23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

New analysis features of the CRExplorer for identifying influential publications

Andreas Thor[*], Lutz Bornmann[**], Werner Marx[***] and Rüdiger Mutz[****]

[*] *thor@hft-leipzig.de*
University of Applied Sciences for Telecommunications Leipzig, Gustav-Freytag-Str. 43-45, 04277 Leipzig (Germany)
[**] *bornmann@gv.mpg.de*
Administrative Headquarters of the Max Planck Society, Division for Science and Innovation Studies, Hofgartenstr. 8, 80539 Munich (Germany)
[***] *w.marx@fkf.mpg.de*
Max Planck Institute for Solid State Research, Information Service, Heisenbergstrasse 1, 70506 Stuttgart (Germany)
[****] *mutz@gess.ethz.ch*
ETH Zürich, Mühlegasse 21, 8001 Zurich (Switzerland)

Introduction

What are the most important publications in a research field? On which shoulders does the research stand? Reference Publication Year Spectroscopy (RPYS) has been developed for identifying the publications (here: cited references, CRs) with the greatest influence in a given paper set (mostly sets of papers on certain topics or fields). "RPYS is based on the analysis of the frequency with which references are cited in the publications of a specific research field in terms of the publication years of these CRs. The origins show up in the form of more or less pronounced peaks mostly caused by individual publications that are cited particularly frequently" (Marx, Bornmann, Barth, & Leydesdorff, 2014, p. 751).

The program CRExplorer[1] (Cited References Explorer) was specifically developed by Thor, Marx, Leydesdorff, and Bornmann (2016) for applying RPYS to publication sets downloaded from Scopus or Web of Science. Publication sets should represent a specific field (e.g., bibliometrics) or deal with a specific topic (e.g., research on Aspirin). Most RPYS publications published hitherto have focused on the history in a scientific field or topic (on the 19th and the first half of the 20th century). In the era of little science (before around 1950, see Marx & Bornmann, 2010) the number of CRs in a field or topic is comparatively low, which facilitates the identification of important contributions.

However, in the big science period, the growth of literature leads to numerous CRs whereby the important contributions are difficult to identify. For purposes of analysing the complete range of contributions, we developed new advanced statistics for the CRExplorer. These statistics are able to identify and characterize the CRs which have been influential across a longer period (many citing years). The new statistics are demonstrated using all the papers

[1] www.crexplorer.net

published in *Scientometrics* between 1978 and 2016. The indicators N_TOP50, N_TOP25, and N_TOP10 can be used to identify those CRs which belong to the 50%, 25%, or 10% most frequently cited publications (CRs) over many citing publication years. In the *Scientometrics* dataset, for example, Lotka's (1926) paper on the distribution of scientific productivity belongs to the top 10% publications (CRs) in 36 citing years.

It is the general objective of the statistics to identify influential papers in the set of CRs on the selected field or topic. The identified cited publications were highly cited over a longer period or at certain time points (shortly or several years after publication). Thus, influence or importance of publications is measured across the publication period of the citing papers.

Methods

At the beginning of 2017, we downloaded 5,506 papers from Scopus (including CR data), which were published in *Scientometrics* between 1978 and 2016. We considered all document types. Since we were interested in the impact of specific CRs over several citing publication years, we decided to use this publication set as an example in this study. *Scientometrics* is the oldest journal dedicated to the field of scientometrics (starting in 1978). Other journals in the field of scientometrics (e.g., *Journal of Informetrics*) offer only significantly shorter periods.

Before we started to analyse the data, we revised the dataset in several steps. First, we restricted the uploaded range of CRs from 1900 to 2005 and cleared the dataset of variants of the same CR using the matching and clustering facilities by CRExplorer. This results in n=44,123 CRs. Second, we deleted all CRs for which the bibliographic information did not match the categorization used by CRExplorer, e.g., CRs without authors or with obviously wrong author information. Furthermore, some variants of the same CRs have been manually aggregated. This procedure led to the final dataset of n=33,812 CRs.

Results

CRExplorer has been initially programmed to identify the most influential reference publication years (RPYs) – the peak years – and the CRs (cited publications) which essentially produced the peaks. Here, the impact of the CRs is measured across all citing publications in the imported dataset. Since as a rule most of the impact is generated in the first three to five years after publication, many publications are important for only a few years after publication. However, some citation classics or landmark papers influence the research in a certain field or on a certain topic over a longer period. Thus, it is additionally interesting to identify those exceptional publications (top publications), which are important (influential) over many citing years. The functionality of the CRExplorer has been extended to facilitate this objective.

We start by explaining the statistics for identifying the period of citing years. Suppose CR X has been published in 1980 and cited at least once in 1980, 1981, 1983, 1984, and 1985 but not in 1982. The first new indicator in the CRExplorer, named **N_PYEARS**, is equal to the number of years in which a CR has been cited. Thus, CR X has been cited in five citing years, N_PYEARS=5. The user of the CRExplorer should be aware that the number of citing years is defined by the publication years (PYs) of the citing publications. For example, a CR from 1990 can only be cited in 10 years (and not in 20 years), if the underlying dataset includes (citing) publications from 2000 to 2009. In order to call the attention of the CRExplorer user to these limitations defined by the range of citing years, the status bar shows not only the

range of the RPYs, but also the range of the citing publication years in the dataset (maximal number of citing years).

The second new indicator in the CRExplorer – named **PERC_PYEAR** – is the percentage of citing years in which the CR has been cited. Thus, N_PYEARS is divided by the maximal number of citing years (i.e., all PYs with at least one citation to a CR in RPY) to yield PERC_PYEAR. In the example above, CR X has been cited in 5 out of 6 citing years, thus, PERC_PYEAR = 5/6 ≈ 83%. PERC_PYEAR highlights those CRs which received at least one citation in many possible citing years.

We are further interested in those CRs which have been cited more frequently in the citing years than other CRs in the dataset. In order to identify these CRs, thresholds are computed which identify the top 50%, top 25%, and top 10% in single citing years. In the first step of the computation, the CRs counts in one citing year are sorted in ascending order. In the second step, the thresholds for the top 50%, 25%, and 10% CRs counts are determined in a given year. In the third step, those CRs are identified which are above the three thresholds. In the fourth step, the numbers of citing years are counted in which the CRs are above the thresholds. These numbers yield **N_TOP50**, **N_TOP25**, and **N_TOP10**.

In the *Scientometrics* dataset (see Table 1), Lotka's (1926) paper on the distribution of scientific productivity and de Solla Price (1963) book "little science, big science" are those publications with the highest number of years in which they have been cited by other publications (N_PYEARS=36). Furthermore, both publications appear at the top of the table in the CRExplorer if the CRs are sorted by the column N_TOP10. It follows Garfield (1979) on citation indexing with N_TOP10=34 at the third position. However, the percentages in the column PERC_PYEAR point out that they have not all been cited in all possible years. The book by de Solla Price (1963) has been cited in 36 out of 38 publication years (PYs) and, thus, PERC_PYEARS=94.74%. Though the overall range of publication years of all citing publication comprises 39 years (1978-2016), there is no publication with PY=1978 citing a CR with RPY=1963. Similarly, cited references from RPY=1979 have not been cited in 1978 and 1979. The value of PERC_PYEARS for Garfield (1979) is therefore 34/37=91.89%. Schubert and Braun (1986) have a higher value for PERC_PYEAR compared to Garfield (1979) but a lower value for N_PYEARS because its maximum range of citing years comprises 31 years only (1986-2016) and it has been cited in all of these years.

Table 1. Ten influential publications on research published in *Scientometrics* with the highest number of citing years in which they belong to the 10% most frequently cited publications (identified by the references cited in *Scientometrics* papers).

Cited reference	Title	Publication medium	N_CR	N_PYEARS	PERC_ PYEARS	N_TOP10
Lotka (1926)	The frequency distribution of scientific productivity	*Journal of the Washington Academy of Sciences*	155	36	100.00	36
de Solla Price (1963)	Little science, big science	Book	216	36	94.74	36
Garfield (1979)	Citation indexing: its theory and application in science, technology, and humanities	Book	151	34	91.89	34

Small (1973)	Co-citation in the scientific literature: a new measure of the relationship between two documents	*Journal of the American Society for Information Science*	162	33	84.62	33
Cole and Cole (1973)	Social stratification in science	Book	74	32	82.05	32
Schubert and Braun (1986)	Relative indicators and relational charts for comparative assessment of publication output and citation impact	*Scientometrics*	120	31	100.00	31
Garfield (1972)	Citation analysis as a tool in journal evaluation: journals can be ranked by frequency and impact of citations for science policy studies	*Science*	109	31	79.49	31
Small and Griffith (1974)	The Structure of Scientific Literatures I: Identifying and Graphing Specialties	*Science Studies*	69	31	79.49	31
Narin (1976)	Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity	Book	48	30	76.92	30
Merton (1968)	The Matthew effect in science	*Science*	115	30	76.92	30

Notes. N_CR=Number of CRs counts, N_PYEARS=Number of years in which the CR has been cited, PERC_PYEAR=Percentage of years in which the CR has been cited, N_TOP10= Number of citing years in which CRs belong to the 10% most frequently cited publications. The "NPCT Range" is set to 0 in the CRExplorer.

The list of influential publications in Table 1 does not only contain publications which are ground-breaking in bibliometrics, such as the paper by Schubert and Braun (1986), the paper by Small (1973) about the introduction of the of co-citation approach, and the first published journal ranking on the basis of the JIF (Garfield, 1972); it also lists classics from the sociology of science. These include the introduction of the Matthew effect (Merton, 1968) and the explanation of the consequences which result from the social stratification system in the scientific community (Cole & Cole, 1973).

Besides the question of identifying exceptionally influential publications it is also of interest to identify the citation dynamic of selected CRs (Bornmann, Ye, & Ye, 2017). Usually, cited publications have a lifetime with the following dynamic: starting with low citations in the first year of publication, growing up to a maximum of citations a few years later, followed by a continuous decrease of citations several years after publication (Redner, 1998). However, other dynamics are also possible: a more or less long period of non-recognition with low citations is followed by a period with high citations after a sudden peak. Such a dynamic is typical for the phenomenon named "sleeping beauty" (van Raan, 2004). These are publications "whose importance is not recognized for several years after publication" (Ke, Ferrara, Radicchi, & Flammini, 2015, p. 7426).

In order to identify statistically the citation dynamics of CRs with CRExplorer, we apply Configural Frequency Analysis (CFA, Stemmler, 2014; von Eye, 2002; von Eye, Mair, & Mun, 2010). CFA is a categorically statistical procedure to reveal configurations in multivariate cross-classifications (i.e., contingency tables). All CRs for a certain RPY and the publication years of the citing publications are cross-classified within a CR×PY matrix. A matrix cell contains the number of CR counts in the citing year PY. In the case of systematic citation dynamics (e.g., at sleeping beauties or lifetime cycles) the cell values in the matrix may deviate strongly from expectations. Expected values are frequencies which would occur if there is no relationship between or independency of the rows (CRs) and the column variable (PY). These expected frequencies are calculated by the CRExplorer. The resulting *z*-values from the comparisons of observed and expected frequencies are standard normally distributed with mean value of zero and standard deviation of 1.0. High positive or negative *z*-values identify cells which strongly deviate from the independency-base model, and thus indicate a certain citation dynamic.

In order to reveal specific **sequences over time**, rows of cells (CR) are considered with average ("0"; $-1 \leq z \leq 1$), above average ("+"; $z > 1$), and below average ("-"; $z < -1$) cells. For example, the sequence of de Solla Price (1963) book starts with "0+00--00" indicating a high number of citations in the second year after publication and comparatively low number of citations in the fifth and sixth year. Based on the sequences, **types of CRs** in terms of different citation dynamics are identified by CRExplorer (see Thor, Bornmann, Marx, & Mutz (2018) for a formal definition including example computations).

- "Sleeping beauties" have low or no citations over a longer initial period and high citations later;
- "Constant performers" have a constant and considerable amount of citations over time;
- "Hot papers" receive a high number of citations directly after the publication and low citations later;
- "Life cycles" have courses of different annual citations across time.

In the Scientometrics dataset, Lotka's (1926) paper is classified as "constant performer" and de Solla Price (1963) book reveals a "life cycle" sequence.

Discussion

In RPYS, CRs of publication sets are analysed to identify the most important contributions in the past. RPYS reveals quantitatively which historical papers are of particular importance for a given publication set. In this study, we have presented some new advanced statistics to identify and characterize CRs which have been influential across a longer period (several citing years). The indicators N_TOP50, N_TOP25, and N_TOP10 can be used in CRExplorer to inspect those CRs with (significantly) higher impact than comparable CRs from the same RPY. The analysis of the example dataset revealed, for example, that the paper by Lotka (1926) entitled "The frequency distribution of scientific productivity" belongs to the 10% most frequently cited publications in 36 citing years. However, papers such as Lotka (1926), are exceptions; many publications show citation distributions which are characterized by changes in citation impact intensities over the citing years. Therefore, CRExplorer analyses the sequence of citations across the given citing years to identify different types. The sequence is used by the program to identify CRs with typical impact distributions. For example, hot papers have early, but not late impact.

References

Bornmann, L., Ye, A. Y., & Ye, F. Y. (2017). Sequence analysis of annually normalized citation counts: An empirical analysis based on the characteristic scores and scales (CSS) method. *Scientometrics, 113*(3), 1665-1680.

Cole, J. R., & Cole, S. (1973). *Social stratification in science*. Chicago, MA, USA: The University of Chicago Press.

de Solla Price, D. J. (1963). *Little science, big science*. New York, NY, USA: Columbia University Press.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation: journals can be ranked by frequency and impact of citations for science policy studies. *Science, 178*(4060), 471-479.

Garfield, E. (1979). *Citation indexing - its theory and application in science, technology, and humanities*. New York, NY, USA: John Wiley & Sons, Ltd.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *PNAS, 112*(24), 7426-7431.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 12*, 317 - 323.

Marx, W., & Bornmann, L. (2010). How accurately does Thomas Kuhn's model of paradigm change describe the transition from a static to a dynamic universe in cosmology? A historical reconstruction and citation analysis. *Scientometrics 84*(2), 441-464.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology, 65*(4), 751-764. doi: 10.1002/asi.23089.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56-63.

Narin, F. (1976). *Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ, USA: Computer Horizons.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B, 4*(2), 131-134.

Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics, 9*(5-6), 281-291.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265-269. doi: 10.1002/asi.4630240406.

Small, H., & Griffith, B. C. (1974). The Structure of Scientific Literatures I: Identifying and Graphing Specialties. *Science Studies, 4*(1), 17-40.

Stemmler, M. (2014). *Person-Centered Methods - Configural Frequency Analysis (CFA) and Other Methods for the Analysis of Contingency Tables*. Heidelberg: Springer.

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016). Introducing CitedReferences Explorer (CRExplorer): A program for Reference Publication Year Spectroscopy with Cited References Standardization. *Journal of Informetrics, 10*(2), 503-515.

Thor, A., Bornmann, L., Marx, W., & Mutz, R. (2018). Identifying single influential publications in a research field: new analysis opportunities of the CRExplorer. *Scientometrics, 116*(1), 591-608.

van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics, 59*(3), 467-472.

von Eye, A. (2002). *Configural Frequency Analysis: Methods, Models and Applications* Mahwah: Lawrence Erlbaum.

von Eye, A., Mair, P., & Mun, E.-Y. (2010). *Advances in Configural Frequency Analysis*. London: The Guilford Press.