KSZI [ξ] AKTÁK

MTAK TTO műhelytanulmányok

2012/5

Measuring the similarity between the reference and citation distributions of journals

Schubert András

schuba@helka.iif.hu

➲ http://www.mtakszi.hu/kszi_aktak/

To be published in *Scientometrics*
DOI: 10.1007/s11192-012-0889-0

Measuring the similarity between the reference and citation distributions of journals

András Schubert
Department of Science Policy and Scientometrics, Library of the Hungarian Academy of Sciences, Budapest, Hungary

Summary

The "Jaccardized Czekanowski index", JCz, an indicator measuring the similarity between the cited and citing journal list of a given journal is proposed in the paper. It is shown that the indicator characterizes the network properties of individual journals and, in aggregated form, also that of subject categories or countries.
For subject categories, JCz appears to be related to the multidisciplinarity of the category. For countries, the multinational or local character of the publishers seems to have determining role.

Introduction

Primarily, citation is an asymmetric relation between documents. Actually, it occurs only in exceptional (in a way, pathological) cases that two papers mutually cite each other (for an extreme example see Rousseau and Small, 2005). As soon as the citation relation is extended to aggregates of documents (such as the papers of given authors or journals), mutual citations become the rule rather than the exception. In such networks the inlink/outlink symmetry of nodes may be an interesting element of characterizing their network properties.
In a simple but useful way, the Journal Citation Reports included basic data for such a dual characterization of journals from the very beginning, "Citing" and "Cited" journal packages provided ranked journal-to-journal citations lists both from the viewpoints of the sources and the targets of citations.
In this paper a simple method is proposed to use these data to characterize the balanced or unbalanced nature of journals as their citation sources and targets are concerned. The method can easily be extended to objects other than journals, such as authors, institutions or countries.

Data and methods

Data were taken from the 2006 Science Citation Index Journal Citation Reports (SCI JCR 2006) database. For each journal (6164 titles) the journal-by-journal distribution of references (Citing Journal Package) and citations (Cited Journal Package) were determined. For a subset of journals having at least 100 references/citations to and from other journals (5037 titles) the similarity of the two distributions was compared.

Similarity measures.
For each journal, the distribution of the references in the given journal over the cited journals, as well as the distribution of the received citations citing journals are categorical distributions without any underlying ordering (i.e., there is no natural order other than, say, alphabetical order among the journals). There are several similarity measures advised in the literature for comparing such distributions (see, e.g., McCune et al., 2002). The Jaccard family of measures can be derived from the classical Jaccard index,

$$J_{A,B} = |A \cap B| / |A \cup B|,$$

where $|A \cap B|$ is the number of non-empty categories (i.e., journals cited/citing at least once) in the intersection of distributions A (say, cited journal distribution) and B (citing journal distribution),

while $|A \cup B|$ is that in their union. If, e.g., a journal cites 100 journals (any number of times) and is cited by 50 journals (any number of times), and 30 titles are present in both lists, then the two lists will contain 100+50-30=120 different titles, and the Jaccard index will be J=30/120=0.25. Thus, the Jaccard index disregards the quantitative aspect of the occupancy of categories (the frequency of citations).

In two recent publications (Schubert, 2010; Schubert & Soós, 2010), the quantitative aspect was incorporated into the Jaccard index by restricting its calculation to the most highly cited subsets of A and B, namely, the h-cores (with citations equal to or higher than the Hirsch index), resulting in a "h-restricted Jaccard index", h-J. There are possibilities to use more refined weighting schemes, as well, shown as follows.

The Sorensen index (also known as the Dice coefficient),

$$So_{A,B} = 2|A \cap B|/[|A \cup B| + |A \cap B|],$$

is directly related to the Jaccard index:

$$So = 2J/(J+1).$$

It can be formulated also as

$$So_{A,B} = 1 - \Sigma_i |\delta_i^A - \delta_i^B| / \Sigma_i (\delta_i^A + \delta_i^B),$$

where δ_i^A and δ_i^B take the value of 0 or 1 depending whether the i-th category is empty or non-empty in distributions A and B, respectively. This latter formulation allows easy extension to account for quantitative differences in occupancy (citation frequency); the Czekanowski index (also called quantitative or relative Sorensen index, proportional similarity index or Bray-Curtis index) is defined as

$$Cz_{A,B} = 1 - \Sigma_i |q_i^A - q_i^B| / \Sigma_i (q_i^A + q_i^B) = 1 - (1/2) \Sigma_i |q_i^A - q_i^B|,$$

where q_i^A and q_i^B are the relative frequencies of category i in the distributions A and B, respectively. Using the relation between the Jaccard and Sorensen indices, one can define the "Jaccardized" Czekanowski index as

$$JCz = Cz/(2 - Cz).$$

This index can be considered the "quantitative" (i.e., occupancy or abundance dependent) version of the Jaccard index.

A substantial amount of experience in the application of the Czekanowski-type indices (under whatever name) has been accumulated in the field of ecology (e.g., Bloom, 1981; Faith et al., 1987, Minchin, 1987a;b). Since our bibliometric model shares several features with those typical in ecology (large number of categories – many of them scarcely populated, strongly skewed, "long tailed" distributions, samples of substantially different size, etc.), these experiences are expected to bear relevance to our model, as well. It is the general opinion that in these cases the Czekanowski-type indices perform better than such time-honored alternatives from the statistical toolkit like the cosine measure,

$$Cos_{A,B} = \Sigma_i q_i^A q_i^B / (\Sigma_i (q_i^A)^2 \Sigma_i (q_i^B)^2)^{1/2},$$

or the chi-squared measure

$$Chi_{A,B} = 1 - (1/2)(\Sigma_i (q_i^A - q_i^B)^2 / (q_i^A + q_i^B))^{1/2}.$$

Preliminary research

In order to get familiarized with the behavior of these indices in the journal sample to be studied, some preliminary studies were made.

Here, like later in the main study, all similarity measures were calculated by leaving out the journal under study from the summation in all formulas. This was motivated by the fact that practically all journals has an extremely high self-share both in the references and in the citations causing a large apparent similarity of the two distributions. By leaving out these self-references/-citations, the indices would measure how references/citations are distributed among the *other journals*.

Furthermore, the study was restricted to journals having at least 100 references/citations to and from *other journals* (5037 titles), granting a fair degree of statistical reliability to the results.

Table 1 shows the linear correlation coefficients (assuming zero intercept) among some of the similarity measures mentioned above. It can be clearly seen that JCz is fairly correlated with all other measures, while the other measures are rather uncorrelated among each other.

Table 1 Linear correlation among some of the similarity measures

	J	h-J	Chi	Cos
JCz	0.5259	0.6035	0.9630	0.6525
J	–.–	0.2410	0.4985	0.2696
h-J	–.–	–.–	0.1821	0.3218

This finding was interpreted as an indication of the fact that JCz is characterizing the journal sample under study in a stable and coherent way, therefore, it is a suitable indicator of cited/citing similarity.

Results and discussion

Overall statistics

Figure 1 shows the frequency distribution of the JCz similarity index calculated for the reference/citation distribution of all journals in the sample (5037 titles; journal self-references/-citations excluded).

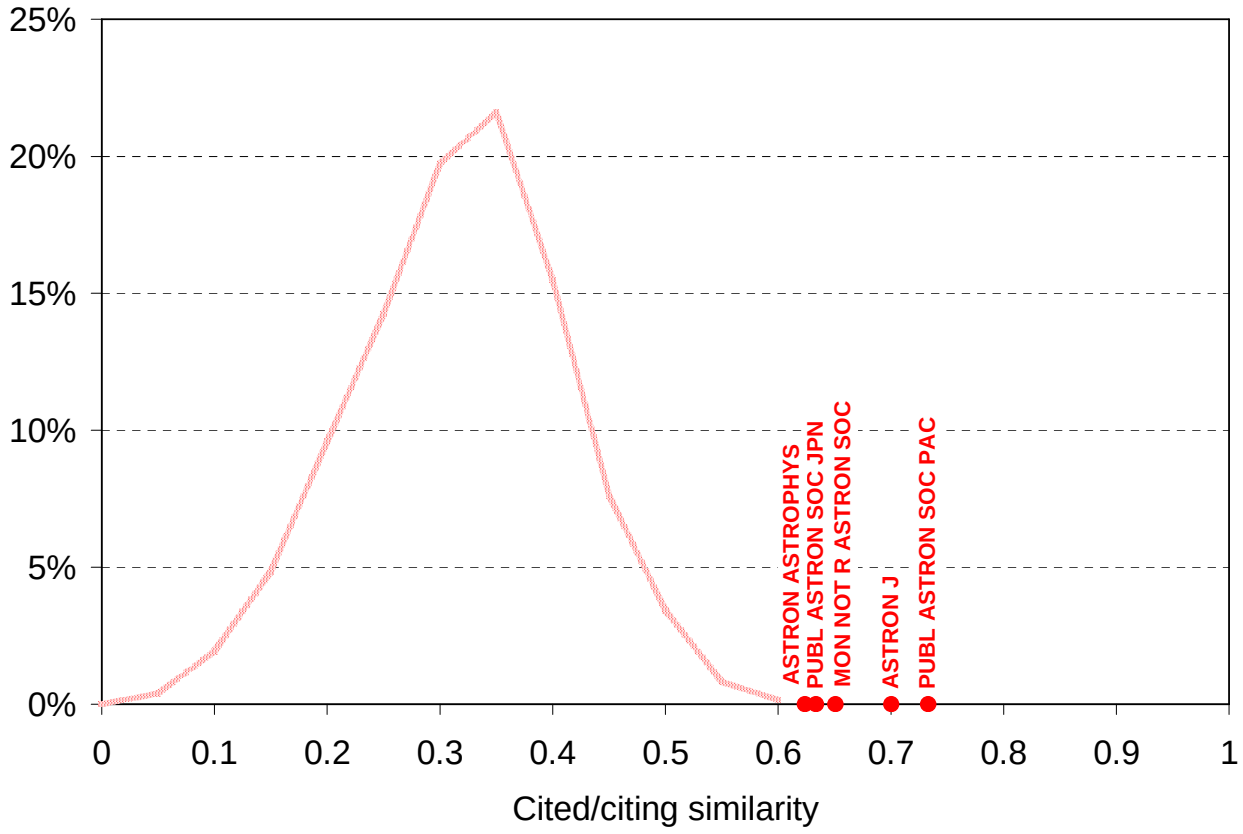

Figure 1 The distribution of the JCz similarity index over journals

The full range extends from 0.012 (INTERNIST) to 0.733 (PUBL ASTRON SOC PAC), the average is 0.294 (standard deviation 0.0946), the median is 0.298.
It is striking that the top five titles individually highlighted in Figure 1 are all astronomy journals. It seemed, therefore, obvious to study in more details the cited/citing similarity of journals by subject categories.

Subject categories

The SCI JCR 2006 classified the journals into 173 subject categories. Figure 2 shows the distribution of the subject category averages and highlights the categories with extremely low and high values.

The subject category averages range between 0.139 (MEDICAL ETHICS) and 0.474 (ASTRONOMY & ASTROPHYSICS) showing that, indeed, there are rather characteristic differences among the categories.

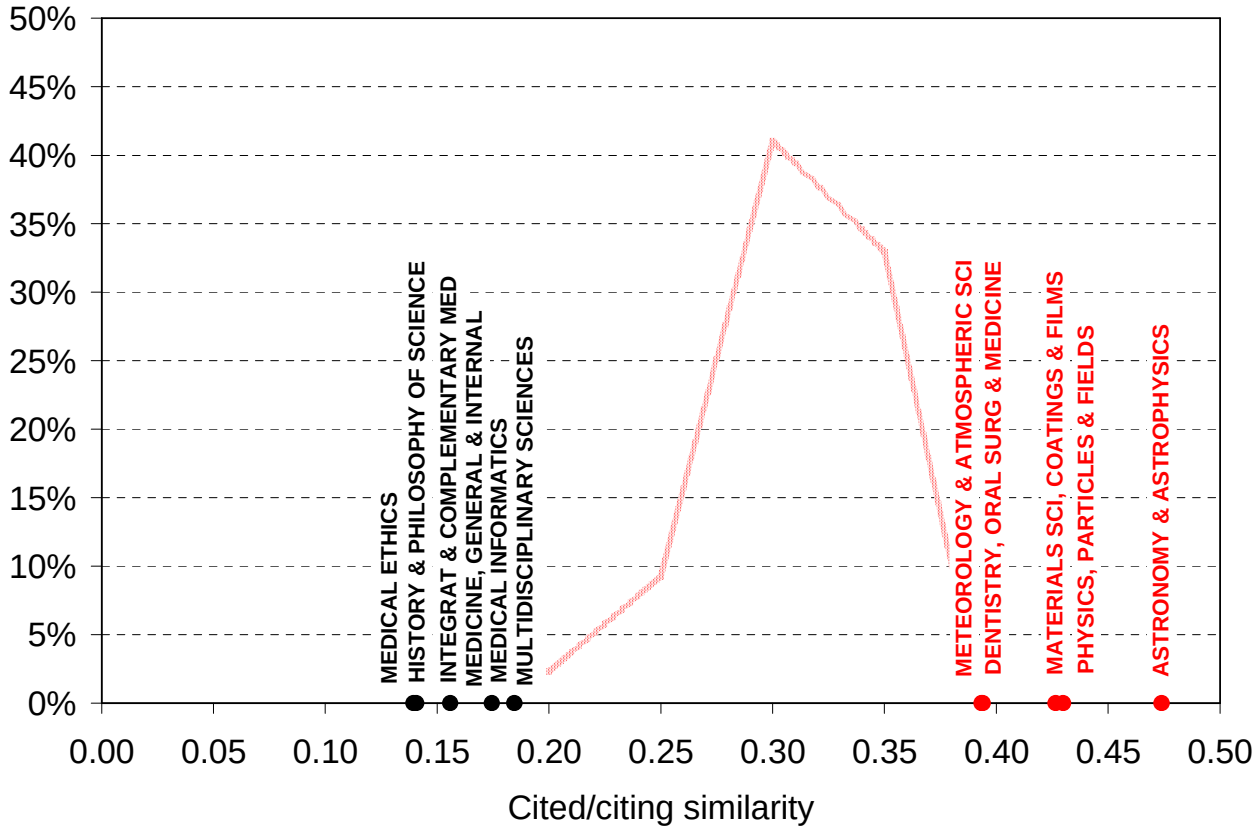

Figure 2 The distribution of the average JCz similarity index over subject categories

A closer look at the highlighted categories also suggests substantive tendencies behind the numerical differences. The categories at the high end appear to be closed, self-contained areas, while at the low end one finds looser, "wide-spectrum" categories. In order to support this impression, the "wideness" of the subject categories were attempted to be measured, as well. For each journal in the sample, the concentration/diversity of the subject categories in their citation traffic (inward and outward citation flows, combined) was characterized by the normalized Gini-Simpson concentration index (introduced by Gini (1912) and adapted by Simpson (1949); in the economic literature also called the normalized Herfindahl-Hirschman index (Hirschman, 1945; Herfindahl, 1950)):

$$GS = ((\Sigma_i \delta_i \Sigma_i q_i^{2)} - 1))/((\Sigma_i \delta_i) - 1),$$

δ_i is 0 or 1 depending whether the i-th category (subject category, in our case) is empty or non-empty (i.e., $\Sigma_i \delta_i$ is the number of non-empty categories), q_i is the relative frequency of category i in the citation traffic distribution. Journal self-references/-citation were again excluded from the calculation. The GS indexes of the journals within the same subject categories were then averaged to characterize the concentrated or diverse nature of the categories. In a sense, this indicator measures the multidisciplinarity of the given subject category

The comparison of the two indicators (JCz and GS) yielded somewhat ambiguous results. Examining the top and bottom 10 categories in both rankings one finds significant similarities, particularly at the low end. (In Table 2, categories having top/bottom positions in both lists are highlighted in bold; only subject categories including more than ten journals are listed.) Not in a single case a category having a top position in one of the list has a bottom position in the other. All these suggest a definite parallelism between the two indicators. At the same time, the correlation coefficient between JCz and GS ($r^2 = 0.0244$) shows total uncorrelatedness.

The situation dramatically changes if the subject categories are partitioned into three groups (see Figure 3). In each group there is an obvious correlation between the two indicators. In the middle group (red solid circles in Figure 3), with ASTRONOMY & ASTROPHYSICS and MEDICAL ETHICS at the extremes, the value of the two indicators is practically identical.
There are only a few points in the group on the right in Figure 3 (blue empty circles), with MATHEMATICS having an extreme position. The categories in this group are characterized by a relatively high subject category concentration index (low multidisciplinarity) with relatively low effect on journal-level reference/citation similarity. E.g., in MATHEMATICS, the sets of cited and the citing journals both are, in large extent from the subject category MATHEMATICS, but yet the two sets do not really overlap. One possible reason is that several MATHEMATICS journals prefer to cite wide-spectrum journals of the category, while they are cited mainly by more narrowly specified topical journals.
In the third group (black crosses in Figure 3), which contains, actually, the majority of the journals, cited/citing similarity is very sensitively (and, of course, inversely) influenced by multidisciplinarity. JCz values are almost double of the GS index.
In summary, the subject category differences between the cited/citing similarity index, JCz, is definitely related to the multidisciplinarity of the categories, but this relation is somewhat concealed. It has to be stressed that while the JCz index itself is completely independent of the choice of the subject category system, the existence and nature of relations between subject-category-level indicators may obviously strongly depend on it. It might even be surmised that irregular behavior of certain subject categories in our study may indicate their ill-defined, incoherent character in the SCI JCR subject category system.

Table 2 Subject categories with the highest and lowest similarity and concentration indices

Subject Category	JCZ	Subject Category	GS
ASTRONOMY & ASTROPHYSICS	0.474	MATHEMATICS	0.633
PHYSICS, PARTICLES & FIELDS	0.430	**DENTISTRY, ORAL SURGERY & MEDICINE**	0.561
DENTISTRY, ORAL SURGERY & MEDICINE	0.394	**ASTRONOMY & ASTROPHYSICS**	0.544
METEOROLOGY & ATMOSPHERIC SCIENCES	0.393	**OPHTHALMOLOGY**	0.537
PHYSICS, NUCLEAR	0.392	STATISTICS & PROBABILITY	0.385
PHYSICS, CONDENSED MATTER	0.389	ENGINEERING, AEROSPACE	0.368
OPHTHALMOLOGY	0.385	POLYMER SCIENCE	0.362
PHYSICS, ATOMIC, MOLECULAR & CHEMICAL	0.379	TELECOMMUNICATIONS	0.358
ELECTROCHEMISTRY	0.378	MATERIALS SCIENCE, PAPER & WOOD	0.351
ORTHOPEDICS	0.368	ORNITHOLOGY	0.351
.		.	
.		.	
.		.	
COMPUTER SCIENCE, HARDWARE & ARCHITECTURE	0.222	THERMODYNAMICS	0.108
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE	0.220	ALLERGY	0.105
GERIATRICS & GERONTOLOGY	0.220	**MEDICAL INFORMATICS**	0.103
COMPUTER SCIENCE, INFORMATION SYSTEMS	0.217	TROPICAL MEDICINE	0.097
BIOLOGY	0.213	**BIOLOGY**	0.096
MEDICINE, RESEARCH & EXPERIMENTAL	0.204	ANATOMY & MORPHOLOGY	0.095
MEDICAL LABORATORY TECHNOLOGY	0.203	**MEDICINE, GENERAL & INTERNAL**	0.095
MULTIDISCIPLINARY SCIENCES	0.185	**MEDICINE, RESEARCH & EXPERIMENTAL**	0.088
MEDICAL INFORMATICS	0.175	**MEDICAL LABORATORY TECHNOLOGY**	0.081
MEDICINE, GENERAL & INTERNAL	0.174	**MULTIDISCIPLINARY SCIENCES**	0.076

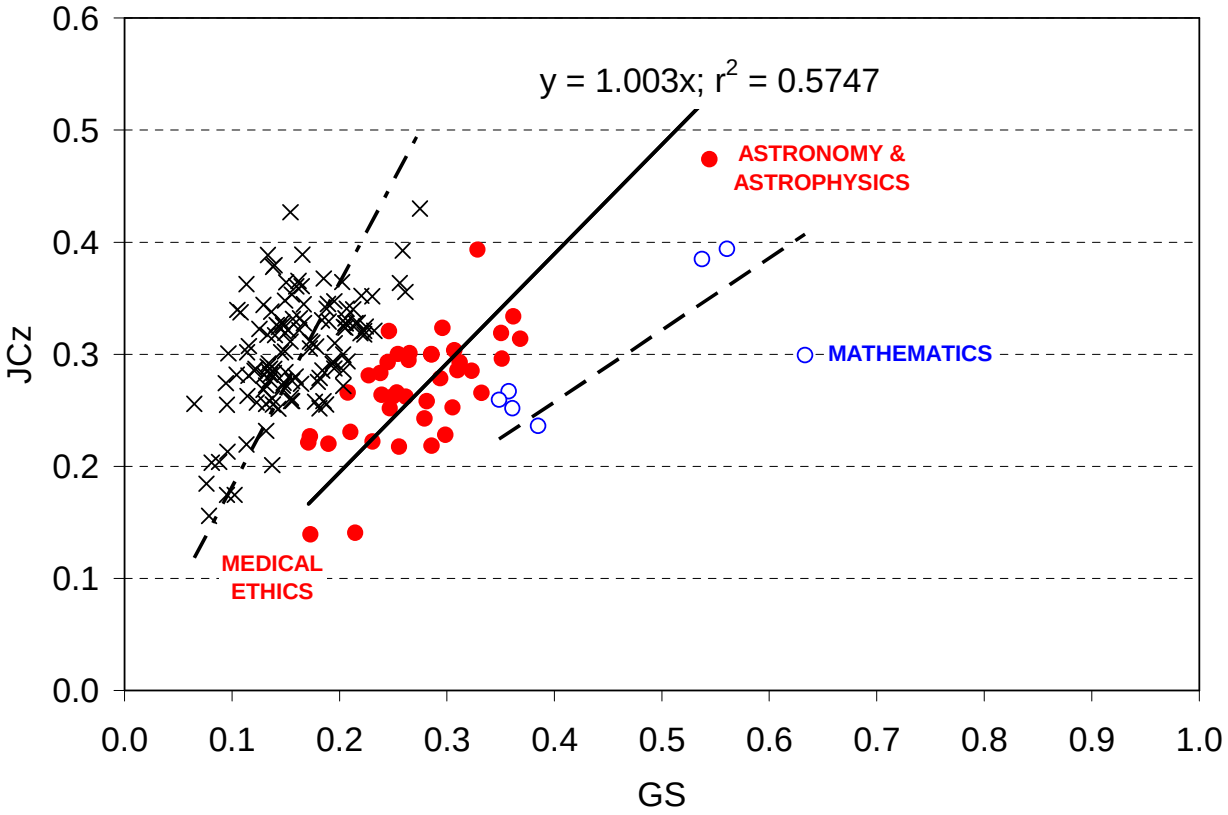

Figure 3 Regression plot of indices JCz and GS in three groups of subject categories

Countries

SCI JCR assigns countries to each journal according to the address of the headquarters of the publisher. It is a rather dubious classification taking into account the real multinational character of all major publishers. Nevertheless, with due reservations, it is worth a try, and the results shown in Figure 4 suggest non-nonsense inferences (only countries with at least 9 journals in the SCI JCR 2006 database are included in the figure).
Countries housing the major multinational publishers are in the top (most of the Austrian journals in the database are published by Springer, Vienna). Journals published in more peripheral countries exhibit, as a rule, significantly less similarity in their reference/citation structure. A typical source of this imbalance is the tendency of these journals to cite "mainstream" international literature and to be more strongly cited by local/regional journals.

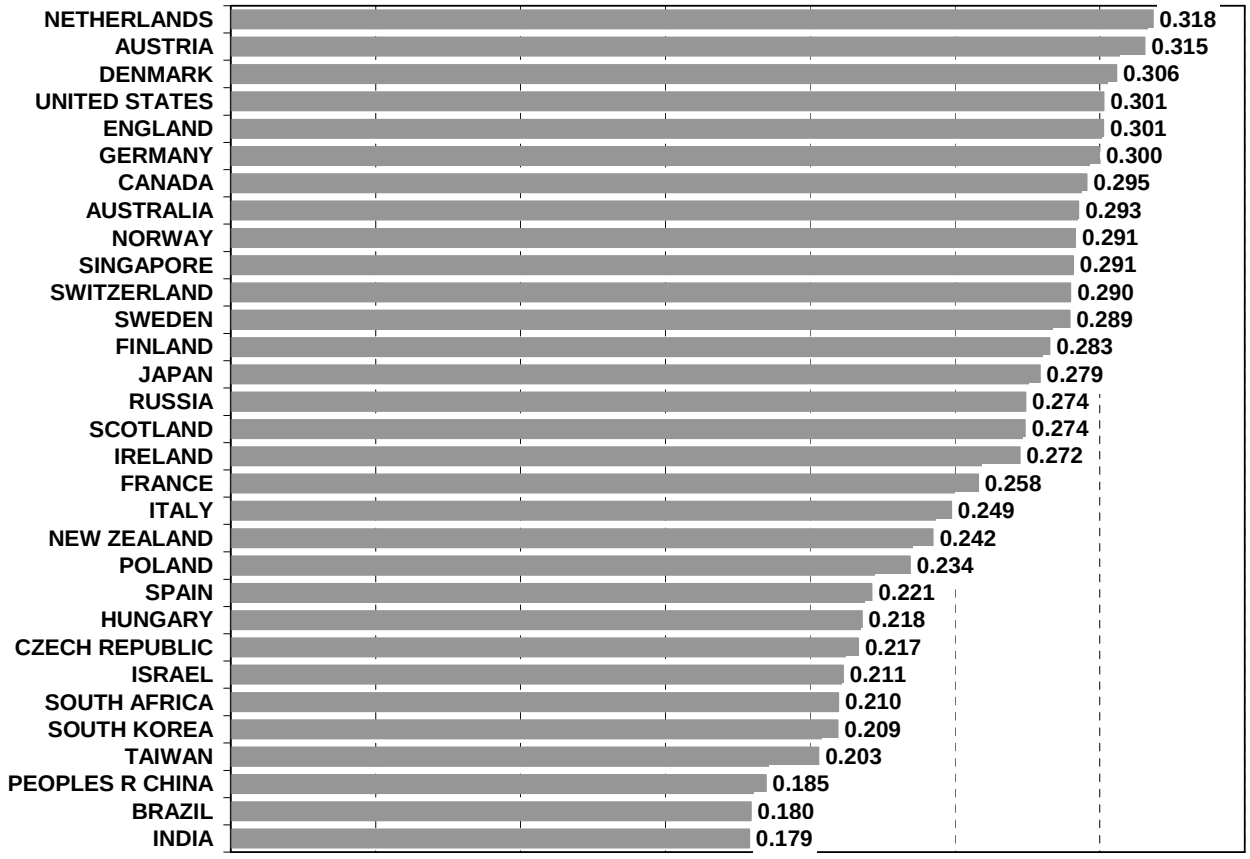

Figure 4 Country averages of the JCz index

Evaluative aspects

The author feels the need to devote a specific paragraph to stress that the similarity index proposed in this paper has no evaluative aspect, whatsoever. Any attempt to find correlation between JCz and some impact factor-like indicator remained unsuccessful whether in the total sample or in selected subsamples (by subject category, country, journal type, etc.).

Conclusions

The "Jaccardized Czekanowski index", JCz, an indicator measuring the similarity between the cited and citing journal list of a given journal was proposed in the paper. It was shown that the indicator characterizes the network properties of individual journals and, in aggregated form, also that of subject categories or countries.

By using a weighting scheme clearly favorizing major cited/citing journals over minor ones, JCz seems to give in this specific study a picture closer to the common-sense concept of similarity/dissimilarity than the binary Jaccard or Sorensen indices or than the chi-squared and the cosine measures, where the nature of weighting is not unambiguous. As compared to the classical Czekanowski index, JCz values scatter over a larger range (within the 0–1 interval) thereby discriminates more clearly among items. Its "normal" behavior is witnessed by the shape of the distribution in Figure 1.

For subject categories, JCz appeared to be related to the multidisciplinarity of the category. For countries, the multinational or local character of the publishers seemed to have determining role.

The similarity or dissimilarity of the cited and citing journals is not a good or bad feature, it is a structural indicator conveying important information of a journal's place and role in the information network. It is definitely not an indicator of evaluative value, but it may help, for example, to outline a fitting editorial policy or publishing strategy.

The reference/citation similarity concept can easily be extended from journals to other bibliometric actors, such as authors, institutions or countries.

References

Bloom, S.A. (1981). Similarity indices in community studies: Potential pitfalls, Marine Ecology – Progress Series, 5, 125–128.
Czekanowski, J. (1909) Zur differential Diagnose der Neandertalgruppe. Korrespondenzblatt der deutschen Gesellschaft fur Anthropologie, Ethnologie und Urgeschichte, 40, 44–47
Faith, D. P., Minchin, P. R., Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. Vegetatio, 69, 57–68.
Gini, C. (1912). Variabilità e mutabilità. In: Pizetti, E., Salvemini, T. Eds., Rome: Libreria Eredi Virgilio Veschi, Memorie di metodologica statistica.
Herfindahl, O. C. (1950) Concentration in the U.S. Steel Industry. Unpublished doctoral dissertation, Columbia University.
Hirschman, A. O. (1945) National power and the structure of foreign trade. Berkeley.
McCune, B., Grace, J.B., Urban, D.L. (2002). Analysis of Ecological Communities. MjM Software Design, Chapter 6, Distance measures. Accessed in August 2012 at http://www.pelagicos.net/BIOL6090/readings/Distance_Measures_Chapter6.pdf
Minchin, P.R. (1987a). An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio, 69, 89–107.
Minchin, P.R. (1987b). Simulation of multidimensional community patterns: towards a comprehensive model. Vegetatio, 71, 145–146.
Rousseau, R., Small, H. (2005) Escher staircases dwarfed. ISSI Newsletter, 1(4), 8–10.
Schubert, A. (2010). A reference-based Hirschian similarity measure for journals. Scientometrics, 84, 133–147.
Schubert, A., Soós, S. (2010). Mapping of science journals based on h-similarity. Scientometrics, 83, 589–600.
Simpson, E.H. (1949). Measurement of diversity. Nature, 163, 688.