Measures of novelty in biomedical literature

Shubhanshu Mishra, Vetle I. Torvik
(smishra8, torvik)@illinois.edu
Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign, Champaign, IL - 61820

We introduce several measures of novelty for a scientific article in MEDLINE based on the concepts associated with it. The concepts associated with an article are identified using the Medical Subject Headings (MeSH) assigned to the article. A temporal profile was computed for each MeSH term (and the combination of pairs of MeSH terms) based on their overall occurrences in MEDLINE, after which papers are labeled by their most novel MeSH [see Figure 1] and pairs of MeSH as measured in years and volume of prior work. Our approach is similar to earlier attempts aimed at measuring novelty of an article, e.g. by using the frequency of co-citations [2] and co-occurrence of keywords [1], however, it differs in its usage of pairwise concepts and a control vocabulary of MeSH terms. We use pair of concepts for quantifying novelty of an article because in principle all scientific publications present some novel concepts, however, it is rare for articles to coin new concepts which are widely adopted by the community. Furthermore, the pairing of existing concepts is quite common in science, this hypothesis is confirmed through our analysis. Across all papers in MEDLINE published since 1985, we find that individual concept novelty is rare (5.4% of papers have a MeSH <= 3 years old; 1.2% have a MeSH <= 20 papers old), while combinatorial novelty is the norm (55% have a pair of MeSH <= 3 years old; 78% have a pair of MeSH <= 20 papers old) [see Figure 2].

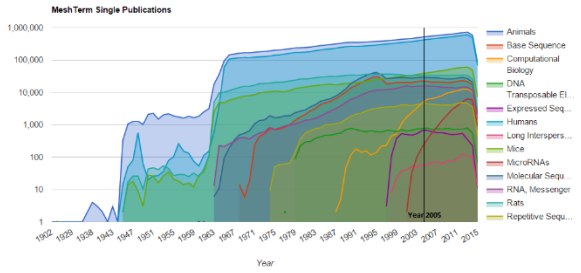

Figure 1 Profile of individual MeSH terms on article with PubMed Id 15922829 published in 2005

In order to operationalize our novelty measures, we model the growth in occurrence of each MeSH term using a logistic model, identifying phases of burn-in, accelerated, decelerated, and constant growth. These growth pattern reflect that 26.6% of the articles in MEDLINE have at least one MeSH term in an acceleration phase where as 73.3% have all their MeSH terms in a deceleration phase.

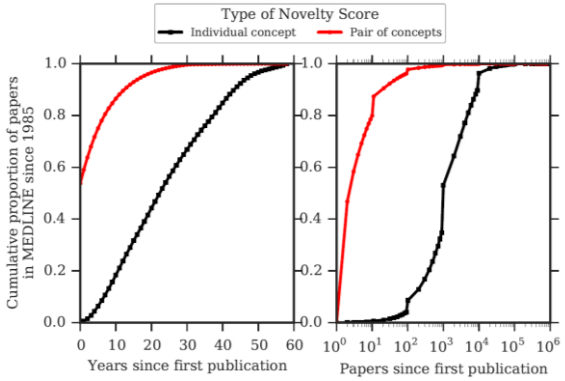

Figure 2 Cumulative distribution of novelty scores for 15.72M (15.71M with at-least a pair of MeSH terms) articles published in MEDLINE since 1985. Lower is more novel.

Our novelty measures are positively correlated with citations rates, after accounting for the journal effect, but they are not strongly predictive. Furthermore, articles on more novel individual concepts are cited more than those which are only novel on pair of concepts. The correlation of our novelty measures with author age is more complex: of authors with > 50 papers about 90% had increasing individual novelty scores over their career on average, but the variability also increased. This probably reflects that a more diverse publication strategy comes with experience. The result also align with the findings of [1] where the authors argue that younger authors publish more novel work. However, the split is nearly 50/50 for combinatorial novelty and there is little, if any, correlation between the author age and the time-point of their most novel work. A web tool is available at http://abel.lis.illinois.edu/gimli/novelty for browsing the temporal profile of MeSH terms on an article, and the change in novelty of an author over their career.

References

[1] PACKALEN, M. and BHATTACHARYA, J., 2015. Age and the Trying Out of New Ideas. *National Bureau of Economic Research Working Paper Series No. 20920*. DOI= http://dx.doi.org/10.3386/w20920.

[2] UZZI, B., MUKHERJEE, S., STRINGER, M., and JONES, B., 2013. Atypical Combinations and Scientific Impact. *Science 342*, 6157 (2013-10-25 00:00:00), 468-472. DOI= http://dx.doi.org/10.1126/science.1240474.