ABSTRACT

FOX, CHRISTOPHER. Following Topics over Time using Epoch Latent Dirichlet Allocation. (Under the direction of Dr. Kristy Elizabeth Boyer.)

As the Internet becomes a pervasive part of daily life, its users find themselves deluged by information. The amount of text online is enormous and growing at an increasing rate. In this context, computational tools that offer insight into large quantities of text are valuable. One such tool that has been studied and extended in recent years is Latent Dirichlet Allocation (LDA), a probabilistic technique for discovering topics in a collection of documents. The topics found by LDA constitute latent probabilistic structures that can provide a useful summary of a document corpus. However, a limitation of LDA is that it cannot model changes in the vocabulary around a topic over time. In many circumstances, such as understanding online news, it is essential to model the time-dependent aspects of a topic. This thesis proposes a novel model, Epoch LDA, that extends the graphical model structure of LDA so that topics evolve over a sequence of time periods, and yet remain coherent through their "core" vocabularies. In experiments with a corpus of online news articles, the proposed model successfully illuminates events surrounding a topic over a time period of several weeks. The results indicate that this new technique holds promise for time-based exploration of large corpora.

Following Topics over Time using Epoch Latent Dirichlet Allocation

by
Christopher Fox

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Computer Science

Raleigh, North Carolina

2012

APPROVED BY:

_____	_____
Dr. Carla Savage	Dr. Dennis Bahler

_____

Dr. Kristy Elizabeth Boyer

Chair of Advisory Committee

DEDICATION

To Leon E. Semke (1909-2006), my grandfather, who taught me to be curious about the world, appreciative of science, and skeptical.

BIOGRAPHY

Christopher Fox was born in Vancouver, Washington, in 1986. He graduated in 2005 from Washougal High School. In 2009, he earned a Bachelors of Science in Mathematics and Economics from the University of Washington.

ACKNOWLEDGMENTS

I would like to thank my parents Dick Fox and Sue Semke-Fox for always supporting me. I would like to thank my advisor Kristy Elizabeth Boyer for her feedback, encouragement, and ideas. I was very fortunate to meet you. I am grateful to my colleagues Chris Mitchell, Aysu Ezen, Joseph Grafsgaard, Fernando Rodriguez, and Joseph Wiggins for their suggestions and copy-editing.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

Now that the Internet has become a nearly ubiquitous tool for learning, entertainment, and communication, users of the Web are deluged by information. Much of this information, such as news, blog posts, emails, social network messages, product descriptions, scholarly articles, and books, is in the form of text. The amount of text online is not only enormous but also growing at an accelerating rate, and there is increasing interest in summarizing, exploring, and searching this text as soon as possible after it has been created. For example, consumers of news want to stay informed about events as they are happening; brands want to respond proactively to consumer feedback; and law enforcement agencies want to detect criminal plots before they are executed. Therefore, automated tools that can assist users in understanding large quantities of text are highly valuable. A familiar tool of this kind is keyword search, which is useful when the user wants to know which documents match a query string. For instance, a company representative monitoring the micro-blogging service Twitter for consumer feedback would not read every message posted to the service but would instead search for posts containing the company's name.

Many times, though, a user is looking not for a specific piece of information but instead wishes to request a broad overview of a set of texts. The user might ask, "What are these documents about?". Perhaps the simplest way to address this question would be to show the most common words in the texts. Blogs sometimes take this approach, displaying a "word cloud," a jumble of the most frequent words, with more frequent words shown at a larger size. A visitor to the blog can glance at this visualization to get a quick overview of the blog's content. However, word clouds give no indication of the structure of a collection of documents.

One way to present a richer overview is to find a natural partitioning, or clustering, of the set of documents. For example, in a set of news articles from many different sources

published in the last few hours, there will usually be multiple articles on each of the major stories. Rather than scanning through all of the headlines, a user might prefer to see a representative headline for each news story. The Google News website[1], by clustering similar articles together and selecting a representative article, provides such an overview of recent news. While document clustering can illuminate document-level aspects of a corpus (e.g., a particular event that several news articles report on, such as a patent lawsuit), it cannot find themes that apply to only some of the words in a document (e.g., one of the companies involved in the lawsuit).

By contrast, a probabilistic model called Latent Dirichlet Allocation (LDA) discovers major themes, or topics, while allowing more than one theme to be present in each document (Blei, Ng, & Jordan, 2003). The strategy for LDA is to cluster instances of words in documents rather than clustering the documents themselves. Words that are of the same *type*[2] or that are in the same document tend to be placed in the same cluster. By displaying the most frequent word types for each cluster, a user can examine what the main topics are in a set of documents, where each cluster of words corresponds to a topic.

A shortcoming of the LDA approach, however, is that it does not capture how the discovered topics change over time. In a number of domains, such as news, a user might be interested in following the evolution of a theme over time. The corresponding question would be, "What are the major themes in these documents and how do they change over time?". For example, after reading an article about a proposed law to regulate greenhouse gas emissions, a reader might wonder what major stories relating to greenhouse gases have been reported on in recent months.

This thesis proposes a novel modeling framework that can answer this question. The technique, Epoch Latent Dirichlet Allocation, incorporates time into the LDA model so that a

---

[1] http://news.google.com

[2] In this thesis, *type* refers to a unique vocabulary item, whereas *word* refers to an instance of a word type in a document. For example, in "two and two makes fours," the first and third words are distinct instances of the word type "two."

topic has both time-dependent and constant aspects. In this way, the extended model is able to represent shifts that occur in a topic while ensuring that the topic remains essentially the same over the span of time that the document collection encompasses. For example, for a Microsoft topic in a news corpus, words such as “Windows” would remain consistently prominent while other words related to product releases would be prominent only around the time of those events. Experimental results indicate that this new technique holds promise as a tool for time-based browsing of large corpora.

The structure of this document is as follows: Chapter 2 reviews the history of techniques leading up to LDA and work related to the novel model proposed in this thesis. Chapter 3 describes an approach to approximate inference in LDA. Chapter 4 introduces the proposed novel technique, Epoch LDA. Chapter 5 presents evaluation results. Finally, Chapter 6 summarizes the work and directions for future research.

CHAPTER 2

Background and Related Work

LDA discovers the main topics in a set of documents and provides a means to compare documents based on these latent topics. By leveraging word co-occurrence data, LDA is able to learn topics from a given corpus in an unsupervised fashion. That is, the model does not require training data in which words are already annotated with topics.

In some cases, words co-occur because they are semantically related, such as "river" and "stream." However, the fact that two words co-occur frequently does not guarantee that their meanings are related. An analysis of recommendation letters, for example, might find the words in the phrase "to whom it may concern" to be highly related because this procedural phrase is common in such letters. Another concern is that words that are frequent in almost every document, namely function words such as "the" and "of", will naturally co-occur frequently with most words. To address this problem, researchers often filter out a predetermined set of "stopwords" or words that occur very frequently in the corpus (Blei, 2012; Steyvers & Griffiths, 2006).

The remainder of this chapter discusses techniques developed prior to LDA that also utilize word co-occurrence. Then, the LDA model is described. Finally, related work on incorporating time into an LDA-like model is considered.

2.1 Predecessors to LDA

Although LDA is currently a widely used topic modeling framework within the computational linguistics community, it was not the first framework of this kind. Several unsupervised techniques have examined patterns in word co-occurrence in order to gain insight into data. These techniques differ in their mathematical underpinnings, but start from the same commonsense observation: that the co-occurrence of two words in the same

document is evidence that the words are related. The more frequently two words occur together, the more likely that there is some connection between them.

2.1.1 Latent Semantic Analysis

Latent Semantic Analysis, or LSA, takes as input an $m \times n$ word-document matrix X, where each entry x_{ij} is the number of occurrences of word type i in document j (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). The singular value decomposition of $X = T_0 S_0 D_0{}'$ is computed, where T_0 and D_0 have orthonormal columns and S_0 is a diagonal matrix. The entries along the diagonal of S_0 are called singular values. The singular value decomposition has a useful property, which is that it readily yields lower rank approximations of X. For $k \leq m, n$, if all but the k largest singular values in S_0 and the corresponding rows in T_0 and columns in $D_0{}'$ are deleted, producing smaller matrices S, T, and D, then the product $\hat{X} = TSD'$ is the optimal rank k approximation to X in the sense of minimizing $\left\| X - \hat{X} \right\|$, as shown in Figure 1.

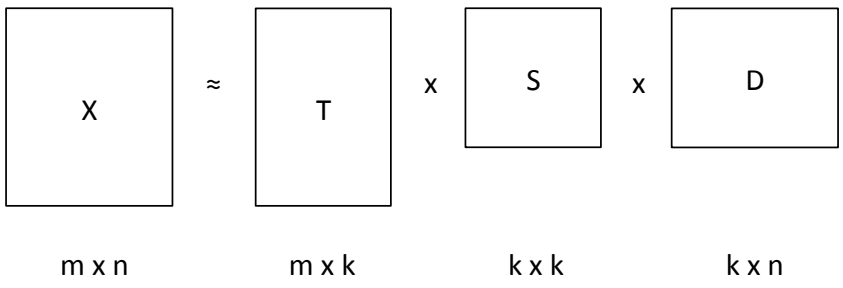

Figure 1. Singular value decomposition of X.

In order to approximate the original term-document matrix X well, $\hat{X}$ exploits patterns in word co-occurrence. Each document j is approximated by a linear combination of the k rows of TS, which are interpreted as latent semantic features of the corpus. The coefficients of the linear combination come from the j-th column of D'. Typically, k is chosen to be much

smaller than the number of words m, forcing the model to choose latent features that efficiently capture the variation in word frequencies among the documents.

However, unlike the word clusters that LDA (section 2.2) finds, the latent feature vectors do not generally map well to concepts. Two factors contribute to the lack of interpretability of the latent features. First, though the word-document matrix has only non-negative entries, the matrix factors in the singular value decomposition of X can have negative entries. That is, the effect of one feature vector can be partly cancelled out by that of another. In addition, there is no penalty for using many features to approximate a document. Consequently, LSA often relies upon complex cancellations among many features to represent groups of frequently co-occurring word types, making the individual features difficult to interpret.

2.1.2 Non-negative Matrix Factorization

Lee and Seung (1999) present non-negative matrix factorization (NMF), in which the factorization of X is constrained to have all non-negative entries. By approximating X as the product of non-negative matrices, as shown in Figure 2, NMF avoids the cancellations that make feature vectors less interpretable in LSA.

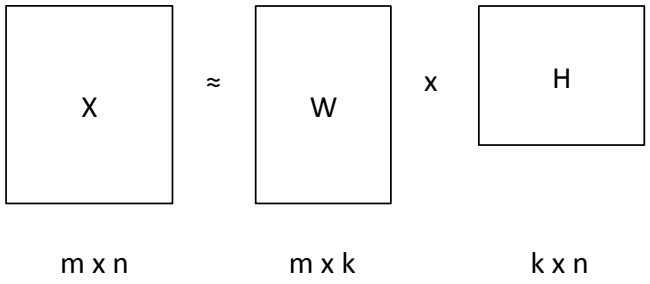

Figure 2. Non-negative matrix factorization.

In order to approximate a document, NMF uses additive combinations of latent feature vectors. The feature vectors have only non-negative coefficients, so adding in a

feature cannot cancel part of the effect of another. As a result, the features NMF chooses tend to be more individually meaningful than those that LSA chooses. Lee and Seung describe the way NMF chooses latent features to approximate documents by saying that "NMF learns a part-based representation." However, NMF does not address the concern that the number of features used to represent a document should be small.

2.1.3 Probabilistic Latent Semantic Analysis

Using a probabilistic approach rather than one based on linear algebra, Hofmann (1999) also addresses the cancellation problem in LSA, where negative entries in the singular value decomposition of the word-document matrix X allow for themes in the document corpus to be represented by cancellations among feature vectors. The technique Hofmann proposes, Probabilistic Latent Semantic Analysis (PLSA), defines a mixture distribution over word observations (d, w), where d is the document label and w is the word type. PLSA posits that the document label and type of a word are conditionally independent given the underlying topic z of the word, as shown in the graphical model in Figure 3. Under this assumption, the mixture distribution over words can be written as:

$$P(d, w) = p(d) \sum_z P(w|z) P(z|d),$$

where $P(w|z)$ is the mixture component and $P(z|d)$ is the mixture weight.

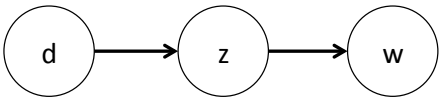

Figure 3. Graphical model for PLSA.

Since the number of latent features is typically much smaller than the size of the vocabulary or the number of documents, PLSA reduces the dimensionality of the data just as LSA and NMF do. Each document is represented as a mixture of a small number of latent features. Much like NMF, only non-negative combinations of features are allowed.

Determining the values of the model parameters (the word-topic probabilities $P(w|z)$ and the topic-document probabilities $P(z|d)$) that are most probable in light of the data will, much like the best low-rank approximate factorization in the case of NMF, uncover groups of related words that may represent the same underlying concept or, at least, be related to the same topic.

Given that PLSA achieves a similar outcome compared to NMF, it is not immediately clear why one should be preferable to the other. The two techniques are very similar. In fact, they are equivalent given a certain choice of objective function for NMF (Ding, Li, & Peng, 2008).

PLSA has two key weaknesses. First, PLSA models the generation of new words in the set of existing documents, but does model the generation of new documents. There is not a natural way, then, to infer the topics that are present in a document outside the training set. Second, since the number of parameters that must be estimated grows linearly with the number of documents, overfitting can be a problem. Hofmann (1999) uses a procedure for approximate inference called tempered Expectation Maximization in order to reduce overfitting. Whereas PLSA models the generation of words, Latent Dirichlet Allocation models the generation of documents.

2.2 Latent Dirichlet Allocation

In Latent Dirichlet Allocation (Blei et al., 2003), a new document is generated by first drawing a topic mixture from a Dirichlet prior. Then, for each word instance in the document, a topic is sampled and, conditioned on the chosen topic, a word type is sampled. Optionally, a Dirichlet prior may also be placed on the topic distributions to produce the smoothed LDA model, a choice other authors have made (Griffiths & Steyvers, 2004). We will focus on smoothed LDA in this paper, and hereafter refer to it as simply LDA. Figure 4 shows the graphical model for LDA.

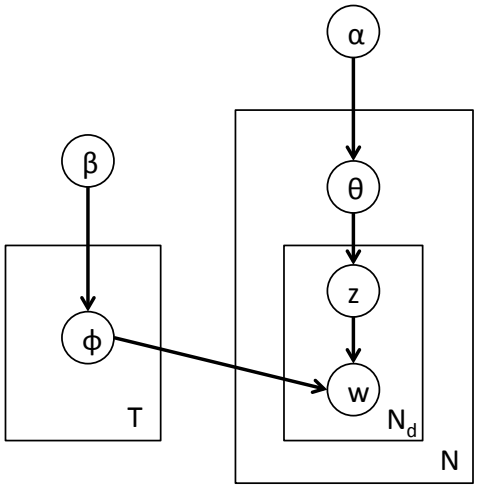

Figure 4. Graphical model for LDA.

Assuming there are N documents, each having length N_d, a vocabulary of W word types, T topics, and parameters α and β, LDA generates a corpus according to the following steps.

1. For each topic j, sample a word distribution ϕ_j from a symmetric Dirichlet distribution with scalar concentration parameter β.
2. For each document d:
    a. Sample a topic mixture θ_d from a symmetric Dirichlet distribution with scalar concentration parameter α.
    b. For each word i in d:
        i. Sample a topic z_{di} from a multinomial distribution with parameter θ_d.
        ii. Sample a word type w_{di} from a multinomial distribution with parameter $\phi_{z_{di}}$.

To see how this procedure relates to the graphical model in Figure 4, note that there are T topic distributions ϕ (one for each topic) that depend on the Dirichlet concentration parameter β, and there are N topic mixtures (one for each document) that depend on the Dirichlet concentration parameter α. The nested plate indicates that there are N_d topic and word type variables for each document d, i.e., a pair for each word in d. For each word, the

topic z depends on the document-level topic mixture θ and the word type w depends on z and the topic distributions ϕ.

The concentration parameters α and β are usually chosen to be less than one, reflecting an expectation that topic mixtures and topic distributions are concentrated, i.e., that each document has a small number of topics and that each topic favors a small subset of the overall vocabulary.

2.3 Modeling Topics Over Time

Previous studies have considered how time can be incorporated into LDA. Some model the evolution of topics over time, while others focus on how the probability that a topic occurs may vary with time. None of this prior work, however, ties time-specific patterns in a document collection (related to an event, for example) to underlying themes that persist through the period spanned by the documents. The technique proposed in this thesis, by contrast, achieves this result.

2.3.1 Dynamic Topic Model

In the Dynamic Topic Model, a topic's word distribution varies from one time period, or *epoch*, to the next in a Markovian fashion (Blei and Lafferty, 2006). By allowing the word distributions in each epoch to vary slightly from the previous epoch, the model captures the way in which the words associated with a topic change with time.

Blei and Lafferty applied their model to a subset of the articles in the journal *Science* published between 1881 and 1999, with word distributions changing on a yearly basis. One of the topics the model found was an "atomic physics" topic that changed substantially during the time spanned by the articles. "Matter" became rarer while "quantum" became more common. Despite the change, by examining the top ten most frequent words for the topic at ten-year intervals, one can see that the topic remains "atomic physics." It is nevertheless possible within this model for a topic's word distribution to change to such an extent that it is no longer recognizable as representing the same topic. In addition, the Dynamic Topic Model favors gradual change from one period to the next (Hall et al., 2008),

whereas the vocabulary around a topic can shift abruptly in domains such as news. The new method proposed by this thesis, by contrast, enforces stability in the core meaning of topics while also permitting substantial changes from one period to the next.

2.3.2 Infinite Dynamic Topic Model

A later study expanded on the Dynamic Topic Model with a nonparametric model, the Infinite Dynamic Topic Model, that allows topics to be "born" and to "die" in each period (Ahmed and Xing, 2010). If a topic survives from one period to the next, its word distribution evolves similarly to the situation in the Dynamic Topic Model. However, the Infinite Dynamic Topic Model also does not guarantee that a topic will be coherent over its lifespan. Again, the proposed approach, Epoch LDA, guarantees that topics remain coherent even as they change over time.

2.3.3 Topics over Time

Wang and McCallum (2006) present the Topics over Time model, in which each topic's word distribution is fixed but the probability that the topic will be involved in a document depends on when the document was created. Along with a distribution over word types, each topic is associated with a Beta distribution over the time interval spanned by the documents. Due to the variety of shapes the Beta distribution can have, the model can account for a topic whose popularity rises or falls gradually or is tightly focused around a particular time.

When the authors applied their model to the text of all 208 U.S. State of the Union addresses, the model successfully learned topics that were localized in time, such as the construction of the Panama Canal, as well as topics that rose gradually over time, such as the Cold War. However, these tightly focused topics are not related to broader topics, in contrast to Epoch LDA.

CHAPTER 3

Approximate Inference

In order to discover latent topics in a document corpus using LDA, it is necessary to estimate the values of the hidden variables in the model that are most likely given the observed set of documents. That is, a choice of topic mixtures $\boldsymbol{\theta}$, topic distributions $\boldsymbol{\phi}$, and per-word topic assignments $\boldsymbol{z}$ is needed that is maximally probable given the data. This is known in Bayesian statistics as a maximum *a posteriori* estimate. Unfortunately, computing such an estimate exactly is impractical because the number of ways to assign topics to words grows exponentially with the number of words. Therefore, a way to approximately estimate the hidden variables given the data is required.

This chapter introduces Gibbs sampling as a method for approximately sampling from a complex probability distribution, describes the application of Gibbs sampling to LDA, and finally discusses convergence in Gibbs sampling for LDA.

3.1 Gibbs Sampling

Gibbs sampling is a method for approximately sampling from a distribution over a set of random variables (Casella & George, 1992). A Markov chain is constructed that has the target distribution as its unique stationary distribution. It is thus guaranteed that the distribution over possible states for the chain at time t becomes arbitrarily close to the target distribution as t increases, regardless of the initial state. After the chain has converged (its distribution is deemed close enough to the target), successive steps in the chain can be used as samples from the target distribution.

To construct the Gibbs sampler, an initial state for the random variables is chosen. To generate the next state, a new value is iteratively sampled for each random variable, one at a time, conditioned on the current values of the rest of the variables. Rather than sampling

from the joint distribution over all the variables, one only needs to sample the (often simpler) conditional probability distribution for each variable given the values of the other variables.

A difficulty with this method is determining when the chain has converged. Unfortunately, there is not a rigorous way to diagnose convergence. In qualitative terms, the state of the chain in the first sampling iterations should be very dependent on the choice of initial state but less dependent as the chain converges. One way to see this is to execute several independent runs of Gibbs sampling and observe whether the chains become less dependent on the initial state after some sufficiently large number of iterations. An example of this qualitative analysis of convergence will be seen in section 3.3.

3.2 Collapsed Gibbs Sampler for LDA

Rather than sampling values for all the hidden variables in LDA, Griffiths and Steyvers (2004) suggest integrating out the topic mixtures and topic distributions and sampling values only for the per-word topic assignments. The continuous variables can be estimated later based on the topic assignments. This technique is known as collapsed Gibbs sampling. By sampling in a subset of the overall state space, the chain can converge more quickly. What is needed, then, is the conditional probability for an individual topic assignment given all the others, which is given by the following formula:

$$p(z_{di} = j | w, z_{-di}) \propto \frac{n_{-di}(w_{di}) + \beta}{n_{-di}(\cdot, j) + W\beta} \frac{n_{-di}(d, j) + \alpha}{n_{-di}(d, \cdot) + T\alpha}$$

The term $n_{-di}(w_{di}, j)$ counts the number of word instances with type w_{di} that have been assigned topic j, excluding word instance i in document d, while $n_{-di}(d, j)$ counts the number of word instances in document d that have topic j, again excluding word instance i in document d. A dot indicates a summation over all possible values of the index. α and β are the Dirichlet parameters.

Given a set of topic assignments $\boldsymbol{z}$, the topic distributions and mixtures can be estimated as follows:

$$\phi_{jw} = \frac{n(w,j) + \beta}{n(\cdot,j) + W\beta}$$

$$\theta_{dj} = \frac{n(d,j) + \alpha}{n(d,\cdot) + T\alpha}$$

3.3 Convergence and Autocorrelation

As mentioned before, it can be difficult to determine when the Gibbs sampler has converged. Though the chain is guaranteed to converge, it is not guaranteed to converge quickly. That is, there is no way to know in advance how many samples must be taken before the distribution of the next sample is acceptably close to the target distribution. In practice, one observes over several sampling runs that the chain wanders from the initial state (which varies between runs) toward a common pattern, which provides qualitative evidence for convergence. In the specific case of LDA, the Gibbs sampler moves quickly to one of several high probability regions in the space of topic assignments and then remains there indefinitely, unless the corpus is trivially small. Beginning from an initial random assignment of topics to words, within a few hundred iterations the topic mixtures and distributions settle and do not vary significantly.

To illustrate this phenomenon, Figure 5 shows the results of an experiment that the author of this thesis carried out, in which Gibbs sampling was applied for a set of online reviews of two distinctly different businesses, a Thai restaurant and a farmers' market. There were 41 reviews of each type. The number of topics was set at 2 in the expectation that LDA would cluster the two kinds of reviews. The topics are labeled 0 and 1. Each line in Figure 5 represents how the proportion of words assigned to topic 0 in the farmers' market reviews evolves over 20 runs of Gibbs sampling. One thousand iterations were performed for each run. Though the proportion of topic 0 words starts at around fifty percent, as a consequence of topics being randomly assigned initially, the proportion changes rapidly and settles near either zero or one hundred percent. Within 100 or 200 iterations, the behavior of each chain follows one of two trends, despite each chain having a different random initial state. In other

words, by 200 iterations it is likely that the chain has converged near the posterior distribution.

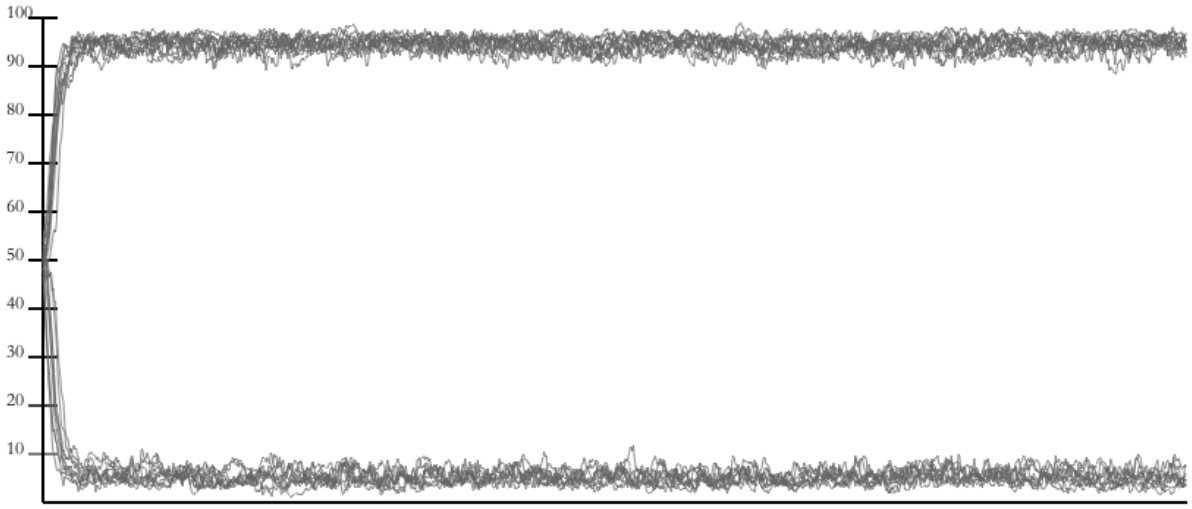

Figure 5. Proportion of words assigned topic 0 in farmer's market reviews over 20 runs.

Note also that there is no run in which the proportion crosses back over 50% after the first few iterations. Whether topic 0 or 1 is the market topic is arbitrary; permuting the topic labels merely labels differently the same clustering of words. The posterior probability of a set of topic assignments in LDA is invariant to permutations of the topic labels. Thus, in a set of independent samples from the posterior over topic assignments, most of the words in the market reviews should have topic 0 about half of the time. In contrast, in all 20 runs shown in this figure, after the chain converges, topic 0 is assigned either to the majority or the minority of market words in every additional sample taken from the Gibbs sampler. This shows that the Gibbs sampler mixes poorly. Once the chain has converged and either topic 0 or topic 1 has become the market topic, the character of the topics tends to stay the same. Gibbs sampling often suffers from poor mixing (Casella & George, 1992). In light of this fact about

Gibbs sampling for LDA, the usual strategy is to use a single sample as the estimate of $\boldsymbol{z}$ once the chain has converged (Griffiths & Steyvers, 2004).

The author of this thesis has implemented collapsed Gibbs sampling for LDA using the Python programming language. This implementation in turn was the basis of an inference implementation for the new framework, Epoch LDA. The next chapter will describe Epoch LDA, as well as the novel inference technique.

CHAPTER 4

Epoch Latent Dirichlet Allocation

In order to find topics that have a time-dependent aspect and yet are coherent across time, this thesis proposes a new technique, Epoch Latent Dirichlet Allocation (Epoch LDA). This chapter presents the model structure and describes an adaptation of collapsed Gibbs sampling to infer the topic of each word and whether the word is time-dependent.

4.1 Model Structure.

Epoch LDA, like LDA, supposes that each word of a document arises from a topic. However, the words can either be associated with a topic independently of time (the topic's *core* words) or in a time-dependent fashion (the topic's *epoch-specific* words). Thus, there are twice as many clusters to which a word can belong than in an LDA model with the same number of topics. The time of creation for each document is assumed known for each document. The time spanned by the documents is partitioned into several epochs. Each topic is associated with a core word distribution as well as an epoch-specific word distribution for each epoch. Each document, in turn, is associated with a distribution over topics. Symmetric Dirichlet priors with parameters α, β, and γ are placed on the topic, core word, and epoch-specific word distributions, respectively. The graphical model structures of LDA and Epoch LDA are shown in Figure 6.

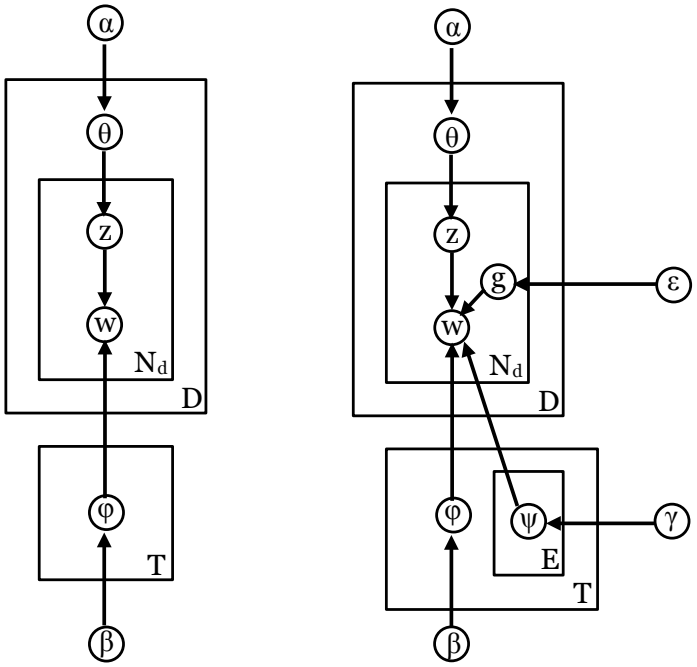

Figure 6. Graphical models for LDA (left) and Epoch LDA (right).

The generative procedure for Epoch LDA is as follows:

1. For each topic j:
    a. Draw a core word distribution ϕ_j from a symmetric Dirichlet prior with parameter β.
    b. For each epoch e, draw an epoch-specific word distribution ψ_{je} from a symmetric Dirichlet prior with parameter γ.
2. For each document d, where e_d is the epoch to which it belongs:
    a. Draw a distribution over topics θ_d from a symmetric Dirichlet prior with parameter α.
    b. For each word i in d:
        i. Draw a topic z_{di} from θ_d.

ii. With probability ϵ, set time indicator g_{di} to 1, which means the word is a core word. Set g_{di} to 0 otherwise.
iii. Draw a word type w_{di} from $\phi_{z_{di}}$ if g is 1, otherwise from $\psi_{z_{di}e_d}$.

4.2 Inference

To approximately infer the values of the hidden variables $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\boldsymbol{\psi}$, **z**, and **g**, the collapsed Gibbs sampling scheme for LDA described in Griffiths and Steyvers (2004) was adapted. After marginalizing out the word distributions and topic mixtures, Gibbs sampling is used to approximately sample from the posterior distribution over the word-level topic assignments. To carry out Gibbs sampling, the conditional probability of a word's topic assignment given the current topic assignments of the rest of the words in the corpus was computed as:

$$p(z_{di} = j, g_{di} = 0 | w, z_{-di}, g_{-di}) \propto (1 - \epsilon) \frac{n_{-di}(w_{di}, e_d, j) + \gamma}{n_{-di}(\cdot, e_d, j) + W\gamma} \frac{n_{-di}(d, j) + \alpha}{n_{-di}(d, \cdot) + T\alpha}$$

$$p(z_{di} = j, g_{di} = 1 | w, z_{-di}, g_{-di}) \propto \epsilon \frac{n_{-di}(w_{di}, j) + \beta}{n_{-di}(\cdot, j) + W\beta} \frac{n_{-di}(d, j) + \alpha}{n_{-di}(d, \cdot) + T\alpha}$$

where $n_{-di}(w_{di}, j)$ counts the number of words with type w_{di} that are core words for topic j, $n_{-di}(w_{di}, e_d, j)$ counts the number of words with type w_{di} that are epoch-specific words for topic j, and $n_{-di}(d, j)$ counts the number of words in document d that are either core or epoch-specific words for topic j, in each case excluding the contribution of word i in document d. A dot indicates a summation over all possible values for that index.

In the experiments with Epoch LDA, $p(g_i = 1) = 0.5$ was selected to reflect that there was no prior assumption regarding whether core or epoch-specific words are more common. For the parameters, the values $\alpha = 1$ and $\beta = \gamma = 0.1$ were chosen, which are consistent with previous studies (Griffiths and Steyvers, 2004).

CHAPTER 5

Evaluation

A corpus of 32,687 texts was collected from 23 well-known online sources of news and commentary about consumer technology.[3] The texts span a period from May 15, 2012, to July 31, 2012. This domain was selected because consumer technology is an area of considerable interest for news audiences, yet exploring the recent history of a particular topic in this area can be difficult. For example, after learning about a verdict in the recent patent infringement case between device manufacturers Apple and Samsung, a reader might be interested to know more about the recent history of patent litigation in the industry. The objective in applying Epoch LDA to this corpus is to enable such exploration by highlighting significant words for topics in different epochs.

The data were prepared using standard techniques (Blei, 2012; Blei et al., 2003; Griffiths & Steyvers, 2004): letters were lowercased, punctuation was replaced with whitespace, and the text was tokenized by splitting on whitespace. Thus, instances of " car ", " Car ", and " car!" in the text all become the token "car." This preprocessing is helpful because each of these strings of characters should map to the same word type. In addition, words shorter than three characters were removed in order to filter fragments of words left over by the removal of punctuation, such as "ve" from "I've." Words occurring in fewer than three documents or in more than thirty percent of all documents were removed, as were words appearing in a list of common English words. The purpose of this step was to remove non-content words like "the," which tend to occur very frequently, and rare words that do not provide much co-occurrence data for the analysis. After this preprocessing, the dataset

[3] The sources were CNET, TechCrunch, All Things Digital, The Verge, Gizmodo, Ars Technica, Wired, BBC, Daring Fireball, CNN, Fox Business, MSNBC, New York Times, LA Times, SlashGear, Mashable, The Next Web, NBC, ABC, CBS, Engadget, Reuters, Forbes, and Yahoo News.

contained 48,041 unique word types and 7,586,857 total words. The overall time period was segmented into seven-day epochs, and the data were grouped according to the epoch in which they occurred.

Preliminary experiments were conducted to determine a suitable number of topics. Selecting the number of topics is often accomplished through qualitative evaluation in the literature on LDA and related models (Blei, 2012; Griffiths & Steyvers, 2004; Steyvers & Griffiths, 2006). In some studies, authors have simply chosen a number of topics without supplying justification (Blei & Lafferty, 2006; Hall et al., 2008; Wang & McCallum, 2006). For this experiment, 20, 50, 100, and 200 were the numbers considered. In each case, the Epoch LDA core topic distributions were evaluated qualitatively in order to assess whether the topics were overly narrow or broad. This judgment was made based on expectations of what kind of topics would be desirable in this context. Since the context is consumer technology, topics relating to, for example, well-known technology companies such as Microsoft and Google were expected. It was observed that 50 is a suitable number of topics for this corpus. Smaller numbers of topics result in topics that are overly broad (e.g., topics focused on major consumer technology companies such as Apple, Google, and Microsoft were not present). On the other hand, larger numbers of topics tend to produce redundant topics (e.g., multiple Apple-related topics). Therefore, the number of topics was set to 50 for further experiments.

The resulting model captures core words for each topic along with epoch-specific words. Figure 7 displays a detailed view of a topic about legal proceedings, whose core signature is *court, patent, case, patents, judge*. The figure details how the epoch-specific topic signatures connect to relevant events that happened during the corresponding weeks. For each of four selected weeks, the figure lists relevant headlines from the corpus and provides a brief manual synopsis of relevant events in order to illustrate how Epoch LDA can aid understanding of the major news of the week for a given topic.

Week/ Epoch	Epoch-Specific Topic Signature	Related Headlines	Manual Synopsis of Week's Events
May 15	*oracle android java alsup itanium*	"11 Jurors Consider Claims of **Android** Patent Infringement" (Wired 5/15/2012) "**Oracle** Flaunts HP Internal Memos in Battle of **Itanium**" (Wired 5/17/2012)	Judge William **Alsup** heard closing arguments from **Oracle** in its lawsuit against Google over alleged infringement of **Java**-related patents in the **Android** operating system. **Oracle** also published documents relating to another lawsuit involving the **Itanium** processor.
June 12	*htc motorola june hearing posner*	"**HTC** Can't Sue Apple With Google's Loaner Patents, Says ITC" (All Things D 6/12/2012) "Judge gives Apple another chance to prove that **Motorola** infringed its patents" (The Verge 6/14/2012)	During this week in **June**, the U.S. International Trade Commission ruled that smartphone manufacturer **HTC** could not use patents borrowed from Google in order to seek an import ban against Apple's iPhone and iPad. In another case involving Apple, Judge Richard **Posner** gave the company a second **hearing** in which to show that **Motorola** infringed several of its patents.
July 10	*dotcom pay aereo software megaupload*	"Kim **DotCom** offers a travel deal to U.S. Justice Department" (CNET 7/10/2012) "Judge rules against broadcasters, denying injunction against **Aereo** TV" (Ars Technica 7/11/2012)	Kim **DotCom**, the founder of file-storage website **MegaUpload**, offered conditions under which he would come to the U.S. to face copyright infringement charges. In a suit brought by U.S. television broadcasters against the company **Aereo**, which streamed broadcast TV live online, a judge refused to grant an injunction ordering the company to halt its service before the trial concluded.
July 31	*samsung apple iphone design koh*	"**Apple** v. **Samsung**: Why the Future of Ideas is at Stake" (Mashable 8/1/2012) "Judge in **Apple-Samsung** trial declines to punish **Samsung** for releasing excluded evidence to media" (CBS News 8/3/2012)	Judge Lucy **Koh** heard opening arguments from **Samsung** and **Apple** in a trial concerning whether the Korean company copied patented elements of the **iPhone**'s software and **design**.

Figure 7. Epoch-specific topic signatures for a "legal" topic.

Epoch LDA was evaluated along several dimensions. First, the author examined the extent to which Epoch LDA achieves its goal of identifying coherent yet dynamic topics

through the use of core and epoch-specific word distributions. Next, the performance of Epoch LDA was compared to a baseline LDA model with automatically computed coherence and perplexity metrics as well as manual evaluation by eight human judges.

5.1 Qualitative Evaluation

The output of Epoch LDA was examined for its topic signatures with regard to both core and epoch-specific word distributions. It was found that the model performed well, identifying epoch-specific topic signatures that included words related to relevant events occurring during that week. For example, in the case of a topic whose core signature is *android, device, galaxy, devices, phone*, the epoch-specific topic signature for the week of June 25 is *google, nexus, tablet, jelly, bean.* This epoch-specific signature is associated with the announcement of the Nexus 7 tablet computer running a version of the Android operating system nicknamed "Jelly Bean." As another example, a topic with core signature *windows, microsoft, phone, nokia, system* has an epoch-specific topic signature for the week of May 28 of *release, preview, apps, skype, metro*. During that week, Microsoft announced availability of the Windows 8 release preview; it was also the one-year anniversary of Microsoft's acquisition of Skype.

Figure 7 provides a more detailed look at the epoch-specific topic signatures during selected weeks for a legal topic. The topic's core signature is *court, patent, case, patents, judge*, suggesting that the topic focuses especially on patent litigation. For each week, the manual synopsis explains the events to which the words in the epoch-specific topic signature correspond. The words for the week of May 15 (*oracle, android, java, alsup, itanium)*, for example, relate to a pair of lawsuits involving Oracle. In the week of July 31 (*samsung, apple, iphone, design, koh*), the epoch-specific words relate to a lawsuit between Apple and Samsung over patents related to the design of the iPhone. The epoch-specific topic signatures serve as a guide to patent litigation news during a given week.

5.2 Coherence and Perplexity

Given the structure of Epoch LDA, it was plausible that the core topics it discovered could be less coherent than general LDA topics, since in Epoch LDA words can also be explained by epoch-specific distributions. To determine whether this was the case, an experiment was conducted comparing the two sets of topic signatures: core topics from Epoch LDA and topics from LDA. An automated topic coherence metric was calculated based on each word pair's frequency of co-occurrence within the corpus (Mimno, Wallach, & Talley, 2011). For a topic t with M most probable words $v_1^{(t)}, \ldots, v_M^{(t)}$, topic coherence is defined as:

$$\sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})},$$

where

- $(v_m^{(t)}, v_l^{(t)})$ is the number of documents in which v_m and v_l co-occur and
- $D(v_l^{(t)})$ is the number of documents in which v_l occurs.

Smaller values (in magnitude) are better. This metric was applied to core topic signatures consisting of the ten most probable words. The average coherence score for Epoch LDA and LDA is shown in Figure 8.The coherence of core topics found by Epoch LDA was not significantly different from those found by LDA.

	Coherence	**SD**
Epoch LDA	-65.14	1.99
LDA	-65.33	0.89

Figure 8. Coherence scores for Epoch LDA and LDA.

An evaluation with respect to perplexity was also performed. Epoch LDA with 50 topics was compared against LDA with 50 topics and LDA with 100 topics, which more closely resembles the number of parameters available to Epoch LDA because it features both core and epoch-specific distributions. For each of these three models, five runs of Gibbs sampling were performed. The average perplexity for each case is shown in Figure 9. Lower

values are better. This perplexity result confirms that the additional epoch-specific distributions do not adversely affect model fit, while providing the additional time-related information for each topic that is not available within LDA.

	Perplexity	SD
Epoch LDA	506	4.35
50-topic LDA	625	5.45
100-topic LDA	500	5.24

Figure 9. Perplexity results for Epoch LDA and LDA.

5.3 Human Ratings

As a final means of comparing the two models, an experiment was conducted with eight human judges, all native speakers of English. The judges were shown the top five most probable core words for each topic and were asked to rate them on a 1 to 5 scale, with 5 being most coherent. The average human coherence rating for Epoch LDA and LDA are shown in Figure 10. Despite the additional flexibility of topics in Epoch LDA, the topics found are equally coherent to those that LDA finds.

	Avg. Human Rating	SD
Epoch LDA	3.52	1.19
LDA	3.64	1.01

Figure 10. Ratings of topic signatures by human judges.

5.4 Discussion

The evaluation described in this chapter demonstrates that Epoch LDA accomplishes the objective of discovering topics in a collection of documents that vary with time and yet

remain coherent. Epoch LDA achieves this result without sacrificing the quality of the core aspects of the topics, in comparison to the topics found by LDA. Thus, the results suggest that Epoch LDA provides topics as coherent as those identified by LDA while also modeling time-related aspects of the topics that LDA is unable to capture.

CHAPTER 6

Conclusion

As the amount of text available online continues to grow, the need for tools that facilitate exploring and analyzing large collections of documents increases. One type of high-level overview that can be useful is a description of the major themes in a document collection and how they vary with time. For example, a user looking at a current news articles might want to know what else has happened recently involving the same topic. However, prominent topic modeling approaches such as LDA are not able to capture variation in topics over time. Previous work has considered allowing topics to change gradually over time or to be localized around a specific time. However, this prior work does not address the question of how to find topics that evolve but remain coherent over time. To answer that question, this thesis has presented a novel technique, Epoch LDA, which models changes in topics over time while retaining each topic's identity through its core word distribution.

The model was applied to a corpus of consumer technology news partitioned into epochs by week. The results indicate that Epoch LDA successfully identifies topics by their core topic signatures while capturing time-specific aspects of those topics through their epoch-specific word distributions. Qualitative evaluation demonstrates that a core topic signature can be associated with an epoch-specific topic signature in a way that directly captures the relationship between the core topic and current events. Additional evaluations indicate that even though the core aspects of topics in Epoch LDA are focused on capturing only the words that do not vary with time, these core topics do not suffer a loss in coherence compared to LDA topics.

There are several promising directions for future work. This novel model could serve as the basis for unsupervised clustering across time of articles based on their core and epoch-specific topic signatures. In the news case, for example, articles from a given week involving a given topic could be clustered using time-specific words, so that each cluster corresponds

to an event during the week connected with that topic. If in addition a representative article were chosen for each cluster, then the user could view a summary of the week's news for that topic by seeing the headline and a brief snippet of each representative article, similar to Google News.

Additionally, a nonparametric version of Epoch LDA, in which the data determine the number of topics, would be worth exploring. Teh et al. (2006) showed how to use Dirichlet processes to produce a nonparametric LDA-like model. A similar approach should be possible for Epoch LDA. A nonparametric variation of the modified Epoch LDA that models events could be especially useful, to avoid having to choose the number of events to find each week. Extensions such as these that build on Epoch LDA hold great promise for automatically extracting the topics that occur within a corpus over time.

REFERENCES

Ahmed, A., & Xing, E. P. (2010). Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. *Proceedings of the 26th International Conference on Conference on Uncertainty in Artificial Intelligence*, 20–29.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4-5), 993–1022.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167–174.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Ding, C., Li, T., & Peng, W. (2008). On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing. *Computational Statistics & Data Analysis*, *52*(8), 3913–3927.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101* (Suppl 1), 5228–5235.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 363–371.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 289–296.

Mimno, D., Wallach, H., & Talley, E. (2011). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

Steyvers, M., & Griffiths, T. (2006). Probabilistic Topic Models. In T. Landauer, D. Mcnamara, S. Dennis, & W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning* (pp. 427-448). Lawrence Erlbaum Associates.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 1566–1581.

Wang, X., & McCallum, A. (2006). Topics over time. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 424-433.