A Hard Science is Good to Find

Textual Similarity as a Measure of Scientific Paradigm Development, A Preliminary Investigation

Susan Biancani
Stanford University School of Education
Stanford, CA 94131
biancani@stanford.edu

Abstract

The notion of a "hierarchy of sciences", in which academic fields can be ordered from "hard" to "soft", is an old one. However, efforts to develop a measure of hardness of a field—the field's level of paradigm development—to date have not been highly successful. Here I explore the possibility of using a text-based similarity measure to quantify the extent of consensus in a field, which is theorized to correlate with hardness.

1 Introduction

Thomas Kuhn's 1962 *The Structure of Scientific Revolutions* advances the notion that a scientific field is characterized by a paradigm—a guiding set of assumptions, methods and values that shape how research in the field is conducted and evaluated, and what constitutes appropriate objects of study[1]. Kuhn argues that different fields are characterized by different levels of paradigm development. In low-paradigm fields, little consensus exists on the important questions in the field or the best methods with which to investigate them; research proceeds in fits and starts, and new findings may not build directly on prior findings. This description tends to fit fields in the social sciences. By contrast, high-paradigm fields—often those in the natural sciences—show much greater agreement on methods and research questions; there is often a race to publish important results, out of fear of getting "scooped." New findings build directly on—or challenge—prior findings, allowing knowledge to accumulate rapidly. High-paradigm fields have elsewhere been described as "high-consensus", "rapid-discovery", and "progressive[2]. Sociologists of science have attempted to characterize fields according to their level of paradigm development, but many efforts to date have not been satisfying. Here I explore the use of a text-based similarity metric as a measure of the level of cohesion—and thus paradigm development—in a field.

2 Prior Work

Auguste Comte first advanced the notion of a "hierarchy of sciences" in the nineteenth century. Since then, much work on this question has come from the field of bibliometrics. Derek de Solla Price developed an "Immediacy Index", which showed faster rates of obsolescence of findings in the natural sciences than in the social sciences [3]. However, this metric was later shown to be an artifact of the differing volumes of work produced in a given time interval in different fields [4]. Cole [5] summarized findings from seven different approaches seeking to find a variable that reliably correlated with widespread perceptions of paradigm development, but found none that did; he concluded, "there are no systematic differences between sciences at the top and at the bottom of the hierarchy in either cognitive consensus or the rate at which new ideas are incorporated" (p. 111).

Promising work comes from Susan Cozzens [6], who demonstrated an intriguing pattern in a detailed qualitative study. Cozzens compared citations of two highly influential papers: one in neuropharmacology, and one in sociology of science. Cozzens finds that citations to the neuropharmacology paper varied over time: early on, citations mentioned either the main finding or other peripheral findings, and frequently commented on experimental technique. Later papers developed a formulaic citation of the main finding, suggesting that this finding had been vetted, and was now taken as fact. In contrast, no such shift was observed in citations of the sociology of science paper. Cozzens attributes this to the fact that very few citing papers at any time referred to the main finding; more often, the paper was abstracted, and was mentioned as an example of a larger trend. Cozzens findings are enlightening, but as a close, qualitative study cannot easily be extended to new fields.

One technique that has successfully distinguished low-paradigm from high-paradigm fields, and which has the potential to be replicated automatically, is the measure of "fractional graph area" (FGA). FGA is a measure of the total fraction of page space in a given article that is taken up by graphs. Smith et al. [7] hypothesized that papers from higher-paradigm fields would be characterized by higher FGA. In doing so, they drew on Latour's assertion that graphs distinguish science from non-science [8]. Graphs are a highly encoded means of communication; they can present a large amount of information in a compact form, because they build on a vast quantity of shared knowledge between the writer and the reader. Much information is embedded in a graph without elaborate explanation; it is assumed that the reader has sufficient prior familiarity with the form of a graph to be able to extract the new finding quickly. Thus, the use of graphs captures much of the nature of a high-paradigm field. In a random sample of 50 articles from each of 30 journals, Smith et al. found that FGA does indeed correlate with scientists' perceptions of the level of paradigm development of seven fields.

Smith et al. relied on prior coding by William Cleveland [9] who measured the FGA of the papers used in the sample. Cleveland describes the process as "detailed and intensive" (261). Clearly, it would be useful to develop an automated measure of paradigm development.

Here, I use the distance to a set of nearest-neighbor papers as an indicator of paradigm. I hypothesize that in a high-paradigm field, a paper will be close to its nearest neighbor: it speaks directly to them, and may share methods or an empirical setting. In a low-paradigm field, published papers may be less closely related to an existing literature. Thus, I expect that papers in harder sciences will be closer, on average, to their nearest neighbors than papers in softer sciences.

3 Data

This study is based on a dataset collected at Stanford University, covering the years 1993-2007. The corpus includes 66,000 abstracts of all papers published by Stanford faculty members. Here, I restrict the study to seven departments: physics, chemistry, biology, medicine, psychology, economics, and sociology. Multiple teams of researchers have found that these fields are widely perceived to be ranked for paradigm development in the above order, with physics showing the highest level of development, and sociology the lowest [7, 8, 9, 10].

Table 1: Descriptive Statistics for the Corpus

Department	People	Publications with Abstracts	Publications with Citations	Keywords	Papers / Person	Keywords / Paper
Physics	43	996	972	42	23.16	0.042
Chemistry	34	1,367	1,073	60	40.21	0.044
Biology	62	1,669	1,681	104	26.92	0.062
Medicine	298	3,931	2,325	139	13.19	0.035
Psychology	47	582	558	95	12.38	0.163

Department	Vocab. Size	Vocab. / Abstracts	Unique Citations	Citations / Publication

Physics	2,942	2.954	1,953	2.009
Chemistry	4,394	3.214	2,786	2.596
Biology	5,720	3.427	3,520	2.094
Medicine	11,911	3.030	4,799	2.064
Psychology	2,327	3.998	1,467	2.629
Economics	1,548	4.324	1,239	3.755
Sociology	670	5.076	680	4.024

A few facts stand out about the distribution of papers in the corpus. First, the departments vary widely in size: both in terms of faculty members and in output. Second, social science departments produced far fewer papers than the natural/physical sciences. The social sciences, and especially Sociology, have a greater diversity of keywords relative to the number of papers produced than the hard sciences. In general, the higher-paradigm fields publish more papers per person than the lower-paradigm fields. Finally, relative to the number of papers they publish, the higher-paradigm fields use a smaller set of keywords, a smaller vocabulary size, and fewer unique citations than the lower-paradigm fields. Interestingly, Medicine, which is rated in the middle on paradigm development, scores lower on these measures than some higher-rated fields.

4 Measuring Paper Similarity

I characterize each paper in two ways: according to the text in its abstract, and according to the references it cites. In both cases, I use a tf-idf approach—giving more weight to rare terms or to rare citations—to generate a term or citation vector for each paper. I then compute the cosine similarity between all possible pairs of papers in a given discipline. For each discipline, I report the average distance from each paper to each of its 50 nearest neighbors.

5 Results: Average Similarity by Department

Figure 1 shows the results from the tf-idf analysis (a), and from citations (b). The tf-idf analysis shows the following ordering: Medicine, Physics & Chemistry (superimposed), Biology, Psychology, Economics, Sociology. The citation analysis shows: Medicine, Chemistry, Economics, Physics, Biology, Psychology, Sociology. It is surprising in both cases that Medicine shows the highest similarity between neighboring papers. It's possible that this effect is in part driven by the vast size of this field relative to the others. With more papers, there's a higher chance of finding a very similar one. The size of the Medicine corpus may also influence some of the measures discussed above: vocabulary size relative to output, relative count of unique citations, and relative count of keywords.

Figure 1: Similarity to 50 Closest Papers using tf-idf of (a) terms, and (b) citations.

6 Conclusion

Text-based and citation-based tf-idf similarity appear to be promising options with which to measure the level of paradigm development of a scientific field. Further work is needed to shed light on how the size of the corpus representing a field influences my measures of interest. In addition, the fact that my data are drawn from a single university is a serious limitation; the departments at this university may be idiosyncratic and may fail to accurately represent the entire discipline. In future work, I plan to extend this analysis to a more complete dataset drawn from the ISI Web of Science database.

7 Acknowledgements

This is paper relies on data collected by the Mimir Project. The Mimir Project is conducted at Stanford University by Daniel A. McFarland, Dan Jurafsky, Chris Manning, and Walter Powell, and has been generously funded by the Office of the President at Stanford University and by NSF Award #0835614. Special thanks also to Kurt Thorn for advice on high-performance computing, and to the staff of Stanford CS229 for suggestions and feedback.

8 References

[1] Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

[2] Collins, Randall. 1994. "Why the social sciences won't become high-consensus, rapid-discovery science." *Sociological Forum*. 9 (2): 155-177.

[3] de Solla Price, Derek J. 1970. "Citation Measures of Hard Science, Soft Science, Technology, and Non Science", in Carnot E. Nelson and Donald K. Pollack (eds), *Communication Among Scientists and Engineers* Lexington, MA: D.C. Heath, 3-22.

[4] Cole, Stephen, Jonathan R. Cole and Lorraine Dietrich. 1978. "Measuring the Cognitive State of Scientific Disciplines." In Yehuda Elkana, Joshua Lederberg, Robert K. Merton, Arnold Thackray and Harriet Zuckerman (eds), *Toward a Metric of Science.* New York: John Wiley & Sons, 209-51.

[5] Stephen Cole. 1983. "The Hierarchy of the Sciences?" *American Journal of Sociology*. 89: 111-39.

[6] Susan E. Cozzens. 1985. "Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science." *Social Studies of Science*. 15 (1): 127-53.

[7] Smith, Laurence D., Lisa A. Best, D. Alan Stubbs, John Johnston, Andrea Bastiani Archibald. 2000. "Scientific Graphs and the Hierarchy of the Sciences: A Latourian Survey of Inscription Practices." *Social Studies of Science*, 30 (1): 73-94.

[8] Latour, Bruno. 1990. "Drawing Things Together", in Michael Lynch and Steve Woolgar (eds), *Representation in Scientific Practice* Cambridge, MA: MIT Press, 1990: 19-68.

[9] Cleveland, William S. 1984. "Graphs in Scientific Publications." *American Statistician.* 38: 261-69.

[10] Lodahl, Janice Beyer and Gerald Gordon. 1972. "The Structure of Scientific Fields and the Functioning of University Graduate Departments." *American Sociological Review*. 37: 57-72.

[11] Biglan, Anthony . 1973. "Relationships Between Subject Matter Characteristics and the Structure and Output of University Departments." *Journal of Applied Psychology*. 57: 204-13.

[12] Ashar, Hanna and Jonathan Z. Shapiro. 1990. "Are Retrenchment Decisions Rational? The Role of Information in Times of Budgetary Stress." *Journal of Higher Education*. 61: 123-41.