UNIVERSITY COLLEGE LONDON

MASTERS THESIS

Dynamic Topic Modeling of PATSTAT Patents Using LDA

Author:
Christopher MARTIN

Supervisor:
Dr. John Shawe TAYLOR

A thesis submitted in fulfillment of the requirements for the degree of Masters of Science

Machine Learning
UCL Dept. of Computer Science

August 18, 2016

Declaration of Authorship

I, Christopher MARTIN, declare that this thesis titled, "Dynamic Topic Modeling of PATSTAT Patents Using LDA" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSITY COLLEGE LONDON

Abstract

Faculty Name
UCL Dept. of Computer Science

Masters of Science

Dynamic Topic Modeling of PATSTAT Patents Using LDA

by Christopher MARTIN

In this paper we evaluate the performance of a time varying family of LDA based topic models meant to capture both the underlying semantic structure of a document collection and the evolution of that structure in time. Such models are useful for illustrating changes in the use of language regarding specialized subject matter and provide a window into the progression of that change. We compare these models to traditional topic models such as LDA as a benchmark and explore the efficacy of such models in a range of applications including document classification, clustering, and influence prediction. Finally, we present results on over 18 years of patent data from the PATSTAT database across X classes of patents demonstrating interpretable trends, better document classification and clustering, and improved topic coherence.

Acknowledgements

I owe the completion of this project to the many people who have helped along the way, either directly or indirectly. To my project supervisors Christopher Grainger and Prof. John Shawe-Taylor for their continued guidance throughout the project, to my colleauges for their advice and discussion, and to my family for their moral support, I express my sincere gratitude, thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to the Thesis Topic

1.1 The need for topic models

Researchers today are faced with a deluge of data. As we continue to digitize and aggregate our collective knowledge we produce ever increasing archives of information. The sheer volume and variety of forms this information may take - text, images, audio, video, social connections etc. - makes it difficult and in most cases impossible to parse manually.

This driving factor of data growth has given rise to internet giants such as Google, who's search tool helps us access and browse pre-indexed swathes of information. However in order to go beyond mere keyword searches, or link analysis, and break into the realm of understanding each document, requires a new approach to data exploration.

A powerful set of computational tools referred to as probabilistic topic models have emerged to meet this challenge. Aimed to discover and annotate large archives of documents with thematic information, topic models identify patterns that reflect the underlying topics which combined to form those documents.

Naturally, it is rare that we would know beforehand exactly what topics a given document contains, and thus topic modeling constitutes an unsupervised task. As a result, topic modeling algorithms are designed to work without prior knowledge of the topic distribution of a given document — that is, the topics are derived from the texts themselves. This makes the organization, summarization and annotation of text corpora possible at an inhuman scale. Consequently, topic models are useful in a variety of settings and have successfully been applied to web archives, news articles (Newman et al., 2006), and academic literature (Steyvers et al., 2004) to elicit insight. In this paper, we focus our experiments on patent data.

1.2 Primer on Latent Dirichilet allocation

Fortunately, the intuition behind LDA topic models is relatively straight forward. To understand how the algorithm infers the topics in relation to documents we first define what constitutes a topic. A topic is a distribution over a fixed vocabulary, where each word has an assigned probability of occurrence. Subsequently, we can take the view that each document is likely a product of one or more topics, a cocktail of themes as it were with different proportions of each ingredient.

Take for example the following document sampled from the August 2015 EPO Worldwide Patent Statistical Database (**PATSTAT**). The patent abstract contained in Figure 1.1 relates to a mechanism for stopping a water wheel. We have taken the liberty of highlighting a selection of words from a few of this document's prominent topics. Words like "pressure", "liquid", and "flow" belong to the **fluids/water** topic and are colored blue. While words relating to the **mechanisms** by which this fluid is directed such as "chamber", "valve", and "guide" are colored red. Finally, words such as "transmission", "speed", and "operated" belong to the topic associated with **signals** and are colored green.

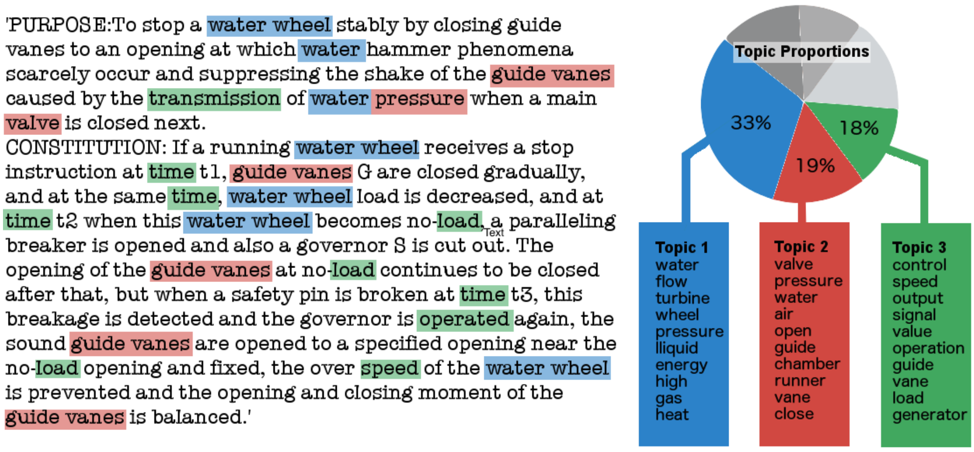

FIGURE 1.1: Topic proportions in a sample patent abstract.

The topics in the previous example were formed not over a single document but over a collection. The grey sections of the pie chart above represent the topics that this patent does not contain strong elements of. This is a key characteristic of LDA topic models, each document has a unique topic 'fingerprint' as a result of a generative process. That process for generating a document word by word is as follows. First we decide, sampling from the distribution of topics, which topic the word will belong to. Then we sample from that topic's distribution to decide what the word itself will be. This process is then simply repeated for each word, and while it works it has the following assumptions:

- Documents can manifest multiple topics (however typically not many)
- Each document is assumed to be the product of a generative process.
- Generative process starts with a topic, i.e. a distribution over a fixed vocabulary.
- Assumes a fixed number of topics

Latent Dirichilet allocation falls into a family of machine learning algorithms called **hidden variable models**. In this family of models, customarily the user "posits a hidden structure in the observed data, and then learns that structure using posterior probabilistic inference" (Blei and Lafferty, 2009). For LDA specifically, the documents are the observed data, the topics and document topic proportions are hidden.

More formally, we may define this process mathematically as a joint distribution over our hidden variables and our observed variables. Namely the distribution over vocabulary β, the topic proportions for document d θ_d , the topic assignment for a word in a document $T_{d,n}$ and of course the observed words themselves $w_{d,n}$.

$$
\begin{aligned}
&P(\beta_{1:K}, \theta_{1:D}, T_{1:D}, W_{1:D}) \\
&= \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \{ \prod_{n=1}^{N} p(T_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:k}, T_{d,n}) \}
\end{aligned}
\tag{1.1}
$$

In Eq 1.1 we see a few dependencies worth noting. Firstly that the topic we assign to a word $T_{d,n}$ depends on the distribution of topics of its document θ_d. Additionally, that the identity of the word itself is dependent on not only the topic we assigned to generate it $T_{d,n}$, but also the vocabulary distributions of each topic $\beta_{1:K}$. Equivalently we can express the dependencies between these variables as a graphical model, illustrated in figure 1.2.

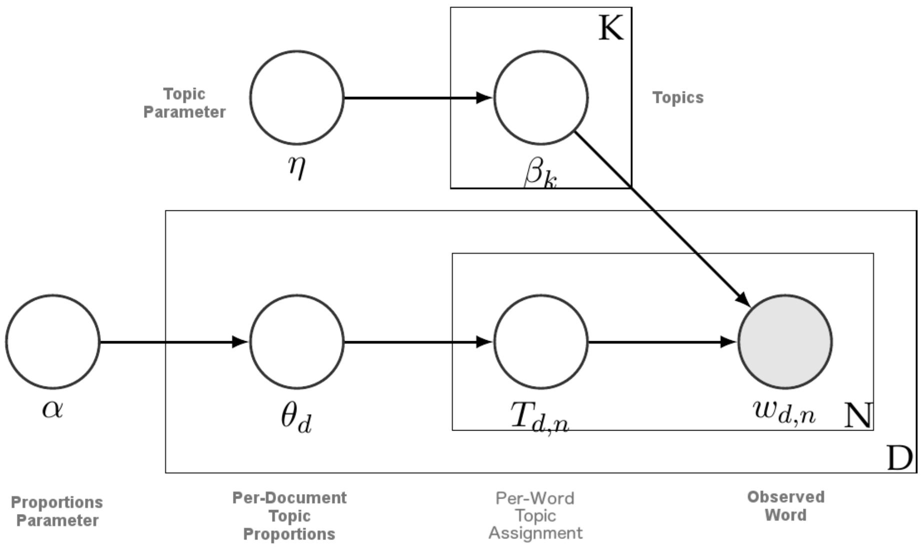

FIGURE 1.2: Graphical model for LDA

So how do we actually obtain our estimates of the hidden parameters? We need to calculate the conditional distribution of our hidden parameters (the topic structure), and the observed words i.e. the posterior distribution described in Eq. 1.2. However the denominator makes this calculation computationally infeasible due to the number of combinations our hidden parameters could take.

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}|W_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}|W_{1:D})}{p(W_{1:D})} \tag{1.2}$$

To move past this, most solutions use either sampling or variational based methods to perform approximate inference and obtain estimates of the hidden parameters. Variational methods allow us to translate the original problem to one of optimization and take advantage of the many optimization techniques available. This in turn allows us to make extensions that are often faster, scale better or allow for different forms of input such as streaming documents.

1.3 Adding a temporal component

One such extension, and the extension we explore in this study, is to relax the implicit assumption of LDA that the order of the documents doesn't matter. By incorporating the order of the documents to the model, a topic is no longer simply a distribution over words but now becomes a *sequence* of distributions over words. This is the jump that allows us not only to identify a theme, as with static LDA, but also track how it progresses in time, giving us the Dynamic Topic Model (**DTM**).

The DTM offers several advantages over traditional LDA including improved predictive performance (Blei and Lafferty, 2006). Primarily though, it facilitates a greater understanding of how each topic developed, and how the ideas therein formed and matured. With it, we can inspect trends of word usage to uncover a richer and more detailed hidden structure. For instance figure 1.3 contains a sample theme from a sub-collection of hydroelectric patents and the progression of word prevalences within it over time.

1.4 Why patents?

Patent data is specifically interesting in this context because of the role patents play in company formation, job growth, economic development, and novel invention. Their history tells a story of technological progression. In an attempt to maintain a competitive edge, many companies large and small spend a considerable amount of energy researching this history to identify technical trends relevant to their industry.

Dynamic topic models have the potential to aid this research by enabling us to track the progression of innovation through language use in patent abstracts. In this paper we look at a number of elements, including

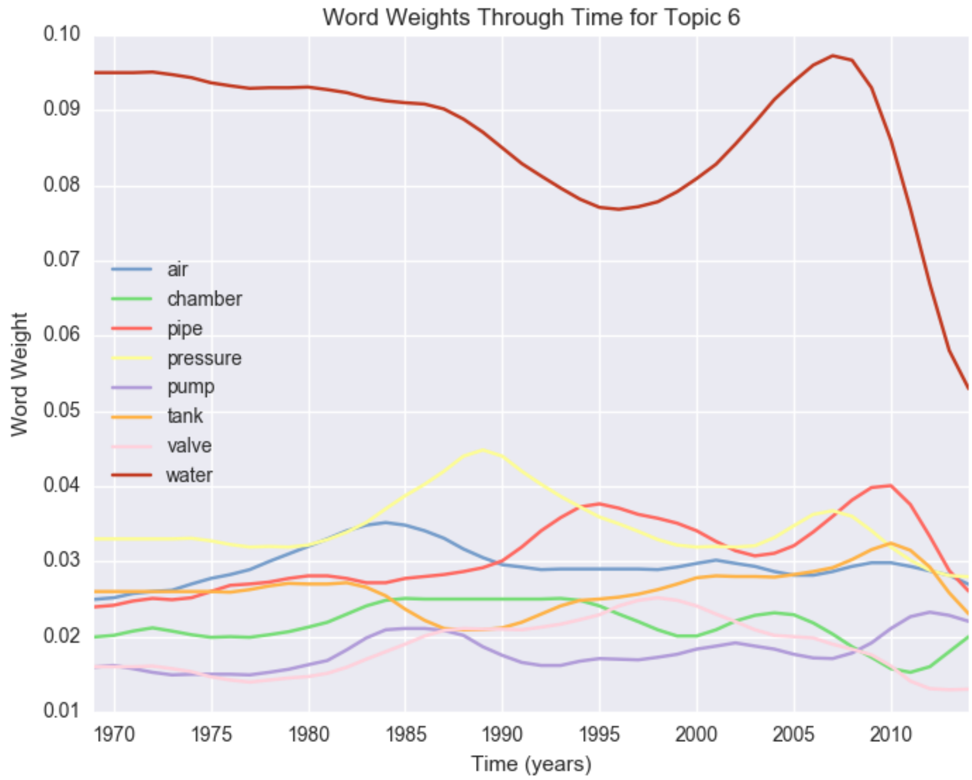

FIGURE 1.3: Distribution over words in a sample hydro-electric topic over time

the evolution of technological themes and their proportions, the origination and progression of language, as well as document influences. Furthermore, the patent corpus and associated International Patent Classification (**IPC**) labels provide a platform for the comparison of various topic modeling algorithms.

1.5 Experiments

At the time of writing this, surprisingly little has been published exploring the effectiveness of both the DTM and the DIM. Much research has evaluated model quality solely based on the likelihood of held-out predictions, however this does not always translate to semantically meaningful topics (Chang et al., 2009). Additionally, the predictive performance of these models is adversely affected by longer time horizons due to an "increase in the rate of specialization in scientific language" (Blei and Lafferty, 2006). Acknowledging the room to explore alternative methods of model evaluation, we implemented the experiments listed below.

Historical Topic Trend Validation

The simplest, but also the most hands-on, method of evaluating the quality of topics produced by the DTM and DIM is simply to validate the inferred topic trends against known industry history. For instance, if in the topic of water purification systems we observe a rise in the usage of words "2D materials" and "lattice membranes" around 2005, we might substantiate this by pointing out graphene's isolation the previous year in 2004.

1.5.1 Topic Coherence

In light of research suggesting that likelihoods and perplexity don't always correlate with human judgement on the interpretibility of topics (Blei and Lafferty, 2006) we borrow several methods of topic coherence suggested by (Rosner et al., 2014). Namely, we evaluated model topic coherences using C_v, C_npmi,C_uci, and U_mass. Using C_v, the metric most correlated with human judgement, DTM achieved the highest with score with XXX compared to static LDA with YYY. For complete results see Table ZZZ in Section 5.

1.5.2 Classification

Unsupervisedly learned word vectors have demonstrated useful as features in a myriad of NLP tasks. We wished to evaluate the proficiency of the word vectors generated by the DTM and DIM at creating an effective feature space for text categorization. To do this we made use of the IPC labels of patents as broad class labels for text content. The resulting topic vectors should then help identify which class a document belongs to. Naturally we tested the efficacy of each model's vector space at correctly classifying the IPC label of documents when fed to a range of classification algorithms. Peak classification performance of the DTM based classifiers was F1 XXX, while LDA was F1 YYY. Text classification results are given in section ZZZ.

1.5.3 Clustering

Another method by which we evaluated the quality of the resulting document vectors was by their ability to define separations in the data relative to the ground truth CPC labels. In order to asses which models yielded vector spaces of the corpus that most effectively clustered the documents we used the following metrics: the adjusted rand index, normalized mutual score info, homogeneity, completeness and the V-measure. Indeed we found that the DTM's vector space tended to outperform that of LDA at clustering with a peak NMI score of XXX compared to YYY. For more detailed results, refer to table ZZZ in section 5.

Chapter 2

Background Information and Theory

2.1 Literature Review

In this section we begin by reviewing a few of the tasks common in topic modeling. Then we describe a handful of the ways the quality of topic models, LDA in particular, are commonly tested. Finally we discuss the specific tasks of document classification and clustering in the context of topic modeling, as well as the corresponding methods for evaluating model performance at these tasks.

2.1.1 Applications of Topic Modeling

The most popular application of topic models is simply summarizing large text collections by mining the topics. This is a task LDA is particularly suited for (Griffiths and Steyvers, 2004; Mei, Shen, and Zhai, 2007). The original LDA paper however (Blei, Ng, and Jordan, 2003) gave promising results on document classification as well. Since then LDA has been used with success not only for document classification, but also for clustering and information retrieval (Wei and Croft, 2006; Nagwani, 2015). This is due to the strength of the topic vectors LDA models provide, which tend to correlate strongly with human judgement.

2.1.2 Ensuring Model Quality

Perplexity Testing

In order to ensure the strength of these topic vectors researchers employ a handful methods to evaluate the topic models. While the most intuitive method is simply to have humans judge the coherence of each topic, this becomes prohibitively time consuming and expensive for large data sets. One commonly used method of automating this process is by evaluating the topic model on a held out set of testing documents and obtaining the log-likelihood perplexity of the unseen documents (Blei, Ng, and Jordan, 2003; Wallach et al., 2009). A higher likelihood on unseen documents, and a lower perplexity score indicates a better model. However this method of evaluating topic model performance has several issues. Firstly, it has been shown that predictive likelihood, or equivalently perplexity, is not always correlated with human judgement, and in some cases is even slightly anti correlated (Chang et al., 2009). Secondly this method of evaluation only acts as a general measure of the entire model. What about the quality of the individual topics?

Coherence Testing

Fortunately several methods of evaluating the coherence of individual topics from topic models exist. For a topic t we define the **Umass** coherence as a sum of the pairwise scores of that topic's top words $W_t = \{w1, ...w_n\}$.

$$\begin{aligned} \text{Umass Coherence } c(t, W_t) &= \sum_{w_i, w_j \in W_t} \text{score}(w_i, w_j) \\ &= \sum_{w_i, w_j \in W_t} log \frac{d(w_i, w_j) + \epsilon}{d(w_i)} \end{aligned} \tag{2.1}$$

Where $d(w_i)$ is the number of documents containing the word w_i and $d(w_i, w_j)$ is the number of documents containing both word w_i and w_j. The ϵ in the numerator is simply to smooth the counts and is typically set to a minimal value such as 1 or .01. Intuitively then, a topic is good if its words cooccur often (Mimno et al., 2011).

The **UCI** measure introduced by (Newman, Bonilla, and Buntine, 2011), operates in the same manner as Umass but with the pointwise mutual information as a scoring function instead, given in eq 2.2.

$$\begin{aligned} \text{UCI Coherence } &= c(t, W_t) = \sum_{w_i, w_j \in W_t} \text{score}(w_i, w_j) \\ &= \sum_{w_1, w_2 \in W_t} log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \end{aligned} \tag{2.2}$$

Where $p(w_i)$ is the probability of seeing word w_i in a random document and $p(w_i, w_j)$ is the probability of seeing both word w_i and word w_j together in a random document. It should be noted that obtaining these probabilities requires empirically estimating them from an external dataset.

Two more noteworthy measures of topic coherence, in addition to those outlined above, were developed by Roder, Both and Hinneberg in their study titled "Exploring the Space of Topic Coherence Measures" (Röder, Both, and Hinneburg, 2015). These measures were the **C_v** and **C_npmi** measures which demonstrated a substantial correlation with human judgement. For brevity we do not replicate their derivations here, but the interested reader will find a detailed description of each in (Röder, Both, and Hinneburg, 2015)

2.1.3 Document Classification

Though the topics produced by topic models are useful in their own right for the qualitative analysis of documents, they are also useful quantitatively when trying to classify documents. For instance a large news organization may want to automatically sort its thousands of articles into the categories "politics", "natural disasters" and "sports". To do this they might use a topic model to get a vector of topic proportions for each document to use as features for a classification algorithm. This process is referred to as document vectorization.

A commonly used baseline for document vectorization is the Term Frequency Inverse Document Frequency (tf-idf) algorithm, prevalent in information retrieval. However it has been shown that in situations where less training data is available LDA outperforms models such as tf-idf, boasting a shorter training time and higher classification accuracy (Li and Zhang, 2010). Additionally when tested against other baseline methods for document vectorization such as the unigram model or probabilistic latent semantic analysis (PLSA), LDA again proved consistenlty more accurate at document classification tasks (Lu, Mei, and Zhai, 2011).

2.1.4 Document Clustering

Another well established task for topic models is document clustering. LDA has been used to successfully cluster a range of documents such as news articles and legal judgements (Lu, Mei, and Zhai, 2011; Xie and Xing, 2013; Kumar and Raghuveer, 2013). As opposed to classification where we want to assign an explicit label to each document, with clustering we wish to evaluate how well the resulting document topic vectors separate the documents into a meaningful structure.

Adjusted Rand Score

On way of accomplishing this is by using the **Adjusted Rand Score** which measures the similarity of two sets of class labels; namely the true labels C and those predicted by a clustering algorithm K. We may calculate the raw (unadjusted) Rand index following equation 2.3 (Hubert and Arabie, 1985).

$$RI = \frac{a+b}{C_2^{n_{\text{samples}}}} \tag{2.3}$$

Where a is the number of pairs of elements in C belonging to the same class, and in K belonging to the same class. Conversely b is the number of pairs of elements in C belonging to different classes, and in k belonging to different classes. Finally, $C_2^{n_{\text{samples}}}$ is the total number of possible pairs in the dataset. In order to ensure that random labelings receive a score of zero we define the adjusted Rand index as

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]} \tag{2.4}$$

Normalized Mutual Info

normalized mutual info

We adopt the normalized mutual information clustering metric used by (Xu, Liu, and Gong, 2003; Cai et al., 2008) which has the benefit of being symmetric and independent of the absolute values of labels.

Homogeneity, Completeness and V-measure

2.2 DTM Model Overview

This section briefly outlines the dynamic topic model (**DTM**), following closely the original derivation found in (Blei and Lafferty, 2006). As this is intended as more of a summary, we recommend the reader examine the original paper for a complete exposition of the mechanics of the DTM. The goal of the DTM is to produce a topic model that takes into account the discrete time evolving nature of topics in a sequential corpus. The general approach is to chain together the underlying topic multinomials and topic proportion distributions, then to make use of the Kalman filter and variational wavelet regression to carry out approximate posterior inference over the latent topics.

With regular latent dirichilet allocation we normally use a Dirichilet distribution to model our uncertainties in word distributions, hence the name. However, this is no longer an option as the Dirichilet distribution does not permit sequential modeling. Instead, we make a statespace model that evolves with Gaussian noise to chain together the natural parameters of each topic $\beta_{z,t}$ such that each topic "evolves" from the last.

$$\beta_{z,t}|\beta_{z,t-1} \sim \mathcal{N}(\beta_{t-1,z}, \sigma^2 I) \tag{2.5}$$

Similarly, document specific topic proportions θ are typically drawn from the Dirichilet distribution. To express the uncertainty over our topic proportions, for the DTM we use a logistic normal with mean α. Then we chain our topic proportions together with the same trick as we did above with word distributions, by using Gaussian noise. This yields the graphical model in figure 2.1.

$$\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I) \tag{2.6}$$

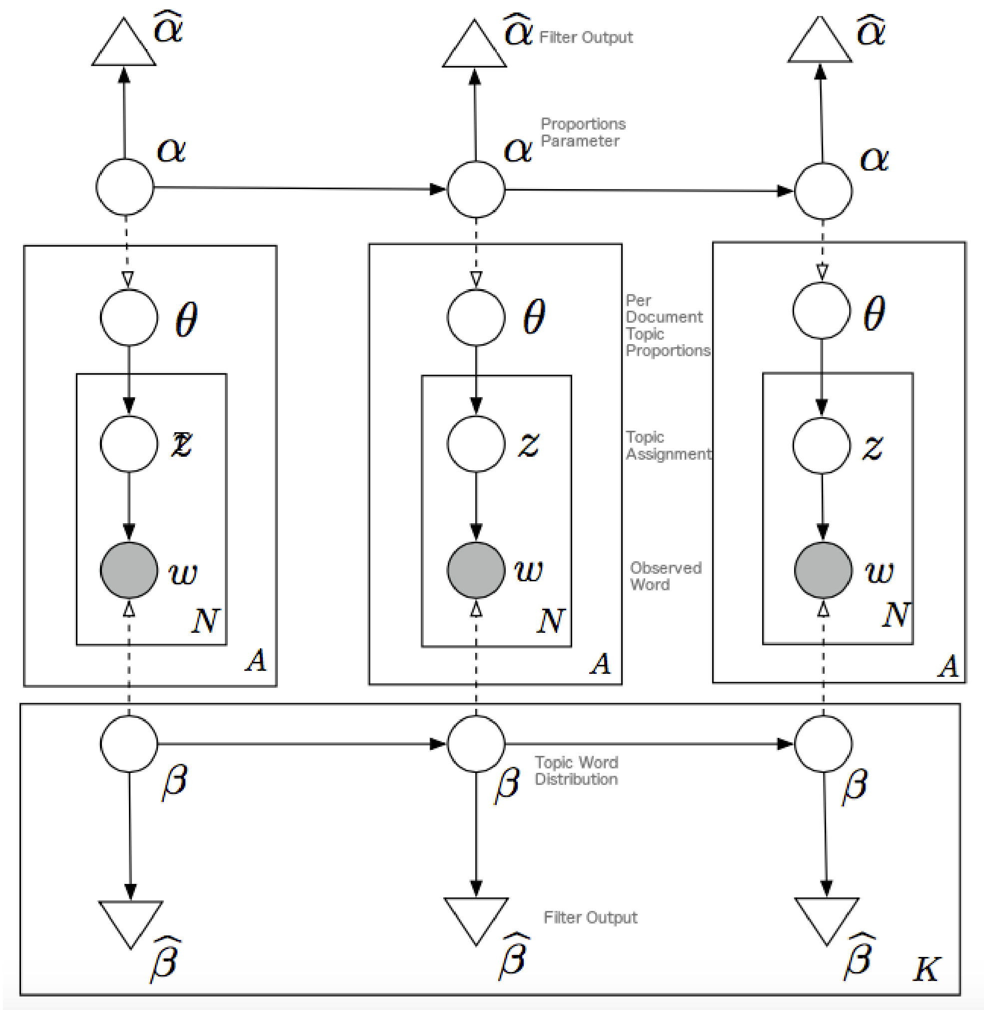

FIGURE 2.1: Graphical model for DTM

Because we have used the Gaussian distribution to model the progression of our parameters, inference becomes intractable due to the non-conjugacy of Gaussian and multinomial models. As before in 1.2 with LDA, we take the variational approach to approximate inference, as opposed to Gibbs sampling, which allows us to handle larger document sets.

The goal here is to use a carefully tuned "dummy" distribution as a substitute for the true posterior. To do this we begin by creating a collection of variational parameters we will optimize over our latent variables. Our latent variables are the topics $\beta_{t,k}$, topic proportions $\theta_{t,d}$, and topic indicators $Z_{t,d,n}$. While we have variational parameters for each topic (consisting of a sequence of multinomial parameters), and for each document (the latent topic proportions). The resulting posterior, again following the notation of (Blei and Lafferty, 2006) is given by equation 2.7.

$$\prod_{k=1}^{K} q(\beta_{k,1}, ..., \beta_{k,T}|\hat{\beta}_{k,1}, ..., \hat{\beta}_{k,T})\times$$
$$\prod_{t=1}^{T} \Big(\prod_{d=1}^{D_t} q(\theta_{t,d})|\gamma_{t,d}) \prod_{n=1}^{N_{t,d}} q(z_{t,d,n}|\phi_{t,d,n}) \Big) \tag{2.7}$$

Here we fit the variational observations $\{\hat{\beta}_{k,1}, ..., \hat{\beta}_{k,T}\}$ according to the KL divergence of our estimated and true posterior. Additionally each topic proportions vector $\gamma_{t,d}$ receives a corresponding free Dirichilet parameter, while each topic indicator $z_{t,d,n}$ receives a corresponding free multinomial parameter $\phi_{t,d,n}$. Subsequently we employ gradient ascent to optimize the topic proportion vectors while the document level parameters simply have a closed form update.

Between time slices the variational parameters $\hat{\beta}$ and $\hat{\alpha}$ can then be predicted via either a Kalman filter or nonparametric wavelet regression. Both provide suitable estimates for tracking the progression of our latent topic parameters through each time step. Derivations for each method are provided in detail in (Blei and Lafferty, 2006) .

Chapter 3

Experimental Set Up

setting the alpha parameter in LDA as recommended by (e.g., as in Steyvers and Griffiths 2007) It acts as a form of regularization. A recommended setting is a = 50/K (Griffiths and Steyvers 2004; Steyvers and Griffiths 2007)

3.1 Data Prep and Considerations

Chapter 4

Experimental Results

4.1 DTM Results and Insights

4.1.1 Topics Through Time

validating topic histories in technology

4.2 DIM Results and Insights

4.2.1 Influence Metric

validating influential patents

correlation with forward citations

correlation with page-rank

4.3 Performance Evaluation

4.3.1 Classification

The clustering result is evaluated by comparing the Normalized mutual information (Xu et al. 2003; Cai et al. 2008) [NEED TO ACTUALLY CITE]

4.3.2 Clustering

Chapter 5

Usefulness in Other Models

5.1 Economic Model

influence can be used as a proxy for forward citations when citations are not available. used as a gamma in model for likelihood of innovation

Chapter 6

Conclusions and Future Work

6.1 Conclusions

6.2 Future Work

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

Blei, D. and J. Lafferty (2009). "Topic Models". In: *Text Mining: Theroy and Applications, Taylor and Francis*.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: ACM, pp. 113–120. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143859. URL: http://doi.acm.org/10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3, pp. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

Cai, Deng et al. (2008). "Modeling Hidden Topics on Document Manifold". In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. Napa Valley, California, USA: ACM, pp. 911–920. ISBN: 978-1-59593-991-3. DOI: 10.1145/1458082.1458202. URL: http://doi.acm.org/10.1145/1458082.1458202.

Chang, Jonathan et al. (2009). "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Neural Information Processing Systems*. Vancouver, BC. URL: docs/nips2009-rtl.pdf.

Griffiths, T. L. and M. Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.Suppl. 1, pp. 5228–5235.

Hubert, Lawrence and Phipps Arabie (1985). "Comparing partitions". In: *Journal of Classification* 2.1, pp. 193–218. ISSN: 1432-1343. DOI: 10.1007/BF01908075. URL: http://dx.doi.org/10.1007/BF01908075.

Kumar, R. V. and K. Raghuveer (2013). "Legal Documents Clustering and Summarization using Hierarchical Latent Dirichlet Allocation". In: *IAES International Journal of Artificial Intelligence* 1.2, pp. 27–35. URL: http://dx.doi.org/10.11591/ij-ai.v2i1.1186.

Li, Lei and Yimeng Zhang (2010). "An empirical study of text classification using Latent Dirichlet Allocation". In:

Lu, Yue, Qiaozhu Mei, and Chengxiang Zhai (2011). "Investigating Task Performance of Probabilistic Topic Models: An Empirical Study of PLSA and LDA". In: *Inf. Retr.* 14.2, pp. 178–203. ISSN: 1386-4564. DOI: 10.1007/s10791-010-9141-9. URL: http://dx.doi.org/10.1007/s10791-010-9141-9.

Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai (2007). "Automatic Labeling of Multinomial Topic Models". In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '07. San Jose, California, USA: ACM, pp. 490–499. ISBN: 978-1-59593-609-7. DOI: 10.1145/1281192.1281246. URL: http://doi.acm.org/10.1145/1281192.1281246.

Mimno, David et al. (2011). "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, pp. 262–272. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145462.

Nagwani, N. K. (2015). "Summarizing large text collection using topic modeling and clustering based on MapReduce framework". In: *Journal of Big Data* 2.1, pp. 1–18. ISSN: 2196-1115. DOI: 10.1186/s40537-015-0020-5. URL: http://dx.doi.org/10.1186/s40537-015-0020-5.

Newman, David, Edwin V. Bonilla, and Wray L. Buntine (2011). "Improving Topic Coherence with Regularized Topic Models." In: *NIPS*. Ed. by John Shawe-Taylor et al., pp. 496–504. URL: http://dblp.uni-trier.de/db/conf/nips/nips2011.html#NewmanBB11.

Newman, David et al. (2006). "Analyzing Entities and Topics in News Articles Using Statistical Topic Models". In: *Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics*. ISI'06. San Diego, CA: Springer-Verlag, pp. 93–104. ISBN: 3-540-34478-0, 978-3-540-34478-0. DOI: 10.1007/11760146_9. URL: http://dx.doi.org/10.1007/11760146_9.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the Space of Topic Coherence Measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Shanghai, China: ACM, pp. 399–408. ISBN: 978-1-4503-3317-7. DOI: 10.1145/2684822.2685324. URL: http://doi.acm.org/10.1145/2684822.2685324.

Rosner, Frank et al. (2014). "Evaluating topic coherence measures". In: *CoRR* abs/1403.6397. URL: http://arxiv.org/abs/1403.6397.

Steyvers, Mark et al. (2004). "Probabilistic Author-topic Models for Information Discovery". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. Seattle, WA, USA: ACM, pp. 306–315. ISBN: 1-58113-888-1. DOI: 10.1145/1014052.1014087. URL: http://doi.acm.org/10.1145/1014052.1014087.

Wallach, Hanna M. et al. (2009). "Evaluation Methods for Topic Models". In: *Proceedings of the 26th International Conference on Machine Learning (ICML)*. Ed. by Léon Bottou and Michael Littman. Montreal: Omnipress, pp. 1105–1112.

Wei, Xing and W. Bruce Croft (2006). "LDA-based Document Models for Ad-hoc Retrieval". In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '06. Seattle, Washington, USA: ACM, pp. 178–185. ISBN: 1-59593-369-7. DOI: 10.1145/1148170.1148204. URL: http://doi.acm.org/10.1145/1148170.1148204.

Xie, Pengtao and Eric P. Xing (2013). "Integrating Document Clustering and Topic Modeling". In: *CoRR* abs/1309.6874. URL: http://arxiv.org/abs/1309.6874.

Xu, Wei, Xin Liu, and Yihong Gong (2003). "Document Clustering Based on Non-negative Matrix Factorization". In: *Proceedings of the 26th Annual*

International ACM SIGIR Conference on Research and Development in Informaion Retrieval. SIGIR '03. Toronto, Canada: ACM, pp. 267–273. ISBN: 1-58113-646-3. DOI: 10.1145/860435.860485. URL: http://doi.acm.org/10.1145/860435.860485.