Application for participation at the Doctoral Forum

Last name: Demarest

First name: Bradford

Institutional affiliation (university/department/link to website): Indiana University, School of Informatics and Computing, Department of Information and Library Science (http://www.ils.indiana.edu/)

Full address of the doctoral student including phone and fax numbers and email address:

503 E. Dixie St.
Bloomington, IN 47401
P: (917) 744-6125
F: (812) 855-6166
Email: bdemares@indiana.edu

Names of the supervisor(s) (links to their websites):
Cassidy Sugimoto (http://ella.slis.indiana.edu/~sugimoto/index.php)
Susan Herring (http://ella.ils.indiana.edu/~herring/index.html)

(in total max 1500 words)
Description of doctoral research project (including research questions, theoretical background, planned methodology, current status)

Scholarly text as a data source has provided scientometrics scholars with a rich source of information about various phenomena. Citation and co-citation analysis provide the quantitative data through which we investigate the influence of scholars, texts, journals, and fields, as well as patterns of the diffusion of ideas; word, co-word, and topic analyses provide similar informational bases for studying the flow and association of ideas to one another, as well as their rise and fall in popularity. Using these measures, information scientists have mapped science (e.g., based upon co-citation, such as Boyack, Klavans, & Börner, 2005, or upon topic, per Leydesdorff & Rafols, 2009).

However, while these two general branches have yielded much research into how the scientific community composes itself through reference, and what the scientific community studies and in association with what else, other kinds of research questions still remain to be pursued, and can still be answered using heretofore untapped aspects of academic texts. Methodologically I draw upon theories of social and epistemological differences among disciplines as developed by sociologists and philosophers of science. These include Becher and Trowler's (2001) synthesis of previous typologies of science into hard-soft and pure-applied as well as considering the population density of research areas, and the effects of such density and centralization upon how research and writing are done, echoing Whitley's (1984) conceptualization of the social and intellectual characteristics of various academic work, as well as Toulmin's (1972) philosophical framework of sciences as evolutionarily driven entities whose characteristics derive from the topic as well as the kind of work able to be done in the area.

Given this theoretical basis, in my current line of investigation I study social and epistemological stance (Biber & Finegan, 1989) in language, or what

has been variously termed metadiscourse (Hyland & Tse, 2004), appraisal, (Martin & White, 2008), and attitude (Halliday, 1985), how these stances differ among cultural groups, and what underlying beliefs, social structures, and values these stances expose in the specific groups being studied. Previous studies of these aspects of language have relied upon the manual collection of data for corpus linguistics; I seek to develop a research method in this area of study I have christened "discourse epistemetrics" (Demarest & Sugimoto, 2014), a quantitative multi-dimensional metric that uses a computational structural approach to both measure differences among cultures' language patterns and expose the specific features (currently words and phrases, although I would like to incorporate syntactic information as well for greater insight) whose relative frequencies designate a given group (up to this point I have been investigating disciplines, although I intend to expand my research to encompass other levels of granularity, such as specializations or disciplinary clusters). Eventually I would also like to investigate social and epistemological stance as it exhibits and varies among non-academic cultural groups whose orientation requires establishing facts (e.g., religious or political groups that base ethical or policy decisions upon some combination of guiding rules and worldly knowledge, as filtered through the knowledge-shaping apparatus of the groups' social structures).

Research Questions. The realms of questions I investigate in the current line of research all center upon investigating disciplines' characteristics (and more specifically the characteristics that are exposed contrastively).

1) How similar/different are academic disciplines as reflected in the socio-epistemic language used in academic writing? While my current line of research has focused on comparing data from contemporary samples of different disciplines, this question could be transposed to pertain to other academic cultural groups (e.g., specializations) or to the same discipline at different points in time.

2) Which specific socio-epistemic terms are most indicative of one discipline's writing in relation to another's, by their increased or decreased frequencies? My research up to this point has leveraged a list of common social and epistemic words and phrases from Hyland's (2005) research, but this question could be expanded based on semantic relations mined from lexical databases such as WordNet (Fellbaum, 1998), among other resources. As with RQ1, although my current line of research has sought to measure these differences among disciplines at a given historical moment, these same measures could be collected and considered at different levels of granularity as well as over time.

Planned Methodology: My current method is to collect an array of relative probabilities for each of 307 social or epistemic terms taken from Hyland (2005) encompassing terms that amplify or mitigate the certainty of an assertion, terms that frame the author or position the reader vis-à-vis an assertion, and that draw in the reader or author explicitly or implicitly to invoke social identity. These arrays are collected for each text under investigation, which data are then used to train a binary support-vector machine (the sequential minimal optimization (SMO) algorithm (Platt, 1998) as implemented in WEKA v.3-6-10 (Hall et al.,

2009)), which is then evaluated using ten-fold cross validation. The accuracy scores from these evaluations are then used as inverse measures of similarity (based on the intuition that disciplines which are more similar in their socio-epistemic term make-up will be hardest to distinguish between, even with an optimized model, based on the chosen features). This similarity measure is then used as arc weights in a network to show relative social and epistemological associations among disciplines (answering RQ1).

RQ2 is addressed by extracting and interpreting feature weights from each of the SVM models. The absolute value of a SVM feature weight reflects the strength of a given term in distinguishing between categories, while the sign (positive vs. negative) of the weight indicates which of the categories is indicated by the increased frequency of the term.

Current Status: I've conducted a study of dissertation abstracts for three disciplines (physics, philosophy, and psychology) whose preliminary results were presented at ISSI 2013 (Demarest & Sugimoto, 2013), and whose final results appear in Demarest and Sugimoto (2014). I'm currently working on an expanded study of research article abstracts for 13 disciplines drawn from the Web of Science; a study of the distributions and clustering characteristics of feature weights across this multiplicity of disciplines is currently under submission to ISSI 2015 as a research-in-progress short paper.

Motivation for student participation at the Doctoral Forum and the issues you wish to receive feedback on from the senior researchers.

I am motivated to participate in the Doctoral Forum at ISSI 2015 because I am interested in the development and application of this method and this area of research as a potential dissertation topic, and would like feedback and guidance on the further development of and investigation of the efficacy of my method with regard to my topic of study, and with regard to its placement in the broader context of scientometrics (i.e., does it belong here? What considerations might I need to take into account to integrate it more fully methodologically?). Although my approach has yielded some promising results preliminarily, I'm concerned that I may be re-inventing the wheel without realizing it. I would also look forward to speaking with senior researchers in the field of scientometrics about what resources or approaches to data collection they might be able to recommend, as I would very much like to be able to expand my corpus of texts from sets of abstracts to sets of full texts.

References

Becher, T., & Trowler, P. R. (2001). Academic Tribes and Territories: intellectual enquiry and the cultures of disciplines (2nd edition). Retrieved September 6, 2012, from http://eprints.lancs.ac.uk/3714/

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

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

Fellbaum, C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT 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.

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

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum International Publishing Group. Retrieved from http://books.google.com/books?hl=en&lr=&id=jgfgHpEqPN8C&oi=fnd&pg=PR8&dq=epistemic+metadiscourse+discipline&ots=60Ofr4zJRL&sig=ZYs8Z8BASLtXs3GGExodyDm07h0

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

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. doi:10.1002/asi.20967

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

Toulmin, S. (1972). *Human understanding*. Oxford: Clarendon Press.

Whitley, R. (1984). *The intellectual and social organization of the sciences*. New York, NY: Clarendon Press.