A Utility Model of Authors in the Scientific Community

Yanchuan Sim, School of Computer Science, Carnegie Mellon University, ysim@cs.cmu.edu
Bryan R. Routledge, Tepper School of Business, Carnegie Mellon University, routledge@cmu.edu
Noah A. Smith, Computer Science & Engineering, University of Washington, nasmith@cs.washington.edu

Introduction

We introduce a probabilistic model of authoring scientific papers where:

- authors have individual preferences,
- writing a paper requires trading off among the preferences of authors as well as extrinsic rewards in the form of community response to their papers,
- preferences (of individuals and the community) and tradeoffs vary over time.

Our model lead to improved predictive accuracy of citations given texts and texts given authors. Further, our model's posterior suggests an interesting relationship between seniority and author choices.

Utility function of an author

Document content Author preferences

$$U(\boldsymbol{\theta}_d) = \kappa_a y_d - \frac{1}{2}\|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2$$

Author trade-offs Response to document

We assume that author a is an optimizer: when writing document d, she seeks to increase response y_d while keeping the contents θ_d, "close" to her preferences, η_a. κ_a is a trade-off between the extrinsic (citation-seeking) and intrinsic (preference-satisfying) objectives:

- If κ_a is large, a might be understood as a citation maximizing agent;
- If κ_a is small, a might appear to care more about writing certain papers than citations.

Assuming a linear model for y_d, the expected utility is

$$\mathbb{E}[U(\boldsymbol{\theta}_d)] = \kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta}_d - \frac{1}{2}\|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2$$

Response largely driven by content

An author's decision will therefore be

$$\hat{\boldsymbol{\theta}}_d = \arg\max_{\boldsymbol{\theta}} \kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta} - \frac{1}{2}\|\boldsymbol{\theta} - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2$$

Optimality implies that θ_d solves the first-order equations

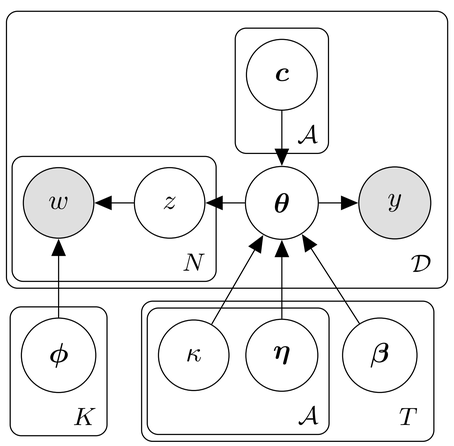
$$\kappa_a \beta_j - (\hat{\theta}_{d,j} - (\eta_{a,j} + \epsilon_{d,a,j})) = 0, \forall 1 \le j \le K$$

A probabilistic model incorporating utility

With multiple co-authors, the joint utility is simply the average of the group's utility, and the "cost" term for each author scaled by $c_{d,a}$, a fractional contribution term. We treat document content, θ_d, as a mixture of topics similar to latent Dirichlet allocation (LDA). Instead, θ_d is constrained as a result of the utility function

$$\boldsymbol{\theta}_d \sim \mathcal{N}\left(\sum_{a \in \boldsymbol{a}_d} \kappa_a \boldsymbol{\beta} + c_{d,a}\boldsymbol{\eta}_a, \|\boldsymbol{c}_d\|_2^2 \mathbf{I}\right)$$

We extend our model to allow variations at different time-steps for $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\kappa}$. To model the temporal dynamics, we place a multivariate Gaussian prior on the variables' differences between consecutive time steps.

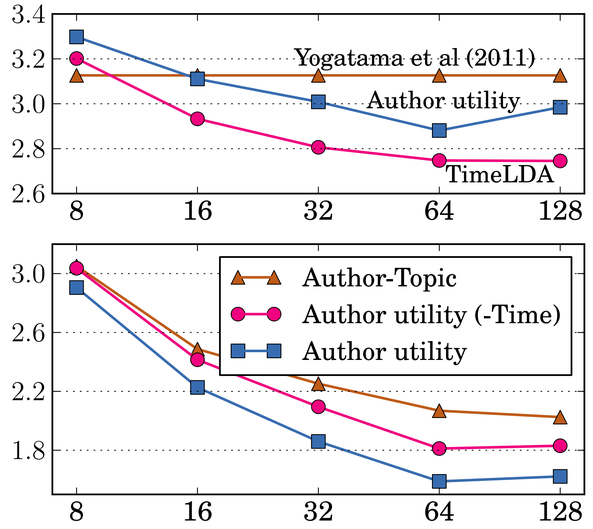


We adopt a Bayesian approach parameter estimation, using Monte Carlo EM to perform approximate inference on our model.

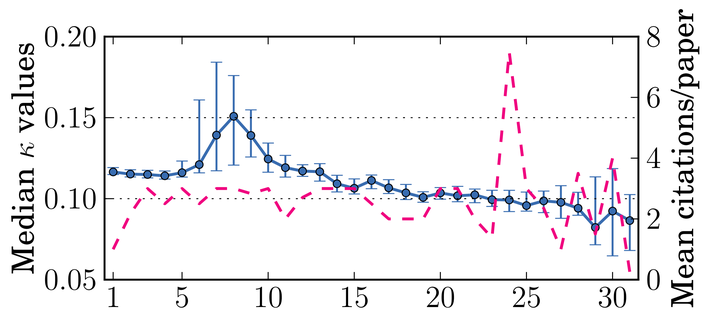
Data. We used a subset of the ACL anthology corpus, focusing on conference papers published between 1980 and 2010. The sub-corpus contains 5,498 documents authored by 5,575 scientists.

Experiments

Predicting citations and words. The mean absolute error (in citation counts, *top*) and held-out perplexity (×10³, *bottom*) plotted against varying number of topics K.



Tradeoffs and Seniority. Plot of authors' median κ (blue, solid) and mean citation counts (magenta, dashed) against their academic age. The Spearman's rank correlation between κ of an author and her age is -0.870, with p-value < 10^{-5}.



References

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The Author-Topic Model for Authors and Documents. In *Proc. of UAI.*

Yanchuan Sim, Bryan Routledge, and Noah A. Smith. 2015. The Utility of Text: The Case of Amicus Briefs and the Supreme Court. In *Proc. of AAAI.*

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. 2011. Predicting a Scientific Community's Response to an Article. In *Proc. of EMNLP.*