Topic Correlations over Time

David Haley(dhaley@), David Hall (dlwh@), and Mike Rodgers (mikepr@cs)

Abstract

Topic models have proved useful for analyzing large clusters of documents. Most models developed, however, have paid little attention to the analysis of the latent topics themselves, particularly with regards to change in their correlation over time. We present a novel, probabilistically well-founded extension to Latent Dirichlet Allocation (LDA) which can explicitly model topic drift over time. Using this extension, we analyze the correlations of topics over time in a corpus of ACL papers.

1 Introduction

A well-documented "Statistical Revolution" in Natural Language Processing (NLP) took place in the early 1990's (Gaz96), leading to a very rapid increase in the use of statistical methods to build natural language systems. Obvious examples include *machine translation* and *parsing*, which have become largely synonymous with their statistical variants. However, no work has been done to analyze this shift, despite the recent proliferation of topic models designed for large document corpora.

In this paper, we develop a model to examine this feature of topic shift over time. We then use the model to analyze the texts of about 12,500 papers from the Association of Computational Linguists, available via the ACL Anthology (Bir). This corpus covers nearly all papers from 1965 until 2006. We first automatically extract a number of topics. We then infer correlation among the most salient of these topics. Finally, we present the results of our analysis, concluding with promising directions for future work.

2 Previous Work

A significant amount of attention has been given to topic modeling, and most recent work in this area has followed

$$
\begin{aligned}
D &= \text{the number of documents in the corpus} \\
K &= \text{the number of topics in the corpus} \\
N_i &= \text{the number of words in document i} \\
\theta_i &\sim \mathrm{Dir}(\alpha), i \in [1, D] \\
\phi_k &\sim \mathrm{Dir}(\eta), k \in [1, D] \\
z_{i,n} &\sim \mathrm{Mult}(\theta_i), i \in [1, D], n \in [1, N_i] \\
w_{i,n} &\sim \mathrm{Mult}(\phi_{z_{i,n}})
\end{aligned}
$$

Figure 1: Latent Dirichlet Allocation

from LDA(BNJ03), which merits a brief review here. (Figure 1.) LDA is a mixture model of latent topics z in documents d, which have words w. Each document is characterized by a multinomial distribution over topics θ_d, and each topic is characterized by a multinomial distribution over words ϕ_z. Finally, both the θ_d's and ϕ_z's have (usually symmetric) Dirichlet priors with hyperparameters α and η respectively. These hyperparameters can be fixed or sampled from a Gamma distribution.

Thus far, this model has no explicit representation of time: all documents are exchangeable. Several models have been proposed to remove this assumption. Most similar to LDA is the Topics over Time model (ToT) (WM06). ToT assigns a time stamp t in the range [0,1] to each document, and repeatedly samples the timestamp according to a topic-dependent Beta distribution ψ_z. To conserve space, we forgo reproducing the entire specification, simply adding the extra parameter here:

$$t_{i,n} \sim \mathrm{Beta}(\psi_{z_{i,n}}), i \in [1, D], n \in [1, N_i]$$

Representation of time as a continuous variable has several useful properties. First, it is a rather minimal addition to the model, making it easy to implement. Moreover, it avoids the problem of Markovization: there is no need to divide time into a specific number of epochs. This negates the need to determine the proper dividing line for documents and the need to decide whether or not to cluster natural epochs. However, it makes finding

$$
\begin{aligned}
T &= \text{\# epochs} \\
D_t &= \text{\# documents in the corpus for the t'th epoch} \\
K &= \text{\# topics in the corpus} \\
N_{t,i} &= \text{\# words in document i for epoch t} \\
\alpha_t &\sim \text{Norm}(\alpha_{t-1}, \sigma_1 I) \\
\phi_t &\sim \text{Norm}(\phi_{t-1}, \sigma_2 I) \\
\theta_{t,i} &\sim \text{Norm}(\alpha_t, \sigma_3 I), i \in [1, D_t] \\
z_{t,i,n} &\sim \text{Mult}(\pi(\theta_{t,i})), i \in [1, D_t], n \in [1, N_{t,i}] \\
w_{t,i,n} &\sim \text{Mult}(\pi(\phi_{t,z_{i,n}}))
\end{aligned}
$$

Figure 2: Dynamic Topic Model

$$
\begin{aligned}
T &= \text{\# epochs} \\
D_t &= \text{\# documents in epoch t} \\
K &= \text{\# topics in the corpus} \\
N_{t,i} &= \text{\# words in document i in epoch t} \\
\alpha_{t,i} &\sim \text{Gamma}(1, \hat{\alpha}), t \in [1, T], i \in [1, K] \\
\theta_{t,i} &\sim \text{Dir}(\alpha_t), i \in [1, D_t] \\
\phi_k &\sim \text{Dir}(\eta), k \in [1, D_t] \\
z_{t,i,n} &\sim \text{Mult}(\theta_i), i \in [1, D_t], n \in [1, N_i] \\
w_{t,i,n} &\sim \text{Mult}(\phi_{z_{i,n}})
\end{aligned}
$$

Figure 3: Unmarkovized Time Model

the change in relationship between any two topics over time much harder.

On the other extreme, the Dynamic Topic Model (DTM) represents documents as points on the topic-distribution simplex about a centroid (BL06). The DTM uses a normal distribution and a normalization function to constrain the points to be on the simplex. Moreover, each epoch's centroid is sampled from the previous year's centroid, enabling "drift" of topics from epoch to epoch. Similarly, the vocabulary distributions for each of the topics are chosen as a centroid. The model is specified fully in Figure 2. This model requires a substantial change to the inference procedures, and it requires renormalizing topic and vocabulary distributions by a transform function π.

Finally, we mention briefly the Correlated Topic Model (BLar), which is more or less identical to the DTM, except that there are no times, only correlations between topics. In principle these models could be combined to give the desired results, but with great mathmatical overhead. In this paper we propose models that accomplish the same result but require slightly less complicated machinations.

3 Temporal Model Specification

In this section we develop an unmarkovized (but still discretized) dynamic model with minimal disruption to LDA. Like the DTM, we split time into discrete epochs (which, for now, are exchangeable), but instead of using a normal distribution, we opt instead to concentrate on the hyperparameter α_t, which we will take as representative of the mean "topic strength" (expectation of θ_t) for a given epoch t. (Strictly, α_t is a multiple of the expectation of the Dirichlet, but normalizing it solves the problem.) First, we remove the symmetry of the Dirichlet distribution for θ, allowing the components of the α's to vary. Finally, we add a prior on these α's, which we call $\hat{\alpha}$. The specification for this non-Markovized model is in Figure 3. $\hat{\alpha}$ is the scale parameter of the αs' prior. This ensures conjugacy, but still maintains the same expectation for the αs.[1] This parameter can be sampled from a Gamma distribution, though in our experiments we found that 0.1 produces good results.

4 Topic Correlations

We considered a number of techniques for determining correlations, but in this section we develop an extension to the UTM (or, simply, LDA) that has support for modeling topic correlations as part of the inference. However, going forward, we will use "correlation" in the informal sense. In particular, we will model "correlation" as "confusability:" with what probability can one topic be "confused" for another topic?

4.1 The Correlated Dirichlet Distribution

In this section, we briefly discuss a new distribution, which we term the Correlated Dirichlet Distribution

[1] We omit the details of deriving the Gibbs sampler in this paper due to space constraints.

$$
\begin{aligned}
T &= \text{\# epochs} \\
D_t &= \text{\# documents in epoch t} \\
K &= \text{\# topics in the corpus} \\
N_{t,i} &= \text{\# words in document i in epoch t} \\
\alpha_{t,i} &\sim \text{Gamma}(1, \hat{\alpha}), t \in [1, T], i \in [1, K] \\
\text{B}^T_{t,i} &\sim \text{Dir}(\beta_i), t \in [1, T], i \in [1, K] \\
\theta_{t,i} &\sim \text{CDir}(\alpha_t, B_t), i \in [1, D_t] \\
\phi_k &\sim \text{Dir}(\eta), k \in [1, D_t] \\
z_{t,i,n} &\sim \text{Mult}(\theta_i), i \in [1, D_t], n \in [1, N_i] \\
w_{t,i,n} &\sim \text{Mult}(\phi_{z_{i,n}})
\end{aligned}
$$

Figure 4: Correlated Unmarkovized Time Model

(CDD). To begin, we note that the Dirichlet has *neutrality*; that is, if $\theta|\alpha \sim Dir(\alpha)$, then $\theta_i \perp \frac{\theta_j}{1-\theta_i}$. Unfortunately, this property indicates that the Dirichlet has no natural means to specify the strength of correlation (formally or informally) between any of the components of its samples. Therefore, we propose the CDD, which explicitly adds a mixture component to the Dirichlet Distribution. Intuitively, we can think of samples from a CDD as samples from a standard Dirichlet Distribution that have been distorted as the result of one step on a random walk specified by some stochastic matrix. Formally, we say that if B is a stochastic matrix, and

$$
\begin{aligned}
\theta^*|\alpha &\sim \text{Dir}(\alpha) \\
\theta &= \theta^* \text{B}
\end{aligned}
$$

then $\theta \sim \text{CDir}(\alpha, \text{B})$. We should note that the CDD bears some resemblance to the Dependent Dirichlet Distribution in (LLWC06). However, the requires that the matrix B take a specific form relating to the Markov transition probabilities defined by their problem. We specify the matrix as a parameter to the Distribution. In this sense, the CDD is a generalization of the DDD.

4.2 The Correlated Unmarkovized Time Model

Integrating the CDD into our model is fairly straightforward. The Correlated Unmarkovized Time Model (CUTM) is specified in Figure 4. In particular, we introduce the CDD and a diagonally-biased prior on the rows of B. Thus, we require that each of the rows of B be a probability distribution. We also introduce a strong diagonal bias in the β hyperparameters to help ensure that topics do not get "washed out" in the mixing. That is, $b_{ii} \gg b_{ij}$ for $b_{ii}! = b_{jj}$ For inference, however, this model becomes slightly more difficult. We cannot use Gibbs sampling to sample the matrix B, though with fixed B it is straightforward to use Gibbs sampling for α. To update B, we use the Metropolis-Hastings algorithm (Has70) to update the columns of B individually. Details of the proposal distributions we used are discussed in the appendix.

5 Results

We implemented both the UTM and CUTM and ran the model on 11,000 papers of the ACL Anthology over a wide range of topics until convergence. For K=100, the UTM ran in about 3/4 the time of the CUTM. In the interest of space, we report only topics from the CUTM. Figure 5 lists the top 20 words chosen by Mutual Information for the top 3 topics (by word count) as well as Topic 3, which we use to illustrate correlations. We caption the topics with titles based on our interpretation of the topics' distribution over the vocabulary.

Figure 6 is a smoothed plot of the α's for those topics over time. As you can see, the graph is incredibly jagged, leading us to conclude that treating epochs as exchangeable should be reconsidered. However, there is a reasonable correspondence between peaks on the graph and the topics themselves. Topic 73 peaks in the early 1980's, and the topics themselves are constrained to dates in the 1970's, and words we associate with papers from that era. Topic 3 has peaks exclusively in the 1980's, which corresponds to the words listed. We see 1987 as a prominent year. (We also mention that 1982, 1984, and 1985 also are in the top 40 words for Topic 3.) Topics 18 and 73 have peaks later on, and have words like "probability", "evaluation", and "performance": words we associate with the Statistical Revolution.

Figure 7 is a smoothed plot of the directional confusability of Topic 3 with Topic 73 and of Topic 73 with itself. Again, we note that there is a considerable amount of variance in both graphs, indicating that the B matrices should also be markovized. However, we see that Topic 3 ("the 1980's") and Topic 73 ("1970's") are most corre-

Topic 3 “1980’s parsing”	Topic 18 “Statistics”
university	features
grammar	model
class	document
trees	task
semantic	data
language	lexical
process	language
constraints	precision
parser	ranking
event	verb
np	score
syntactic	relation
set	probability
elements	research
1987	semantic
wordnet	evaluation
number	labeled
utterance	goal
book	dialogue
Topic 73 “1970s”	**Topic 74 “Stat. Parsing”**
computer	discourse
object	target
science	lexicon
speech	information
1975	verb
program	constraints
network	machine
discourse	language
report	state
structure	similar
objects	rule
federal	entity
act	clause
1973	feature
1976	head
information	performance
research	definite
semantic	module
location	function

Figure 5: Topic 20 words for 4 topics, chosen by Mutual Information

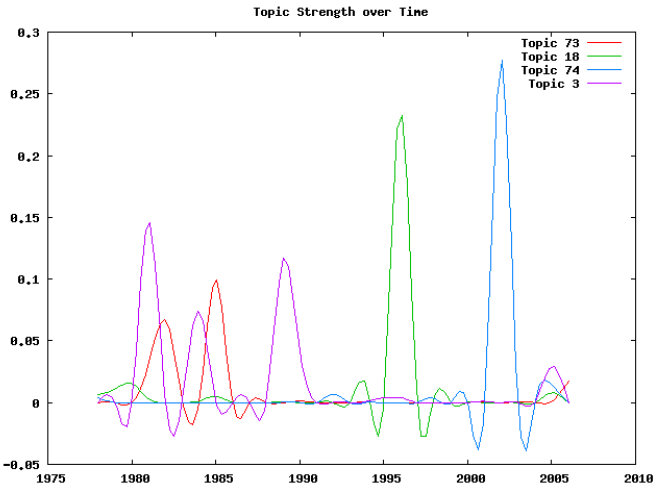

Figure 6: $\alpha_{t,i}$ over time for 4 topics

lated in the 1990’s. Again, this matches our expectation: the older two topics become, the more confusable they are. These topics tend to arise in historical papers, or they are used in papers that cite older work.

6 Conclusion and Future Directions

We have developed an unmarkovized model for directly learning topic prominence and correlation over time using an additional level of hyperparameters and a stochastic matrix, respectively. We then examined the results of fitting that model, and observed that the topics and their strengths were in fact meaningfully related. We also looked at confusability of two topics and found a similar relation there.

However, much more work is needed. The assumption of the exchangeability of epochs is clearly incorrect: the jaggedness seen in the graphs is too pronounced. On the other hand, that we saw localized peaks and not several disjoint eras indicates that the topics being extracted were in fact located in certain eras, so this model is still valuable. Regardless, for practical purposes, it is probably best to condition epochs on their ancestors.

More work should also be done to compare this model with LDA and the DTM in terms of log likelihood. Moreover, as with all topic models, a better metric of performance is necessary: log likelihood is useful for training, but still we could use a better method of evaluating these systems.

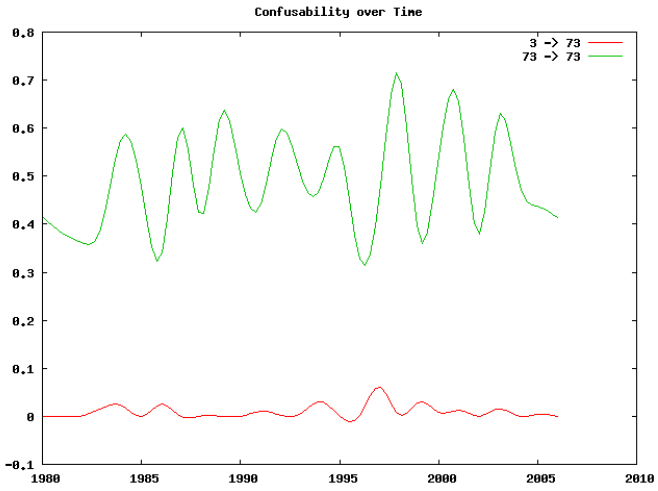

Figure 7: $\beta_{t,ik}$ (Confusability) over time for 2 topics

Finally, there are still other ways to look at correlations. One could argue that correlations should be inferred *post hoc*, and thus an approach like Independent Component Analysis should be useful. Some initial analyses (which do not fit here), seem to find reasonable *post hoc* topics.

Acknowledgements

We would like to thank Dan Ramage, Chris Manning and Dan Jurafsky for their help in thinking through the issues here. We would also like to thank Hal Daumé for the Hierarchical Bayesian Compiler (Dau), which we used to implement early prototypes of the UTM.

References

Steven Bird. Association of computational linguists anthology. http://www.aclweb.org/anthology-index/.

David Blei and John D. Lafferty. Dynamic topic models. *ICML*, 2006.

David Blei and John D. Lafferty. A correlated topic model of *science*. *Annals of Applied Science*, To Appear.

David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Hal Daumé. HBC: Hierarchical Bayes Compiler. http://hal3.name/HBC. 2007.

Gerald Gazdar. *Paradigm merger in natural language processing*, pages 88–109. Cambridge University Press, 1996.

W.K. Hastings. Monte carlo sampling methods using markov chains and their applications, 1970.

Xuejun Liao, Qiuhua Liu, Chunping Wang, and Lawrence Carin. Neighborhood-based classification. http://www.ee.duke.edu/∼lcarin/DDD_talk_yale_3.pdf, 2006.

Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. 2006.

A Derivation of Proposal Distributions for Metropolis Hastings Sampling

A.1 Resampling θ^*

A highly effective proposal distribution for θ^* can be computed fairly easily. We proceed by relaxing the posterior. Let $\vec{z}$ be the topic counts for a document.

$$\begin{aligned} p(\vec{z}|\theta^*)p(\theta^*|\alpha) &\propto \prod_k (\theta^* \mathrm{B}_k)^{z_k} \prod_k (\theta^*_k)^{\alpha_k} \\ &= \prod_k (\sum_i \theta^*_i \mathrm{b}_{ki})^{z_k} (\theta^*_k)^{\alpha_k - 1} \\ &\approx \prod_k (\theta^*_k \mathrm{b}_{kk})^{z_k} (\theta^*_k)^{\alpha_k - 1} \\ &\propto \prod_k (\theta^*_k)^{z_k + \alpha_k - 1} \\ &= \mathrm{Dir}(\alpha + \vec{z}) \end{aligned}$$

The step in which we dropped the summation takes advantage of the assumption that $\mathrm{b}_{kk} \gg \mathrm{b}_{kj}$, for $k \neq j$, which we enforce on our priors. This result matches our intuition: draws from a Dirichlet distribution should be roughly the same as draws from a Correlated Dirichlet distribution with a strong diagonal bias.

A.2 Resampling B_k^T

Resampling the rows of B is a bit different. While a similar calculation could be used, in practice we find that is it is often better to choose a weak diagonally-biased prior and weight the current sample more highly when conditioning the Dirichlet. That is,

$$\mathrm{B}_k^{\mathsf{T}'} \sim \mathrm{Dir}(\beta_k + C\mathrm{B}_k^{\mathsf{T}})$$

where C is large.