2018

Automated Text Analysis of Document and Reference Similarities - An Application of LDA Topic Modeling

Xiang Liu
Teachers College, Columbia University, xl2438@tc.columbia.edu

Zhou Zhou
Teachers College, Columbia University, zhou.eye8@gmail.com

Hui Soo Chae
Teachers College, Columbia University, hsc2001@columbia.edu

Gary Natriello
Teachers College, Columbia University, gjn6@columbia.edu

Automated Text Analysis of Document and Reference Similarities

- An Application of LDA Topic Modeling

Introduction

Since its inception, Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003), a hierarchical Bayesian model, has been widely used to model text topics. LDA has also been modified to incorporate collaborative filtering information in the application of building recommendation system (C. Wang & Blei, 2011). Some examples of other extensions include a hierarchical topic model (Blei, Griffiths, Jordan, & Tenenbaum, 2004), a supervised learning topic model (Mcauliffe & Blei, 2008), an author-topic model (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004), and a chronological topic model (X. Wang & McCallum, 2006).

In the field of education, LDA also recently has gained some attention. For example, Y. Wang, Bowers, & Fikis (2016) analyzed topic trends of publications in an educational leadership journal over time. In the current paper, we propose an application of the LDA topic model in examining the relevance between publications and their references. The outline of the paper is as follows - we'll first briefly review LDA model, its inference algorithm, and a cosine measure for document similarities. Next, we'll describe the data and the analysis procedure. Finally, we'll present the results and discuss some implications and future directions.

LDA Topic Model

A review of LDA

LDA is a three-level hierarchical Bayesian model for analyzing discrete data. In the context of topic modeling, the interested variables are topics, topic proportions, topic assignments, and observed words. In this paper, we use the same notations as Blei (2012). Let - $\beta_{1:K}$ denote K topics, which are distributions over the vocabulary. Each element $\beta_{k,:v}$ is the proportion of the

v^{th} word in the k^{th} topic. $\theta_{1:D}$ are topic distributions associated with D Documents. Each element $\theta_{d,k}$ in the topic distribution vector describes the topic proportion for the k^{th} topic in the d^{th} document. $z_{1:D}$ are topic assignments for D documents. Each $z_{d,n}$ is a categorical random variable which denotes the topic assignment of the n^{th} document. These quantities are all latent. In other words, we don't directly observe them. Instead, we observe words in the documents. $\omega_{1:D}$ are vectors of words in D documents. Each $\omega_{d,n}$ is the n^{th} word in the d^{th} documents. LDA assumes the following generative process:

1. For the d^{th} document, the topic proportion arise from a Dirichlet distribution with a concentration parameter α_θ, i.e. $\theta_d \sim Dir(\alpha_\theta)$. Note that θ_d as well as α are K dimension vectors. The number of topic K is a modeling choice which are chosen a priori.
2. For the n^{th} word in the d^{th} document, the topic assignment follows a categorical distribution with parameter θ_d, i.e. $z_{d,n} \sim Cat(\theta_d)$.
3. Finally, n^{th} word in the d^{th} document is chosen according to the vocabulary distribution for the $z_{d,n}^{th}$ topic, i.e. $\omega_{d,n} \sim Cat(\beta_{z_{d,n}})$. The vocabulary distribution is assumed to follow a Dirichlet distribution with a concentration parameter α_β, i.e. $\beta_{z_{d,n}} \sim Dir(\alpha_\beta)$.

It is clear that LDA is indeed a three-level hierarchical Bayesian model. Alternatively, the above generative process could also be represented using graphical model notations (see Figure 1). This representation explicitly shows that the only observed variable in LDA is words in documents. Other variables are all latent which we don't directly observe. According to the generative process, the joint distribution of the latent and observed variables enjoys the following factorization,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \omega_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} \left(p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(\omega_{d,n}|\beta_{1:K}, z_{d,n}) \right). \quad (1)$$

The factorization implies the dependence structure of the variables. For a particular observed word in a document, it depends on all topics and the topic assignment of that word. The topic assignment of a word in a document depends on the topic distribution of the document.

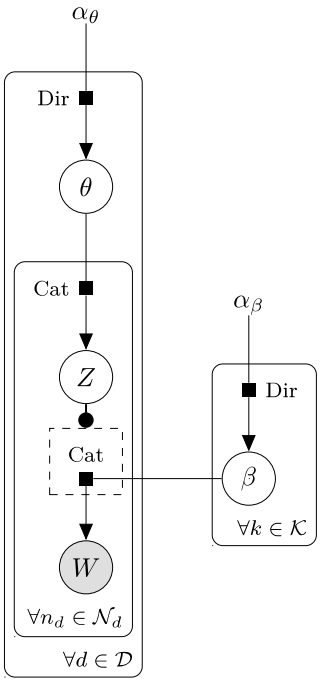

Figure 1. Latent Dirichlet Allocation as a directed factor graph

Inference Algorithm

The goal of the inference in LDA is to estimate latent variables conditional on the observed variables. In other words, we are estimating the vocabulary distribution of each topic, the topic distribution of each document, and the topic assignment of each word. To achieve this goal, we need to compute the posterior distribution, i.e.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|\omega_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \omega_{1:D})}{p(\omega_{1:D})} \quad (2)$$

$$= \frac{\prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} \left(p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(\omega_{d,n}|\beta_{1:K}, z_{d,n})\right)}{\int\int\int \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} \left(p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(\omega_{d,n}|\beta_{1:K}, z_{d,n})\right)_{d\beta d\theta dz}} \quad (3)$$

In general, the high dimensional integral in the denominator of Equation 3 is intractable. Consequently, the posterior distribution cannot be computed exactly. Instead, we rely on approximating the posterior distribution and make inference based on the approximation. There are two classes of approaches - Markov Chain Monte Carlo and variational inference. In the following section, we'll briefly describe a MCMC algorithm.

A Collapsed Gibbs Sampler. One popular algorithm within the MCMC approach is the Gibbs sampling (Geman & Geman, 1984). Even though the exact form of the posterior distribution as in equation 3 cannot be analytically derived, conjugacy exists for full conditional distributions (Bishop, 2006). Intuitively, Gibbs sampler reduces the global computation to local computations by capitalizing on the existence of conjugacy for full conditional distributions. The sampler iteratively draws from the full conditional distributions. These samples correspond to draws from states of a Markov Chain. As the chain runs long enough, the samples are eventually draws from the posterior distribution.

Now we describe a collapsed Gibbs sampler (Liu, 1994) for LDA. Recognizing the topic proportions θ_d and word distributions β_k are nuisance parameters when we estimate topic assignments for each word, we can integrate (collapse) them out. As a result, the Markov chain usually converges faster. After other parameter estimates are obtained, θ_d and β_k can then be estimated in turn. LDA in its probabilistic model form is

$$\omega_{d,n}|z_{d,n}, \beta_{z_{d,n}} \sim Cat(\beta_{z_{d,n}})$$

$$z_{d,n}|\theta_d \sim Cat(\theta_d)$$

$$\beta_{z_{d,n}} \sim Dir(\alpha_\beta)$$

$$\theta_d \sim Dir(\alpha_\theta).$$

After integrating out nuisance parameters, the full conditional distribution of each word assignment (Griffiths, 2002) is given by

$$p(z_i = j|z_{-i}, \omega) \propto p(\omega\_i|z_i = j, z_{-i}, \omega_{-i})p(z_i = j|z_{-i}). \quad (4)$$

Notice the above expression does not depend on θ_d or β_k which have been integrated out. In addition, each word no longer needs to be indexed by documents. This is also the result of integrating out the document level parameters – topic proportions θ_d. Furthermore, the first term in the right-hand side of equation 4 is the result of integrating out the vocabulary distribution β_j, i.e.

$$p(\omega_i|z_i = j, z_{-i}, \omega_{-i}) = \int p(\omega_i|z_i = j, \beta_j)p(\beta_j|z_{-i}, \omega_{-i})d\beta_j, \quad (5)$$

where

$$p(\beta_j|z_{-i}, \omega_{-i}) \propto p(\omega_i|\beta_j, z_{-i})p(\beta_j). \quad (6)$$

$p(\beta_j)$ is Dirichlet, and $p(\omega_i|\beta_j, z_{-i})$ is categorical. Following the conjugacy results for distributions within the exponential family (Bishop, 2006), the posterior $p(\beta_j|z_{-i}, \omega_{-i})$ is a Dirichlet distribution with a concentration parameter $\alpha_\beta + n_{-i,j}^{(\omega)}$, where $n_{-i,j}^{(\omega)}$ is the number of counts that word ω is assigned to topic j excluding the current word. It is then clear that the posterior of the topic only depends on the words that have been assigned to this topic. Evaluating the integral in equation 5 with this result, we can obtain

$$p(\omega_i|z_i = j, z_{-i}, \omega_{-i}) = \frac{n_{-i,j}^{(\omega_i)} + \alpha_\beta}{n_{-i,j}^{(\cdot)} + W\alpha_\beta}, \quad (7)$$

Where $n_{-i,j}^{(.)}$ is the total number of words assigned to topic j excluding the current word, and W is the total number of words. $p(z_i = j|z_i)$ is similarly obtained by integrating over the topic distribution for the document where ω_i is drawn, i.e.

$$p(z_i = j|z_{-i}) = \int p\left(z_i = j\middle|\theta_{d_i}\right)p\left(\theta_{d_i}\middle|z_{-i}\right)d\theta_{d_i} \tag{8}$$

$$= \frac{n_{-i,j}^{d_i} + \alpha_\theta}{n_{-i}^{d_i} + T\alpha_\theta}, \tag{9}$$

Where $n_{-i,j}^{d_i}$ is the number of words assigned to topic j in document i excluding the current word, $n_{-i}^{(d_i)}$ is the number of words in document i excluding the current word, and T is the total number of topics.

The above results lead to the full conditional distribution of each word assignment, i.e.

$$p(z_i = j|z_{-i}, \omega) \propto \frac{n_{-i,j}^{(\omega_i)} + \alpha_\beta}{n_{-i,j}^{(.)} + W\alpha_\beta} \frac{n_{-i,j}^{d_i} + \alpha_\theta}{n_{-i}^{d_i} + T\alpha_\theta}. \tag{10}$$

The sampler initializes each word to a topic assignment between 1 and T. Then it cycles through and update assignment for each word based on equation 10. After burn-in cycles, the draws are approximately samples from the posterior distribution, which can be summarized to provide posterior inference.

Cosine Similarity

The cosine similarity is a common measure of topic similarity between documents. Geometrically, it measures the angle between two non-zero vectors, i.e.

$$\cos\theta = \frac{v_1 \cdot v_2}{||v_1|| \cdot ||v_2||}\ . \tag{11}$$

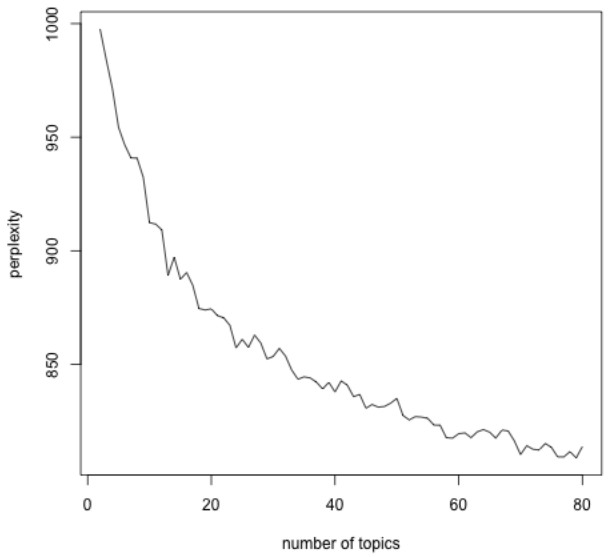

Figure 2. Cross-validation of number of topics

The cosine function is bounded between -1 and 1. If the two vectors are in the exactly same direction, $\cos\theta = 1$; $cos\theta = -1$ if they are of the opposite directions. Between the two extremes, $\cos\theta$ is monotonically decreasing as θ.

In the context of topic modeling, the vectors are topic distributions for each document. Thus, the entries of the vectors are necessarily non-negative and sum to one. As a result, the angle between any two topic distribution vectors are bounded between 0 and π. Consequently, the cosine similarity is between 1 and 0 with $cos\theta = 1$ being the most similar and $cos\theta = 0$ being the least similar.

Data analysis

Data and Model Choice

For the data analysis, we picked an article (Baran, Correia, & Thompson, 2013) from *Teachers College Record*. There are 51 references used in this article. We extracted the full texts of these articles as well. We fitted an LDA topic model to these data.

When doing topic modeling, the number of topics K is usually a modeling choice which we have to specify a priori. A common procedure of choosing K is cross-validation (Blei et al., 2003). The idea is to find K such that it minimizes some measure of model fit while taking the complexity of the model into consideration. One particularly popular choice is *perplexity* (Blei et al., 2003). It has an inverse relationship with likelihood. In other words, the model with lower perplexity fits better. We did a 10-fold cross validation on the extracted data. The results are in Figure 2. The cross-validation results did not conclusively suggest the best choice of K as the perplexity continued to decrease as the number of topics increased. However, the improvement in perplexity was minimal from $K = 70$ to $K = 80$. For this reason, we decided to fit an LDA topic model with the number of topics $K = 70$. Then cosine similarity measures (see equation 11) are calculated based on the posterior distribution of topics for each document.

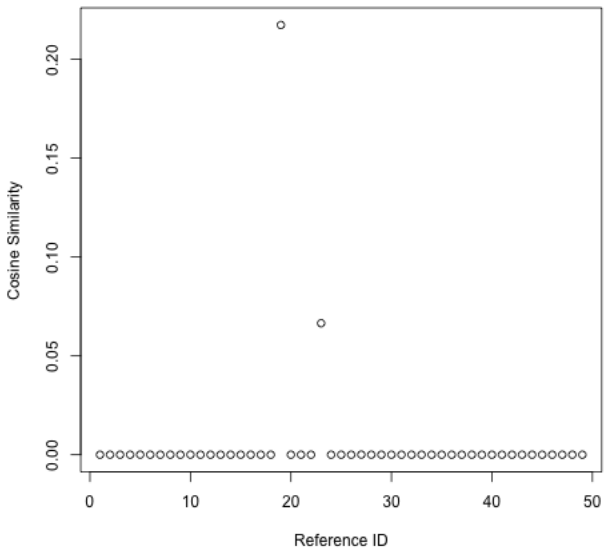

Figure 3. Similarity distribution

sim_rank	sim	article
1	0.2173	Laat, Lally, Lipponen, & Simons, 2007
2	0.0665	Anderson, Liam, Garrison, & Archer, 2001

Table 1. References with highest cosine similarity

Results

Figure 3 shows the distribution of cosine similarity measures between the article and its references. While most references have cosine similarity measures around 0, there are a couple exceptions. One reference scored a similarity score above 0.20, and the other scored around 0.06. We listed these two references with their similarity scores in table 1. A deeper investigation reveals that these two references and the article all have their posterior distribution of topics heavily concentrated on a single topic. Some terms with highest probability within the topic include: online, teachers, student, interview, feedback, presence, exemplary, social, and pedagogy.

Discussion

In this article, we presented an application of LDA topic modeling in determining the degrees of relevance between a journal article and its references. The cosine similarity measure could potentially be a useful metric in gauging the quality of an article. Even though we chose this metric in this paper, it is certainly not the only choice. A comparison between different distance metrics could be interesting.

One difficulty of this application is perhaps the relatively small number of text corpus (Yin & Wang, 2014). Fitting an LDA to a corpus of 50 documents might not lead to a good estimation of topics. The cross-validation in this paper did not clearly show the best number of topics. This could due to this very reason. Another difficulty is fitting an LDA to a topically homogeneous

group of documents. Generally, an article along with its references are likely about similar topics. LDA might not lead to stable classification of documents. This is also an interesting research topic that worth further exploring.

Care should be take when interpreting the cosine similarity measure. The interpretation ties closely to the topics extracted. Consider the scenario where an LDA topic model with three topics is fitted. If an article talks about the three topics evenly, the posterior distribution would be an uniform discrete distribution, i.e. (0.33, 0.33, 0.33). However, since the same distribution is chosen as a non-informative prior, another article not talking about any of the three topics will also have the same posterior distribution. As a result, the cosine similarity measure between these two articles will be 1.0 even though they are talking about different topics. Thus it is important that the practitioners check the posterior distributions carefully and see on which topic(s) each article has their posterior distribution mass concentrated. If the posterior distribution is uniformly distributed as described earlier, further information should be brought into consideration. The interpretation of the cosine similarity measure is also relative. It provides some information about the relative relevance among the documents. However it does not have an absolute interpretation. Whether two article with a cosine similarity of 0.5 should be considered similar remains subjective and largely depends on the specific circumstances (e.g. number of topics fitted, terms associated with each topic, posterior distributions).

The relevance between an article and its reference is a feature that could be potentially useful in predicting the quality of an article. The future development along this line could lead to applications in academic publishing and/or other industries.

References

Anderson, T., Liam, R., Garrison, D. R., & Archer, W. (2001). Assessing Teaching Presence in a Computer Conferencing Context. *Journal of the Asynchronous Learning Network*, *5*(2).

Baran, E., Correia, A.-P., & Thompson, A. (2013). Tracing Successful Online Teaching in Higher Education: Voices of Exemplary Online Teachers. *Teachers College Record*, *115*(3).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. *Pattern Recognition*. https://doi.org/10.1117/1.2819119

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.1984.4767596

Griffiths, T. (2002). *Gibbs sampling in the generative model of Latent Dirichlet Allocation*.

Laat, M. De, Lally, V., Lipponen, L., & Simons, R.-J. (2007). Online teaching in networked learning communities: A multi-method approach to studying the role of the teacher. *Instructional Science*, *35*(3), 257–286. https://doi.org/10.1007/s11251-006-9007-0

Liu, J. S. (1994). The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a

Gene Regulation Problem. *Journal of the American Statistical Association*, *89*(427), 958–966. https://doi.org/10.1080/01621459.1994.10476829

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised Topic Models. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in Neural Information Processing Systems 20* (pp. 121–128). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/3328-supervised-topic-models.pdf

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The Author-topic Model for Authors and Documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 487–494). Arlington, Virginia, United States: AUAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=1036843.1036902

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11* (p. 448). New York, New York, USA: ACM Press. https://doi.org/10.1145/2020408.2020480

Wang, X., & McCallum, A. (2006). Topics over Time: A non-Markov Continuous-time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 424–433). New York, NY, USA: ACM. https://doi.org/10.1145/1150402.1150450

Wang, Y., Bowers, A. J., & Fikis, D. J. (2016). Automated Text Data Mining Analysis of Five Decades of Educational Leadership Research Literature: Probabilistic Topic Modeling of EAQ Articles From 1965 to 2014. *Educational Administration Quarterly*, 0013161X16660585. https://doi.org/10.1177/0013161X16660585

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text

clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 233–242). New York, New York, USA: ACM Press. https://doi.org/10.1145/2623330.2623715