Machine-learning based classification of research grant award records

Christina Freyman*, John Byrnes, and Jeffrey Alexander

*christina.freyman@sri.com
SRI International (USA)

Background

Policy makers frequently ask agencies involved in scientific research and technology development to report how much money they are spending on research and development activities in specific fields or topics. As noted by Wallace and Rafols (2015), these sets of investments are characterized frequently as "research portfolios," where an agency's budget is classified into groups based on key interests: socio-economic objective, discipline, program area, etc. Therefore, federal R&D funding agencies have a specific interest in performing portfolio analysis to answer these questions. In addition, interest in portfolio analysis has increased recently due to a growing consensus to make the management of research portfolios more "scientific" and the growing ability of technology and data to enable a more quantitative approach to portfolio analysis and management (Srivastava et al. 2007).

Keyword searches are a common way some agencies find awards on a specific topic, and could therefore be a classification tool for R&D portfolio analysis. However, as Srivastava et al. point out, "in the absence of a government-wide ontology to describe research and development activities, the use of keyword searches to identify individual projects that meet a given criteria (e.g., nanotechnology) may return inconsistent results"(Srivastava et al. 2007). A tagging system, structured and applied in a consistent manner, may be a better system for retrieving information relevant to different queries about government research and development activity. One way to assign documents to a controlled vocabulary is to label documents when they are created. In this approach, however, the individual applying the tags may not be aware of all possible stakeholder interests in the document, and therefore may not select the most salient tags. There would also likely be problems of inaccuracy and inconsistency across multiple taggers. In principle, automated tagging using a single methodology would have the advantage of high consistency when compared to tagging by many different people.

This work explores how machine-learning techniques could be used to automatically classify NSF awards using pre-determined tagging schemes for scientific disciplines or for socioeconomic objectives using the words contained in the abstracts of the grant record. We use the metadata (Directorate, Division, and Program) to validate the results, and do not access the metadata as part of the automated tagging process.

Method

As described above, our goal here was to classify NSF grant abstracts automatically into standard classifications for scientific disciplines and socioeconomic objectives. The automation used unsupervised machine learning, so we started without any tagged abstracts. We classify the abstracts into two different taxonomies: by socio-economic objective, and by scientific discipline. We do so by taking the terms in each taxonomy and creating a "language model" around each term, using descriptors or associate terms taken from external reference standards. We then take the abstracts and cluster them into 200 term clusters and 200 document clusters using a method called "Association-Grounded Semantics" (Byrnes and Rohwer, 2005).

We use the pointwise mutual information to formalize the measurement of how specific a set of terms is to a set of documents. The classification proceeds by calculating a distance (technically, a divergence) between each cluster of abstracts and each language model. We treat each abstract cluster as a probability

distribution over term clusters, and we treat each language model as a probability distribution over these same term clusters. Terms in the language model that do not appear in the abstracts are dropped.

Results

The results show that in the case of scientific disciplines, where our language models were well-formed and we had a reliable comparison set for manual classification, the machine assigned tags were a reasonable and reliable means for describing the research conducted under each grant. In assigning socio-economic objectives to grants, we saw relatively poor precision and recall in classification, due to the poorly-formed and sparse language models available for those terms. Our analysis suggests that this approach can be used to classify large corpora of scientific awards into desired categories.

References

Byrnes, J. and Rohwer, R. (2005), 'Text Modeling for Real-Time Document Categorization', *Aerospace Conference, 2005*. IEEE, 1-11.

Srivastava, Christina Viola, Towery, Nathaniel Deshmukh, and Zuckerman, Brian (2007), 'Challenges and opportunities for research portfolio analysis, management, and evaluation', *Research Evaluation*, 16 (3), 152-56.

Wallace, Matthew L. and Rafols, Ismael (2015). 'Research portfolio analysis in science policy: moving from financial returns to social benefits', *Minerva*, April. Published online 4 April 2015.