# Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models

**Michael Paul and Roxana Girju**
University of Illinois at Urbana Champaign
Urbana, IL 61801
{mjpaul2, girju}@illinois.edu

## Abstract

This paper presents preliminary results on the detection of cultural differences from people's experiences in various countries from two perspectives: tourists and locals. Our approach is to develop probabilistic models that would provide a good framework for such studies. Thus, we propose here a new model, **ccLDA**, which extends over the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and cross-collection mixture (ccMix) (Zhai et al., 2004) models on blogs and forums. We also provide a qualitative and quantitative analysis of the model on the cross-cultural data.

## 1 Introduction

In today's society, people from different cultural backgrounds have to understand each other, interact on a daily base and travel to or work in more than one country. Understanding cultural diversity, as well as addressing the need to communicate effectively across cultural divides, have become imperative in almost every aspect of life. These constitute an important language aspect since the lack of such cultural awareness can lead to misinterpretations.

This paper presents preliminary results on the detection of cultural differences from people's experiences in various countries from two perspectives: tourists and locals. Since the advent of Web 2.0, user-generated data in the form of blogs and newsgroup messages have reached high proportions. In this paper we take advantage of such resources of blogs and forums to perform various cross-cultural analyses.

Our approach is to develop probabilistic models that would provide a good framework for such studies. Thus, we propose here a new model, **ccLDA**, which extends over the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and cross-collection mixture (ccMix) (Zhai et al., 2004) models. Our contribution is as follows:

(1) Unsupervised topic models such as LDA are elegant and flexible approaches to clustering large collections of unannotated data. 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.

(2) We improve on similar previous work by crafting a model that can better generalize data and is less reliant on user-defined parameters.

(3) We apply our new model on blogs and forums to identify cross-cultural differences.

Thus, different models can be compared to reflect different hypotheses about the data.

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 parameter estimation. Experimental results are presented in Section 5, followed by discussion, future work, and conclusions.

## 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 (CTM). Given news articles from different sources (about the same event), ccMix can extract what is common to all the sources and what is unique to one specific source.

Our model improves over ccMix by replacing their probabilistic latent semantic indexing (pLSI) (Hofmann, 1999) framework with that of LDA.

*Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1417,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

Under the ccMix model, the probability of generating the $i$th word in a document belonging to collection $c$ is:

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

where each topic is denoted $z$. $\lambda_B$ is the probability of choosing a word from the background word distribution and is user-defined. $\lambda_C$ 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 $\lambda_C$ – 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.

In computational linguistics, topic models have been used in various applications, such as predicting response to political webposts (Yano et al., 2009), analyzing Enron and academic emails (McCallum et al., 2007a), analyzing voting records and corresponding text of resolutions from the U.S. Senate and the U.N. (McCallum et al., 2007b), as well as studying the history of ideas in various research fields (Hall et al., 2008; Paul and Girju, 2009). To our knowledge, the application of topic models to identifying cross-cultural differences is novel.

## 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 $i$th 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 $\phi_z$ from Dirichlet($\beta$) for each topic $z$

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

The Dirichlet parameters $\alpha$ and $\beta$ 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.

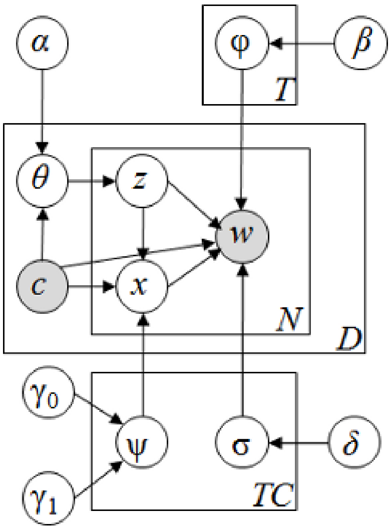

Figure 1: Graphical representation of ccLDA. $C$ is the number of collections, $T$ is the number of topics, $D$ is the number of documents, and $N$ is the length of each document.

The generative process is thus:

1) Draw a collection-independent multinomial word distribution $\phi_z$ from Dirichlet($\beta$) for each topic $z$

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

3) Draw a Bernoulli distribution $\psi_{z,c}$ from Beta$(\gamma_0, \gamma_1)$ 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($\alpha_c$). 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 $\phi_z$;
     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

[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.

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 $\lambda_C$. Since it is not clear how to set this parameter[2], in our model, we learn this probability automatically. Furthermore, the nature of the $\lambda_C$ 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 $\alpha_c$) 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.

## 4 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 approximates a global maximum (whereas EM algorithms may only converge to a local maximum).

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:

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

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

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

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

Because of the conjugacy of the Beta/Dirichlet and binomial/multinomial distributions, we can integrate out $\theta$, $\phi$, $\sigma$ and $\psi$ 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.

$\alpha_c$ 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 Experimental Results

Our experiments focus on discovering cultural differences by running our model on text from or about three countries: the UK, India, and Singapore. We explore the notion of *perspective* by experimenting 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*).

### 5.1 The Data

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

[2] If needed, one can effectively set this probability manually in ccLDA as well by using a large prior.

our model improves upon previous work on such datasets. 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[3]. We show how this is a solid step toward automatic discovery of cultural differences.

Moreover, we compare the two perspectives on the topic of food. We show that there are some strong similarities between the topic in each dataset (thereby enforcing our inferences from each experiment individually), but we also show some differences in the foods tourists find interesting and what locals actually eat.

In all of 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 $\gamma_0 = \gamma_1 = 1.0$.

Our implementation is loosely based on the LDA Gibbs sampler[4] by Phan and Nguyen (2008).

### 5.2 Analysis Along the *Tourists* Dimension

In the first experiment we consider data about three destination countries. Using the data provided by lonelyplanet.com, 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.

Since we are not interested in the thread discussions on a particular travel topic, we treated each thread or discussion of multiple messages as a single document. We were able to use simple pattern matching to extract only the discussion text. We removed HTML tags, stop words, and words with a corpus frequency less than 10. There were 703,551 tokens after preprocessing.

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.

<table>
<tr><td colspan="3">weather time day going rain<br>summer month high days thanks</td></tr>
<tr><th>UK</th><th>India</th><th>Singapore</th></tr>
<tr><td>wind</td><td>leh</td><td>hot</td></tr>
<tr><td>waterproof</td><td>monsoon</td><td>humid</td></tr>
<tr><td>ending</td><td>road</td><td>humidity</td></tr>
<tr><td>rolling</td><td>manali</td><td>heat</td></tr>
<tr><td>walkers</td><td>ladakh</td><td>degree</td></tr>
<tr><td>rochdale</td><td>trekking</td><td>equator</td></tr>
<tr><td>layers</td><td>trek</td><td>sweat</td></tr>
<tr><td>snow</td><td>season</td><td>bring</td></tr>
<tr><td>footwear</td><td>rains</td><td>rain</td></tr>
<tr><td>ankle</td><td>monsoons</td><td>umbrella</td></tr>
</table>

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

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).

### 5.3 Analysis Along the *Locals* Dimension

The results of the first experiment offer an unsupervised aggregation of factual information that is important to travelers such as a destination's climate, law, and infrastructure; however, the data did not offer much in terms of cultural information. We would now like to see if we can get better insight into this problem by modeling text *authored* by residents of these same countries. In doing this we can compare what they talk about and in what manner they talk about certain topics.

For this experiment 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

[3] The dataset is available for download at http://apfel.ai.uiuc.edu/resources.html

[4] http://gibbslda.sourceforge.net

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.

We removed HTML tags from the documents, but we made no attempt to segment the documents into article text – there are efficient methods of doing this (Pasternack and Roth, 2009) and this may be worth experimenting with, but we found that noise such as navigation menus and advertisements would mostly get grouped into their own topics. We removed stop words and words with a corpus frequency less than 20. All punctuation was treated as word separators. There were 8,599,751 tokens in the end.

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 1 is about fashion, and we can compare which fashions are popular in each country. Shoes are popular in the UK; leather and jewelry are more popular in India. Singapore bloggers seem to focus on prices and the shopping aspect of apparel.

From Topic 2 (about pets) it seems that Britons slightly prefer dogs and Singaporians slightly prefer cats. In general, it seems that Singaporians have an affinity for small animals, considering the presence of *hamster* and *rabbit* in their word distribution.

Topic 3 is about religion, in which we see that Christianity is common to all of them, but Hinduism is prominent in India as well.

There are many topics not shown here including politics, gardening, health, etc. The health topic is interesting in that homeopathy and herbal medicines are discussed in Indian blogs. Smoking is a bigger topic in the UK than the others.

It is also interesting to compare what technologies and web services people use. Twitter and Facebook are popular in the UK whereas Orkut is more popular in India. Blogging services like Wordpress are popular in Singapore.

From the travel topic, shown in Table 5, we see that people travel close to 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.

### 5.4 Differences in Perspective: *Tourists* vs. *Locals*

Having modeled the same countries from two different perspectives (that of travelers and that of locals), it would be interesting to see how topics compare between the two perspectives.

Do people have the same view of themselves as outsiders see them? Are locals interested in the same things as tourists?

We hope to answer these questions by examining related topics within these two datasets. While the two datasets consist of mostly different topics, there are a few that would be interesting to compare. In particular, we examine the topic of food and eating. The top words from this topic are shown in Table 3.

We first examine this topic from the blog data (that is, from the perspective of residents). By looking at each collection-specific word distribution we can see which foods are more 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.

Looking now at the topics induced from the lonelyplanet.com forums (that is, from the perspective of travelers), we see some interesting similarities. Most notably, the Indian distribution again consists of words related to cooking, affirming our observation that dining out is not as popular in India. The Singapore distribution also matches that in the other dataset – the common words include seafood and noodles. The UK distribution, however, shows that tourists are mostly interested in local specialties (such as *fish and chips* and *haggis*).

To see where these perspectives on food differ the most, we computed the ratio of the probability of each word given the topic between the two datasets. That is, if $p = P(w|z)$ in the locals data and $q = P(w|z)$ in the tourists data, then $\lambda = p/q$ gives us a measure of how much more (or less) prominent that word is among locals than it

| **Topic 1** | | | **Topic 2** | | | **Topic 3** | | |
|---|---|---|---|---|---|---|---|---|
| fashion style look dress wear new collection accessories black | | | dog dogs pet animals animal comments cat like food plant | | | god jesus lord life faith holy man christ church love | | |
| **UK** | **India** | **Singapore** | **UK** | **India** | **Singapore** | **UK** | **India** | **Singapore** |
| shoes | fashion | price | garden | water | cat | church | krishna | god |
| fashion | women | posted | dog | energy | cats | god | religion | sin |
| clothing | indian | earrings | pet | carbon | dog | john | religious | john |
| high | designer | length | cat | earth | pet | todd | spiritual | spirit |
| designer | sarees | item | dogs | green | training | bentley | guru | things |
| style | leather | sgd | pets | solar | pets | jesus | lord | lamb |
| love | girls | silver | gardening | jai | hamster | christ | sri | exodus |
| london | china | clothes | cats | climate | cute | luke | shri | suffering |
| shirts | jewellery | shop | puppy | environment | hamsters | bible | baba | cross |
| bag | jewelry | code | flowers | warming | rabbit | 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.

| **Perspective of Locals** | | | **Perspective of Tourists** | | |
|---|---|---|---|---|---|
| food add chicken recipe cooking taste rice recipes sugar soup | | | food eat restaurant restaurants tea cheap meal eating cafe drink | | |
| **UK** | **India** | **Singapore** | **UK** | **India** | **Singapore** |
| food | recipe | coffee | chips | cooking | hawker[a] |
| wine | recipes | cup | haggis | spices | satay |
| restaurant | powder | oil | fish | sick | stalls |
| coffee | indian | comments | respectability | flour | noodles |
| cheese | salt | fried | decent | tomato | roti |
| soup | tsp | add | veggie | batter | stall |
| eat | rice | restaurant | pudding | ate | seafood |
| chef | masala | rice | photoblog | cook | malay |
| english | oil | tea | sausages | olive | rochester |
| drink | coriander | seafood | sandwiches | recipe | noodle |

[a] A hawker centre is an open-air complex with many food stalls, commonly found in Singapore and Malaysia.

Table 3: A comparison of the food topic from two different datasets, one of which comes from a travel forum and the other of which consists of blogs authored by residents of each respective country.

is among tourists in the food topic. Table 4 shows the words with the highest (left) and lowest (right) values of $\lambda$.

| **Preferred by Locals** | | | **Preferred by Tourists** | | |
|---|---|---|---|---|---|
| recipe bowl lemon tomato simple spring spoon vanilla stir pour | | | street cheap couple yeah crowd old road floor run locals | | |
| **UK** | **India** | **Singapore** | **UK** | **India** | **Singapore** |
| food | indian | cup | pubs | mother | quay |
| healthy | recipes | comments | music | ate | coast |
| shop | cup | tea | lane | tree | parkway |
| favorite | chicken | mins | brick | party | reasonably |
| wine | minutes | pot | fish | fields | air |
| icing | kitchen | note | jazz | base | sultan |
| coffee | mustard | nice | pints | rock | tum |
| leeds | fried | salt | dancing | toilet | views |
| duck | ginger | tarts | arms | bottled | plenty |
| extra | salt | fish | recommend | olive | rochester |

Table 4: This table shows words in the food topic that are more popular in the tourists data than the locals data or vice versa.

The prominent trend, which is largely a logistical matter, is that travelers are more interested in restaurants and locals talk more about cooking. Most of the words that are more prominent from the tourist perspective have to do with eating locations. We also noticed that wine and coffee rank more prominently among the locals, whereas travelers are more likely to ask about beer and liquor.

### 5.5 Model Evaluation

In this subsection 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.

#### 5.5.1 Cluster Coherence

Because our research relies 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 $\lambda_B$ (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 $\lambda_C$, analogous to $P(x = 0)$, to 0.6, which is the average value learned by ccLDA on this data, and it seems quite 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 first experiment above 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[5] 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 5 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.

### 5.5.2 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 forum dataset from the first experiment 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.

## 6 Discussion and Future Work

While there are obvious limitations of the unigram approach used here, our system was nevertheless able to capture some interesting details. It is important, however, to point out some limitations for possible future extensions.

Consider Topic 2 in Table 2. The UK and Singapore word distributions are both clearly pertinent

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

| **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 5: The topic of *travel* as discovered by the 3 different models.

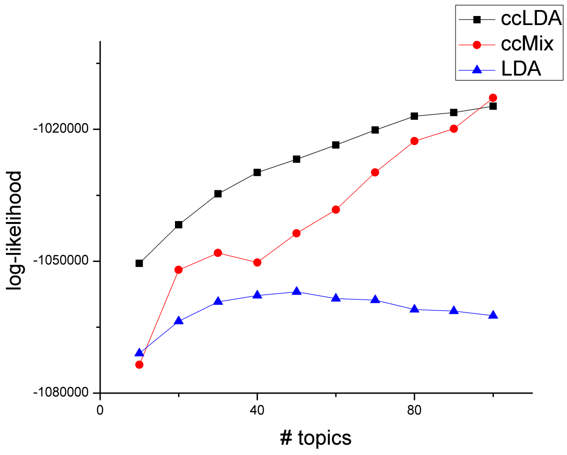

Figure 2: Comparison of the log-likelihood of held-out data with the 3 models.

to the topic of pets, but the India distribution seems entirely unrelated, being about energy and the environment. This could be because the environment topic was statistically too strong to ignore, but not found in other collections, so it made its way into a largely unrelated topic. (In fact, the formation of the environment cluster within this topic is not entirely random, as the pets topic also includes some words related to gardening, including “water” and “plant”, which are likely to also co-occur with environmental words.)

This is perhaps the main weakness of the model. If an emerging topic is not shared among all collections, it will either form as a primary topic that is unique to only a subset of collections (and thus some of the collection-specific distributions will be noisy), or it will form as a collection-specific distribution that is not strongly related to the main collection-independent distribution. This can make the results difficult to interpret, although an automated solution would be to remove or flag topics that are not evenly shared, which could be done by comparing the learned collection-dependent Dirichlet parameters $\alpha_c$.

This is also a matter of how the model performs with different numbers of collections. It would be interesting to see what results we would get by modeling UK-India, UK-Singapore, and India-Singapore as only a pair at a time. The performance should not degrade with larger numbers of collections if the collections are fully comparable, but in practice, with more collections there are likely to be more topics that are difficult to fit across all collections.

In future work, we would like to enrich the model and/or feature set to move beyond the limitations of a bag-of-words analysis. For example, by considering negation and word polarity, we can better capture the opinions of the authors, which is an important component of such cultural analysis.

Certainly, there are many other possible applications of this model, including product comparison, media bias detection, and interdisciplinary literature analysis. Cultural awareness is also important in marketing and we can use this model to investigate, for example what products and what aspects of life people in different regions focus on.

## Acknowledgments

We would like to thank ChengXiang Zhai for thoughtful discussions and for providing an implementation of the ccMix model. We would also like to thank the reviewers for their constructive comments.

## References

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

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*.

Tl Griffiths, M. Steyvers, and Jb 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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 363–371.

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.

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

A. McCallum, X. Wang, and A. Corrada-Emmanuel. 2007a. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research (JAIR)*, 30:249–272.

A. McCallum, X. Wang, and N. Mohanty. 2007b. Joint group and topic discovery from relations and text. In *Statistical Network Analysis: Models, Issues and New Directions - Lecture Notes in Computer Science 4503*, pages 28–44.

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.

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

J. Pasternack and D. Roth. 2009. Extracting article text from the web with maximum subsequence segmentation. In *The International World Wide Web Conference*, April.

M. Paul and R. Girju. 2009. 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)*.

X. Phan, L. Nguyen, and S. Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 91–100, 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.

T. Yano, W. Cohen, and N. Smith. 2009. Predicting response to political blog posts with topic models. In *The 7th Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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.