Coming to Terms: A Discourse Epistemetrics Study of Article Abstracts from the Web of Science

Bradford Demarest[1], Vincent Larivière[2], Cassidy R. Sugimoto[3]

[1] *bdemares@indiana.edu*
Indiana University – Bloomington, School of Informatics and Computing, Bloomington, IN (United States)

[2] *vincent.lariviere@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, Montreal, QC (Canada)

[3] *sugimoto@indiana.edu*
Indiana University – Bloomington, School of Informatics and Computing, Bloomington, IN (United States)

Abstract
This study investigates the relative power and characteristics of a set of social and epistemic terms to distinguish among disciplines of research article abstracts, using a corpus of 928,572 abstracts from 13 disciplines indexed by Web of Science in 2011. Applying the machine-learning approach to discourse epistemetrics using a sequential minimal optimization (SMO) algorithm, and a feature set of terms derived from Hyland's (2005) metadiscourse studies per Demarest and Sugimoto (2014), the current paper reports subsets of terms that best (and least) distinguish among disciplines, finding that the terms least able to distinguish among disciplines are rarely used and overwhelmingly adjectival or adverbial markers of authorial attitude, reflecting personal positioning, while terms best able to distinguish disciplines are mostly verbs frequently used as engagement markers, framing the generation of knowledge for the readership in ways that are standardized within disciplines (while varying among them). We plan to analyze the findings of the current research-in-progress from discipline-based as well as term-based perspectives, incorporating both into a two-mode network, as well as incorporating finer grained data for specific specializations to compare with the current higher-level disciplinary findings.

Conference Topic
Methods and techniques, altmetrics

Introduction
Understanding and depicting the relationships among different academic realms (whether disciplines, fields, specialisms, or a host of other divisions using some combination of social, epistemological, and institutional aspects) is a well-studied subarea of scientometric (Leydesdorff & Rafols, 2009). Initial forays into modeling disciplinary differences based on a core set of social and epistemic terms have yielded potentially promising results (Demarest & Sugimoto, 2013; Demarest & Sugimoto, 2014). However, no studies to date have used computational approaches to compare the abilities of specific social and epistemic terms to distinguish among disciplines. The current work-in-progress seeks to enact such a comparison, using a machine-learning approach to derive term differences between pairs of disciplines and by extension between a given discipline and all other disciplines under study. In finding the social and epistemic terms that best distinguish among academic disciplines, we hope to open new dimensions of analysis of the sciences through their texts.

Literature Review
There have been very few previous attempts to map the relatedness of academic disciplines based upon common social and epistemic terms. However, previous research of social and epistemic discourse usage in different academic disciplines as well as previous studies of document, journal, author, and discipline similarity or relatedness based on a variety of other measures guide the current study.

Differences in how academic disciplines employ language that positions the author in relation to the reader, the text itself, and previous scholars and works have been studied under various monikers, including stance (Biber & Finegan, 1989), metadiscourse (Hyland & Tse, 2004), appraisal (Martin & White, 2008), and attitude (Halliday, 1985). For the most part these differences have not been studied using automated quantitative methods (although cf. Argamon and Dodick, 2004), and in no cases have the resulting metrics been used as a basis for mapping the relatedness of disciplines. The current study draws upon Hyland's (2005) study of metadiscourse in a number of different disciplines, leveraging a set of words and phrases that Hyland (2005) found to be widely occurring in academic writing as our feature set for machine learning-based modeling of term differences among disciplines.

Previously, scholars have sought to map science based upon patterns of co-citation (Boyack, Klavans, & Börner, 2005) as well as topic, via ISI subject headings (e.g., Leydesdorff & Rafols, 2009). Other studies of similarity or relatedness have sought to compare multiple kinds of networks, including "bibliographic coupling, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks" (Yan & Ding, 2012, p. 1313). While the current work-in-progress focuses on a single type of similarity, it is with the intention of eventually adding to and comparing with these previously established measures of comparison. Furthermore, in order to create results that are comparable to previous work, we will also draw our data from the Web of Science, focusing specifically on the genre of scholarly articles, and use the high-level subject categories (although in future iterations of this study we hope to look at both higher and lower-level subject categories).

Methods

The current study analyzes all journal article abstracts from 13 disciplines contained in the Web of Science from 2011, totaling 928,572. Table 1 provides an overview of disciplines and counts of abstracts in the data corpus.

Table 1. Counts of abstracts by discipline.

Discipline	Abstracts
Engineering and Tech	172949
Biomedical Research	153166
Chemistry	129685
Physics	121702
Biology	93765
Earth and Space	70018
Mathematics	42685
Social Sciences	40463
Professional Fields	34590
Health	28343
Psychology	25802
Humanities	13673
Arts	1731
TOTAL	**928572**

For each abstract, relative frequencies were computed for 307 words or phrases taken from Hyland (2005). These terms fall into one or another of the following categories: hedges, boosters, attitude markers, engagement markers, and self-mentions. Hedges (e.g., "perhaps", "possible", "approximately") mitigate the certainty of an assertion, while boosters (e.g., "clearly", "obvious") amplify it. Attitude markers, such as "unexpectedly" or "unfortunately", frame assertions affectively, expressing the author's emotion regarding the

asserted facts, as distinct from their assurance of the facts' certainty. Engagement markers (such as "the reader" and "you", but also imperative verbs such as "consider" or "observe") address the reader explicitly or implicitly, and guide the reader to specific social and epistemic framing of an assertion (e.g., as an externally observable fact or as an idea intended for mental simulation). Finally, self-mentions, such as "I", "we", or "the author", serve as means for authors to insert themselves into the text, either as subjective actors or as social players (whether alone or as part of an authorial cohort).

After preparing the data, the Sequential Minimal Optimization algorithm (SMO) (Platt, 1998), a support-vector model classifier implemented in the WEKA v3.6.6 tool (Hall et al., 2009), was employed to create models distinguishing between each pair of disciplines based on the socio-epistemic features' relative frequencies. The resulting term weights for each model of discipline pairs were then normalized across the model, such that the absolute values of weights for a given discipline pair model would sum to 1. Model-normalized weights for each term were then averaged for each discipline across all discipline pairs for which the given discipline was a pair member. For the sake of standardization, negative term weights indicate a positive correlation with a given discipline (i.e., the more frequently the term appears in a text, the more likely this text belongs to the given discipline), while positive term weights indicate a negative correlation (i.e., the more frequently the term appears in the text, the less likely this text belongs to the given discipline).

Results

Due to space limitations, we eschew reporting the full 307 term set of results, focusing instead on the terms that most and least distinguish among disciplines. We discern these terms based upon the standard deviation of model-normalized average weights, as terms that discern well among disciplines will result in strong positive as well as negative weights, depending on which discipline is being modeled, while terms whose weights have small absolute values will in turn have smaller standard deviations, as all weights approach the 0 point.

Table 2 reports the 20 terms with the highest standard deviations of model-normalized average weights, as well as the 20 terms with the lowest standard deviations. While the results might at first blush suggest that the terms with the lowest standard deviations are part of a universal academic discourse, it is worth noting that many of the terms in the Bottom 20 list are exceedingly rare in the sample – out of 928,572 abstracts, "unbelievable" appears in 3 of them (although "shockingly" also appears in 3 abstracts; however, "unbelievable" is found in 2 engineering abstracts and one humanities abstract, suggesting that the scant data that exists shows no distinction between two otherwise fairly different disciplines). Also worth noting is that any terms that appeared in no abstracts at all are eschewed from the reported results.

However, the bottom 20 terms do provide some information about scholarly writing across the disciplines – the vast majority of these terms (19 out of 20) act as attitude markers; given the wide range of adjectives and adverbs available to describe the affective state of the author (and given that adjectives and adverbs are linguistic "open classes", i.e., new words can and are generated for these classes regularly), it is not surprising that such terms would be diffuse, rare, and not strongly indicative as individual terms.

Pivoting to consider the top 20 terms, the first notable characteristic is that where the bottom 20 terms tend toward adjectives and adverbs (as well as attitude markers), 19 of the top 20 terms are either self-mentions or engagement markers (and the latter for the most part are verbs). While nouns and verbs are also linguistic open classes, the use of verbs to describe the epistemic frame of scientific work here as well as the terms with which scientific authors refer to themselves can be seen to be more standardized within disciplinary communities, whereas the attitude markers of the bottom 20 terms are more personalized. The indicative

strength of self-mentions such as “we”, “my”, and “author”, as well as verbs like “argue” and “measure” also resonates with previous findings of Demarest and Sugimoto (2014), with “argue” and “my” serving as a strong indicator of philosophy and “measure” and “we” a better indicator of psychology and physics in dissertation abstracts as well.

Table 2. The top and bottom 20 social and epistemic terms for distinguishing among disciplines (ranked by standard deviation).

Top 20		Bottom 20	
Term	Standard Deviation	Term	Standard Deviation
we	0.009848	shockingly	0.0009166
argues	0.009686	view	0.0008793
prove	0.009614	disappointed	0.0008707
argue	0.009098	astonishingly	0.0008043
author	0.009063	!	0.0007801
showed	0.008494	incontestable	0.0007541
about	0.008138	knowledge	0.0007406
let	0.008044	incontrovertible	0.0007283
proved	0.008019	presumable	0.0007005
my	0.007908	unclearly	0.0006577
recall	0.007684	desirably	0.0006524
estimate	0.007646	amazed	0.0006068
review	0.007592	disappointingly	0.0006046
measure	0.007268	uncertainly	0.0004573
pay	0.007173	undisputedly	0.0003956
thought	0.007102	unbelievably	0.0003247
claims	0.006978	incontrovertibly	0.0002968
consider	0.006879	incontestably	0.0002821
shown	0.006687	astonished	0.0002649
set	0.006672	unbelievable	0.0001121

Another aspect of the findings to consider is that while the standard deviation values derive from the full set of model-normalized average weights, in some circumstances high standard deviation values can derive from a single outlier, while in others it derives from a more uniform spread of weights. Figure 1 depicts the model-normalized average weights for the top 20 terms ranked by standard deviation. Visual inspection reveals terms whose weights are more uniformly distributed (e.g., “author”), which suggest that they may serve as robust terms to distinguish among a variety of disciplines, while other terms (e.g. “let”, “prove”, and “proved”) serve as strong indicators of a single outlier discipline, with all other disciplines much more tightly clustered. As it happens, the terms “let”, “prove”, and “proved” provide a strong indication of mathematics as they occur more frequently in a text, in contrast to all other disciplines.

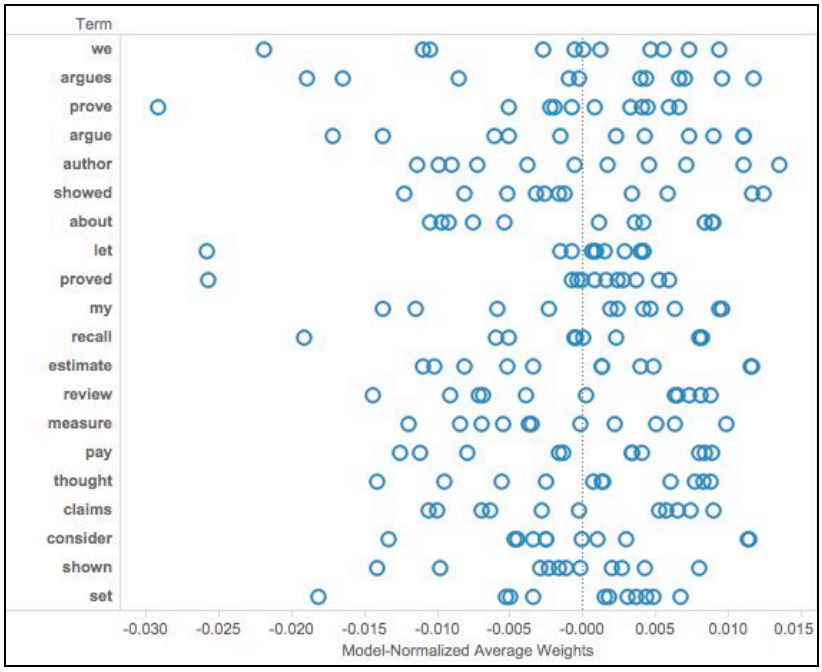

Figure 1. Model-normalized average weights (Top 20, ranked by standard deviation).

Future Directions

While the results of the current study-in-progress have focused on summary ranking and overall patterns of distribution of weights per term, our next goals in the near term are to more deeply tease apart trends as they appear for single disciplines as well as groups of disciplines, including the traditional groupings of soft vs. hard and pure vs. applied (Biglan, 1973). Further, we can derive overall measures of similarity among disciplines from the overall accuracy measures of the machine-learning models from which these terms are taken (per Demarest & Sugimoto, 2014), or more ambitiously we could seek to cast disciplines and terms in a bipartite network, to more fully grasp the interplay between different disciplinary communities and the words they use.

More distantly, we intend to use this same approach, in light of patterns and trends perceived at the current level of aggregations, to consider specializations, so that we may ask questions such as how broad the social and epistemic spread of specialized areas of study are within disciplines – are some disciplines more socially or epistemically diverse, and others more centralized? Do these degrees of variety reflect patterns of fragmentation and specialization in subject area? It is questions such as these that compels the current research-in-progress.

References

Argamon, S. & Dodick, J. (2004). Conjunction and modal assessment in genre classification: A corpus-based study of historical and experimental science writing. In *AAAI Spring Symposium on Attitude and Affect in Text*. Retrieved from http://www.aaai.org.proxyiub.uits.iu.edu/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-001.pdf?origin=publication_detail

Biber, D. & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, *9*(1), 93–124.

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology*, *57*(3), 195.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374. doi:10.1007/s11192-005-0255-6

Demarest, B. & Sugimoto, C. R. (2013). Interpreting epistemic and social cultural identities of disciplines with machine learning models of metadiscourse. *In Proceedings of ISSI 2013 (Vol. 2, pp. 2027–2030)*. Vienna. Demarest, B., & Sugimoto, C. R. (2014). Argue, observe, assess: Measuring disciplinary identities and

differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23271

Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2011). Language and Ideology in Congress. *British Journal of Political Science*, *42*(01), 31–55. doi:10.1017/S0007123411000160

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum International Publishing Group.

Hyland, K. & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics*, *25*(2), 156–177.

Klavans, R. & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455–476.

Leydesdorff, L. & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362.

Martin, J. R. & White, P. R. R. (2008). *Language of Evaluation: Appraisal in English* (First Edition.). Palgrave Macmillan.

Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. In *Advances in Kernel Methods - Support Vector Learning*. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4376

Yan, E. & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326. doi:10.1002/asi.22680