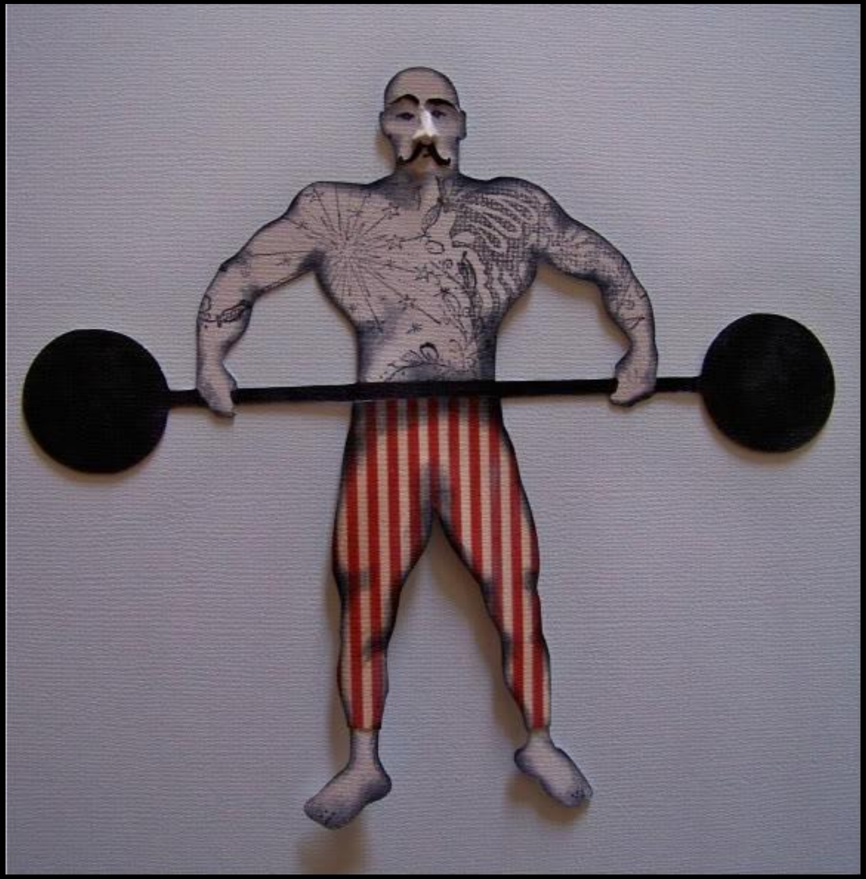
Comparative discourse epistemetrics: Research article abstracts and full texts

Bradford Demarest
METRICS 2015



Discourse epistemetrics

- A quantitative approach to studying social and epistemic differences between knowledge-oriented communities
  - Different to what degree? In what ways?

- Previously used with abstracts for dissertations and articles (Demarest & Sugimoto, 2015; Demarest, Larivière, & Sugimoto, 2015)

In short

Academic communities have different social structures, and different ways of making and testing knowledge.

These differences are reflected in their written language (previously measured in article and dissertation abstracts).

What about full texts of articles?

Full text is the new black

- to measure paper and disciplinary affiliation via semantic profiles of papers (Knoth & Herrmannova, 2014)

- to discern multiply-authored papers from singly-authored, using stylometric indicators (Rexha et al., 2015)

- to analyze the semantic frames of verbs and other terms in citation contexts (Small, 2011; Bertin et al., 2015)

Research Questions

Using machine learning models with social and epistemic term frequencies as features:

1. How do accuracy rates for full article texts compare to article abstracts for pairwise comparisons of physics, psychology, and philosophy?

2. What features distinguish best between:
    - each pair of disciplines
        - for full texts of articles
        - for abstracts of articles
    - each genre for a single discipline?

3. What do these differences imply about differences between genres, disciplines, and disciplinary genres?

Sample

- From Cogprints, paired abstracts and full texts filtered by presence of processable PDF:
  - Philosophy: 977 -> 458
  - Psychology: 1714 -> 679
- Texts extracted from PDFs, with abstracts removed and processed separately.

Data Analysis

- Support vector machines (SMOs) via Weka determine a hyperplane that most cleanly separates classes based on n-dimensional feature arrays, assigning weights to terms.

- This model is then tested for accuracy via 10-fold cross-validation.

Features

Features from Hyland (2005):

- Hedges (perhaps, approximately)
- Boosters (decidedly, clear)
- Self-mentions (the author, we)
- Attitude markers (surprisingly)
- Engagement markers (the reader, ?)

307 terms and phrases in total, collected as relative frequencies (presence/absence for cross-genre).

Features

Not keywords.
Not topical.
Not nouns.

Findings - Accuracy

- Cross-discipline (baseline: 59.8%):
  - Abstract: 68.49%
  - Full text: 80.79%
- Cross-genre (baseline: 50%):
  - Psychology: 95.12% (74.5% w/ relative frequencies)
  - Philosophy: 95.35% (86.16% w/ relative
  - frequencies)

Cross-Discipline Features (Abstracts)

Psychology

terms	weights
regard	-1.39
showed	-1.38
likely	-1.35
the writer	-1.30
input	-1.23
surprising	-1.15
found	-1.15
demonstrated	-1.11
striking	-1.03
compare	-1.02

Philosophy

terms	weights
argue	3.7284
review	2.4997
state	2.2191
analyse	2.1498
my	2.0965
realized	2.0948
thought	2.0001
indicates	1.9404
in general	1.8579
remarkable	1.8421
we	1.7839
possible	1.7295
argued	1.7062
key	1.6675
about	1.6624
argues	1.6434
interesting	1.6345
agrees	1.5845
us	1.5834
our	1.5698
certain	1.5652
claim	1.5139
prove	1.5025

Cross-Discipline Features (Full texts)

Psychology

terms	weights
likely	-1.91
found	-1.77
demonstrated	-1.76
indicated	-1.56
shows	-1.47
surprisingly	-1.41
assess	-1.31
relatively	-1.30
appeared	-1.29
approximately	-1.25
expected	-1.23
develop	-1.18
surprised	-1.16
probable	-1.15
demonstrate	-1.14
evaluate	-1.14
you	-1.11
estimate	-1.09
often	-1.09
show	-1.07
determine	-1.01
showed	-1.01

Philosophy

terms	weights
claim	2.0665
us	1.8489
interesting	1.6715
refer	1.5957
certain extent	1.5776
TRUE	1.3144
argues	1.284
must	1.2623
my	1.241
realize	1.2092
astonished	1.2041
believes	1.1382
from our perspective	1.1024
look at	1.0792
consider	1.0788
proved	1.0715
feels	1.071
order	1.069
our	1.0355
certain	1.0214
indisputable	1.0193

Cross-Genre Features (Psychology)

Abstracts

terms	weights
define	-0.6971
employ	-0.5542
need to	-0.5399
must	-0.5125
contrast	-0.4803
sometimes	-0.4525
evident	-0.4108
interesting	-0.4049
notice	-0.3749
preferred	-0.3627
integrate	-0.3241
dramatically	-0.3137
certainly	-0.3108
in my opinion	-0.2993
in fact	-0.2985
suggest	-0.2732
astonishingly	-0.2655
believed	-0.2536

Full texts

terms	weights
!	2
analyse	1.464
mount	1.2525
?	1.178
thinks	1
note	0.8174
indicates	0.8047
key	0.7679
see	0.7557
probable	0.7354
you	0.7309
calculate	0.7242
correctly	0.7025
establish	0.692
essentially	0.6878
definite	0.6716
probably	0.6117

Cross-Genre Features (Philosophy)

Abstracts

terms	weights
know	-0.5977
integrate	-0.522
somewhat	-0.4101
argued	-0.3857
really	-0.3745
usually	-0.3659
quite	-0.3342
sometimes	-0.3206
tended to	-0.3062
tend to	-0.3013
may	-0.2991
show	-0.2969

Full texts

terms	weights
?	1.652
me	1.248
!	1.0563
input	1
probable	1
add	0.9337
analyze	0.8678
your	0.825
perhaps	0.8187
feels	0.7518
increase	0.7495
analyse	0.708
state	0.6912
review	0.6854
regard	0.6445
clearly	0.6333
compare	0.609

Some Implications

- Cross-disciplinary comparisons show similar differentiating terms in each genre (with differences in ranking).

- Cross-genre comparisons within disciplines find drastic discipline-specific differences.

- Abstracts frame a paper in the briefest, strongest terms, while articles have more allowance for nuance.

- Abstracts describe their affiliated articles; articles report (psychology) or enact (philosophy) the underlying study.

Next

Physics!

Thank you!
Questions?

Bradford Demarest
bdemares@indiana.edu

References

Bertin, M., Atanassova, I., Larivière, V., & Gingras, Y. (2015). Mapping the linguistic context of citations. *Bulletin of the American Society for Information Science and Technology*, 41(2), 26–29. http://doi.org/10.1002/bult.2015.1720410208

Demarest, B., Larivière, V., & Sugimoto, C. R. (2015). Coming to terms: A discourse epistemetrics study of article abstracts from the Web of Science. In *Proceedings of ISSI 2015 Istanbul*. Istanbul, Turkey.

Demarest, B., & Sugimoto, C. R. (2015). Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*, 66(7), 1374–1387. http://doi.org/10.1002/asi.23271

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

Knoth, P., & Herrmannova, D. (2014). Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication’s Contribution. *D-Lib Magazine*, 20(11), 8–.

Rexha, A., Klampfl, S., Kröll, M., & Kern, R. (2015). Towards authorship attribution for bibliometrics using stylometric features. In *Proceedings of the First Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics*. Istanbul, Turkey.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, 87(2), 373–388.