MEASURES OF NOVELTY IN BIOMEDICAL LITERATURE

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

Introduction

Research Question:
How to quantify conceptual novelty of articles published in the biomedical literature?

Data:

- **18M** MEDLINE articles annotated with MeSH (a controlled vocabulary)
- **25k** unique MeSH terms
- **1B** unique MeSH pairings yearly

Methods

Measures of novelty and growth:
Novelty scores **capture the age of a concept** (or pair of concepts) as measured in **years (or number of prior articles) since its first appearance** in MEDLINE.

- **Time novelty** is based on the number of years since the first appearance of a MeSH term (or a pair of MesH) in a MEDLINE
- **Volume novelty** is based on the number of articles since the first appearance of a MeSH term (or a pair of MesH) in MEDLINE

For a given article, the minimum age (in year and papers) across its MeSH terms is assigned.

Modeling temporal profile of individual concepts:
A logistic growth model $f(t)$ can capture 4 typical phases of a concept: novel, accelerated growth, decelerated growth, and saturation:

$$f(t) = \left[\frac{b}{1 + e^{-(c+dt)}}\right]$$

where instantaneous log-normalized-volume = $f(t)$, growth = $f'(t)$ and acceleration = $f''(t)$.

We identified the following phases in a concept's profile:

- **Burn-In Phase**: Topic is new, publication rate is small, and growth is marginal.
- **Accelerating Growth Phase**: Topic is bursting, publication rate is rapidly increasing.
- **Decelerating Growth Phase**: Publication rate is still increasing but is starting to stabilize.
- **Constant Growth Phase**: Growth is marginal and publication rate has stabilized.

Results

Pairwise scores are better at capturing novelty

- **Very few articles are on a novel individual concept:** (5.4% ≤ 3 year old and 1.2% ≤ 20 papers old)
- **The majority of articles are on novel pair of topics:** 55% ≤ 3 year old and 78% ≤ 20 papers old.
- **The majority papers are on decelerating growth phase of their most novel individual concept:** **27%** on accelerated growth and **73%** on decelerated growth.

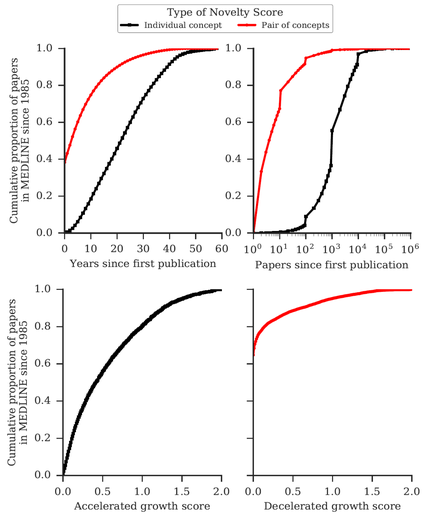


Fig 2. Distribution of novelty (top) and growth (bottom) scores for articles published since 1985.

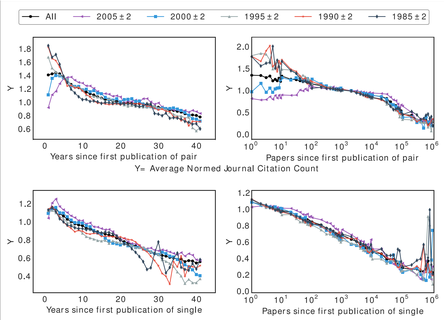


Fig 4. Novelty scores correlated with average journal normalized citation count.

Novelty is positively correlated with impact

- Novel articles are cited slightly more than other articles published in the same journal and the same year
- The effect is higher for articles on individual than pair of concepts.

The average individual novelty of the majority of authors decline as their career progresses

- An author's most novel work can come at any point in their career but is slightly less likely to occur very early.
- However, most authors (> 90%) have a declining novelty on average.
- For the data on careers we used authors with more than 50 papers during 1985-2009 years in the Author-ity 2009 dataset [1]

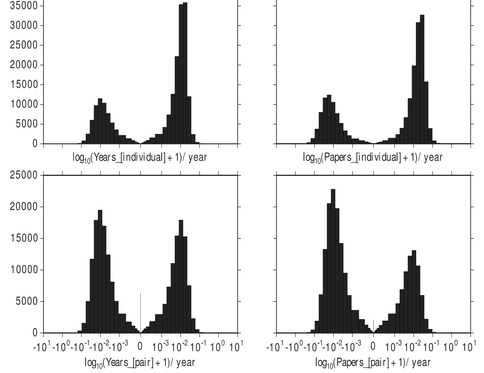


Fig 3. Distribution of model slope of the minimum novelty score versus author's career age. (Authors with more than 50 articles)

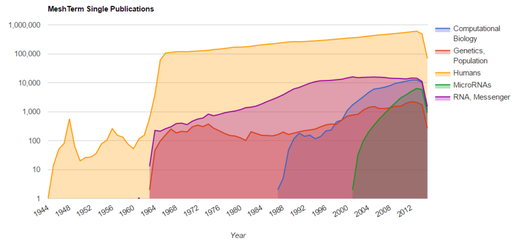
References:
[1] Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, *56*(2), 140–158

Table 1. Estimated parameters of various models fitted to predict the log(Citations) of an article.

	Full Model		Time Model		Volume Model	
Features	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**
Log(Individual concept age (year))	-0.1634	0.002	-0.1676	0.002	-	-
Log(Pair concept age (year))	-0.0307	0.001	-0.0048	0.001	-	-
Log(Individual concept age (year))^2	-0.046	0.001	-0.0475	0.001	-	-
Log(Pair concept age (year))^2	-0.0151	0.001	-0.0077	0.001	-	-
Log(Individual concept age (volume))	0.0021	0.006	-	-	-0.1007	0.005
Log(Individual concept age (volume))^2	-0.0011	0.001	-	-	0.0053	0.001
Log(Pair concept age (volume))	0.0393	0.001	-	-	0.0344	0.001
Log(Mean journal citation in year)	0.9222	0.001	0.9227	0.001	0.9338	0.001
(Number of MeSH terms)^0.5	0.0415	0	0.0405	0	0.0442	0
Intercept	-0.642	0.009	-0.6107	0.003	-0.7097	0.008

Discussion

An interface for inspecting novelty scores for all PubMed articles and identifying their most novel MeSH terms across various categories.
http://abel.lis.illinois.edu/gimli/novelty



Category	Age in years	Age in papers
Chemicals	MicroRNAs (3)	MicroRNAs (331)
InfoSci	Computational Biology (17)	Computational Biology (15,412)
Organisms	Humans (60)	Humans (8M)

Category-Category	Age in years	Age in papers
Chemicals-Chemicals	MicroRNAs - RNA, Messenger (1)	MicroRNAs - RNA, Messenger (50)
Chemicals-InfoSci	Computational Biology - MicroRNAs (1)	Computational Biology - MicroRNAs (11)
Organisms-Chemicals	Humans – MicroRNAs (1)	Humans – MicroRNAs (63)
Organisms-InfoSci	Computational Biology -Humans (16)	Computational Biology -Humans (2,470)

Fig 5. The temporal distribution of all MeSH terms from a particular MEDLINE article (top) and the most novel (individual and pair) MeSH terms present in the article (bottom).

Acknowledgement

Research reported in this publication was supported in part by the National Institute on Aging of the NIH (Award Number P01AG039347) and the Directorate for Education & Human Resources of the NSF (Award Number 1348742).