Algorithmic Procedure for Finding Semantically Related Journals

Alexander I. Pudovkin
Institute of Marine Biology, Far East Branch, Russian Academy of Sciences, Vladivostok 690041, Russia. E-mail: aipud@online.ru

Eugene Garfield
Chairman Emeritus, Institute for Scientific Information™ (ISI™), 3501 Market Street, Philadelphia, PA 19104-3389. E-mail: garfield@codex.cis.upenn.edu

Using citations, papers and references as parameters a relatedness factor (RF) is computed for a series of journals. Sorting these journals by the RF produces a list of journals most closely related to a specified starting journal. The method appears to select a set of journals that are semantically most similar to the target journal. The algorithmic procedure is illustrated for the journal *Genetics*. Inter-journal citation data needed to calculate the RF were obtained from the 1996 *ISI Journal Citation Reports on CD-ROM*©. Out of the thousands of candidate journals in JCR©, 30 have been selected. Some of them are different from the journals in the *JCR* category for genetics and heredity. The new procedure is unique in that it takes varying journal sizes into account.

Introduction

The classification of scientific and scholarly journals is a problem well known to scientists and librarians for decades. Traditional classification relies on subjective analysis which for one reason or another proves inadequate and is subject to the vagaries of time. Quantitative methods have been proposed for overcoming these problems. This was greatly facilitated with the introduction of citation indexes in the 1960's and the later introduction of the *ISI Journal Citation Reports*. *JCR*'s for science and social science are produced annually. In the seventies, *JCR*'s in print were issued as the last volume of the *Science Citation Index*© or *Social Sciences Citation Index*©. Later microform and CD-ROM editions were introduced and more recently it appeared on the Internet.

JCR reports inter-journal citation frequencies for thousands of journals. In addition to an alphabetic listing, journals are grouped by categories. Journals are assigned to categories by subjective, heuristic methods.[1] In many fields these categories are sufficient but in many areas of research these "classifications" are crude and do not permit the user to quickly learn which journals are most closely related.

JCR provides, for each journal, a set of its most closely related journals based on citation relationships. These are the journals it cites most heavily (cited journals) and also the journals which cite it most often (citing journals). These are extremely useful and provide a crude classification, but unfortunately due to the variations in the sizes of journals one only obtains a superficial perception of the relatedness between two or more specific journals.

Various authors have studied journal-to-journal citation rates, mostly for the purposes of hierarchical clustering of the journals and delineation of specialty fields (Narin et al., 1972; Narin & Carpenter, 1973; Leydesdorff, 1994; Narin et al., 2000). However, they do not deal with the key problem of varying journal sizes. In this paper we have described a method which takes size into account. The method has its origins in earlier works by Pudovkin and Elizabeth Fuseler (Pudovkin, 1992, 1993; Pudovkin & Fuseler, 1995). They attempted to visualize citation relationships of core marine

[1] One of the referees asked for a description of the procedures used by ISI in establishing journal categories for *JCR*. These procedures are followed by the ISI editorial group in charge of journal selection and are similar to those used for the *SCI* and *Current Contents*® journal categories. This method is "heuristic" in that the categories have been developed by manual methods started over 40 years ago. Once the categories were established, new journals were assigned one at a time. Each decision was based upon a visual examination of all relevant citation data. As categories grew, subdivisions were established. Among other tools used to make individual journal assignments, the Hayne-Coulson algorithm is used. The algorithm has never been published. It treats any designated group of journals as one macrojournal and produces a combined printout of cited and citing journal data.

Received October 15, 2001; revised May 21, 2002; accepted May 21, 2002

and freshwater biology journals. For that purpose indexes of citation relatedness were used. This enabled journals to be clustered and then displayed in a two-dimensional diagram. The resultant "map" of journal relatedness was quite meaningful: a tight group of multi-disciplinary marine biology journals occurred in the center of the diagram, journals more narrow in scope were situated on the periphery, topically similar journals being grouped close to each other. The more meaningful visualization of marine journals was the result of using the indexes of citation relatedness, which took into account the variation in journal sizes.

Recently, Egghe and Rousseau have developed a theory for quantifying language preferences in journal citations (Egghe et al., 1999; Rousseau, 1999; Egghe & Rousseau, 2000). The measures suggested by them are similar to the indexes of citation relatedness suggested by Pudovkin (1992, 1993): the measures they developed also take into account the number of citations from one journal to another and the sizes of the journals. However, our approach is more pragmatic than theoretical. We wished to develop a procedure that would, through quantitative evaluation of citation relatedness, allow one to automatically find topically similar journals, that is, without considering the titles of papers or journal content.

The algorithm described here uses the indexes of citation relatedness, suggested by Pudovkin (1992, 1993). The process appears to approximate the subjective, that is, semantic judgment of experts. We have illustrated the procedure using one core journal in the field of genetics and heredity, the well-known *Genetics*, published by the Genetics Society of America.

Journal Relationship Measures

Let journal relatedness of two journals, "i" and "j" be symbolized by $R_{i>j} = H_{i>j} * 10^6 / (Pap_j * Ref_i)$, where $H_{i>j}$ is the number of citations in the current year from journal "i" to journal "j" (to papers published in "j" in all years of 'j'), Pap_j and Ref_i are the number of papers published and references cited in the j-th and i-th journals in the current year. An arbitrary multiplier of 10^6 makes the values of the relatedness index more easily perceived and handled. For example, the 1996 issues of *Genetics* cited all years of *Heredity* 351 times. The number of references cited in *Genetics* was 21,060, and the number of papers published in *Heredity* was 146. Substituting these numbers in the formula we get $R_{G>H} = 351{*}10^6 /(146{*}21{,}060) = 114.2$ (where G stands for *Genetics* and H stands for *Heredity*). Figure 1 visualizes these calculations.

Journals	Citations (C) from G to H and H to G	Papers (Pap)	Cited references (Ref)
Genetics (G)	351	448	21,060
Heredity (H)	149	146	4,869

$$R_{G>H} = \frac{C_{G>H} \times 10^6}{Pap_H \times Ref_G} = \frac{351 \times 10^6}{146 \times 21{,}060} = 114.2$$

$$R_{G\&H}max = \frac{C_{H>G} \times 10^6}{Pap_G \times Ref_H} = \frac{149 \times 10^6}{448 \times 4{,}869} = 68.3$$

$R_{G>H}$ is the maximum of the two.

FIG. 1. Calculation of *Genetics-Heredity* Relatedness Factor.

The rationale for the formulation of the indexes follows. The number of citations from one journal to another journal should be (on average) proportional to the number of papers published in the cited journal and to the number of cited references in the citing journal. Thus, a journal publishing 1,000 papers a year will tend to receive 10 times as many citations as a journal publishing only 100 papers, all other conditions being the same. Similarly, a journal which has cumulatively cited 10,000 references will tend to cite another journal ten times more often than a topically similar journal that cumulatively cites 1,000 references. Thus, these numbers, which reflect the sizes of citing and cited journals, are placed into the denominator of the formula. The number of citations a journal receives depends on the cumulative number of papers published in the journal during all the years of its existence. Since an annual *JCR* does not provide this historical information, it was decided to use the number of papers published in the current year. It was understood, of course, that this convention introduces a fortuitous error in the estimation of citation relatedness, as journal sizes change differently from year to year. Though, for the majority of journals their sizes are relatively stable over the years (Garfield, 1996). It was considered unwise to use the number of citations to the papers of the current year because of the time lag in getting citations, which is quite significant in less than hot research fields. Besides, yearly citation scores are rather low for many journals, hence they would be too subject to chance fluctuations.

If we consider a pair of journals, A and B, there may be two indexes: $R_{A>B}$ and $R_{B>A}$. These can be very different. Consider the above mentioned journals, *Genetics* and *Heredity*, $R_{G>H} = 114.2$ and $R_{H>G} = 68.3$. It is noteworthy that the citation relatedness of a journal to itself (that is "self-relatedness") may be lower than its relatedness to some other journals. For instance, *Journal of Genetics* has both citing and cited relatedness indexes with *Genetics* that are higher than the self-relatedness of *Genetics.* The latter, $R_{G>G} = 301.9$; the former $R_{JG>G} = 338.3$; and $R_{G>JG} = 503.7$. The same is true for *Genetics* and *Genetical Research* relationship: $R_{GR>G} = 393.0$; and $R_{G>GR} = 306.0$. It is interesting to note the very high self-related-

TABLE 1. Thirty journals ranked by relatedness factor to *Genetics* in 1996.

Journal title	A	B	C	D	E	F	G
Journal of Genetics	0.278	8	390	88	**49**	503.7	**1**
Genetical Research	2.102	45	1562	290	**18**	393.0	**2**
Journal of Neurogenetics	1.235	4	165	33	**94**	391.7	**3**
Annual Review of Genetics	9.741	22	3225	163	**26**	351.8	**4**
Genetics	4.928	448	21060	2848	**1**	301.9	**5**
Theoretical Population Biology	1.609	31	1065	121	**35**	253.6	**6**
Advances in Genetics	1.773	10	999	70	**56**	227.9	**7**
Heredity	2.014	146	4869	351	**16**	160.9	**8**
Molecular Biology and Evolution	5.969	133	6065	386	**13**	142.1	**9**
Cell	40.997	451	20305	1327	**3**	139.7	**10**
Molecular and Cellular Biology	10.727	270	42307	763	**5**	134.2	**11**
Theoretical and Applied Genetics	2.313	336	10322	616	**7**	133.2	**12**
Molecular Ecology	2.799	91	3392	194	**24**	127.7	**13**
Genetics Selection Evolution	0.902	35	1042	58	**66**	124.2	**14**
Annual Review of Biochemistry	38.966	25	5312	65	**58**	123.5	**15**
Genome	1.792	154	5114	281	**19**	122.7	**16**
Chromosoma	2.633	62	2367	154	**30**	117.9	**17**
Journal of Heredity	1.443	82	2147	96	**45**	99.8	**18**
Microbiological Reviews	19.526	30	9128	62	**60**	98.1	**19**
Genetica	1.243	62	2382	104	**43**	97.5	**20**
Genes & Development	18.810	259	15476	468	**10**	85.8	**21**
Animal Genetics	1.235	151	1434	54	**70**	84.1	**22**
Evolutionary Biology	3.000	8	1321	19	**135**	83.1	**23**
Evolution	3.203	248	13739	507	**8**	82.4	**24**
Genes and Genetic Systems	n/a	29	701	24	**121**	76.4	**25**
Annual Review of Ecology and Systematics	3.964	20	3446	44	**82**	76.0	**26**
Maydica	0.557	29	772	26	**112**	75.2	**27**
Roux Archives of Developmental Biology	1.681	31	1077	47	**77**	72.0	**28**
Current Genetics	1.802	132	4723	145	**32**	68.5	**29**
Trends in Ecology and Evolution	6.252	83	3180	96	**46**	67.4	**30**

Data based on JCR, 1996. A: Impact Factor; B: number of 1996 papers; C: number of cited references; D: maximal number of citations (to or from the journal); E: rank by "D"; F: relatedness factor to *Genetics,* $R_{G\&i}$max; G: rank by "F". Journals in JCR "Genetics & Heredity" category in **bold.**

ness of the *Journal of Genetics*, $R_{JG>JG} = 961.5$ and *Genetical Research*, $R_{GR>GR} = 1693.0$.

As was mentioned above, each pair of journals may be characterized with a pair of indexes, that quantifies their reciprocal citation levels: "A" citing "B", and "B" citing "A". How should one integrally characterize the citation relatedness of a pair of journals? Previously, Pudovkin (1993) and Pudovkin and Fuseler (1995) used the arithmetic average of the two indexes where $R_{A\&B} = (R_{A>B} + R_{B>A})/2$. Now it is suggested we use the larger of them, $R_{A\&B}\text{max} = \max(R_{A>B}, R_{B>A})$, which we shall call the relatedness factor (RF). A similarly sounding term, Relationship Factor, was recently introduced by Shama et al. (2000), though it refers to the relationship between disciplines rather than journals. It takes into account the impact factors of journals and the number of citations from journals of one discipline to the journals of another.

Consider the pair of journals *Genetics* and *Genetika* (*Russian Journal of Genetics*). The latter is the title of the low circulation cover-to-cover translation in English that is published simultaneously with the original. Both *Genetics* and *Genetika* are very similar in content, publishing papers on all aspects of *Genetics.* But being a Russian language journal *Genetika* receives few citations from *Genetics*, while it cites *Genetics* quite often. The citation relatedness indexes for them are $R_{A>B} = 49.7$ and $R_{B>A} = 1.6$ (where A stands for *Genetika* and B stands for *Genetics*). Similar situations are observed with other national journals, even those published in English: e.g. the French English language journal *Genetics Selection Evolution* and *Genetics*, $R_{GSE>G} = 124.2$, $R_{G>GSE} = 25.8$. Two other examples: Scandinavian *Genetica* and *Genetics*, $R_{1>2} = 97.5$, $R_{2>1} = 42.9$; the British journal *Heredity* and *Genetics*, $R_{1>2} = 160.9$, $R_{2>1} = 48.5$. The analogous situation applies when the pair of journals involves one which is an older, established journal and the other is a recently launched one, e.g.,. *Genome* and *Genetics*, $R_{1>2} = 122.7$ and $R_{2>1} = 29.9$. Another example: *Molecular Ecology* and *Genetics*, $R_{1>2} = 127.7$ and $R_{2>1} = 16.7$. Thus, the maximal value of the two indexes seems to better reflect the topical similarity of the journals.

The asymmetry of citation relationships in some journal pairs discussed above has some similarity to the language preferences studied by Egghe and Rousseau (2000), though the asymmetry revealed by us is certainly a different phenomenon, as it is often seen in journal pairs in the same language.

Illustration: Finding the Journals Most Related to *Genetics*

For each journal *JCR* provides two lists: citing and cited journals. The cited and citing citation scores were retrieved

TABLE 2. Thirty journals giving or receiving the highest number of citations to or from *Genetics.*

Journal title	A	B	C	D	**E**	F	**G**
Genetics	4.928	448	21060	2848	**1**	301.9	**5**
Proceedings of the National Academy of Sciences of the USA	10.244	2790	101511	1343	**2**	22.9	**86**
Cell	40.997	451	20305	1327	**3**	139.7	**10**
Nature	28.417	885	24642	844	**4**	45.3	**47**
Molecular and Cellular Biology	10.727	270	42307	763	**5**	134.2	**11**
Science	23.605	1025	36553	617	**6**	28.6	**69**
Theoretical and Applied Genetics	2.313	336	10322	616	**7**	133.2	**12**
Evolution	3.203	248	13739	507	**8**	82.4	**24**
EMBO Journal	13.255	725	37262	480	**9**	31.4	**61**
Genes & Development	18.810	259	15476	468	**10**	85.8	**21**
Nucleic Acid Research	4.448	726	26422	408	**11**	26.7	**73**
Journal of Biological Chemistry	7.452	4949	209095	394	**12**	3.8	**197**
Molecular Biology and Evolution	5.969	133	6065	386	**13**	142.1	**9**
Journal of Bacteriology	3.889	1062	42375	369	**14**	16.5	**108**
Molecular & General Genetics	2.601	317	11809	352	**15**	62.0	**33**
Heredity	2.014	146	4869	351	**16**	160.9	**8**
Development	9.182	435	21054	332	**17**	35.2	**53**
Genetical Research	2.102	45	1562	290	**18**	393.0	**2**
Genome	1.792	154	5114	281	**19**	122.7	**16**
Journal of Molecular Biology	5.195	697	32974	251	**20**	17.1	**105**
Journal of Cell Biology	12.680	483	27458	244	**21**	24.0	**79**
Gene	1.931	747	16654	209	**22**	13.3	**118**
Journal of Molecular Evolution	3.052	151	6370	199	**23**	62.6	**32**
Molecular Ecology	2.799	91	3392	194	**24**	127.7	**13**
Developmental Biology	4.963	340	16485	168	**25**	23.5	**82**
Annual Review of Genetics	9.741	22	3225	163	**26**	351.8	**4**
American Naturalist	3.525	137	7018	161	**27**	55.8	**39**
Bioessays	6.227	126	6478	161	**28**	55.5	**40**
Trends in Genetics	10.781	123	3019	159	**29**	61.4	**34**
Chromosoma	2.633	62	2367	154	**30**	117.9	**17**

Data based on JCR, 1996. A: Impact Factor; B: number of 1996 papers; C: number of cited references; D: maximal number of citations (to or from the journal); E: rank by "D"; F: relatedness factor to *Genetics,* $R_{G\&i}$max; G: rank by "F". Journals in JCR "Genetics & Heredity" category in **bold.**

for those journals that cited *Genetics* or were cited by it seven or more times. Also retrieved were journals with lesser citation scores (two and more), which seemed "genetical" judging from their titles. There were 271 such journals. Thirty journals with the highest RF (with *Genetics*) are given in Table 1. Table 2 lists 30 journals which give to or receive from *Genetics* the highest number of citations (raw citation scores). Journal titles in boldface are included in the "Genetics & Heredity" ("G & H") category of *JCR*.

It is evident that the new algorithmic approach selected the journals that are similar in content to *Genetics*: Twenty one journals listed in Table 1 are in the "G & H" category while only 13 journals in Table 2 are in this category. This difference is due to the weighting (or filtering) property of the citation relatedness indexes and the RF, which will be discussed below. The algorithm located some other journals that should be included in the "G & H" category (or genetics should be indicated as subcategory for them), that are not now included: 1) *Molecular Biology and Evolution*, 2) *Molecular Ecology*, 3) *Maydica*. The first journal is categorized by *JCR* as "biochemistry & molecular biology," though it mostly covers population and evolutionary genetics. The second journal publishes population and evolutionary genetics papers, touching on ecology. *JCR*'s category for it is "ecology" without any mention of "genetics." The third journal is characterized by *JCR* as "agriculture; plant science." Consideration of the journal paper titles shows that twenty papers of 42 published in *Maydica* in 1996 dealt with genetics or genetic improvement in cultivated plants. Also noteworthy, the subcategory of "genetics" is not indicated in the *JCR* category for *Annual Review of Ecology and Systematics,* which publishes many highly cited papers on population and evolutionary genetics. It ranks 26th in Table 1.

It is interesting to note the high citation relatedness to *Genetics* of journals dealing with developmental and cell biology. These disciplines are much "geneticized" now. This is reflected in Table 1. The journals *Cell*, *Molecular and Cellular Biology*, *Rouxs Archives of Developmental Biology* are among 30 journals most related to *Genetics*.

An important feature of the suggested approach is the calculation of SPECIFIC citation relatedness, that is, the new indexes take into consideration the sizes of citing (through the number of references) and cited (through the number of published papers) journals. The word SPECIFIC is used as are terms in physics such as "specific weight," "specific density," etc. If one ignores journal size in con-

TABLE 3. Some core genetics journals ranked by relatedness factor to *Genetics,* $R_{G\&i}$max or raw citation scores.

Journal title	A	B	C	D	E	F	G
Journal of Genetics	0.278	8	390	503.7	88	**1**	**49**
Journal of Neurogenetics	1.235	4	165	391.7	33	**3**	**94**
Theoretical Population Biology	1.609	31	1065	253.6	121	**6**	**35**
Advances in Genetics	1.773	10	999	227.9	70	**7**	**56**
Genetics Selection Evolution	0.902	35	1042	124.2	58	**14**	**66**
Journal of Heredity	1.443	82	2147	99.8	96	**18**	**45**
Genetica	1.242	62	2382	97.5	104	**20**	**43**
Animal Genetics	1.235	151	1434	84.1	54	**22**	**70**
Evolutionary Biology	3.000	8	1321	83.1	19	**23**	**135**
Genes and Genetic Systems	n/a	29	701	76.4	24	**25**	**121**
Annual Review of Ecology and Systematics	3.964	20	3446	76.0	44	**26**	**82**
Trends in Ecology and Evolution	6.252	83	3180	67.4	96	**30**	**46**
Biochemical Genetics	0.813	30	915	56.1	34	**38**	**91**
Fungal Genetics and Biology	n/a	28	823	54.2	20	**41**	**133**
Annals of Human Genetics	3.491	43	1210	480	26	**45**	**113**
Human Heredity	1.611	56	1038	34.4	16	**56**	**160**
Hereditas	0.545	40	945	32.1	27	**60**	**109**
Human Biology	1.474	58	2177	29.7	29	**65**	**102**
Development Genes Evolution	n/a	39	1481	28.6	19	**70**	**138**
Silvae Genetica	0.491	42	1232	27.2	15	**72**	**163**

Data based on JCR, 1996. A: Impact factor, B: number of 1996 papers, C: number of cited references, D: relatedness factor to *Genetics,* $R_{G\&i}$max; E: raw citation score (maximal of "to" or "from"), F: rank by D, G: rank by E. Journals in JCR "Genetics & Heredity" category in **bold.**

sidering citation scores, the pattern of relatedness is quite different. Table 2 includes 30 journals that give or receive the highest number of citations to or from *Genetics*. It is important to note the high ranks of multidisciplinary journals such as *Proceedings of the National Academy of Sciences of the USA*, *Nature*, *Science* and of very large non-genetics journals such as *Journal of Biological Chemistry*, *Journal of Bacteriology*, *Journal of Molecular Biology*. Among the journals in Table 2 one does not find smaller journals that are highly related to *Genetics* and included in the *JCR* "G & H" category such as *Journal of Genetics, Journal of Neurogenetics, Genetics Selection Evolution*, *Evolutionary Biology*, *Genes & Genetic Systems*. The proposed method is further illustrated when one compares the data for a few other journals included in *JCR*'s "G & H" category, by raw citation scores and by the RF (Table 3). It can be seen that all the journals have much higher ranks when sorted by RF rather than by raw citation scores. The differences in ranks of three "genetical" journals are noteworthy. These are not included in the "G & H" category. They are *Fungal Genetics and Biology—133 and 41 (the first number is the rank by raw citation score, the second is the rank by RF), Development Genes and Evolution*—138 and 70, *Silvae Genetica*—163 and 72. The RF ranks these genetics journals closer to *Genetics* than raw citation scores do. To illustrate the low information content of the latter, compare the data for two journals that are very different in size: *Journal of Biological Chemistry* and *Molecular Biology and Evolution* (Table 4). Though they give to and receive from *Genetics* similar numbers of citations, they are very different in relevance to *Genetics*, which is clearly reflected in the relatedness indexes: $R_{G>j}$ and $R_{j>G}$ are 3.8 & 2.6 and 109.2 & 142.1, respectively.

The small *Journal of Genetics* published in India is an interesting case. It is a journal with a low impact factor of 0.278. In 1996 it published only eight papers containing 390 cited references. It ranks 1st in the Table 1, but when sorted by raw citation score it ranks 49th. Of its 390 cited references 88 are to *Genetics* (that is 22.6%, while self-citation of *Genetics* is only 13.5%). It probably means that Indian scientists publishing in the *Journal of Genetics* frequently publish in *Genetics* as well and in their papers in *Genetics* frequently cite the papers they publish in *Journal of Genetics*. Evidently, this is not true for authors in other national journals such as the French *Genetics Selection Evolution*,

TABLE 4. Citation relatedness of two journals of different sizes, which cite or are cited by *Genetics* with similar numbers of citations.

Journal title	Citation scores: To	Citation scores: From	Number of papers	Number of references	$R_{j>G}$	$R_{G>j}$
Journal of Biological Chemistry	243	394	4949	209095	2.6	3.8
Molecular Biology and Evolution	386	306	133	6065	142.1	109.2

Data based on JCR, 1996. $R_{j>G}$ and $R_{G>j}$ are indexes of citing and cited relatedness of a journal "j" and *Genetics.*

TABLE 5. Citation relatedness of journals on human and medical genetics to *Genetics.*

Journal title	A	B	C	D	E	F	G
Human Molecular Genetics	6.512	278	10312	7.6	35	**152**	**89**
Immunogenetics	3.348	178	4756	6.1	13	**168**	**171**
Genetic Epidemiology	1.094	42	1230	5.7	5	**173**	**216**
Human Genetics	2.455	322	7857	5.1	18	**182**	**151**
Cancer Genetics and Cytogenetics	1.405	235	5359	2.1	5	**230**	**222**
Journal of Medical Genetics	2.263	226	5442	0.8	4	**266**	**245**
Clinical Genetics	0.996	158	2963	0	0	**271**	**271**

Data based on JCR, 1996. A: Impact Factor; B: number of 1996 papers; C: number of cited references; D: relatedness factor to *Genetics,* $R_{G\&i}max$; E: raw citation score (maximal of "to" or "from"); F: rank by D; G: rank by E. Journals in JCR "Genetics & Heredity" category in **bold.**

the Scandinavian *Genetica*, the British *Heredity* and the Russian *Genetika*.

It seems unexpected that *Genetics* is so weakly related to journals on human and medical genetics (see Table 5). Medical and clinical genetics journals probably should be listed in a separate *JCR* category to differentiate them from those in "Genetics & Heredity."

Conclusions and Future Work

Here we summarize the results of our study.

1. The new algorithmic approach enables one to find thematically related journals out of a multitude of journals.
2. Weighting citation data by journal size allows identifying journals that are similar in content better than unweighted raw citation data.
3. In the case of the starting journal *Genetics* the method identified those journals which are significantly genetic in content, but were not included in the "Genetics & Heredity" category of the *JCR*.
4. Journals included in the "G & H" category are rather heterogeneous in content. Some are highly related to *Genetics*, while others, as for example journals on medical genetics are poorly related to its content. There is a significant difference between subjects such as plant, animal, human and other aspects of genetics.
5. *JCR* has become an established world wide resource but after two or more decades it needs to reexamine its methodology for categorizing journals so as to better serve the needs of the research and library community.
6. Using the methods described *JCR* could provide additional options for its web version. *JCR*'s listings for cited and citing journals could provide a column with relatedness indexes ($R_{A>j}$ and $R_{j>A}$) and provide the option to sort by raw citation scores, relatedness indexes and relatedness factor just as it does now for the impact factor.

One might speculate on further usage of the suggested procedure, when it is computerized. Three applications come easily to mind.

1. Searching for relevant journals to form a small laboratory library. Specify a small set of journals (say, 3 to 5), which are undoubtedly relevant to the Lab's research profile. Pool up all the references contained in these journals and sum up numbers of papers in each of them, thus forming a pooled-up "macrojournal". [The idea was used by Cozzens & Leydesdorff, 1993, but had been earlier used by Garfield, 1986). Earlier journal citation studies too numerous to mention had used terms such as core, unit, group, or category, and coincided with the appearance of the first *JCR* in 1975 (Garfield, 1975).] Count the number of citations given to the macrojournal by all other available journals and received from the macrojournal by each of them. Calculate the RF of the macrojournal with all other journals.

   Sort the journals by the RF. Select a reasonable number of the journals with the highest ranks. These will constitute the desired set of journals most relevant to the Lab's research profile.
2. Determining the subject category for a journal, when it is not evident from the journal title. Perform the procedure for the journal under categorization, identical to that described above for the journal *Genetics*. The journals with the highest ranks (after sorting by RF) will characterize the semantic category of the journal under categorization.
3. Algorithmic categorization of journals according to a pre-specified set of subject categories. Set up the desired set of categories. Select for each category a small set of undoubtedly relevant (diagnostic) journals. Form a macrojournal for each category (as described above in item 1). For each journal to be categorized calculate RF. Sort the diagnostic macrojournals by RF (in relation to the journal under categorization). If the value difference of RF for the 1st and 2nd ranks is substantial, ascribe to the journal under categorization the category of the macrojournal ranked 1st. If the difference in RF values is not substantial, ascribe to the journal the categories of macrojournals ranked 1st and 2nd.

Acknowledgments

We wish to thank ISI™ for permission to use the *JCR* data for this study, and three anonymous referees for useful comments.

References

Cozzens, S.E., & Leydesdorff, L. (1993). Journal systems as macro-indicators of structural change in the sciences, In A.F.J. Van Raan, R.E.

de Bruin, H.F. Moed, A. J. Nederhof, & R.W.J. Tijssen (Eds.), Science and Technology in a Policy Context (Leiden: DSWO Press), 219–233.

Egghe, L., & Rousseau R. (2000). Partial orders and measures for language preferences. Journal of the American Society for Information Science, 51 (12), 1123–1130.

Egghe, L., Rousseau, R., & Yitzhaki, M. (1999). The “own-language preference”: Measures of “relative language self-citation.” Scientometrics, 45, 217–232.

Garfield, E. (1975). No-growth libraries and citation analysis; or, pulling weeds with ISI’s Journal Citation Reports. Current Contents No. 26, 5–8. Reprinted in Essays of an Information Scientist, Volume 2, pp. 300–303 (1975). Philadelphia: ISI Press. Available: http://www.garfield.library.upenn.edu/essays/v2p300y1974-76.pdf

Garfield, E. (1986). Journal Citation Studies. 46. Physical chemistry and chemical physics journals. Part 2. Core journals and most-cited papers. Current Contents No. 2, 3–10 (January 13, 1986). Reprinted in Essays of an Information Scientist, Volume 9, pp. 9–16 (1988). Available: http://www.garfield.library.upenn.edu/essays/v9p009y1986.pdf

Garfield, E. (1996). The significant scientific literature appears in a small core of journals. Scientist, 10 (17), 13–16. Available: http://www.the-scientist.com/yr1996/sept/research_960902.html

ISI Journal Citation Reports. http://www.isinet.com/isi/products/citation/jcr/index.html

Leydesdorf, L. (1994). The generation of aggregated journal-journal citation maps on the basis of the CD-ROM version of the Science Citation Index. Scientometrics, 31: 59–84.

Narin F., & Carpenter, M.P. (1973). Clustering of scientific journals. Journal of the American Society for Information Science and Technology, 24, 425–435.

Narin, F., Carpenter, M.P., & Berlt N. (1972). Interrelationships of scientific journals. Journal of the American Society for Information Science and Technology, 23: 323–331.

Narin F., Hamilton, K.S., & Olivasto, D. (2000). The development of science indicators in the United States. In B. Cronin & H. B. Atkins (Eds). The Web of Knowledge: A Festschrift in Honor of Eugene Garfield (pp. 337–360). Medford, NJ: Information Today.

Pudovkin, A.I. (1992). Citation links of the journal *Biologiya Morya* (Soviet Journal of Marine Biology). Biologiya Morya, No. 5–6, 83–92. (in Russian)

Pudovkin, A.I. (1993). Citation relationships among marine biology journals and those in related fields. Marine Ecology Progress Series, 100, 207–209.

Pudovkin, A.I., & Fuseler, E.A. (1995). Indices of journal citation relatedness and citation relationships among aquatic biology journals. Scientometrics, 32, 227–236.

Rousseau, R. (1999). Temporal differences in self-citation rates of scientific journals. Scientometrics, 44 (3), 521–531.

Shama, G., Hellgardt K., & Oppenheim C. (2000). Citation footprint analysis. Part I: UK and US chemical engineering academics. Scientometrics, 49 (2), 289–305.