NATURE | NEWS

Researchers aim to chart intellectual trends in arXiv

'Culturomics' team pivots from Google Books to scientific preprints.

Eric Hand

24 February 2012

When physicist Paul Ginsparg goes to next week's American Physical Society meeting in Boston, Massachusetts, he plans to take with him a 64-gigabyte flash drive containing all 740,000 or so articles from arXiv, the preprint repository he founded in 1991 that is managed by Cornell University in Ithaca, New York.

He will pass the data on to researchers from the Cultural Observatory at Harvard University in Cambridge, Massachusetts. They want to break down the full text of the articles into component phrases to see how often a particular word or phrase appears relative to others — a measure of how 'meme-like' a term is. Their goals: to give arXiv a new tool for identifying original source papers in physics, mathematics and computer science — and to enable historians to spot trends from the 20 years that the repository has existed.

Kris Snibbe/Harvard University

Jean-Baptiste Michel (front) wants to use arXiv to track how scientific language has changed.

“How do you find the moment when a given scientific transformation occurred?” asks Jean-Baptiste Michel, co-director of the Cultural Observatory and a postdoctoral researcher in psychology at Harvard. “You can help the reader figure out where in time the most relevant papers were located, which has always been difficult to do.”

Word games

The Cultural Observatory team has already won acclaim for applying a similar approach to 5 million books in the Google Books database. Using a tool the researchers call the *n*-gram viewer, they calculated the rate at which irregular verbs regularize, and charted how 'corporate-speak' made its way into the vernacular (see 'Culturomics: Word play'). But the Google Books database carries with it a major limitation: because many of the works are under copyright, users cannot be pointed to the actual source material.

As a result, the researchers are concentrating on adapting the interface to analyse freely available data sets, such as arXiv. They have applied the new interface, which they call Bookworm, to about 1 million copyright-free books collected by the Open Library, adding an ability to screen for books by genre and place of publication. They have already tested the tool using one month of arXiv data, but plan to add the full arXiv data set in the coming weeks. Michel is excited not only because a new group of users will be testing the tool, but also because the knowledge embodied in arXiv is different. “It might show different patterns than pop culture,” he says.

One of Bookworm’s creators, Benjamin Schmidt, a graduate student in history at Princeton University in New Jersey, wants to mine arXiv’s articles on quantitative finance to see if the adjectives surrounding the Black–Scholes equation — used to set prices for financial derivatives — changed before and after the 2008 market crash.

Ginsparg says that he would use the tool to chart trends he might expect in high-energy physics: a quickening pulse of papers citing the Higgs boson, for example, or a peak in papers about supersymmetry, a theory which may soon be waning (see 'Beautiful theory collides with smashing particle data'). He says that the tool could advance the nascent field of “information genealogy” by spotting the moments when neologisms are invented, and when scientific fields merge and diverge.

“Some historians will be unbelievably fascinated to look at this retrospectively,” Ginsparg says. “What you’re going to be registering are intellectual movements in the community.” He suggests that science policy-makers could even use Bookworm to identify new fields that are in need of funding, or moribund fields that might require fewer grants.

Once they have analysed the arXiv repository, Schmidt and his colleagues won't be short of new data sets to mine next. Examples include not only general-interest sources such as newspapers, but also scientific ones such as PubMedCentral, an online repository containing some 2.3 million biomedical articles.