Web Appendix

The *Journal of Consumer Research* at Forty: A Historical Analysis

XIN (SHANE) WANG
NEIL BENDLE
FENG MAI
JUNE COTTE

Topic Models

Topic modeling is used to automatically discover the index of ideas contained in the documents and identify which documents are about the same kinds of ideas (Blei and Lafferty 2009). The statistical topic model we use, latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003), assumes that the procedure of producing a JCR abstract can be decomposed into a number of simple probabilistic steps. The statistical inference based on hierarchical Bayesian analysis can then uncover the semantic structures in the texts and discover patterns of word usage. LDA has been shown to be a powerful technique for gaining insight into the content of academic documents (Griffiths and Steyvers 2004).

There are many advantages provided by the LDA model compared to more traditional text analytic models such as naïve Bayes classification and Latent Semantic Analysis (LSA). For example, LDA is built upon a rigorous foundation of Bayesian statistical inference and therefore has more principled model fitting and selection procedures. It provides "soft" classification for documents and therefore allows each document to be a multi-membership mixture of different topics. LDA also extends the ideas of probabilistic latent semantic analysis (PLSA) (Blei and Lafferty 2009) and can automatically learn contexts of word usage without recourse to a dictionary or thesaurus (Hofmann 2001).

There are several statistical assumptions inherent in the LDA model. The first major assumption of the LDA is the "bag of words" model, which means that the words appearing in the abstracts are assumed to be exchangeable. Therefore, when applying statistical topic models such as LDA, we represent each abstract as a vector of word counts and neglect the order of the words. Although the order of the words is important for human readers to comprehend a document, Blei et al. (2003) have argued, using De Finetti (1977)'s exchangeability theorem, that this simple representation can result in computationally efficient methods while preserving semantic themes in the collection of documents.

Given the “bag of words” assumption, the LDA model further assumes that: 1) words contained in each abstract are generated from a mixture of topics; 2) each topic has a probability distribution over a fixed word vocabulary; 3) the topics are shared by all of the JCR abstracts, but the topic proportions differ across abstracts. Formally, LDA can be described using a generative process. It assumes that there are K different topics (the parameter K can be chosen using model selection techniques) and the vocabulary size is V. Each topic is associated with a Dirichlet distribution over all words in the vocabulary with parameters $\boldsymbol{\beta}$. For all topics $k \in 1 \dots K$ the process first draws a vocabulary mixture $\boldsymbol{\phi}_k$ for the topic from Dirichlet ($\boldsymbol{\beta}$). Then, each JCR abstract $m \in 1 \dots M$ is assumed to be produced from the following generative process:

1. Sample length of the abstract N_m from a Poisson distribution with parameter ξ.
2. Sample topic proportions $\boldsymbol{\theta}_m$ from a Dirichlet distribution with parameters $\boldsymbol{\alpha}$**.**
3. For each of the $n \in 1 \dots N$ words in m:
    a. Sample a topic assignment $z_{m,n}$ from Multinomial ($\boldsymbol{\theta}_m$), where $z_{m,n}$ is a topic index between $1 \dots K$.
    b. Choose a word $w_{m,n}$ from Multinomial ($\boldsymbol{\phi}_{z_{m,n}}$).

The objectives of topic modeling can be viewed as reversing the above generative process using Bayesian inference (Blei 2012). We wish to infer the topic mixture of each abstract $\boldsymbol{\theta}_m$, and the word distributions of each topic $\boldsymbol{\phi}_k$. The former parameters indicate which topic(s) are covered in a given abstract, while the latter parameters tell us the representative words for each topic. Researchers have developed approximate inference algorithms such as Gibbs sampling (Steyvers and Griffiths 2006) and variational methods (Blei et al. 2003; Teh, Newman, and Welling 2006) as exact inference is intractable for the model. Heinrich (2005) presents a detailed discussion of various parameter estimation methods for LDA.

Implementation

We downloaded the abstracts of the 1875 JCR articles from JSTOR Data for Research (dfr.jstor.org). We completed sentence and word segmentation, part-of-speech (POS) tagging, and word lemmatization using Stanford CoreNLP (Manning et al., 2014).

We used the Stanford Topic Modeling Toolbox (ver. 0.4.0) developed and distributed by the Stanford Natural Language Processing Group (Ramage et al. 2009) for the inference task. To account for the power-law of word usage, and avoid the domination of common words across topics (McCallum, Mimno, and Wallach 2009), we excluded common stopwords. In addition, we excluded the top 40 most frequent words, outside of the stopwords list, that are most common in the JCR abstracts. These words include methodology related words such as *examine*, *method* and *study*, and words that are less topic-specific but are widely used in many topics such as *behavior*, *brand*, and *people*. We take these preprocessing steps to reduce noise and improve the interpretability of the resulting topics. A robustness check was performed by fitting the model without excluding these words and the analysis resulted in qualitative similar topics.

We used the collapsed variational Bayes (CVB0) approximation outlined in Asuncion et al. (2009) to approximate the posterior distribution. Two hyperparameters α and β in the LDA model control the smoothing for document-topic distributions and topic-term distributions respectively. A smaller β generates more fine-grained topics and a smaller α tends to assign fewer topics to a document. We used $\beta = 0.01$ and $\alpha = 50/K$ following the recommendation of Griffiths and Steyvers (2004)[1]. The optimal number of topics K was chosen to minimize perplexity, a widely-used performance metric that gives useful characterization of the predictive quality of a language model and correlates with other measures well (Asuncion et al. 2009). More specifically, we trained LDA models with $K = 2$ to 24 on half of the data with 2,000 iterations of CVB0 algorithm and evaluated the perplexity on the other half of the data. We chose $K = 16$ because it offered the minimal perplexity on the test set (Figure 1). The final results presented in the paper were produced using 10,000 iterations of the CVB0 algorithm with $K = 16$ on all the data.

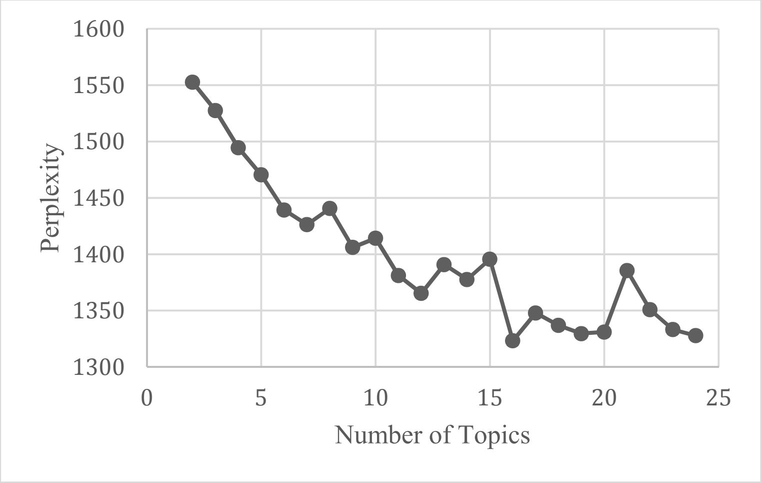

Figure 1. Perplexity as a function of number of topics

Lastly, we estimated the consumer research topics distribution over time using the Hall model (Hall, Jurafsky, and Manning. 2008), in which *post hoc* calculations are performed based on the observed probability of topics over the years. The Hall model is non-restrictive on the overall trend of the topics and thus offers the flexibility needed for our exploratory analysis. The empirical probability that an arbitrary abstract d written in year y was about topic z is:

$$\hat{p}(z|y) = \sum_{d:t_d=y} \hat{p}(z|d)p(d|y)$$

where $\hat{p}(z|d)$ is the estimated document-topic distribution, and $p(d|y)$ is the proportion of the articles written in year y.

[1] α and β are parameter vectors for Dirichlet distribution. We follow the convention of using symmetric Dirichlet priors, i.e., each entry is the same scalar.

References

Asuncion, Arthur, Max Welling, Padhraic Smyth, , and Yee Whye Teh (2009). *On Smoothing and Inference for Topic Models.* Paper presented at the Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence.

Blei, David M. (2012). Probabilistic Topic Models. *Communications of the ACM, 55*(4), 77-84.

Blei, David M, and John D. Lafferty (2009) Topic models. *Text Mining: Classification, Clustering, and Applications, 10*, 71.

Blei, David M, Andrew Y. Ng, and Michael I. Jordan (2003). Latent Dirichlet Allocation. *the Journal of Machine Learning Research, 3*, 993-1022.

De Finetti, Bruno. (1977). *Theory of Probability, volume I.*

Griffiths, Thomas L, and Mark Steyvers (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America, 101* (Suppl 1), 5228-235.

Hall, David, Daniel Jurafsky, and Christopher D. Manning (2008). Studying the History of Ideas Using Topic Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 363-371.

Heinrich, Gregor. (2005). Parameter Estimation for Text Analysis.

Hofmann, Thomas. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning, 42*(1-2), 177-196.

Manning, Christopher D, Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J, & McClosky, David. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics System Demonstrations*, 55-60.

McCallum, Andrew, David M. Mimno, and Hanna M. Wallach (2009). *Rethinking LDA: Why Priors Matter.* Paper presented at the Advances in Neural Information Processing Systems.

Ramage, Daniel, Evan Rosen, Jason Chuang, Christopher D. Manning, and Daniel A. McFarland, (2009). *Topic Modeling for the Social Sciences.* Paper presented at the NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond.

Steyvers, Mark, and Tom Griffiths (2006). Probabilistic Topic Models. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *Latent Semantic Analysis: A road to Meaning*: Lawrence Erlbaum.

Teh, Yee W, David Newman, and Max Welling (2006). *A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation.* Paper Presented at the Advances in Neural information Processing Systems.