Cross-Collection Topic Models: Automatically Comparing and Contrasting Text

Michael Paul
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
mjpaul2@illinois.edu

Abstract

This paper describes cross-collection latent Dirichlet allocation (ccLDA), a probabilistic topic model that captures meaningful word co-occurrences across multiple text collections. The model is applied to three different applications: discovering cultural differences in blogs and forums from different countries, discovering research topics across multiple scientific disciplines, and comparing editorial differences between multiple media sources. A variety of qualitative and quantitative evaluations of ccLDA are performed, including log-likelihood measurements and performance measurements of the model used as a generative classifier. Improvements over previous work are demonstrated. Finally, possible extensions and modifications to the model are presented with promising results.

1 Introduction

Unsupervised topic models such as Latent Dirichlet Allocation (LDA) are increasingly popular approaches to clustering large amounts of unannotated data. Topic models capture meaningful co-occurrences of words and can uncover the underlying semantic structure of a collection. They can be used to facilitate efficient browsing of a collection as well as large-scale analyses of text (Blei and Lafferty, 2009).

These models, however, have conceptually focused on one single collection of text, which is inadequate for comparative analyses of text. We thus develop an LDA-based model that can not only discover topics but also model their similarities and differences across multiple text collections.

This paper describes a new model, cross-collection LDA (**ccLDA**), which extends over the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and cross-collection mixture (ccMix) (Zhai et al., 2004) models. We improve on similar previous work by crafting a model that can better generalize data and is less reliant on user-defined parameters.

The paper is organized as follows. In Section 2 we summarize relevant previous work and give a detailed description of the model in Section 3. Section 4 details the model's inference and parameter estimation. Experimental results of various applications are presented in Section 5, followed by various model evaluations in Section 6. Finally, we show possible extensions and modifications to the model in Section 7, followed by a conclusion.

2 Previous Work

A topic model for comparing text collections (ccMix) was previously introduced by Zhai et al. (2004) for a problem called comparative text mining. For example, given a set of reviews for different laptops, ccMix can extract what is notable about each specific laptop while organizing this information by topic such as battery life and notebook design.

Our model improves over ccMix by replacing their probabilistic latent semantic indexing (pLSI) (Hofmann, 1999) framework with that of LDA. Under the ccMix model, the probability of generating a word in a document belonging to collection c is:

$$\begin{aligned} P(w) &= (1-\lambda_B)\sum_{z\in Z} P(z)(\lambda_C P(w|z) + \\ &\quad (1-\lambda_C)P(w|z,c)) + \lambda_B P(w|B), \end{aligned}$$

where each topic is denoted z. λ_B is the probability of choosing a word from the background word distribution and is user-defined. λ_C is also defined by the user and is the probability of drawing a word from the collection-independent word distribution instead of the collection-specific distribution. The parameters can be estimated using the Expectation-Maximization algorithm (Dempster et al., 1977).

However, in addition to the advantages of LDA over pLSI such as the incorporation of Dirichlet priors and a natural way to deal with new documents, our model avoids the limitations of using a single user-defined parameter λ_C – this probability is learned automatically under our model. Furthermore, we allow this probability to depend on the collection and topic, which is a less restrictive assumption.

Our model, ccLDA, shares with the LDA-Collocation (Griffiths et al., 2007) and Topical N-Grams (Wang et al., 2007) models the assumption that each word can come from two different word distributions, one of which depends on another observable variable. In these models, a word can come from either its topic's word distribution, or it can come from a word distribution associated with the previous word, in the case that the word is determined to be part of a collocation. The key difference here is that in these models, the alternative word distribution depends on the word preceding a token, while in ccLDA, this depends on the document's collection.

The model is also related to hierarchical variants of LDA, in particular the hierarchical Pachinko allocation (hPAM) (Mimno et al., 2007) model, in which both a topic and hierarchy depth are chosen, and there is a different word distribution at different levels in the hierarchy. A natural way to view our model is as a two-level hierarchy where the top level represents the collection-independent distributions and the bottom level represents the collection-specific distributions. One of the main differences here is that the discovered hierarchies in hPAM can be arbitrary, whereas the graphical structure of our model is pre-determined such that each topic has exactly one "sub-topic" representing each collection.

Wang et al. recently introduced Markov topic models (MTM) (2009), a family of models which can simultaneously learn the topic structure of a single collection while discovering correlated topics in other collections. This is promising in that this type of model makes no assentation that each topic is in some way shared across all collections. However, it does not explicitly model the similarities and differences between collections as we do in this research.

3 The Model

In this section we first review the basic pLSI and LDA models. We then introduce our extension to LDA: *cross-collection LDA* (ccLDA).

3.1 Basic Topic Modeling

The most basic generative model that assumes document topicality is the standard Naïve Bayes model, where each document is assumed to belong to exactly one topic, and each topic is associated with a probability distribution over words (Mitchell, 1997).

While this single-topic approach can be sufficient for classification tasks, it is often too limiting for unsupervised grouping of semantically related words into topics. A better assumption is that each document is a mixture of topics. For example, a news article about a natural disaster may include topics about the causes of such disasters, the damage/death toll, and relief aid/efforts. Probabilistic latent semantic indexing (pLSI) (Hofmann, 1999) is one such model. Under this model, the probability of seeing the ith word in a document is:

$$P(w_i|d) = \sum_{z \in Z} P(w_i|z)P(z|d)$$

One of the main criticisms of pLSI is that each document is represented as a variable d and it is not clear how to label previously unseen documents. This issue is addressed by Blei et al. with latent Dirichlet allocation (2003). Furthermore, the probabilities under this model have Dirichlet priors, which results in more reasonable mixtures and less overfitting. In LDA, a document is generated as follows:

1) Draw a multinomial distribution of words ϕ_z from Dirichlet(β) for each topic z

2) For each document d[1], draw a topic mixture distribution $\theta^{(d)}$ from Dirichlet(α). Then for each word w_i in d:

[1] One should also assume that a document length is sampled from an arbitrary distribution, but this does not affect the derivation of the model, so we ignore this here and elsewhere.

a) Sample a topic z_i from $\theta^{(d)}$
b) Sample a word w_i from ϕ_z

The Dirichlet parameters α and β are vectors which represent the average of the respective distributions. In many applications, it is sufficient to assume that these vectors are uniform and to fix them at a value pre-defined by the user. In this case, the Dirichlet priors simply function as smoothing factors.

3.2 Cross-Collection LDA

In this subsection we introduce our extension of LDA for comparing multiple text collections, which we refer to as cross-collection LDA (ccLDA). Under this model, each topic is associated with two classes of word distributions: one that is shared among all collections, and one that is unique to the collection from which the document comes. For example, when modeling reviews of different laptops, the topic describing the preloaded software contains the words "software", "application", "programs", etc. in its shared distribution with high probability, and the Apple-specific word distribution contains the words "itunes", "appleworks", and "iphoto".

When generating a document under this model, one first samples a collection c (which is observable in the data), then chooses a topic z and flips a coin x to determine whether to draw from the shared topic-word distribution or the topic's collection-specific distribution. The probability of x being 1 or 0 comes from a Beta distribution (the bivariate analog of the Dirichlet distribution) and is dependent on the collection and topic of the current token.

The generative process is thus:

1. Draw a collection-independent multinomial word distribution ϕ_z from Dirichlet(β) for each topic z

2. Draw a collection-specific multinomial word distribution $\sigma_{z,c}$ from Dirichlet(δ) for each topic z and each collection c

3. Draw a Bernoulli distribution $\psi_{z,c}$ from Beta(γ_0, γ_1) for each topic z and each collection c

4. For each document d, choose a collection c and draw a topic mixture $\theta^{(d)}$ from Dirichlet(α_c).

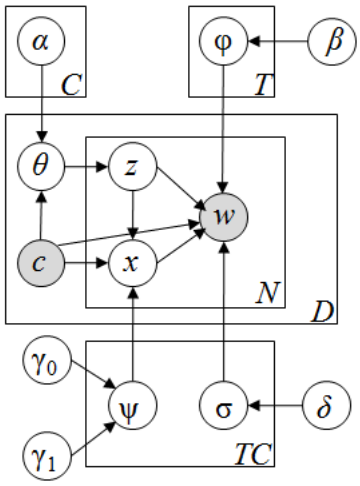

Figure 1: Graphical representation of ccLDA. C is the number of collections, T is the number of topics, D is the number of documents, and N is the length of each document.

Then for each word w_i in d:

a) Sample a topic z_i from $\theta^{(d)}$
b) Sample x_i from $\psi_{z,c}$
c) If $x_i = 0$, sample a word w_i from ϕ_z; else if $x_i = 1$, sample w_i from $\sigma_{z,c}$

As mentioned in section 2, this model is in some respects an LDA-based analog of the Zhai et al. (2004) model (ccMix), and thus it offers the same improvements that LDA offers over pLSI (described in the previous subsection), but there are some other differences. An obvious structural difference between the models is that ccMix has a special topic for background words, whereas we simply address this by removing stop words during preprocessing, which seems to give reasonable performance in this respect. This could easily be incorporated into our model such that x can take a third value that designates that a word comes from the background, but removing stop words hugely reduces the number of tokens in the data, and thus very significantly improves the time needed to estimate the model.

In the ccMix model, the probability that a word comes from the collection-specific distribution versus the shared distribution depends on a single user-defined parameter λ_C. Since it is not clear how to set this parameter[2], in our model, we learn this proba-

[2]If needed, one can effectively set this probability manually

bility automatically. Furthermore, the nature of the λ_C parameter is quite restrictive in that it is the same regardless of the topic and collection. In our model, this probability depends on the collection and topic, which should allow for a more accurate fitting of the data, as some topics may be shared across the collections to a different degree than others.

Additionally, our model allows the topic distributions for each document to come from non-uniform Dirichlet priors (parameterized by the vector α_c) that depends on the document's collection. Because the learned Dirichlet parameters can be interpreted as the average mixing level of each topic in the different collections, we can easily determine if a topic is not shared among all collections, and thus we can automatically remove or set aside such topics. We discuss this in detail in subsection 6.4.

4 Inference and Parameter Estimation

Exact inference is often intractable in complex Bayesian models and approximate methods must be used. Blei et al. (2003) offer a variational EM algorithm for LDA. Griffiths and Steyvers (2004) show how Gibbs sampling can be used for approximate inference in LDA. Gibbs sampling is a type of Markov chain Monte Carlo algorithm and is what we employ in this paper, as it is simple to derive, comparable in speed to other estimators, and it does not get trapped in local minima as easily as EM algorithms.

In a Gibbs sampler, one iteratively samples new assignments of hidden variables by drawing from the distributions conditioned on the previous state of the model (Gilks et al., 1995). In each Gibbs sampling iteration we alternately sample new assignments of z and x with the following equations:

$$P(z_i|x_i = 0, \mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta) \propto (n^d_{z_i} + \alpha_{cz}) \times \frac{n^{z_i}_{w_i} + \beta}{n^{z_i}_{\cdot} + W\beta} \quad (1)$$

$$P(z_i|x_i = 1, \mathbf{z}_{-i}, \mathbf{w}, \alpha, \delta) \propto (n^d_{z_i} + \alpha_{cz}) \times \frac{n^{z_i,c}_{w_i} + \delta}{n^{z_i,c}_{\cdot} + W\delta} \quad (2)$$

$$P(x_i = 0|\mathbf{x}_{-i}, \mathbf{z}, \mathbf{w}, \gamma, \beta) \propto \frac{n^{z,c}_{x=0} + \gamma_0}{n^{z,c}_{\cdot} + \gamma_0 + \gamma_1} \times \frac{n^{z_i}_{w_i} + \beta}{n^{z_i}_{\cdot} + W\beta} \quad (3)$$

$$P(x_i = 1|\mathbf{x}_{-i}, \mathbf{z}, \mathbf{w}, \gamma, \delta) \propto \frac{n^{z,c}_{x=1} + \gamma_1}{n^{z,c}_{\cdot} + \gamma_0 + \gamma_1} \times \frac{n^{z_i,c}_{w_i} + \delta}{n^{z_i,c}_{\cdot} + W\delta} \quad (4)$$

in ccLDA as well by using a large prior.

Because of the conjugacy of the Beta/Dirichlet and binomial/multinomial distributions, we can integrate out θ, ϕ, σ and ψ to obtain these equations, a technique known as "collapsed" Gibbs sampling (Heinrich, 2008).

n^b_a denotes the number of times a has been assigned to b, excluding the assignment of the current token i. W is the size of the vocabulary. x should be initialized as 0 for all tokens; that is, we initially assume that everything comes from the shared word distributions, otherwise the collection-specific word distributions will form independently.

α_c is a non-uniform vector that is collection-specific. A simple and efficient way to approximate this is through moment-matching such that $\alpha_{cz} \propto \frac{1}{N_c} \sum_d \frac{n^d_z}{n^d_{\cdot}}$, where d belongs to collection c and N_c is the number of documents in c (details in (Minka, 2003); (Li and McCallum, 2006)). The other hyperparameters can be updated similarly, although in our research we simply keep that at fixed, uniform values, as they do not largely affect the sampling procedure at small values.

5 Applications

In this section we present examples of three applications of ccLDA.

5.1 Cross-Cultural Analysis

Our experiments focus on discovering cultural differences by running our model on text from or about three countries: the UK, India, and Singapore. We experiment with datasets with two distinctly different perspectives: one in which the text is about each country (*tourists*), and one in which the text is authored by residents of each country (*locals*).

A thorough study of this application can be found in (Paul and Girju, 2009a).

5.1.1 Experimental Setup

In our first experiment, we model 3,266 discussions from the forums at lonelyplanet.com, the largest blog website for travelers with a forum for nearly every potential travel destination. We crawled 1,108 threads from the UK forum, 1,112 from the India forum, and 1,046 from the Singapore forum. Messages are predominantly written by people who have traveled or plan to travel to that country. We

show how this can be used for comparative content aggregation and summarization.

In the second experiment, we compare by authorship (blogs written by *locals*), and we run our model on 7,388 English-language weblogs from the same set of three different countries. We downloaded 2,715 blogs from the UK, 2,630 blogs from India, and 2,043 blogs from Singapore. We found these English-language blogs through blogcatalog.com, a blog directory which lists a blog's language and country of origin. We downloaded only the front page of each blog, which usually included multiple articles or postings, and treated each such page as a single document.

In both datasets, we removed HTML tags, stop words, and words with a corpus frequency less than 20. All punctuation characters were treated as word separators.

In our experiments, we ran the Gibbs sampler for a burn-in period of 3000 iterations, then we collected and averaged 15 samples, each separated by a 100-iteration lag. We used $\beta = \delta = 0.01$ and $\gamma_0 = \gamma_1 = 1.0$.

5.1.2 Perspective of Tourists: Topics in Travel Forums

We modeled this dataset with 25 topics. General topical words were grouped into the shared word distribution of each topic, but each collection-specific distribution contained words in the topic that best describe that country. For example, the topic on weather is characterized by words like *weather*, *rain* and *snow*, but each collection's distribution might give one a sense of the weather in each country. Table 1 shows that travelers in India, for example, should be aware of monsoon season, and travelers to Singapore can expect to be hot and sweaty. The UK distribution suggests that campers should prepare for potentially hazardous weather with the appropriate clothing and gear.

As another example, let's consider the topic whose shared words are *english*, *school*, *language*, and *speak*. The results show that English is common to all three, but the collection-specific word distributions indicate that Irish language is found in the UK region, Hindi is common in India, and Mandarin is common in Singapore.

Other common topics include immigration requirements, monetary issues, air and rail travel, etc., all containing information specific to each country. This could be used for automatic summarization by topic which would be useful either to travelers who are visiting multiple destinations, or for a potential traveler in the process of choosing where to go. Someone interested in shopping for music should go to the UK while someone interested in electronics should go to Singapore, for example (at least according to one of the topics discovered).

weather time day going rain summer month high days thanks		
UK	**India**	**Singapore**
wind	leh	hot
waterproof	monsoon	humid
ending	road	humidity
rolling	manali	heat
walkers	ladakh	degree
rochdale	trekking	equator
layers	trek	sweat
snow	season	bring
footwear	rains	rain
ankle	monsoons	umbrella

Table 1: The topic of weather, modeled across travel forums for three different countries.

5.1.3 Perspective of Locals: Topics in Blogs

Table 2 shows 3 topics induced from modeling this data with 50 topics. By looking at these we can see some clear differences between the three groups of native bloggers. For example, Topic 2 is about food, and we can compare which foods are popular in each country - cheese and soup in the UK, curry in India, and seafood in Singapore. We also noticed that tea and coffee are more popular in Singapore, wine and beer are more popular in the UK, while in Indian blogs beverages are not commonly mentioned. Perhaps a less trivial observation is that the words restaurant and chef are frequent in UK blogs, but the Indian word distribution is dominated by words pertaining to recipes. From this one might infer that people in the UK (and to a lesser extent in Singapore) eat out more often than people in India, who do more home cooking.

There are many topics not shown here including politics, gardening,and health. From the travel topic, shown in Table 7, we see that people travel close to

Topic 1			Topic 2			Topic 3		
fashion style look dress wear new collection accessories black			food add chicken recipe cooking taste rice recipes sugar soup			god jesus lord life faith holy man christ church love		
UK	**India**	**Singapore**	**UK**	**India**	**Singapore**	**UK**	**India**	**Singapore**
shoes	fashion	price	food	recipe	coffee	church	krishna	god
fashion	women	posted	wine	recipes	cup	god	religion	sin
clothing	indian	earrings	restaurant	powder	oil	john	religious	john
high	designer	length	coffee	indian	comments	todd	spiritual	spirit
designer	sarees	item	cheese	salt	fried	bentley	guru	things
style	leather	sgd	soup	tsp	add	jesus	lord	lamb
love	girls	silver	eat	rice	restaurant	christ	sri	exodus
london	china	clothes	chef	masala	rice	luke	shri	suffering
shirts	jewellery	shop	english	oil	tea	bible	baba	cross
bag	jewelry	code	drink	coriander	seafood	christian	hindu	lives

Table 2: A sample of topics induced on a set of blogs from 3 countries. Shown are the top 10 words from the shared topic-word distribution $P(word|x = 0, topic)$ and the top 10 words from $P(word|x = 1, topic, class)$ for each collection.

home, so to speak. Britons travel around Europe, especially Spain, Paris and London, while Singaporians travel to popular destinations in that part of the world, such as Hong Kong, Thailand and Bali.

5.2 Interdiscplinary Research Analysis

A common application of topic models like LDA is the discovery of topics in scientific literature (Griffiths and Steyvers, 2004), and this has been studied in extended models that include documents' authors (Rosen-Zvi et al., 2004) and date of publication (Wang and McCallum, 2006). In this domain, topic models can also be used to assign research papers to reviewers (Karimzadehgan et al., 2008) (Mimno and McCallum, 2007). In computational linguistics, (Hall et al., 2008) use LDA to model topics in this field and study their history.

These studies, however, have ignored the multifaceted and interdisciplinary nature of many scientifc topics. (Paul and Girju, 2009b) model scientific literature from multiple disciplines such as computational linguistics and linguistics, however, the fields are modeled independently using LDA. We show how ccLDA can improve upon this by incorporating multiple disciplines directly into the model.

5.2.1 Experimental Setup

Our corpus consists of approximately 11,100 abstracts from the ACL Anthology[3] and 6,000 abstracts from Linguistics journals. The exact distribution is shown in Table 3. We chose to include journals based on the following criteria: the journal is considered a "top" journal, the journal covers topics in areas that are pertinent to this project, and the journal covers a timespan of at least a decade.

We removed a standard set of stop words as well as words with a corpus frequency less than 10. All punctuation was treated as a word separator.

In each experiment we ran the Gibbs sampler for a burn-in period of 2000 iterations, then we collected and averaged 15 samples, each separated by a 100-iteration lag. We used $\beta = \delta = 0.01$ and $\gamma_0 = \gamma_1 = 1.0$.

5.2.2 Topic Discovery

Automatic discovery of scientific topics is an important part of modern literature analysis, and topic models like LDA can be used to aid trend analysis and browsing of related literature (Griffiths and Steyvers, 2004). Our contribution is to discover scientific topics that cross disciplines and to see how they compare and differ across fields.

For this, we modeled our corpus of 2 scientific fields – computational linguistics (CL) and linguistics (LING). We used 20 topics, determined after some empirical experimentation. If the number of topics is too large, then most of the topics that form are not shared across both collections.

Table 4 shows an example of a topic related to

[3] http://www.aclweb.org/anthology-new/

Field	Venue	No. of docs.	Year range
CL	ACL Journal	943	80-06
CL	ACL Workshops	4,122	80-07
CL	ACL	1,826	79-08
CL	EACL	517	83-06
CL	NAACL	543	01-07
CL	Applied NLP	262	83-00
CL	COLING	1,549	65-08
CL	HLT	872	86-05
CL	IJCNLP	471	05-08
CL	**Total**	**11,105**	**65-08**
LING	Language	379	93-08
LING	Linguistics	152	97-08
LING	Linguistic Inquiry	448	99-08
LING	Int. Journal of American Linguistics	449	93-08
LING	Int. Journal of Sociology of Lang.	1,778	76-08
LING	Language & Speech	1,385	58-08
LING	Natural Language & Ling. Theory	558	83-08
LING	Ling. & Philosophy	847	77-08
LING	**Total**	**5,996**	**58-08**

Table 3: Dataset - number of tokens and documents per field and publication venue. CL stands for Computational Linguistics; LING - Linguistics.

communication. We see that in CL, this is strongly relevant to dialogue systems; in linguistics, this topic is more focused on human behavior and social interaction.

Another example is the topic of grammar and structure. The LING distribution features words like *clause* and *subject*, while the CL distribution is dominated by words pertaining to *parsing*.

speech spoken interaction human discourse paper understanding task	
CL	**LING**
dialogue	social
user	communication
systems	verbal
information	women
utterances	speaker
dialogues	speakers
utterance	relationship
agent	interaction
plan	ways
recognition	means
agents	behavior

Table 4: The topic of communication as it appears in the computational linguistics (CL) and linguistics (LING) datasets.

5.2.3 Topic Evolution Over Time

An interesting analysis we can do with research topics is to consider how the topics change over time. For example, (Mei and Zhai, 2005) model the evolution of topics by partitioning the data into time periods and modeling topics in each time period. They discover related topics across time periods using KL-divergence, a measure of the similarity of two probability distributions. However, while this is a way to model a topic over time, it does not draw attention to the key changes between time intervals.

ccLDA can by applied to this task in a similar manner – by partitioning the data into time intervals – but because ccLDA will explicitly model what is unique to each interval, it may give additional insights into this problem.

For this experiment, we partitioned the computational linguistics documents into two collections: "new" publications (year $\geq$ 2000) and "old" publications (everything else). The results of ccLDA will show which words are notable in the two time periods.

For example, in the machine learning topic, we see *neural networks* at the top of the old distribution and *support vector machines* at the top of the new distribution, suggesting a shift in the type of classifiers that are commonly used.

We also see that unsupervised and semi-supervised approaches have gained prominence in the recent decade. An example of the machine translation topic is shown in Table 5 – we see that automatic alignment methods have become important recently.

5.3 Media Analysis

It is reasonable to expect that different news articles from different outlets will present the same topics, but that these topics may be emphasized differently depending on the source. ccLDA can naturally model such differences, and we briefly show how it could be used for the detection of media bias and editorial differences.

5.3.1 Experimental Setup

We collected 623 news articles from August 2008 from two American media outlets, msnbc.com and foxnews.com. We chose these outlets because they

translation machine english source target languages language statistical	
Old	**New**
transfer	alignment
mt	improvements
automatic	improvement
structural	comparable
lexical	trained
translated	alignments
match	resources
differences	novel
aided	nist
correspondence	score
translator	training

Table 5: The topic of machine translation found in computational linguistics. The data is partitioned into two time intervals: old (published before the year 2000) and new ($\geq$ 2000).

are anecdotally said to lean politically left and right, respectively, in their coverage.

We used website-specific pattern matching to extract the article text, then removed HTML tags, stop words, and words with a corpus frequency less than 10. All punctuation characters were treated as word separators.

The Gibbs sampler is run with the same parameters as in 5.1.1.

5.3.2 Editorial Differences

Because we modeled a fairly small number of documents from a short time period, we cannot say that these results are representative of long-term trends in the two venues. Nonetheless, we do see some content differences between the two in this dataset.

For example, according to the topic of economics, it seems that MSNBC extensively covers business and finance, with words like *stocks* and *trades* at the top of its distribution. The FOX distribution focuses on other economic issues such as oil drilling and poverty.

As another example, Table 6 shows the topic of cars. While the FOX distribution is fairly broad, we see that MSNBC focuses on alternative fuels and vehicles, such as diesel and hybrids.

car vehicle cars fuel drive	
MSNBC	**FOX News**
diesel	mazda
says	gallardo
autos	chrysler
camaro	minivan
tax	horsepower
credit	lamborghini
smaller	mph
mileage	sports
hybrid	lp

Table 6: The topic of cars/vehicles as found in two different news sources.

6 Model Evaluation

In this section we evaluate ccLDA against ccMix and LDA both qualitatively, through blind judgments of cluster quality, and quantitatively, by measuring the likelihood of held-out data with each model and by testing the performance of the model when used as a generative classifier.

6.1 Cluster Coherence

Because all of the above applications rely on analyses of discovered topics, it is important that we use a model that gives the best empirical quality of word clusters. We compare against ccMix (Zhai et al., 2004), the only related model that is naturally suited to our task. Using blind human judgments we show that ccLDA unquestionably delivers topics that are more coherent than those obtained with the ccMix model.

A direct comparison with ccMix is tricky because it incorporates a model for background words, whereas our model expects stop words to be removed during preprocessing. So that they are fully comparable, we set the parameter λ_B (the probability that a word comes from the background) to 0 and fed the model the same input as we did ccLDA. We set the parameter λ_C, analogous to $P(x = 0)$, to 0.6, which is the average value learned by ccLDA on this data, and intuitively it seems reasonable. Using an implementation provided by the authors of ccMix, we ran the EM procedure for 20 trials and saved the model with the best log-likelihood.

We performed human judgments of the 25 topics induced by ccLDA in the travel forum dataset

and by the ccMix model with the number of topics set again to 25. We aligned the topics automatically using a symmetric KL-divergence score computed on the collection-independent distributions – specifically, $D(P||Q) + D(Q||P)$ where $D(P||Q)$ is the KL-divergence[4] of the distributions P and Q.

Each aligned pair of topics (ordered randomly for each topic to avoid bias) was presented to two natural language processing researchers who were asked to choose which one was better, based on the following criteria: (1) semantic coherence of the topic as a whole (e.g. are the words in the clusters related?) and (2) coherence across collections, that is, are the collection-specific distributions related to each other and to the common one? The judges were also given the option to rate a pair as "no opinion" in the case that the aligned topics were too dissimilar to compare (because the two models did not discover the same topic), or that the topics did not carry enough semantic information to judge (i.e. topics composed mostly of function words).

Of the 25 pairs, there were 10 that both judges rated. Of these 10, the judges disagreed on 3. The other 7 were all rated in favor of ccLDA.

Similarly, the 50 topics from the second experiment were judged against 50 topics formed using ccMix. There were 22 topics that both judges rated. Among these, they disagreed on only 3; of the remaining topics they voted in favor of ccMix for 1 topic and in favor of ccLDA for 18 topics.

It has been observed that the performance of a model can largely depend on the estimator used (Girolami and Kabán, 2003), so it may be that the weaker performance of ccMix is because the EM algorithm is getting stuck in local maxima, even after several trials.

Table 7 shows the topic of travel compared with both ccMix and LDA. To compare against LDA, we performed a post-hoc estimation of the topic's word distribution for each collection by considering topic assignments of documents within each collection. We see that the ccLDA distributions are much more coherent than that of ccMix. Furthermore, the advantage over LDA is clear – with LDA, we do not get a separation of the words that are common to all of the collections, and thus it is hard to detect the important differences at a glance.

[4]Kullback-Leibler divergence is a commonly used measurement of the similarity of two probability distributions.

6.2 Likelihood Comparison

To measure how well our model can generalize unseen documents, we compute the likelihood of held-out data using ccLDA compared with ccMix and LDA. We partitioned the travel forum dataset into a subset of 80% of the data on which the models are learned, and an evaluation set of the remaining 20%.

To calculate the likelihood of the held-out documents with ccMix, we use the "fold-in" method (Hofmann, 1999) in which the mixing proportions except for $P(z|d)$ are fixed during the EM process. As with our cluster evaluation above, we set $\lambda_B = 0$ and $\lambda_C = 0.6$. With LDA and ccLDA, we approximate $P(z|d)$ through another Gibbs sampling procedure, by averaging 10 samples collected after 100 iterations with a 10-iteration lag in between each sample.

The log-likelihood of the three models is shown at various numbers of topics in Figure 2. As expected, ccLDA generally achieves a higher likelihood than ccMix, although the difference between them diminishes at higher numbers of topics. This appears to be because the pLSI-based ccMix does not regularize the topic mixtures and can thus achieve higher values of $P(z|d)$, and the smoothing of ccLDA has a greater effect at higher numbers of topics.

Both cross-collection models achieve a higher likelihood than LDA, which is not too surprising, given that these models utilize extra information (specifically, the document's collection) to assign a higher probability to words more likely to appear in a document given that information.

It should be noted that even though the likelihood of both cross-collection models increases with the number of topics up to 100, we observed empirically that the best cluster quality in this dataset occurs around 20 to 30 topics; more than that results in clusters that are repeated and are largely specific to only one collection. This agrees with the observation of (Boyd-Graber et al., 2009) that the likelihood of a model may not correlate with the quality of the topics as interpreted by humans.

ccLDA			ccMix			LDA		
travel hotel hotels city best place holiday visit trip world			travel hotel comments hotels city posted road trip labels airport			travel city hotel park holiday hotels place beach road visit		
UK	**India**	**Singapore**	**UK**	**India**	**Singapore**	**UK**	**India**	**Singapore**
holiday	india	singapore	yang	india	yang	travel	travel	travel
holidays	delhi	kong	train	delhi	dan	holiday	city	hotel
hotels	indian	hong	london	tourism	ini	hotel	beach	city
spain	mumbai	spa	saya	dubai	dengan	city	place	park
london	bangalore	hotel	nie	indian	untuk	london	hotel	place
great	tour	beach	travel	tour	itu	park	temple	beach
surf	air	chinese	flight	bangalore	saya	hotel	road	trip
breaks	dubai	pictures	luxury	mahindra	orang	place	park	hotels
train	city	restaurant	dan	hotels	tidak	holidays	hotels	spa
ski	mahindra	bangkok	advert	marathi	dalam	hall	tourism	visit

Table 7: The topic of *travel* as discovered by the 3 different models.

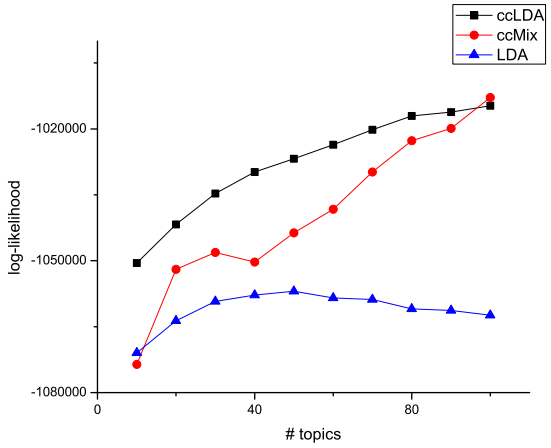

Figure 2: Comparison of the log-likelihood of held-out data with the 3 models.

6.3 Document Classification

The main thing we would like to glean from the analyses in the above applications is the set of terms within each topic that are good descriptors of what is unique to each collection. We can quantitatively evaluate the model's ability to do this by applying it to the task of collection prediction, which will give us a measure of how discriminative the collection-dependent word distributions are.

Because ccLDA gives a document likelihood that depends on the document's collection or class, it is naturally suited for this task. Classification of an unlabeled document d thus becomes the problem of choosing the c that maximizes the formula:

$$P(c) \prod_{w \in d} \sum_{z} P(z|d) \qquad [P(x=0|c,z)P(w|z,x=0) + P(x=1|c,z)P(w|z,c,x=1)]$$

These probabilities are obtained when the model is learned on a training set, except for $P(z|d)$, which depends on the new document. We can learn this through another Gibbs sampling procedure, treating the document as if c is known and doing this for all values of c, however, the ability to quickly label a new document is necessary for many classification tasks, so we instead use a simple approximation from the learned Dirichlet prior for each collection, which represents the average topic mixture within that collection. That is, $P(z|d) \approx \frac{\alpha_{cz}}{\sum_z \alpha_{cz}}$.

To see how important $P(z|d)$ is to the performance, we also experimented with approximating this as a uniform constant, $P(z|d) = \frac{1}{Z}$ where Z is the number of topics.

In our experiment, we classified documents as old vs new using the computational linguistics dataset, described in 5.2.3. Table 8 shows the 5-fold cross-validation accuracy compared against that of naive Bayes and an optimally tuned SVM. In each cross-validation iteration, the data is partitioned in the same way for each classifier; that is, they are evaluated with the same training/test sets. We used the SVM^{light} kit[5] using a linear kernel, with the regu-

[5] http://svmlight.joachims.org

larization factor (that is, the trade-off between margin size and training error) set to the default[6] $\frac{1}{x^2}$. ccLDA was run with 50 topics.

	NB	SVM1	SVM2	ccLDA1	ccLDA2
Accuracy	0.679	0.793	0.754	0.792	0.781

Table 8: The accuracy obtained by various classifiers during 5-fold cross-validation on the "new" vs. "old" dataset. NB stands for Naive Bayes. SVM1 refers to a support vector machine using the regularization parameter $C = \frac{1}{x^2}$; SVM2 uses $C = 1.0$. ccLDA1 uses the method described above with an approximation for $P(z|d)$ based on α_c. ccLDA2 uses a topic-independent, uniform approximation of $P(z|d)$.

Likely due to its ability to separate out less-discriminative words by way of the collection-independent model, ccLDA achieves comparable performance to the SVM.

We also tried using ccLDA as a classifier using different priors for $P(x)$. Intuitively, if we can force a high value of $P(x = 0)$, then more words will be pulled into the shared distribution, giving more weight to the most discriminative words. Table 9 shows the accuracy using different settings. We see that it does indeed perform much better with $P(x = 0)$ slanted toward 1 rather than 0, and the performance is comparable when using a negligible prior (the left column).

	$\gamma_0 = 1.0$; $\gamma_1 = 1.0$	$\gamma_0 = 80000.0$; $\gamma_1 = 20000.0$	$\gamma_0 = 20000.0$; $\gamma_1 = 80000.0$
Accuracy	0.792	0.792	0.726

Table 9: The accuracy obtained during 5-fold cross-validation on the "new" vs. "old" dataset with ccLDA using different priors for $P(x)$.

6.4 Identifying Problematic Topics

The ccLDA model assumes that all topics are represented in all collections in the corpus, and it struggles if this is not the case. Table 10 shows an example of this in the research papers dataset. The model tries to fit the topic of information retrieval across both the linguistics and computational linguistics collections, but because it is not really found in linguistics, a linguistics-specific word distribution forms which is coherent but completely unrelated. We offer two suggestions of how to identify topics such as these:

document information documents text retrieval web search topic content	
CL	**LING**
query	yiddish
summarization	hebrew
automatic	jewish
queries	jews
articles	israel
news	israeli
extraction	judeo
summary	fifteenth
automatically	sorbian
extract	slavic

Table 10: An example of a topic induced when modeling CL and LING that is not well-shared across both collections. The topic, at least as it is described by its shared word distribution, is really only found in CL, and an unrelated word cluster is formed in LING.

(Method A) Because the learned α_c parameters reflect the average topic mixtures for each collection c, we can check if they are exceptionally uneven. After normalizing the α_c values, if there is an i such that $\alpha_{iz} > \mu_A \alpha_{jz} \forall_{j \neq i}$ then we say the topic z is not well-shared. That is, we flag topics that are at least μ_A times more likely to appear in one collection than any of the others.

(Method B) Because $P(x|z, c)$ reflects the likelihood of using the shared vs. collection-specific word distribution, and because it depends on both the topic and document collection, we can use this to identify topics that are not well-shared. For a topic z, if $P(x = 1|z, c) > \mu_B$ for any collection c, then we flag this topic. That is, we flag topics such that there is a high probability that a collection will not use the topic's shared distribution.

It is difficult to offer guidance on how to set the μ parameters, as a good value will depend on the data in use as well as on one's needs. We evaluated these two procedures using various settings for μ.

Two judges familiar with these fields were asked to label the 20 topics induced from the CL-LING

[6] We tried many other settings and found this to consistently give the best performance.

corpus as "shared" or "not shared" across both collections based on what the topic's main word distribution would indicate. For example, if a topic's word distribution featured words such as *information*, *retrieval* and *query*, the judges would rate it as "not shared" because this is a topic that really only applies to the CL collection.

The judges could not confer with each other, and they were only shown the shared, collection-independent word distribution of each topic. This was done so that topics would be judged as "shared" if, semantically speaking, the topic is known to appear in both fields (i.e. "is this topic pertinent to both research fields?"), and not necessarily on how well the model was able to fit each topic across both collections. In the case of borderline decisions, such as with topics composed primarily of function rather than content words, the judges were told to label them "shared".

	P	**R**	**A**
$\mu_{\mathbf{A}} = \mathbf{2}$	.454	1.0	.625
$\mu_{\mathbf{A}} = \mathbf{5}$	.556	1.0	.75
$\mu_{\mathbf{A}} = \mathbf{10}$	1.0	.4	.688
$\mu_{\mathbf{B}} = \mathbf{0.3}$	.313	1.0	.313
$\mu_{\mathbf{B}} = \mathbf{0.4}$	.333	.4	.5
$\mu_{\mathbf{B}} = \mathbf{0.5}$	0.0	0.0	.5

Table 11: The precision, recall, and accuracy of our two proposed methods to identify uneven topics at various parameter settings.

Of the 20 topics, the judges agree on 16 of them. Of the 16 topics on which they agreed (11 "shared" and 5 "not shared"), we computed the precision (the percentage of topics flagged as "not shared" by the automated system were labeled that way by the judges), recall (the percentage of topics labeled as "not shared" by the judges were flagged by the system), and accuracy (the percentage of labels by the judges match that of the system).

We see in Table 11 that method B performs quite poorly, and $P(x)$ does not seem to be a good predictor for this problem. We thus recommend method A – we achieved the best precision/recall tradeoff around $\mu_A = 5$, but of course this may depend on the particular problem. It is surprising that method B does not work well, but it seems that it is quite possible for $P(x = 0|z, c)$ to appear normal but for there to be very few assignments of z in documents belonging to c, and thus it works better to consider the presence of a topic in a collection, as done in method A.

7 Model Extensions

In this section we present two extensions and modifications to ccLDA. Results from preliminary experimentation of these extensions show promise.

7.1 Modeling Background Words

A common extension to topic models is to incorporate a topic-independent language model for common "background" words, as an alternative to or an augmentation to removing stop words in preprocessing, and this is indeed a part of the ccMix model. (By "background", we mean words that do not belong to any particular topic.) We suggest removing stop words regardless, as this greatly reduces the number of tokens in the data and reduces the time needed to learn the model. However, it is hard to define which words are "stop" words and it is even harder to construct an exhaustive list of them, and consequently, many topics will form that are not topical in the semantic sense.

A simple way to add this to ccLDA is to change it such that x can take on a third value that determines that the word is drawn from a topic-independent background word distribution. The generative process thus becomes:

1. Draw a collection-independent multinomial word distribution ϕ_z from Dirichlet(β) for each topic z
2. Draw a collection-specific multinomial word distribution $\sigma_{z,c}$ from Dirichlet(δ) for each topic z and each collection c
3. Draw a topic-independent multinomial word distribution π_c from Dirichlet(λ) for each collection c
4. Draw a multinomial distribution $\psi_{z,c}$ from Dirichlet(γ) for each topic z and each collection c
5. For each document d, choose a collection c and draw a topic mixture $\theta^{(d)}$ from Dirichlet(α_c). Then for each word w_i in d:

a) Sample a topic z_i from $\theta^{(d)}$

b) Sample x_i from $\psi_{z,c}$

c) If $x_i = 0$, sample a word w_i from π_c; else if $x_i = 1$, sample a word w_i from ϕ_z; else if $x_i = 2$, sample w_i from $\sigma_{z,c}$

Something that may seem unintuitive is that a topic is assigned to a token even if the word comes from the background model and that the probability of the word coming from the background model depends on this topic. The topic assignment of background words will depend on the document's topic mixture, and it is possible that different topics will co-occur with background words with varying frequencies, thus, we feel this is a reasonable approach.

The reader may also notice that there are actually C background models, one for each collection. We did this because we noticed that different collections will have different common words, and sometimes these words ended up in collection-specific topics. Variations on these ideas would be something to experiment with in varying applications. An example of these background distributions is shown in Table 12.

To see if this model improves topic quality, we compared 20 topics induced by this model with 20 topics induced by the basic ccLDA model on the CL-LING set. In both cases, we set the γ values to 1.0. We also experimented with the background model such that $\gamma_0 = 50000$, $\gamma_1 = 50000$, and $\gamma_2 = 30000$. We presented the sets of topics to two judges who were asked to label each topic as "topical" or "not-topical". This is of course subjective, but a topic is considered topical if it is composed of meaningful content words – the idea is to see if we can induce fewer "noisy" topics when we have incorporated a background word model.

The judges agreed on 18 of the 20 topics induced by the basic ccLDA model. Of these, 4 were rated as non-topical. The judges agreed on 17 topics from the improved model, and rated 3 as non-topical. The judges agreed on 17 topics from the improved model with the large prior, and rated only 2 as non-topical.

This shows that the clusters are mostly topical, even without the background model, but it does seem that the background model removes one or two noisy topics.

CL	LING
paper	language
introduction	article
based	introduction
language	english
information	languages
using	linguistic
approach	study
first	work
work	paper
different	different
present	analysis
natural	first
processing	presents
systems	based
problem	discusses

Table 12: The top "background" words for the CL and LING collections when modeling the two with our extended version of ccLDA.

7.2 A Hierarchical Structure

We mentioned in section 2 that a natural way to conceptualize ccLDA is as a 2-level hierarchy where the top level is shared among all collections and topics in the lower level are specific to each collection. In this subsection we will formalize such a model and briefly present some results.

Imagine a model where each word is associated with both a super-topic and a sub-topic, where the sub-topic is hierarchically associated with the super-topic. Under this model, a super-topic T is first chosen according the probability of seeing the super-topic T in the document. Then a sub-topic t is chosen according to $P(t|T)$. Finally, it must be decided whether to draw a word from the super-topic distribution or the sub-topic distribution over words, and a word is chosen from this distribution.

Formally, the generative process is:

1. Draw a collection-independent multinomial word distribution ϕ_T from Dirichlet(β) for each super-topic T

2. Draw a collection-specific multinomial word distribution $\sigma_{t,c}$ from Dirichlet(δ) for each sub-topic t and each collection c

3. Draw a multinomial topic distribution $\pi_{T,c}$ from Dirichlet(λ) for each super-topic T and collection c

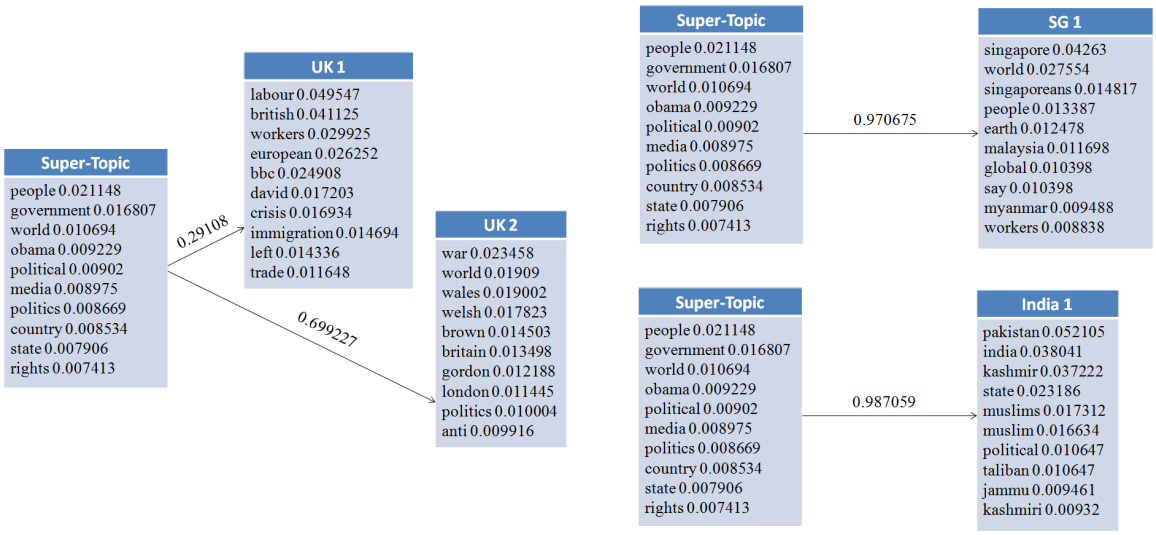

Figure 3: An example of topics discovered by the hierarchical variant of ccLDA. The directed arrow indicates the probability of the collection-specific sub-topic occuring with the collection-independent super-topic.

4. Draw a binomial distribution $\psi_{T,t,c}$ from Beta(γ_0, γ_1) for each super-topic T, sub-topic t, and collection c

5. For each document d, choose a collection c and draw a super-topic mixture $\theta^{(d)}$ from Dirichlet(α_c). Then for each word w_i in d:

   a) Sample a super-topic T_i from $\theta^{(d)}$
   b) Sample a sub-topic t_i from $\pi_{T,c}$
   c) Sample a hierarchy level ℓ_i from $\psi_{T,t,c}$
   d) If $\ell_i = 0$, sample a word w_i from ϕ_T; if $\ell_i = 1$, sample a word w_i from $\sigma_{t,c}$

This model is actually just a generalization of ccLDA. ccLDA is a special case constrained such that for each super-topic $T = j$ there is exactly one sub-topic t such that $P(t = j|T = j) = 1$ and $P(t = i|T = j) = 0, \forall i \neq j$.

As an experiment, we applied this alternative model to the blog dataset described in 5.1.3, which contains three collections: blogs from the UK, India, and Singapore. We find that the topics discovered by the model are mostly the same, and the sub-topics under each super-topic are analogous to the collection-specific word distributions of ccLDA.

Figure 3 shows the topic of politics as discovered using this model. The super-topic has words that are common in all collections, while each collection has its own sub-topic(s). As one would expect, we see that political articles from each country focus on that country's region of the world – India and Pakistan in the India collection, and Malaysia and Myanmar in the Singapore (SG) collection. This super-topic associates with two sub-topics in the UK collection: one that seems generally about UK politics, and another that seems to focus on the Labour Party.

The ability for multiple sub-topics to belong to the same super-topic is something this type of model offers over ccLDA. In the same vein, the model has the option to assign no sub-topics to a super-topic for a particular collection, which could alleviate the problem described in 6.4, which is that ccLDA may struggle to find topics that fit across all collections. A full analysis of this model is left for future research.

8 Conclusion

We have described cross-collection latent Dirichlet allocation (ccLDA), a probabilistic topic model that captures meaningful word co-occurences across multiple text collections. Three possible applications of the model are demonstrated: discovering cultural differences in blogs and forums from different countries, discovering research topics across multiple scientific disciplines, and comparing editorial differences between multiple media sources.

A variety of qualitative and quantitative evaluations of ccLDA are performed, including log-likelihood measurements and performance measurements of the model used as a generative classifier. Improvements over previous work are demonstrated. Finally, possible extensions and modifications to the model are presented and the preliminary results are promising.

Acknowledgments

The research in this paper was conducted with my advisor, Roxana Girju, to whom I am grateful for her generous support. I would also like to thank ChengXiang Zhai for thoughtful discussions and for providing an implementation of the ccMix model, as well as the students in the Semantic Frontiers group for their useful feedback.

References

D. Blei and J. Lafferty. 2009. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.

J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. 1995. *Markov Chain Monte Carlo in Practice*. CRC Press.

M. Girolami and A. Kabán. 2003. On an equivalence between plsi and lda. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 433–434, New York, NY, USA. ACM.

T. Griffiths and M. Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*.

T. Griffiths, M. Steyvers, and J. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

D. Hall, D. Jurafsky, and C. Manning. 2008. Studying the history of ideas using topic models. In *Empirical Natural Language Processing Conference*.

G. Heinrich. 2008. Parameter estimation for text analysis. Technical report, University of Leipzig.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA. ACM.

M. Karimzadehgan, C. Zhai, and G. Belford. 2008. Multi-aspect expertise matching for review assignment. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1113–1122, New York, NY, USA. ACM.

W. Li and A. McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *International Conference on Machine Learning*.

Q. Mei and C. Zhai. 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the Knowledge Discovery and Data Mining (KDD) Conference*.

D. Mimno and A. McCallum. 2007. Expertise modeling for matching papers with reviewers. In *KDD '07: Proceedings of the 13th ACM SIGKDD international*

conference on Knowledge discovery and data mining, pages 500–509, New York, NY, USA. ACM.

D. Mimno, W. Li, and A. McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *International Conference on Machine Learning*.

T. Minka. 2003. Estimating a dirichlet distribution. Technical report, Microsoft Research.

T. Mitchell. 1997. *Machine Learning*. McGraw-Hill, Boston.

M. Paul and R. Girju. 2009a. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1417, Singapore, August. Association for Computational Linguistics.

M. Paul and R. Girju. 2009b. Topic modeling of research fields: An interdisciplinary perspective. In *Proceedings of the the International Conference on Recent Advances in Natural Language Processing (RANLP) (to appear)*.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. 2004. The author-topic model for authors and documents. In *UAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, Arlington, Virginia, United States. AUAI Press.

X. Wang and A. McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, New York, NY, USA. ACM.

X. Wang, A. McCallum, and X. Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 697–702. IEEE Computer Society.

C. Wang, B. Thiesson, C. Meek, and D. Blei. 2009. Markov topic models. In *The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 583–590.

C. Zhai, A. Velivelli, and B. Yu. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of KDD 04*, pages 743–748.