Evaluating the outcomes of government funded research programs: measuring interdisciplinarity through text analysis of abstracts of award-derived publications

John Chase, Christina Freyman and John Byrnes

Many government programs fund scientific research with a specific goal of increasing the interdisciplinarity of research. Since 2002, the Directorate for Geosciences and the Division of Mathematical Sciences, within the Directorate for Mathematical and Physical Sciences, have made about 300 awards totaling almost $100 million in the crosscutting Collaboration in Mathematical Geosciences (CMG) program. The stated purpose of the CMG program is to enable collaborative, cutting-edge research at the intersection of mathematical sciences and geosciences, and to encourage cross-disciplinary training of researchers with skills in both the mathematical (and/or statistical) sciences and the geosciences. Proposals were required to include at least one geoscientist and one mathematical scientist, and the research topic was to include an intrinsic need for a non-trivial collaboration for research on geoscience topics.

To analyze bibliographic data, the SRI team applied two different and independent analysis techniques to abstracts, co-authors, and cited references to evaluate the CMG program. Many program evaluations analyze Web of Science journal classifications to measure the interdisciplinarity of program-related publications' cited references. We used this technique to quantify and measure interdisciplinarity by analyzing the subject category diversity of cited references' journals of publications derived from CMG awards. This technique found that derived publications' references are from a wider variety of Web of Science journal subject categories than publications produced by the same researchers before and after the CMG awards.

To complement this well-accepted bibliometric method, the SRI team leveraged its expertise in artificial intelligence and topic co-clustering to analyze the text of publication abstracts. SRI used topic co-clustering methods to analyze publication abstracts supported by NSF's Collaboration in Mathematical Geosciences program. The term co-clustering method of association-grounded semantics was used to analyze the abstracts of the analysis groups. This method first generates term clusters, and then represents each abstract as a probability distribution over those term clusters. To create a "geoscience standard," AGS was used to create a probability distribution based on the abstracts of all of the geoscience comparison award-supported researchers. The geoscience standard is the expected probability distribution of term clusters across all of these publications. The "math standard" was created in the same way. Once the math and geoscience standards were calculated, the probability distributions for each analysis group's publications in the before, derived, and after periods were compared to the standard using the Kullback-Leibler divergence technique, a method for measuring the difference between probability distributions. The divergence of two probability distributions, P and Q, of a discrete random variable is defined to be:

$$D_{KL}\left(P\|Q\right) = \sum_{i} P\left(i\right)\ln\frac{P(i)}{Q(i)}$$

With this technique, we expect that the mathematical scientists' abstracts will have a smaller divergence from the math standard than from the geoscience standard.

The co-clustering algorithm compared the terms in each analysis group's publication abstracts to the clusters of terms of the "standard". One standard was defined by a corpus of math publications from researchers supported by the math comparison awards; and one was defined on the corpus of geoscience publications from researchers supported by the geoscience comparison awards. Each publication was compared to that standard and the divergence in the distributions of terms was measured. Analysis with respect to before-award, derived, and after-award found that publications produced by CMG-supported geoscientists and attributed to CMG awards used more terms associated to the math standard than did publications produced by the same researchers before and after the CMG awards. In addition, the publications produced by the CMG-supported mathematical scientists attributed to CMG awards used more terms associated with the geoscience standard than did publications produced by the same researchers before and after the CMG awards.

This technique was developed for this evaluation but could be applied to many subjects, and adds an additional dimension to a purely citations-based bibliometric analysis by showing the change in the terms researchers use in their abstracts.