reading, publication and citation rates are related in the scientific process. In particular it has been found that reading rates predict publication and citation rates (BRODY & HARNAD, 2004). What’s being read today can therefore serve as an indication of what will be cited tomorrow.

Unfortunately, readership data is difficult to obtain and validate; how can we know whether someone has indeed read a paper? For that reason most investigations of reader-related scholarly phenomena focus their efforts on the analysis of the more general class of *usage* data. The term usage refers to a class of information consumption and interest indicators recorded in the framework of digital information services which includes, but is not limited to, downloading the full-text version of a document, requesting bibliographic data, accessing a service pertaining to a particular document, etc. Each instance of usage data can with a varying degree of reliability be interpreted as an indication of user interest, and thus as an indicator of on-going scholarly work on the corresponding subject.

Figure 1 shows an overview of which aspects of the scientific process usage data can capture (EGGHE & ROUSSEAU, 2000; WOUTERS, 1997). Since usage data can be recorded in real-time, it allows us to study all facets of the scientific process well before its results are manifested in the publication record. Usage data can furthermore relate to a wider range of scholarly communication items then publication data, e.g. raw data sets and multimedia documents (VAN DE SOMPEL et al., 2004).

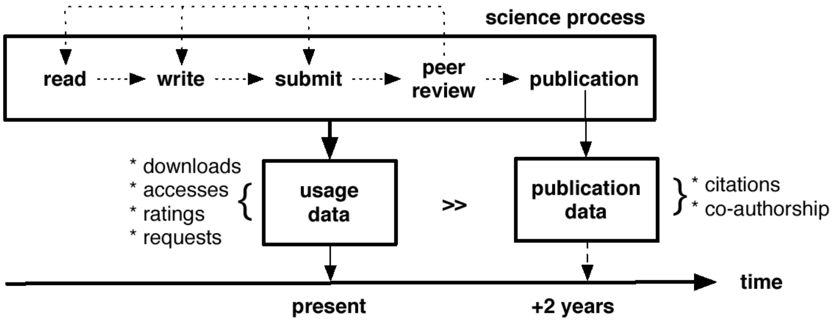


Figure 1. Usage data represents scientific trends as they occur today

It is for those reasons that the analysis of usage data has emerged in efforts to detect early scientific trends. BOLLEN & LUCE (2002) and BOLLEN et al. (2005) discuss methods to derive metrics of journal and article impact from recorded digital library usage data. KURTZ et al. (2004a, 2004b) study the temporal relation between reading and citation rates and formulate a model for the prediction of researcher productivity from present readership rates.

Such analysis of usage data reveals scientific trends as they occur in the present but are limited to one-dimensional rankings; they do not produce the structural maps of science we commonly find in bibliometrics. As an example of the latter, BOYACK (2004) produces a geographical mapping of journals using the VxInsight knowledge visualization tool (BOYACK et al., 2002) and uses k-means clustering to generate subject groupings. Such visual mappings are compelling instruments to study the structure of science, but they are entirely lacking in the domain of bibliometric usage studies.

As a proof of concept, following a similar methodology as BOYACK et al. (2002), we propose to create maps of science on the basis of usage data. A directed, weighted network of journal relationships was constructed from user access data recorded at the Los Alamos National Laboratory (LANL). The resulting journal relationships were mapped in a 2-dimensional plane according to the results of a Principal Components Analysis (PCA). A k-means clustering was used to determine journal subject groupings which are overlayed with the PCA map to visualize the relations of subject domains in LANL usage.

Methodology

Our objective is to demonstrate that usage data can be validly used to map and visualize the structure of science, much like the mappings that have previously been performed on the basis of citation data. We do so by the following three-phase methodology:

Creation of journal usage networks. Directed, weighted journal networks are created from usage data recorded at the LANL research library.

PCA mapping of journal relations. A PCA analysis reduces the dimensionality of journal relationships to a set of 2-dimensional map positions. Results are validated by a visual inspection of the resulting journal positions, an analysis of the generated principal components and the degree to which a subsequent k-means clustering overlaps with the produced visual grouping of journals.

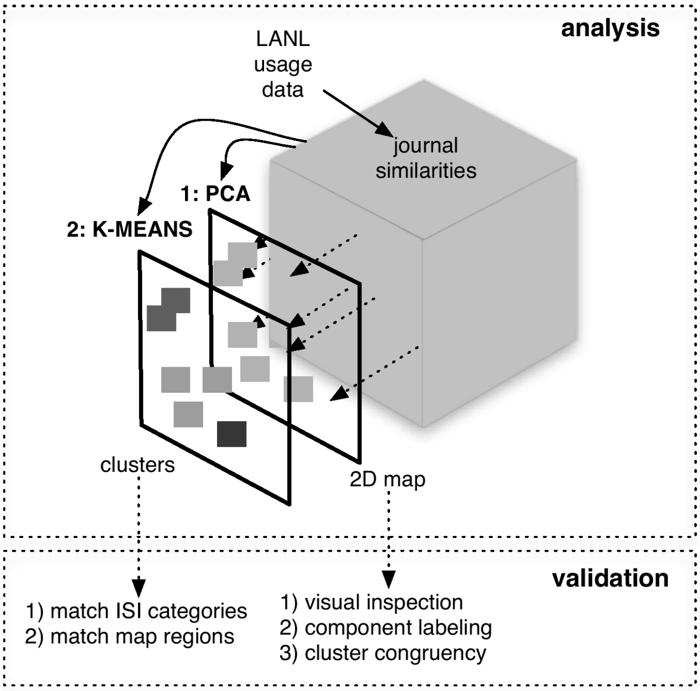


Figure 2. Mapping methodology using Principal Component Analysis overlayed with k-means cluster analysis

Subject domain grouping. A k-means cluster analysis groups journals according to their subject domains and is overlayed with the generated PCA map. The generated clusters are validated by a χ^2 analysis of how well they match Thomson ISI subject categories.

An overview of this procedure is shown in Figure 2. We further explain the details of this methodology below.

Creation of usage networks

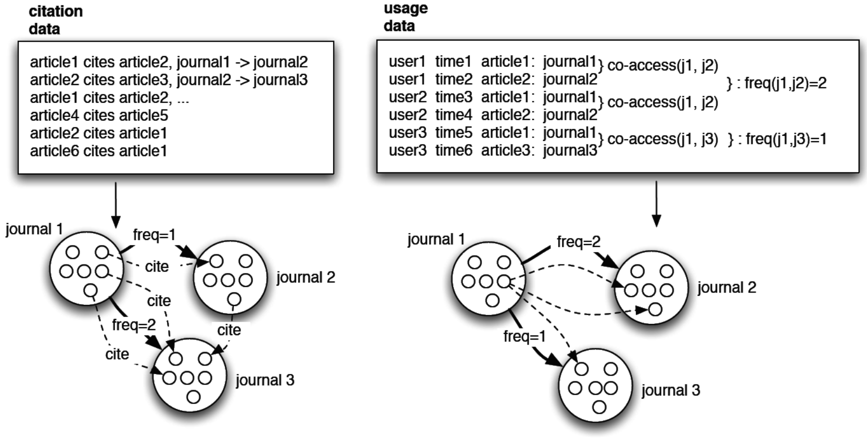
We extracted a network of journal relationships from usage data recorded at the LANL Research Library (RL) in the period February 2004 to April 2005. This particular data set reflected the usage of 5,866 library users who accessed services pertaining to 330,109 articles published in a set of 10,696 journals. The total number of recorded user accesses was 392,455. This data set will be referred to as LANL04 for brevity.

Certain challenges are associated with applications of usage data in particular with regards to validity and user privacy. To prevent the biases and distortions often found in web server logs, the LANL04 data was recorded by the LANL RL linking server which connects users to library services and can thus centrally record usage across a wide variety of different services (VAN DE SOMPEL & BEIT-ARIE, 2001). In addition, to maintain the privacy of LANL users, all user IDs were replaced by a unique, but anonymous numerical code. Finally, to capture the widest possible range of expressions of user interest, our usage data pertained to a range of usage types including requests for full bibliographic information, full-text downloads, and requests for additional services relating to a particular document as recognized by the linking server (VAN DE SOMPEL & BEIT-ARIE, 2001). An individual instance of such usage, i.e. a triplet formed by a user, document and recorded time of usage, will be referred to as an “access”. A set of such accesses can be generally referred to as “usage”.

Journal relationships were generated from the LANL04 data by a methodology outlined in BOLLEN et al., (2005). The central assumption of this methodology is that when users frequently access a pair of journal articles in a given sequence this indicates the degree to which the articles, and consequently their journals, are related. The sequence of accesses in log data induces a directionality of relationships which relates to the degree to which an access of one article, or journal, is contingent upon the other. The analogy to citation data is that rather than to assume two journals are related because their articles cite each other frequently, we assume that two journals are related if their articles are frequently co-accessed in a particular sequence. Figure 3 provides a graphical overview of this process.

Deriving item relationships from access or usage patterns is common in commercial recommender systems, e.g. amazon.com which track purchases and extract product relationships from “co-purchases” (AGRAWAL et al., 1993). Such systems often apply a technique referred to as item-based collaborative filtering (SARWAR et al., 2001; KIM et al.,

2004) which in addition to association rule learning has been applied to web usage data mining (BRIN et al., 1997; SPILIOPOULOU, 1998; CHAN, 1999; SRIVASTAVA et al., 2000; MOBASHER et al., 2001). These approaches have been strongly validated in the literature as well as in commercial applications. Although the results outlined in this paper will provide support for the validity of our methodology to derive journal graphs from usage data, we further refer the reader to the cross-validation undertaken in (BOLLEN et al., 2005).



```
article1 cites article2, journal1 -> journal2
article2 cites article3, journal2 -> journal3
article1 cites article2, ...
article4 cites article5
article2 cites article1
article6 cites article1
```

```
user1  time1  article1:  journal1} co-access(j1, j2)
user1  time2  article2:  journal2                      } : freq(j1,j2)=2
user2  time3  article1:  journal1} co-access(j1, j2)
user2  time4  article2:  journal2
user3  time5  article1:  journal1} co-access(j1, j3)  } : freq(j1,j3)=1
user3  time6  article3:  journal3
```

Figure 3. Extracting journal relationship data from digital library access logs

A directed network of journal relationships was extracted from LANL04 data as described above. The resulting journal graph was represented by the matrix R which contained 33,256 non-zero entries. The density of the reader-based journal relationship graph was very low: only 3 out of 10,000 possible edges had non-zero weights. The directionality of the network, and thus the usage it was derived from, is indicated by the low correlation coefficient between opposite edge weights, namely Pearson's $r = 0.45$, $p<0.001$.

Mapping of journal relationships

A mapping of the journal relationships in our usage-derived journal networks was achieved by means of a Principal Component Analysis (PCA). PCA and Multi-Dimensional Scaling (MDS) are well-established in scientometric research (EVERETT & PECOTICH, 1991) and have a rich tradition in cognitive science and the social sciences (JOLLIFFE, 2002). Whereas MDS is intended to produce a graphical map corresponding to the similarities among a set of items, PCA is used to extract the factors or components that best explain the variation between the items in a collection. It allows highly dimensional data sets to be reduced to a two or three dimensional representation based on its strongest components thereby rendering it amenable to 2D or 3D plotting. In addition, the extracted components can be interpreted in terms of how they organize the original data set.

PCA mapping. It is our objective to map each journal in the LANL04 data set to positions in a 2-dimensional plane, so that similar journals are spatially close and the plane's dimensions correspond to the factors which account for the largest degree of variation in the original data set. The similarity of each pair of journals v_i and v_j was defined as the Spearman rank-order correlation coefficient ρ of their row vectors in matrix R. The Spearman rank-order correlation coefficient was chosen as a more robust, non-parametric alternative to Pearson's r in light of the derivation of matrix R from usage data. A correlation matrix R_c was subsequently created so that each of its entries $r_c(i,j) \in [-1,1]$ corresponded to the similarity of journals v_i and v_j as given by their rank-order correlation coefficient $\rho(i,j)$. This similarity principle is related to the notions of co-citation and bibliographic coupling (SMALL, 1973; KESSLER, 1963).

A PCA was then performed on the matrix R_c by performing an eigenvector analysis. The two most significant components, i.e. eigenvectors with highest eigenvalues, were retained. The original journal row vectors were then projected upon these components to reduce the original $n \times n$ journal similarity matrix to a $n \times 2$ matrix denoted R'_c thus mapping each journal v_i to set of (x,y) coordinates. As such each journal could be mapped to a particular position in a Euclidean plane spanned by the correlation matrix's 2 primary eigenvectors.

PCA validation. The resulting PCA map was then validated as follows:
Visual inspection. We visually investigated whether the 2D positioning of journals was meaningful.
Journal PCA density contour plot. We visualized journal density values in the PCA plot by means of a density contour plot to determine in which regions of the map journals were clustered most densely.
Factor labeling. We examined journals scoring highly on either one of the 2 principal components to determine a semantic interpretation of the PCA mapping.
K-means cluster overlap. We determined whether journals within the same k-means generated clusters were positioned in each other's proximity.

K-means clustering

The generated PCA map functions as a geographical map of how journals are related according to LANL usage. To augment the map we performed a k-means clustering of journals (SPAETH, 1980) which delineated map regions containing highly related journals. K-means clustering consists of a greedy-algorithm which iteratively assigns items to a pre-determined number of clusters to optimize both inter-cluster distance and intra-cluster cohesion. It therefore belongs to a class of unsupervized clustering algorithms which includes Kohonen self-organizing maps (KOHONEN, 1995) and automated probabilistic classifiers such as decision-tree learners (TUFEKCI, 1993).

K-means cluster method. The k-means cluster analysis was performed on the journal correlation matrix R_c resulting in the assignment of each journal in the LANL04 usage networks to a particular cluster. Each journal was automatically assigned a cluster code indicating the cluster it had been assigned to, e.g. cluster 1 → "x". These codes were overlayed on top of the 2D PCA mapping of journals so that each journal had both an (x,y) coordinate in the plane and a cluster code corresponding to its cluster assignment. To demarcate the regions occupied by a particular cluster a convex hull was defined which followed the outer edges of each cluster in the PCA defined plane. A cubic spline interpolation function rounded the generated regions.

The k-means cluster analysis was applied to the full, unreduced set of journal correlations, represented by matrix R_c. Cluster regions were thus defined in the original n dimensional journal space. As a consequence, when projected onto the 2D PCA map, cluster regions could visually overlap.

Cluster interpretation and validation. To interprete and validate the k-means clustering of journals, we extracted Thomson ISI journal subject categories[*] for all journals assigned to a particular cluster following BOYACK et al., (2005). The distribution of journal subject categories could then point to an interpretation of the cluster's subject domain. The generated clusters were interpreted and validated as follows:

Category TFIDF weighting. How strongly do particular Thomson ISI journal subject categories uniquely occur within a particular cluster?
Category distribution entropy. How sharply focused or diffuse are category weights within a cluster?
χ^2 analysis. How strongly do cluster assignments correlate with Thomson ISI journal subject categories?

These phases are discussed in more detail below.

Not all Thomson ISI subject categories are equally relevant to the interpretation of a particular cluster's subject domain. Some occur more frequently in general than others and will be highly frequent within any given cluster; they are therefore poor indicators of an individual cluster's subject domain. We therefore need to balance how frequently a particular category code occurs within a given cluster, i.e. its intra-cluster frequency, with how frequently it occurs over all clusters, i.e. its between-cluster frequency.

[*] Although it is only one among many subject classification systems and a particularly coarse one, the ISI category system has a number of advantages: categories are manually vetted and constitute a commonly accepted standard.

Following the principle of TFIDF index term weighting in Information Retrieval (IR) (SALTON, 1988; BAEZA-YATES & RIBERIO-NETO, 1999) we define a category *i*'s weight for a given cluster *j*, denoted *w*(*i*,*j*), as the ratio of the category's within-cluster frequency and its overall between-cluster frequency as follows:

$$w(i, j) = \frac{f(i, j)}{n(j)} \times \log\left(\frac{N_a}{n_c(i)}\right)$$

where *f*(*i*,*j*) represents the category's raw within-cluster frequency, *n*(*j*) the total number of categories within a cluster *j*, N_a represents the total number of clusters, and $n_c(i)$ the total number of clusters that carry the specific category, i.e. its between-cluster frequency. The resulting category weight values are thus the product of two factors; a category's normalized within-cluster frequency *f*(*i*,*j*)/*n*(*j*) vs. its normalized, inverse between-cluster frequency, log ($N_a/n_c(i)$). This weighting scheme reduces the importance of frequent, non-cluster specific categories and increases the significance of highly cluster-specific categories.

Once we have established category weight values, the "sharpness" of their distribution within a cluster can provide additional information on the cluster's degree of domain focus. If all categories within a cluster have equal weights, the cluster's subject domain is ill-defined. If only few categories have high weights, the cluster's subject domain is well-defined. Following the seminal definition of entropy by SHANNON (1948) we define the entropy of a cluster's category weight distribution *H*(*j*) as follows:

$$H(j) = -\sum_i w(i, j) \log_2(w(i, j))$$

The entropy of a cluster's category distribution can be interpreted in a manner similar to the traditional definition of entropy in information theory: if all categories are equally strongly weighted within a particular cluster, its entropy will be high indicating a low degree of subject domain focus. Vice versa, if a cluster's category weights are highly unequally distributed, its entropy will be low indicating the cluster's domain focus is high.

Finally, we used a χ^2 analysis to determine how well the generated k-means cluster assignments match Thomson's ISI subject categories. This serves as a cross-validation of the generated clusters: does a clustering of journals according to LANL local usage match the Thomson's ISI subject categories?

Results

Figures 4 and 5 shows the results of the above described PCA and k-means analysis for the LANL04 data.

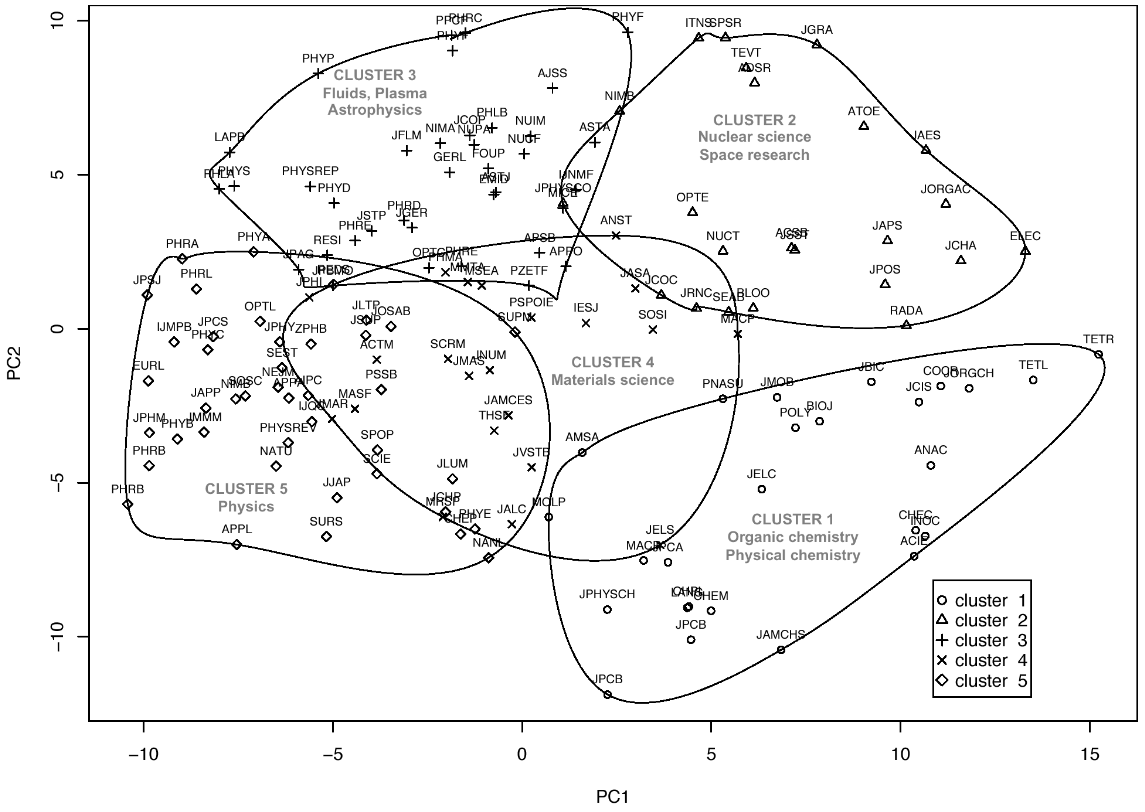


Figure 4. PCA and k-means mapping of LANL04 data

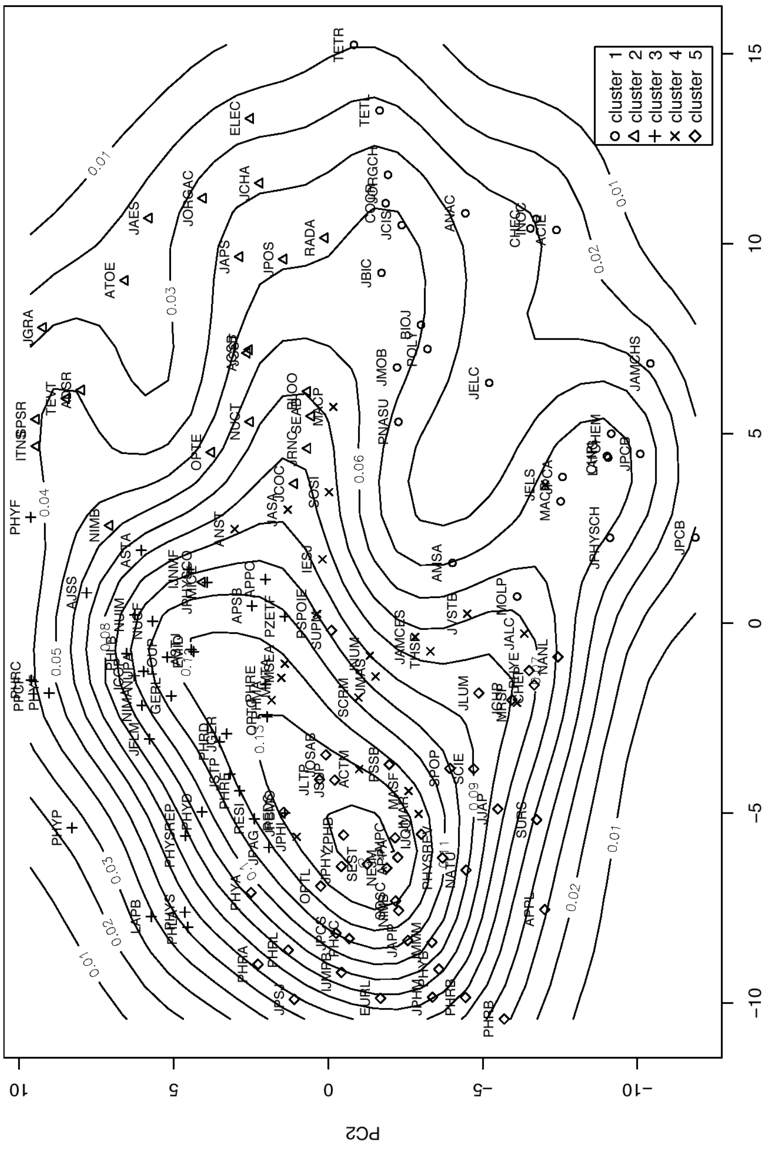


Figure 5. Contour map of spatial journal placement density in LANL04 PCA analysis

Journal names are abbreviated to a four letter code to reduce clutter in the graph. Abbreviations can be looked up in Table 7 in the Appendix. To further increase readability the map includes only the 150 most used journals. Each cluster has been labeled with the subject domains derived later in this section. The PCA map and contour plot reveals a highly meaningful organization of journals, with a focus on the subject domains considered characteristic for the LANL research community. We find groupings of journals relating to physical chemistry, organic chemistry, material sciences, applied physics, plasma physics, space research and nuclear science. The structure of the generated PCA map and its validation will be discussed in more detail in the following sections.

Principal components and validation

A PCA generates a set of components (or map dimensions) and ranks them according to the amount of variation in item correlations they account for, i.e. the components' factor loadings. A highly skewed factor loading distribution in our PCA analysis will indicate that journal correlations can be modeled by a few, powerful components. Vice versa, a flat distribution of factor loadings indicates journal correlations vary according to a larger set of weaker components.

The distributions of factor loadings for the LANL usage data reveals that LANL usage data varies according to a complex set of local constraints. Table 1 lists the factor loadings. We find that the factor loadings are highly distributed; PC1 has a loading of 25% and PC2 of 17%. PC1 and PC2 combined explain only 41% of the total amount of variation among journal correlations. Adding PC3 we find a cumulative loading of 51%. Following the scree test criterion, the selection of 2 or 3 factors was justified; loading values taper off considerably after the 3rd factor.

Table 1. Factor loadings of PCA performed for LANL04 data

Component	1	2	3	4	5
Loading (%)	25.07	15.92	10.44	4.74	3.69
Cumulative (%)	25.07%	40.99%	51.43%	56.17%	60%

To interprete the possible meaning of the first two components we examined the sets of journals found on the extremes of the PC1 and PC2 generated by the LANL04 PCA. These sets are listed in Table 2.

Table 2. Journals on negative and positive extremes of PC1 and PC2 resulting from LANL04 PCA

LANL04 - PC1 (horizontal)		LANL04 - PC2 (vertical)	
PC1 < -8	PC1 > 10.4	PC2 < -7.2	PC2 > 7.7
PHYS REV B	INORG CHEM	CHEM PHYS LETT	PHYS PLASMAS
PHYS REV LETT	TETRAHEDRON LETT	J PHYS CHEM B	PHYS FLUIDS
PHYS REV A	TETRAHEDRON	J PHYS CHEM A	IEEE T NUCL SCI
PHYSICA B	J ORGANOMET CHEM	CHEM MATER	ADV SPACE RES
J PHYS-CONDENS MAT	ELECTROPHORESIS	J AM CHEM SOC	PHYS FLUIDS
J APPL PHYS	ANAL CHEM	LANGMUIR	J GEOPHYS RES ATMOS
PHYSICA C	J ORG CHEM	J PHYS CHEM	SPACE SCI REV
J MAGN MAGN MATER	J AEROSOL SCI	ANGEW CHEM INT EDIT	PLASMA PHYS CONTR F
EUROPHYS LETT	COORDIN CHEM REV	J PHYS CHEM B	PHYS REV C
PHYS LETT A	J COLLOID INTERF SCI	MACROMOLECULES	ASTROPHYS J SUPPL S

On the negative end of PC1 we find a preponderance of journals relating to physics, in particular those relating to condensed matter. On the positive extreme, we find a majority of journals relating to chemistry, in particular organic chemistry. This matches the clusters that are located on the extremes of PC1 as shown in Figure 4. In particular, on the positive extreme of PC1 we find cluster 1 and 2 (organic chemistry and physical chemistry) and on its negative extreme cluster 5 (physics). The extremities of PC1 suggests an interpretation in terms of a split between natural and life sciences, except that in this case the split seems to be related to “inorganics” vs. “organics”, indicative of research community with a narrow focus on the natural sciences, in particular physics and chemistry.

The second component, PC2, seems to represent a more elusive dimension of the LANL04 data. The journals on the negative extreme of PC2 correspond to chemistry but with a particular focus on chemical physics, colloids and surfaces. The positive extreme of PC2 corresponds to plasma physics, space research (including geophysics) and nuclear physics. In fact, along the axis of PC2 we see a gradual shift from physical chemistry, followed by material science to a set of journals on the subjects of plasma, fusion, fluids and astrophysics. This seems to suggest a transition from the chemical “micro” constituents of matter, note the journal *Nano Letters*, to studies of “macro” phenomena relating to geophysics, space research, and astrophysics. Unfortunately, an adequate semantic label for this shift eludes us.

Of particular interest is the positioning of journals relating to materials science, namely at the intersection of PC1 and PC2. Material science can indeed be situated on the intersection of chemistry and physics, and the “micro” world of physical chemistry to the “macro” domain of fusion, plasma research, and astrophysics.

Usage clusters: category weighting and entropy

The k-means clustering of LANL usage was set to yield 5 distinct clusters. Journal cluster assignments are shown in Figure 4 and listed in Table 7 (Appendix). Information entropy values were calculated for each cluster's category distributions to determine the degree to which a particular cluster corresponded to a particular well-defined domain.

Most category distributions are characterized by a sharp decline of category weight values after the first two or three categories and thus correspond narrowly to a particular subject domain as shown in Figure 8 (Appendix). Using the three or four most highly valued journal categories, we subjectively label the generated cluster content as shown in Table 3. Entropy values are listed after the cluster labels as an indication of the cluster's domain focus.

Table 3. Labels assigned to LANL04 clusters on the basis of most highly weighted cluster ISI subject categories

LANL04			
Cluster	Label	ISI Category codes	Entropy (*H*(*j*))
1	Organic chemistry and biology:	EI, EE, DY, CQ, UY	3.411
2	Nuclear Science:	RY, EA, EC, CO	4.365
3	Fluids, plasma and astrophysics:	UF, UI, UR, UN, PU, SY, BU	4.029
4	Materials science:	PM, PZ, UP, UK	3.261
5	Condensed matter and physics:	UK, UB, UH, UI, SY	3.517

Again, we find a pattern that separates physics and related domains from chemistry and biology. As indicated on the left-hand side of the contourplot in Figure 5, the LANL04 clusters 3, 4, and 5 form a tight conglomerate of natural science research focused on materials science, applied physics, condensed matter and fluids and plasma. On the right of the LANL04 PCA map, shown in Figure 5, we find a less cohesive aggregation of the journals in clusters 1 and 2 which relates to nuclear science, inorganic and organic chemistry, and biology.

Of particular interest is a group of journals in cluster 2 relating to ongoing research on nuclear propulsion and space applications. This cluster is associated with the highest category entropy and is thus thought to have least domain focus. In addition, in cluster 3 we find a group of journals on the subject of applications of nuclear physics to astronomy and geophysics indicating the existence of a research group involved with astrophysical models for space observation and modeling. Cluster 4, material science, overlaps with all other clusters thereby indicating its multi-disciplinary focus at the intersection of physics, chemistry, and to a lesser degree organic chemistry and nuclear science. Surprisingly, this cluster has the lowest entropy values and thus the strongest domain focus.

Cluster validation

To validate the generated k-means clusters we performed a χ^2 analysis (SHESKIN, 2004) which tested whether k-means cluster assignments and journal ISI categories were significantly related, i.e. did LANL04 usage clusters overlap with journal ISI categories? In case the χ^2 analysis indicated a significant relationship, we calculated Cramer's Phi coefficient (CRAMER, 1946), denoted ϕ_c, to assess the strength of that relationship.[*] Table 4 lists the results of this analysis.

Table 4. Chi-square analysis for cluster assignment and ISI categories in JCR03 and LANL04 data sets

LANL04		
χ^2	df	p
293.610	140 $\phi_c = 0.699$	<0.001

The χ^2 analysis indicated the relationship between k-means cluster assignments and ISI journal categories was statistically significant, i.e. $p<0.001$, meaning that the clustering of journals in the LANL04 data matched ISI categories. In addition, the ϕ_c value was found to be 0.699 indicating a strong relationship between LANL04 journal clustering and ISI journal categories.

How about citation?

We applied the discussed methodology to the 2003 Thomson ISI *Journal Citation Reports* (Science Edition) to validate its ability to map the structure of scientific domains. In this case, however, the resulting mapping would reflect the structure of science for the global community of authors publishing in the set of Thomson ISI selected journals as opposed to the local community of LANL library users. As a shorthand we'll refer to the 2003 Thomson ISI *Journal Citation Reports* data set as JCR03.

A journal relationship matrix was derived from the JCR03 data for all pairs of journals in the collection (8624). For each pair of journals *A* and *B* we obtained a citation count which corresponds to the frequency with which articles published in journal *A* in the year 2003 cited articles published in journal *B* during the two preceding years (2002 and 2001). A directed, weighted 8,624 journal graph resulted which we

[*] Cramers's ϕ_C varies in the [0,+1] interval where 0 indicates the absence of a relationship, +1 indicating a srtong relationship.

represented by the matrix C which contained 1,004,289 non-zero entries, indicating a highly sparse journal citation graph: only 1% of all possible entries of matrix C had non-zero weights.

The resulting PCA mapping is shown in Figures 6 and 7. Journal abbreviation codes can be found in Table 8. Table 5 lists the cluster labels generated on the basis of the discussed category frequency weighting procedure and the corresponding entropy values. The respective distributions of category weights are shown in Figure 9. In addition, the χ^2 analysis results for the comparison of JCR03 k-means cluster assignment and Thomson ISI journal domain subject categories are listed in Table 6.

Table 5. Labels assigned to clusters on the basis of most highly weighted cluster ISI subject categories for JCR03 data

JCR03			
Cluster	Label	Category Codes	Entropy ($H(j)$)
1	Astronomy and astrophysics:	BU, UP, UN, UI, SY	2.524
2	Biochemistry and cell biology:	CQ, DM, RU, NI, CU	4.276
3	Physics and geophysics:	UK, UB, UI, UH, LE, FI	3.218
4	Medicine and microbiology:	MA, DE, WE, YQ, YP, QU	4.557
5	Chemistry and materials science:	EE, DY, EC, EI, EA, UY	3.232

Table 6. Chi-square analysis for cluster assignment and ISI categories in JCR03 data set

JCR03		
χ^2	df	p
469.000	172	<0.001
	$\phi_c = 0.884$	

We observe a more dense, clustered placement of journals in the JCR03 PCA mapping than was found for LANL04 PCA. This is clearly shown in the contour plot which reveals a highly dense placement of journals in particular regions of the graph. The particular journal clusters listed in Table 5 correspond to more general, less institution-driven subject categories such as “astronomy and astrophysics”, whereas the LANL04 clustering is strongly driven by institutional foci such as “applications of plasma physics to astrophysics”. In spite of the denser arrangement of journals, the JCR03 clusters are not more strongly focused than the LANL04 ones. Entropy values for the JCR03 clusters do no significantly differ from entropy values for the LANL04 cluster as evidenced by a t-test ($df = 5$, $p = 0.73$).

As is the case for the LANL04 data, the JCR03 k-means clustering is statistically significantly related to Thomson’s ISI subject categories. However, the JCR03 ϕ_c is slightly higher, i.e. 0.884 vs. 0.699, indicating citation patterns correspond better to existing subject categories.

These results indicate that the applied methodology can identify valid journal maps and subject groupings on the basis of global citation data, thereby validating its use on local LANL04 usage data.

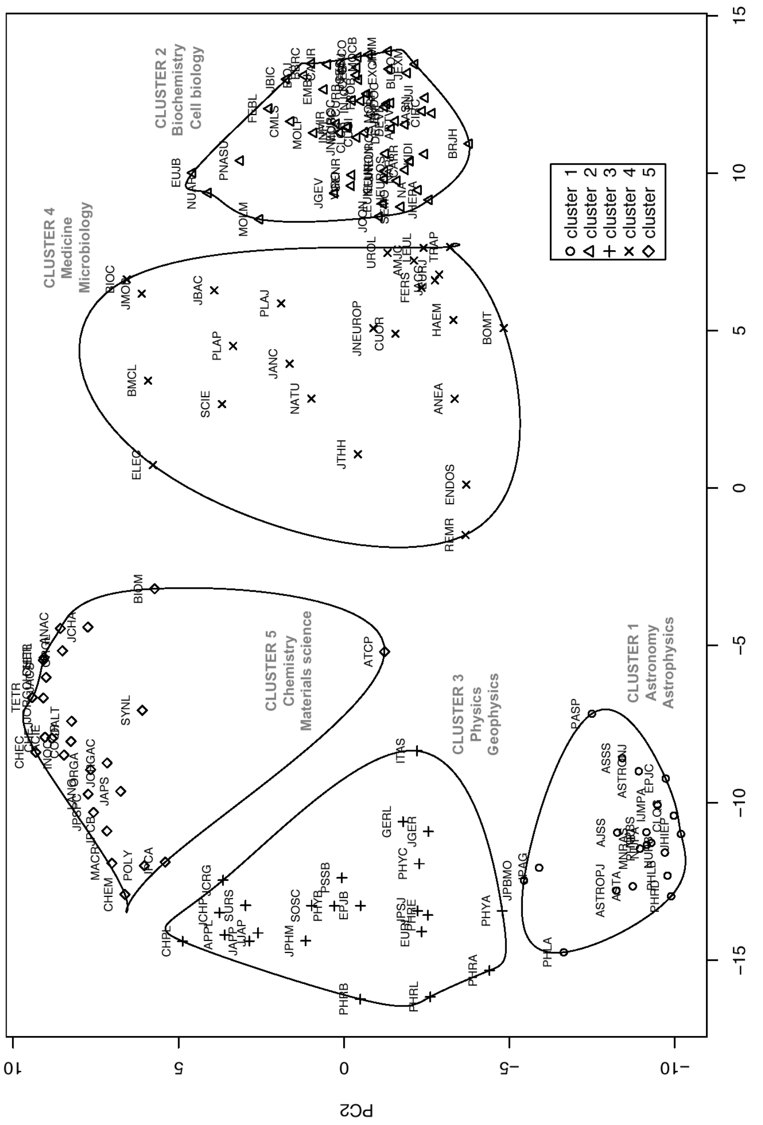


Figure 6. PCA and k-means cluster model of 2003 *Journal Citation Reports* data

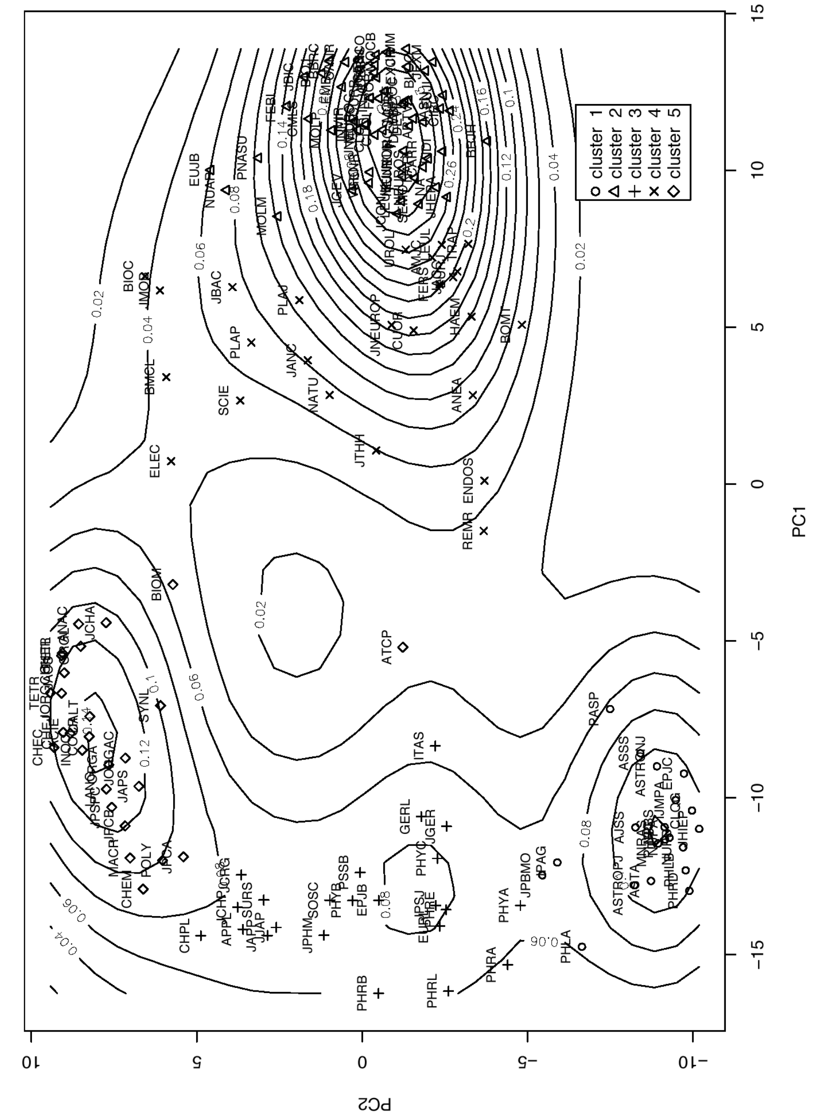


Figure 7. Contour map of spatial journal placement density in 2003 Thomson ISI *Journal Citation Reports*

Conclusion

Although there exist many methodologies to map the structure of science, most rely on the use of citation and authorship data, i.e. they examine the structure of the body of published material. Due to publication delays and citation biases any investigation of the structure of science on the basis of publication data is essentially studying science as it was 2 to 3 years ago. Since present usage rates have been shown to predict future citation rates, it has been speculated that usage data can be used as a viable, more contemporary proxy to scientific trends.

We have formulated a methodology to reconstruct networks of document relationship from DL usage data. We have done so for a set of LANL RL usage logs recorded through 2004 and 2005. A network of 10,696 journals was constructed. A PCA overlayed with a 5 cluster k-means analysis was used to map and cluster journal relationships as they occurred in this usage network.

We conclude the following from these results. First, the combined PCA and k-means clustering analysis indicates that a meaningful mapping of journal relationships can be derived from usage data. In other words, not only can usage data be applied to the ranking of journals and articles on the basis of usage frequency, the temporal patterns of usage contain information from which reliable and meaningful relational data can be inferred.

Second, when we examine the differences between the LANL04 usage and JCR03 citation mappings we observe that:

1. The geographical distribution of journals according to the PCA performed on usage data is more diffuse than the distribution of journals according to the PCA performed on citation data.
2. LANL04 usage clusters are equally strongly focused as JCR03 citation clusters, but on a set of different subject domains.
3. Usage and citation clusters do not strongly overlap in terms of subject domains. Usage defines a particular grouping of journals and subjects which is shaped by local institutional foci and contingencies.

Third, the distributions of factor loadings seem to suggest that citation data is focused on fewer and stronger subject dimensions such as “life sciences” vs. “natural sciences” then usage data. Indeed, journal correlations in the JCR03 PCA can be largely explained by the first two components (cumulated 87%) while the LANL04 PCA’s first two components explain only a cumulative 42% of all journal correlations. Citations furthermore adhere to commonly accepted subject groupings, such as physics, chemistry, astronomy and biology whereas usage seems to be more diffuse and interdisciplinary, and less easily explained by common scholarly subject categories.

As an example of such interdisciplinary focus we refer to a LANL04 journal cluster related to applications of nuclear energy to space technology, and a material science cluster which overlaps with a range of clusters related to subjects as varied as organic chemistry, plasma physics and condensed matter.

This paper has presented a preliminary attempt to compare the structure of usage and citation behavior in the scholarly community. In terms of our data samples and methodology additional work can be done. Although the LANL04 usage data provided an extensive sample of journal usage patterns, it does not represent usage in the global scientific community. Therefore our results confound both LANL specific patterns and general characteristics of usage. To fully represent the scientific community and the characteristics of usage as compared to citation we need to expand our log data sample to a wide range of representative institutions. Although certain technical hurdles exists in the large-scale aggregation of log data, VAN DE SOMPEL et al. (2003) discusses a number of practical solutions.

We should, however, note that similar limitations of sample size and coverage apply to the JCR03 citation data. The 2003 Thomson ISI *Journal Citation Reports* apply to a small subset of all published literature, i.e. a selection of journals for which Thomson's ISI has decided to publish citation and impact data. Possible future extensions of this work may rely on public sources of citation data such as Citeseer or arXiv to increase the relevance and scope of this work as shown by BOYACK (2004). In addition, the discussed analysis should be extended to the level of individual articles for a finer grained map of science.

Since usage data is generally more current than citation data, an interesting future research area would be to study the evolution of usage clusters and components over time. Given sufficient longitudenal log data reflecting the usage patterns of a representative sample of the scientific community, it is entirely feasible to study the temporal evolution of interest clusters. It may then be possible to use such data to predict future scientific trends. The factor loadings of our PCA analysis suggests furthermore that more accurate 3D models of usage and citation could be constructed by including additional components.

We conclude from these promising results that usage data is a viable, if not essential, part of the future scientometric and informetric instrumentarium. The discussed mapping of science on the basis of DL log data offers tantalizing clues to the possibility of studying local and global scientific trends as they occur in the present, free from the biases, delays and proprietary decision-making that generally accompanies citation and authorship data.

*

We want to thank Marko Rodriguez (Los Alamos National Laboratory) for his relentless proof-reading and contributions to the paper. Furthermore, we wish to extend our gratitude to the R developers (http://www. r-project.org) whose statistical software allowed us to perform our analysis.

References

ADAMS, J. (2005), Early citation counts correlate with accumulated impact. *Scientometrics*, 63 (3) : 567–581.

AGRAWAL, R., IMIELINSKI, T., SWAMI, A. (1993), Mining association rules between sets of items in large databases. In: *ACM SIGMOD International Conference on Management of Data,* Washington, DC, pp. 207–216.

BAEZA-YATES, R. A., RIBEIRO-NETO, B. A. (1999), *Modern Information Retrieval*. ACM Press/AddisonWesley.

BOLLEN, J., LUCE, R. (2002), Evaluation of digital library impact and user communities by analysis of usage patterns. *D-Lib Magazine*, 8 (6).

BOLLEN, J., VAN DE SOMPEL, H., SMITH, J., LUCE, R. (2005), Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management,* 41 (6) : 1419–1440.

BOYACK, K. W. (2004), Mapping knowledge domains: Characterizing PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (Suppl. 1).

BOYACK, K. W., KLAVANS, R., BOERNER, K. (2005), Mapping the backbone of science. *Scientometrics,* 64 (3) : 351–374.

BOYACK, K. W., WYLIE, B. N., DAVIDSON, G. S. (2002), Domain visualization using VxInsight(R) for science and technology management. *Journal of the American Society for Information Science and Technology,* 53 (9) : 764–774.

BRAAM, R. R., MOED, H. F., RAAN, A. F. J. VAN (1991a), Mapping of science by combined co-citation and word analysis. 1: Structural aspects. *Journal of the American Society for Information Science,* 42 (4) : 233–251.

BRAAM, R. R., MOED, H. F., RAAN, A. F. J. VAN (1991b), Mapping of science by combined co-citation and word analysis. 2: Dynamical aspects. *Journal of the American Society for Information Science,* 42 (4) : 252–266.

BRIN, S., MOTWANI, R., SILVERSTEIN, C. (1997), Beyond market baskets: generalizing association rules to correlations. In: *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data,* ACM Press, pp. 265–276.

BRODY, T., HARNAD, S. (2005), *Earlier Web Usage Statistics as Predictors of Later Citation Impact* (Eprint No. 10647). ECS, Intelligence, Agents, and Multimedia Group: University of Southampton.

CHAN, P. K. (1999), Constructing web user profiles: a non-invasive learning approach. In: B. MASAND, M. SPILIOPOULOU (Eds), *Web Usage Analysis and User Profiling* LNAI 1836. San Diego, CA: Springer.

CHEN, C. M., PAUL, R. J. (2001), Visualizing a knowledge domains intellectual structure. *Computer,* 34 (3).

CRAMER, H. (1946), *Mathematical Models of Statistics. Princeton*, NJ: Princeton University Press.

EGGHE, L., ROUSSEAU, R. (2000), The influence of publication delays on the observed aging distribution of scientific literature. *Journal of the American Society for Information Science,* 51 (2) : 158–165.

EVERETT, J. E., PECOTICH, A. (1991), A combined loglinear/MDS model for mapping journals by citation analysis. *Journal of the American Society for Information Science*, 42 (6) : 405–413.

HE, S., SPINK, A. (2002), A comparison of foreign authorship distribution in JASIST and the Journal of Documentation. *Journal of the American Society for Information Science and Technology,* 53 (11) : 953–959.

JOLLIFFE, I. T. (2002), *Principal Component Analysis*. New York: Springer Verlag.

KESSLER, M. M. (1963), Bibliographic coupling between scientific papers. *American Documentation,* 14 : 10–25.

KIM, D.-H., IM, I., ADAM, N., ATLURI, V., BIEBER, M., YESHA, Y. (2004), A clickstream-based collaborative filtering personalization model: Towards a better performance. In: *Proceedings of the 6th ACM International Workshop on Web Information and Data Management* (WIDM 2004). Washington, DC.

KING, M. F., BRUNER, G. C. (2000), Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing,* 17 (2) : 79–103.

KOHONEN, T. (1995), *Self-Organizing Maps.* Berlin: Springer.

KURTZ, M. J., EICHHORN, G., ACCOMAZZI, A., GRANT, C. S., DEMLEITNER, M., MURRAY, S. S. (2004a), The bibliometric properties of article readership information. *JASIST,* 56 (2) : 111–128.

KURTZ, M. J., EICHHORN, G., ACCOMAZZI, A., GRANT, C. S., DEMLEITNER, M., MURRAY, S. S. (2004b), Worldwide use and impact of the NASA Astrophysics Data System digital library. *JASIST,* 56 (1) : 36–45.

LEYDESDORFF, L. (2004a), Clusters and maps of science journals based on bi-connected graphs in journal citation reports. *Journal of Documentation,* 60 (4) : 371–427.

LEYDESDORFF, L. (2004b), Top-down decomposition of the journal citation report of the social science citation index: graphand factor-analytical approaches. *Scientometrics,* 60 (2) : 159–180.

LIU, X., BOLLEN, J., NELSON, M. L., VAN DE SOMPEL, H. (2005), Co-authorship networks in the digital library research community. *Information Processing and Managemen*t, 41 (6) : 1462–1480.

LIU, X., BOLLEN, J., NELSON, M. L., VAN DE SOMPEL, H., HUSSELL, J., LUCE, R., MARKS, L. (2004), Toolkits for visualizing co-authorship graph. In: *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL)*, Tuscon, AZ, p. 404.

LUWEL, M., MOED, H. F. (1998), Publication delays in the science field and their relationship to the ageing of scientific literature. *Scientometrics,* 41 (1–2) : 29–40.

MCCAIN, K. (1991), Mapping economics through the journal literature: An experiment in journal cocitation analysis. *Journal of the American Society for Information Science,* 42 (4) : 290–296.

MOBASHER, B., DAI, H., LUO, T., NAKAGAWA, M. (2001), Effective personalization based on association rule discovery from web usage data. In: *WIDM01: Proceedings of the 3rd international workshop on Web Information and Data Management,* ACM Press, pp. 9–15.

NAGPAUL, P. S. (2002), Visualizing cooperation networks of elite institutions in India. *Scientometrics,* 54 (2) : 213–228.

NEDERHOF, A. J. (1985), Methods of coping with social desirability bias. A review. *European Journal of Social Psychology,* 15 (3) : 263–280.

NEWMAN, M. E. J. (2001a), Scientific collaboration networks. II. Clustering and preferential attachment in growing networks. *Physical Review E,* 64 (2) : 016132/1–7.

NEWMAN, M. E. J. (2001b), Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E,* 64 (1), 016131/1–8.

RINIA, E. J., LEEUWEN, T. N. VAN, BRUINS, E. E. W., VUREN, H. G. VAN, RAAN, A. F. J. VAN (2001), Citation delay in interdisciplinary knowledge exchange. *Scientometrics,* 51 (1) : 293–309.

SALTON, G. (1998), Term-weighting approaches in automatic text retrieval. *Information Processing and Management,* 24 (5) : 513–523.

SARWAR, B., KARYPIS, G., KONSTAN, J., REIDL, J. (2001), Item-based collaborative filtering recommendation algorithms. In: *Proceedings of the Tenth International Conference on World Wide Web,* ACM Press, pp. 285–295.

SHANNON, C. E. (1948), A mathematical theory of communication. *Bell System Technical Journal*, 27 : 379–423.

SHESKIN, D. J. (2004), *Parametric and Nonparametric Statistical Procedures*. New York, NY: Chapman and Hall.

SMALL, H. (1973), Co-Citation in the scientific literature: a new measure of the relationship between documents. *Journal of the American Society for Information Science*, 42 : 676–684.

SPATH, H. (1980), *Cluster Analysis Algorithms*. Chichester, UK: Ellis Horwood.

SPILIOPOULOU, M. (1999), The laborious way from data mining to web mining. *Computer Systems Science and Engineering*, Special Issue on "*Semantics of the Web*", 14 (2) : 113–126.

SRIVASTAVA, J., COOLEY, R., DESHPANDE, M., TAN, P.-N. (2000), Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explorations Newsletter*, 1 (2) : 12–23.

TUFEKCI, S. (2003), Generalized decision trees: methodology and applications. *Computers and Industrial Engineering*, 24 (1) : 109–124.

VAN DE SOMPEL, H., BEIT-ARIE, O. (2001), Open linking in the scholarly information environment using the OpenURL framework. *D-Lib Magazine*, 7 (3).

VAN DE SOMPEL, H., PAYETTE, S., ERICKSON, J., LAGOZE, C., WARNER, S. (2004), Rethinking scholarly communication. Building the system that scholars deserve. *D-Lib Magazine,* 10 (9).

VAN DE SOMPEL, H., YOUNG, J. A., HICKEY, T. B. (2003), Using the OAI-PMH ... differently. *D-Lib Magazine*, 9 (7–8).

WAGNER, C. S., LEYDESDORFF, L. (2003), Mapping global science using international co-authorships: a comparison of 1990 and 2000. In: *Ninth International Conference on Scientometrics and Informetrics.* Beijing: ISSI.

WOUTERS, P. (1997), Citation cycles and peer review cycles. *Scientometrics*, 38 (1) : 39–55.

Appendix

Usage data: distributions of ISI journal classification code in clusters

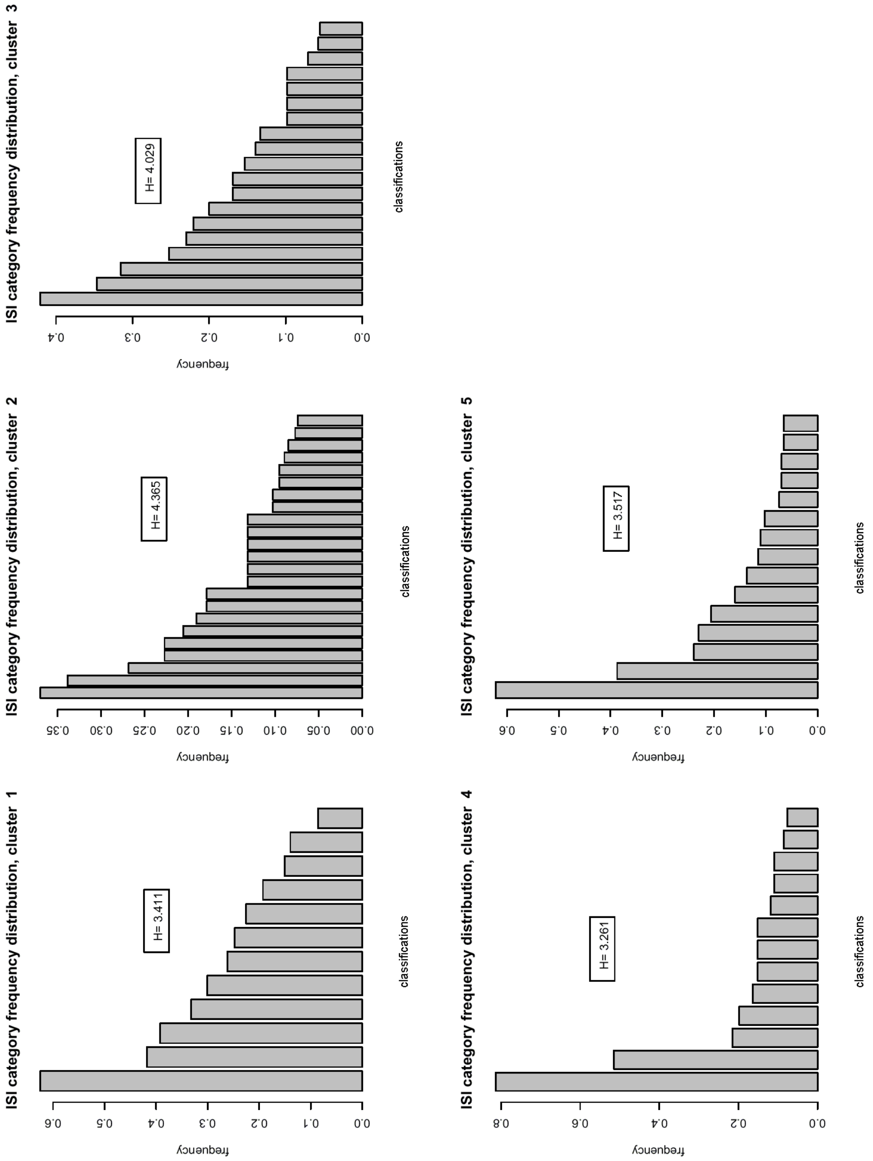


Figure 8. Cluster specific journal classification distributions for LANL04 data

Citation data: distributions of ISI journal classification code in clusters

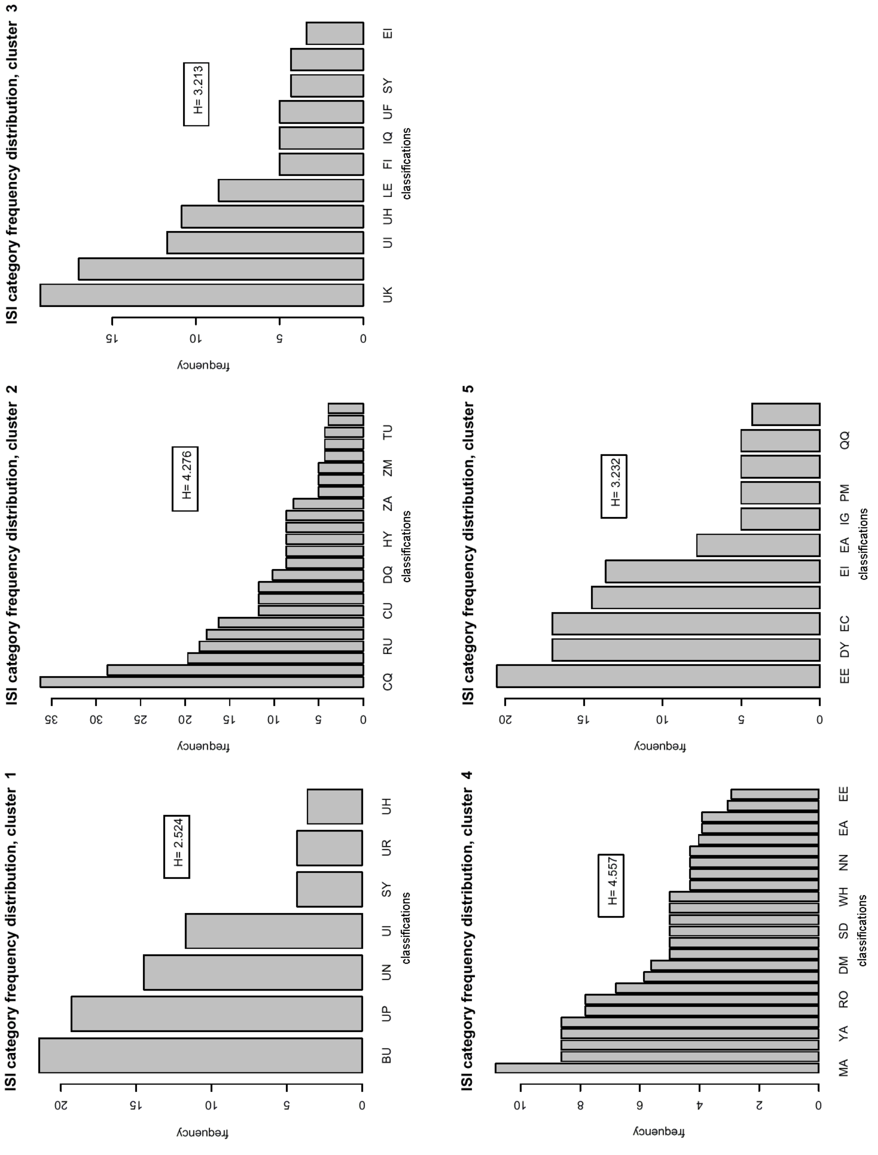


Figure 9. Cluster specific journal classification distributions for JCR03 data

Usage data: Journal abbreviations

Table 7. Journal title abbreviations for LANL04 PCA graphs

Number	Abbreviation	Short title	ISSN	Cluster
1	ACIE	ANGEW CHEM INT EDIT	1433–7851	1
2	AMSA	AMS ABSTR	0065–7727	1
3	ANAC	ANAL CHEM	0003–2700	1
4	BIOJ	BIOPHYS J	0006–3495	1
5	CHEC	CHEM COMMUN	1359–7345	1
6	CHEM	CHEM MATER	0897–4756	1
7	CHPL	CHEM PHYS LETT	0009–2614	1
8	COCR	COORDIN CHEM REV	0010–8545	1
9	INOC	INORG CHEM	0020–1669	1
10	JAMCHS	J AM CHEM SOC	0002–7863	1
11	JBIC	J BIOL CHEM	0021–9258	1
12	JCIS	J COLLOID INTERF SCI	0021–9797	1
13	JELC	J ELECTROANAL CHEM	0022–0728	1
14	JMOB	J MOL BIOL	0022–2836	1
15	JORGCH	J ORG CHEM	0022–3263	1
16	JPCA	J PHYS CHEM A	1089–5639	1
17	JPCB	J PHYS CHEM B	1089–5647	1
18	JPCB	J PHYS CHEM B	1520–6106	1
19	JPHYSCH	J PHYS CHEM	0022–3654	1
20	LANG	LANGMUIR	0743–7463	1
21	MACR	MACROMOLECULES	0024–9297	1
22	MOLP	MOL PHYS	0026–8976	1
23	PNASU	P NATL ACAD SCI USA	0027–8424	1
24	POLY	POLYMER	0032–3861	1
25	TETL	TETRAHEDRON LETT	0040–4039	1
26	TETR	TETRAHEDRON	0040–4020	1
27	ACSB	AM CERAM SOC BULL	0002–7812	2
28	ADSR	ADV SPACE RES	0273–1177	2
29	ATOE	ATOMN ENERG	0004–7163	2
30	BLOO	BLOOD	0006–4971	2
31	ELEC	ELECTROPHORESIS	0173–0835	2
32	ITNS	IEEE T NUCL SCI	0018–9499	2
33	JAES	J AEROSOL SCI	0021–8502	2
34	JAPS	J APPL POLYM SCI	0021–8995	2
35	JCHA	J CHROMATOGR A	0021–9673	2
36	JCOC	J COMPUT CHEM	0192–8651	2
37	JGRA	J GEOPHYS RES ATMOS	0747–7309	2
38	JORGAC	J ORGANOMET CHEM	0022–328x	2
39	JPHYSCO	J PHYS COLLOQ	0449–1947	2
40	JPOS	J POWER SOURCES	0378–7753	2
41	JRNC	J RADIOANAL NUCL CH	0236–5731	2
42	JSST	J SOL–GEL SCI TECHN	0928–0707	2
43	NIMB	NUCL INSTRUM METH B	0168–583x	2
44	NUCT	NUCL TECHNOL	0029–5450	2
45	OPTE	OPT ENG	0091–3286	2

Table 7. (continued)

Number	Abbreviation	Short title	ISSN	Cluster
46	RADA	RADIOCHIM ACTA	0033–8230	2
47	SEAB	SENSOR ACTUAT B–CHEM	0925–4005	2
48	SPSR	SPACE SCI REV	0038–6308	2
49	TEVT	TEPL VYS TEMP	0040–3644	2
50	AJSS	ASTROPHYS J SUPPL S	0067–0049	3
51	APPO	APPL OPTICS	0003–6935	3
52	APSB	APS BULL	0003–0503	3
53	ASTA	ASTRON ASTROPHYS	0004–6361	3
54	ASTJ	ASTRON J	0004–6256	3
55	EMID	EMERG INFECT DIS	1080–6040	3
56	FOUP	FOUND PHYS	0015–9018	3
57	GERL	GEOPHYS RES LETT	0094–8276	3
58	IJNMF	INT J NUMER METH FL	0271–2091	3
59	JCOP	J COMPUT PHYS	0021–9991	3
60	JFLM	J FLUID MECH	0022–1120	3
61	JGER	J GEOPHYS RES	0148–0227	3
62	JPAG	J PHYS A–MATH GEN	0305–4470	3
63	JPBMO	J PHYS B–AT MOL OPT	0953–4075	3
64	JSTP	J STAT PHYS	0022–4715	3
65	LAPB	LASER PART BEAMS	0263–0346	3
66	MICE	MICROELECTRON ENG	0167–9317	3
67	NIMA	NUCL INSTRUM METH A	0168–9002	3
68	NUCF	NUCL FUSION	0029–5515	3
69	NUIM	NUCL INST MET	0029–554X	3
70	NUPA	NUCL PHYS A	0375–9474	3
71	OPTC	OPT COMMUN	0030–4018	3
72	PHLA	PHYS LETT A	0375–9601	3
73	PHLB	PHYS LETT B	0370–2693	3
74	PHRC	PHYS REV C	0556–2813	3
75	PHRD	PHYS REV D	0556–2821	3
76	PHRE	PHYS REV E	1063–651X	3
77	PHRE	PHYS REV E	1539–3755	3
78	PHYD	PHYSICA D	0167–2789	3
79	PHYF	PHYS FLUIDS	0031–9171	3
80	PHYF	PHYS FLUIDS	1070–6631	3
81	PHYP	PHYS PLASMAS	1070–664X	3
82	PHYS	PHYS SCRIPTA	0281–1847	3
83	PHYSREP	PHYS REP	0370–1573	3
84	PPCF	PLASMA PHYS CONTR F	0741–3335	3
85	PZETF	PISM ZHU EKSP TEOR FIZ	0370–274X	3
86	RESI	REV SCI INSTRUM	0034–6748	3
87	ACTM	ACTA MATER	1359–6454	4
88	ANST	AM NUCL SOC TRANS	0003–018X	4
89	IESJ	IEEE SENS J	1530–437X	4
90	JALC	J ALLOY COMPD	0925–8388	4

Table 7. (continued)

Number	Abbreviation	Short title	ISSN	Cluster
91	JAMCES	J AM CERAM SOC	0002–7820	4
92	JASA	J ACOUST SOC AM	0001–4966	4
93	JELS	J ELECTROCHEM SOC	0013–4651	4
94	JMAR	J MATER RES	0884–2914	4
95	JMAS	J MATER SCI	0022–2461	4
96	JNUM	J NUCL MATER	0022–3115	4
97	JPHI	J PHYS IV	1155–4339	4
98	JVSTB	J VAC SCI TECHNOL B	1071–1023	4
99	MACP	MATER CHEM PHYS	0254–0584	4
100	MASF	MATER SCI FORUM	0255–5476	4
101	MMTA	METALL MATER TRANS A	1073–5623	4
102	MRSP	MAT RES SOC PROC	0272–9172	4
103	MSEA	MAT SCI ENG A–STRUCT	0921–5093	4
104	PHMA	PHILOS MAG A	0141–8610	4
105	PSPOIE	PROC SOC PHOTO OPT INSTRUM ENG	0361–0748	4
106	SCRM	SCRIPTA MATER	1359–6462	4
107	SOSI	SOLID STATE IONICS	0167–2738	4
108	THSF	THIN SOLID FILMS	0040–6090	4
109	AIPC	AIP CONF	0094–243X	5
110	APPA	APPL PHYS A–MATER	0947–8396	5
111	APPL	APPL PHYS LETT	0003–6951	5
112	CHEP	CHEM PHYS	0301–0104	5
113	EURL	EUROPHYS LETT	0295–5075	5
114	IJMPB	INT J MOD PHYS B	0217–9792	5
115	IJQC	INT J QUANTUM CHEM	0020–7608	5
116	JAPP	J APPL PHYS	0021–8979	5
117	JCHP	J CHEM PHYS	0021–9606	5
118	JJAP	JPN J APPL PHYS	0021–4922	5
119	JLTP	J LOW TEMP PHYS	0022–2291	5
120	JLUM	J LUMIN	0022–2313	5
121	JMMM	J MAGN MAGN MATER	0304–8853	5
122	JOSAB	J OPT SOC AM B	0740–3224	5
123	JPCS	J PHYS CHEM SOLIDS	0022–3697	5
124	JPHM	J PHYS–CONDENS MAT	0953–8984	5
125	JPHY	J PHYS	0302–0738	5
126	JPSJ	J PHYS SOC JPN	0031–9015	5
127	JSUP	J SUPERCOND	0896–1107	5
128	NANL	NANO LETT	1530–6984	5
129	NATU	NATURE	0028–0836	5
130	NEJM	NEW ENGL J MED	0028–4793	5
131	NIMB	NUCL INSTRUM METH B	0168–583X	5
132	OPTL	OPT LETT	0146–9592	5
133	PHRA	PHYS REV A	1050–2947	5

Table 7. (continued)

Number	Abbreviation	Short title	ISSN	Cluster
134	PHRB	PHYS REV B	0163–1829	5
135	PHRB	PHYS REV B	1098–0121	5
136	PHRL	PHYS REV LETT	0031–9007	5
137	PHYA	PHYSICA A	0378–4371	5
138	PHYB	PHYSICA B	0921–4526	5
139	PHYC	PHYSICA C	0921–4534	5
140	PHYE	PHYSICA E	1386–9477	5
141	PHYSREV	PHYS REV	0031–899X	5
142	PSSB	PHYS STATUS SOLIDI B	0370–1972	5
143	REDS	RADIAT EFF DEFECT S	1042–0150	5
144	SCIE	SCIENCE	0036–8075	5
145	SEST	SEMICOND SCI TECH	0268–1242	5
146	SOSC	SOLID STATE COMMUN	0038–1098	5
147	SPOP	SPIE OPT PROC	0277–786X	5
148	SUPM	SUPERLATTICE MICROST	0749–6036	5
149	SURS	SURF SCI	0039–6028	5
150	ZPHB	Z PHYS B	0722–3277	5

Citation data: Journal abbreviations

Table 8. Journal title abbreviations for JCR03 PCA graphs

Number	Abbreviation	Short title	ISSN	Cluster
1	AJSS	ASTROPHYS J SUPPL S	0067–0049	1
2	ASSS	ASTROPHYS SPACE SCI	0004–640X	1
3	ASTA	ASTRON ASTROPHYS	0004–6361	1
4	ASTRONJ	ASTRON J	0004–6256	1
5	ASTROPJ	ASTROPHYS J	0004–637X	1
6	CLQG	CLASSICAL QUANT GRAV	0264–9381	1
7	EPJC	EUR PHYS J C	1434–6044	1
8	IJMPA	INT J MOD PHYS A	0217–751X	1
9	JHIEP	J HIGH ENERGY PHYS	1029–8479	1
10	JPAG	J PHYS A–MATH GEN	0305–4470	1
11	JPBMO	J PHYS B–AT MOL OPT	0953–4075	1
12	MNRAS	MON NOT R ASTRON SOC	0035–8711	1
13	NPBS	NUCL PHYS B–PROC SUP	0920–5632	1
14	NUPA	NUCL PHYS A	0375–9474	1
15	NUPB	NUCL PHYS B	0550–3213	1
16	PASP	PUBL ASTRON SOC PAC	0004–6280	1
17	PHLA	PHYS LETT A	0375–9601	1
18	PHLB	PHYS LETT B	0370–2693	1
19	PHRC	PHYS REV C	0556–2813	1
20	PHRD	PHYS REV D	0556–2821	1

Table 8. (continued)

Number	Abbreviation	Short title	ISSN	Cluster
21	ARTV	ARTERIOSCL THROM VAS	1079–5642	2
22	BBRC	BIOCHEM BIOPH RES CO	0006–291X	2
23	BIOJ	BIOCHEM J	0264–6021	2
24	BLOO	BLOOD	0006–4971	2
25	BRJH	BRIT J HAEMATOL	0007–1048	2
26	CANR	CANCER RES	0008–5472	2
27	CARR	CARDIOVASC RES	0008–6363	2
28	CIRC	CIRCULATION	0009–7322	2
29	CIRR	CIRC RES	0009–7330	2
30	CLCR	CLIN CANCER RES	1078–0432	2
31	CMLS	CELL MOL LIFE SCI	1420–682X	2
32	CUOI	CURR OPIN IMMUNOL	0952–7915	2
33	CURB	CURR BIOL	0960–9822	2
34	DEVB	DEV BIOL	0012–1606	2
35	DEVE	DEVELOPMENT	0950–1991	2
36	EMBJ	EMBO J	0261–4189	2
37	ENDOC	ENDOCRINOLOGY	0013–7227	2
38	EUJB	EUR J BIOCHEM	0014–2956	2
39	EUJI	EUR J IMMUNOL	0014–2980	2
40	EUJN	EUR J NEUROSCI	0953–816X	2
41	EXCR	EXP CELL RES	0014–4827	2
42	FASJ	FASEB J	0892–6638	2
43	FEBL	FEBS LETT	0014–5793	2
44	FROB	FRONT BIOSCI	1093–9946	2
45	GENR	GENOME RES	1088–9051	2
46	INFI	INFECT IMMUN	0019–9567	2
47	INJC	INT J CANCER	0020–7136	2
48	JASN	J AM SOC NEPHROL	1046–6673	2
49	JBIC	J BIOL CHEM	0021–9258	2
50	JCEB	J CELL BIOL	0021–9525	2
51	JCES	J CELL SCI	0021–9533	2
52	JCON	J COMP NEUROL	0021–9967	2
53	JEXM	J EXP MED	0022–1007	2
54	JGEV	J GEN VIROL	0022–1317	2
55	JHEPA	J HEPATOL	0168–8278	2
56	JIMM	J IMMUNOL	0022–1767	2
57	JNEUROC	J NEUROCHEM	0022–3042	2
58	JNEUROS	J NEUROSCI	0270–6474	2
59	JVIR	J VIROL	0022–538X	2
60	KIDI	KIDNEY INT	0085–2538	2
61	LEUK	LEUKEMIA	0887–6924	2
62	MOBC	MOL BIOL CELL	1059–1524	2
63	MOCB	MOL CELL BIOL	0270–7306	2
64	MOLE	MOL ENDOCRINOL	0888–8809	2
65	MOLM	MOL MICROBIOL	0950–382X	2
66	MOLP	MOL PHARMACOL	0026–895X	2
67	NA	NA	0008–543X	2
68	NARI	NAT REV IMMUNOL	1474–1733	2

Table 8. (continued)

Number	Abbreviation	Short title	ISSN	Cluster
69	NEURON	NEURON	0896–6273	2
70	NEUROS	NEUROSCIENCE	0306–4522	2
71	NUAR	NUCLEIC ACIDS RES	0305–1048	2
72	ONCO	ONCOGENE	0950–9232	2
73	PNASU	P NATL ACAD SCI USA	0027–8424	2
74	SEMO	SEMIN ONCOL	0093–7754	2
75	VIRO	VIROLOGY	0042–6822	2
76	APPL	APPL PHYS LETT	0003–6951	3
77	CHPL	CHEM PHYS LETT	0009–2614	3
78	EPJB	EUR PHYS J B	1434–6028	3
79	EURL	EUROPHYS LETT	0295–5075	3
80	GERL	GEOPHYS RES LETT	0094–8276	3
81	ITAS	IEEE T APPL SUPERCON	1051–8223	3
82	JAPP	J APPL PHYS	0021–8979	3
83	JCHP	J CHEM PHYS	0021–9606	3
84	JCRG	J CRYST GROWTH	0022–0248	3
85	JGER	J GEOPHYS RES	0148–0227	3
86	JJAP	JPN J APPL PHYS	0021–4922	3
87	JPHM	J PHYS–CONDENS MAT	0953–8984	3
88	JPSJ	J PHYS SOC JPN	0031–9015	3
89	PHRA	PHYS REV A	1050–2947	3
90	PHRB	PHYS REV B	1098–0121	3
91	PHRE	PHYS REV E	1063–651X	3
92	PHRL	PHYS REV LETT	0031–9007	3
93	PHYA	PHYSICA A	0378–4371	3
94	PHYB	PHYSICA B	0921–4526	3
95	PHYC	PHYSICA C	0921–4534	3
96	PSSB	PHYS STATUS SOLIDI B	0370–1972	3
97	SOSC	SOLID STATE COMMUN	0038–1098	3
98	SURS	SURF SCI	0039–6028	3
99	AMJC	AM J CARDIOL	0002–9149	4
100	ANEA	ANESTH ANALG	0003–2999	4
101	BIOC	BIOCHEMISTRY–US	0006–2960	4
102	BMCL	BIOORG MED CHEM LETT	0960–894X	4
103	BOMT	BONE MARROW TRANSPL	0268–3369	4
104	CUOR	CURR OPIN RHEUMATOL	1040–8711	4
105	ELEC	ELECTROPHORESIS	0173–0835	4
106	ENDOS	ENDOSCOPY	0013–726X	4
107	EURJ	EUR RESPIR J	0903–1936	4
108	FERS	FERTIL STERIL	0015–0282	4
109	HAEM	HAEMATOLOGICA	0390–6078	4
110	JACC	J AM COLL CARDIOL	0735–1097	4
111	JANC	J ANTIMICROB CHEMOTH	0305–7453	4
112	JBAC	J BACTERIOL	0021–9193	4
113	JMOB	J MOL BIOL	0022–2836	4

Table 8. (continued)

Number	Abbreviation	Short title	ISSN	Cluster
114	JNEUROP	J NEUROPHYSIOL	0022–3077	4
115	JTHH	J THROMB HAEMOST	1538–7933	4
116	LEUL	LEUKEMIA LYMPHOMA	1042–8194	4
117	NATU	NATURE	0028–0836	4
118	PLAJ	PLANT J	0960–7412	4
119	PLAP	PLANT PHYSIOL	0032–0889	4
120	REMR	REV MAL RESPIR	0761–8425	4
121	SCIE	SCIENCE	0036–8075	4
122	TRAP	TRANSPLANT P	0041–1345	4
123	UROL	UROLOGY	0090–4295	4
124	ACIE	ANGEW CHEM INT EDIT	1433–7851	5
125	ANAC	ANAL CHEM	0003–2700	5
126	ATCP	ATMOS CHEM PHYS	1680–7324	5
127	BIOM	BIOMATERIALS	0142–9612	5
128	CHEC	CHEM COMMUN	1359–7345	5
129	CHEJ	CHEM–EUR J	0947–6539	5
130	CHEM	CHEM MATER	0897–4756	5
131	CHER	CHEM REV	0009–2665	5
132	COCR	COORDIN CHEM REV	0010–8545	5
133	DALT	DALTON T	1477–9226	5
134	INOC	INORG CHEM	0020–1669	5
135	JACS	J AM CHEM SOC	0002–7863	5
136	JAPS	J APPL POLYM SCI	0021–8995	5
137	JCHA	J CHROMATOGR A	0021–9673	5
138	JORGAC	J ORGANOMET CHEM	0022–328X	5
139	JORGCH	J ORG CHEM	0022–3263	5
140	JPCA	J PHYS CHEM A	1089–5639	5
141	JPCB	J PHYS CHEM B	1520–6106	5
142	JPSPC	J POLYM SCI POL CHEM	0887–624X	5
143	LANG	LANGMUIR	0743–7463	5
144	MACR	MACROMOLECULES	0024–9297	5
145	ORGA	ORGANOMETALLICS	0276–7333	5
146	ORGL	ORG LETT	1523–7060	5
147	POLY	POLYMER	0032–3861	5
148	SYNL	SYNLETT	0936–5214	5
149	TETL	TETRAHEDRON LETT	0040–4039	5
150	TETR	TETRAHEDRON	0040–4020	5