Journal of Informetrics 8 (2014) 1–12

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Classification of individual articles from all of science by research level

Kevin W. Boyack[a,*], Michael Patek[b], Lyle H. Ungar[c], Patrick Yoon[c], Richard Klavans[d]

[a] SciTech Strategies, Inc., Albuquerque, NM 87122, USA
[b] SciTech Strategies, Inc., Keene, NH 03431, USA
[c] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA
[d] SciTech Strategies, Inc., Berwyn, PA 19312, USA

ARTICLE INFO

Article history:
Received 9 September 2013
Received in revised form 16 October 2013
Accepted 16 October 2013
Available online 9 November 2013

Keywords:
Research level
Basic science
Applied science
Multinomial logistic regression model
Article-level classification

ABSTRACT

A system of four research levels, designed to classify scientific journals from most applied to most basic, was introduced by Francis Narin and colleagues in the 1970s. Research levels have been used since that time to characterize research at institutional and departmental levels. Currently, less than half of all articles published are in journals that been classified by research level. There is thus a need for the notion of research level to be extended in a way that all articles can be so classified. This article reports on a new model – trained from title and abstract words and cited references – that classifies individual articles by research level. The model covers all of science, and has been used to classify over 25 million articles from Scopus by research level. The final model and set of classified articles are further characterized.

1. Introduction

There are a variety of means that have been introduced to characterize the scientific literature and the actors who produce it along different axes. For example, output is reported using publication counts while impact is reported using citation counts and other citation-based measures. Disciplinary profiles are used to characterize the content produced by institutions and department. Keywords and controlled vocabularies (such as MeSH), while intended primarily for information retrieval, are used to characterize the detailed topic space for individual researchers. Metrics of different types abound.

One perhaps underutilized way of characterizing scientific literature is classification of journals or articles as basic or applied. Narin, Pinski, and Gee (1976) introduced a system of four research levels, ranging from most applied to most basic, classified journals by research level, and used those classifications to characterize institutional research. Research levels have the potential to aid in the characterization of translational pathways in medicine and other sciences (Cambrosio, Keating, Mercier, Lewison, & Mogoutov, 2006). Currently, less than half of all articles published are in journals that been classified by research level. There is thus a need for the notion of research level to be extended in a way that all articles can be so classified. We have thus undertaken this study to develop a new model that will allow individual articles to be classified by research level across all of science.

* Corresponding author. Tel.: +1 505 856 1267.
E-mail addresses: kboyack@mapofscience.com (K.W. Boyack), mpatek@gmail.com (M. Patek), ungar@cis.upenn.edu (L.H. Ungar), me@patyoon.com (P. Yoon), rklavans@mapofscience.com (R. Klavans).

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.10.005

Table 1
CHI research levels and exemplars.

RL	Biomedical definition	Non-biomedical definition	Example journals
1	Clinical observation	Applied technology	British Medical Journal Journal of Urology Energy and Fuels IEEE Transactions on Industrial Electronics International Journal of Product Research Journal of Environmental Management
2	Clinical mix	Engineering-technological mix	New England Journal of Medicine International Journal of Cardiology Key Engineering Materials Materials Science and Engineering A Computers & Industrial Engineering Scientometrics
3	Clinical investigation	Applied research	Blood Journal of Immunology Applied Physics Letters Journal of Physical Chemistry B Journal of Chemical Education Journal of the American Statistical Society
4	Basic research	Basic scientific research	Journal of Biological Chemistry Nature Journal of the American Chemical Society Physical Review E Social Studies of Science

This article proceeds as follows. First, the history and use of research levels is set forth. A short description of multinomial logistic regression models, the type of model used in this study, is then presented. This is followed by a description of the various detailed models that were trained along with characteristics of the result sets. The article concludes with a characterization of the final model and of the classification of all Scopus documents by research level using the model, along with potential implications associated with their use.

2. Background

Narin et al. (1976) and CHI Research introduced a classification scheme representing the basic-to-applied spectrum when they classified 900 biomedical journals into four research levels (RL). Journals were assigned to a RL based on a combination of expert knowledge and citation patterns. The citation pattern portion was based on the assumption that clinical research would cite basic research, but that the reverse would not be true. For example, given the types in Table 1, journals in RL1 would cite journals in RL2, RL3, and RL4, but journals in RL4 would only cite other RL4 journals.

The CHI research level classification system was expanded in the 1980s to include journals in the physical sciences (Carpenter et al., 1988). Although additional journals have been added to the research level list at various times, of the nearly 20,000 source titles (journals, conference proceedings, etc.) available in Scopus, only around 4200 have assigned research levels (Boyack & Klavans, 2011).

Since their introduction, RL have been used to characterize research along the basic-to-applied continuum for academic institutions (Carpenter et al., 1988; McAllister & Narin, 1983), departments and teams (Bordons & Zulueta, 1997), and the pharmaceutical industry (Narin & Rozek, 1988). They have been correlated to the attraction of funding (with basic research favored) in Australia (Butler, Biglia, & Bourke, 1998), and have been used to characterize arthritis-related articles with different funding acknowledgment types (Lewison & Devey, 1999). Bordons, Gomez, Fernandez, Zulueta, and Mendez (1996) correlated research level with collaboration type (local, domestic, international) for a number of biomedical research areas, finding that research level was slightly higher (more basic) for international collaborations. It has also been shown that the majority of the biomedical papers cited by industrial patents are from the basic science category (McMillan, Narin, & Deeds, 2000), while most papers cited by clinical guidelines are from the two most applied categories (Grant, Cottrell, Cluzeau, & Fawcett, 2000).

When using the CHI set of research level classifications, three deficiencies must be acknowledged – these deal with drift, uniformity, and coverage.

- Drift: Most of the CHI classification system dates to 20 years or more. It is not only possible, but likely, that some journals have shifted their focus over time. However, the CHI system does not account for drift.
- Uniformity: All papers in a single journal are assumed to be of the same RL regardless of their actual level. It is highly likely that there is a mix of research levels in most journals.