Undirected and Interpretable Continuous Topic Models of Documents

Abstract

We propose a new type of undirected graphical model suitable for topic modeling and dimensionality reduction for large text collections. Unlike previous Boltzmann machine and harmonium based methods, this new model represents words using Discrete distributions akin to traditional 'bag-of-words' methods. However, in contrast to directed topic models such as latent Dirichlet allocation, each word is drawn from a distribution that takes into account all possible topics, as opposed to a topic-specific distribution. Furthermore, our models use positive continuous valued latent variables and learn more interpretable latent topic spaces than previous undirected techniques. As other undirected models, once such models have been learned, inference required for representing a document in the latent space is fast. We present document retrieval experiments showing the benefits of our new approach.

1. Introduction

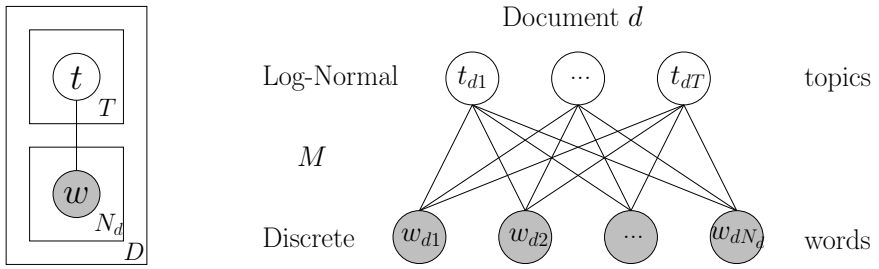
Research in statistical machine learning models of co-occurrence has led to the development of a variety of useful *topic models* — mechanisms for discovering latent, low-dimensional, multi-faceted summaries of documents or other discrete data. In these models, graphical model structures are carefully-designed, often by employing latent variables, to capture the relevant structure and co-occurrence dependencies in the data.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Graphical models can be categorized into two fundamental classes: directed (aka Bayesian networks) and undirected (aka Markov random fields). The first category include models of words alone, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Griffiths & Steyvers, 2004), of relations between entities (Nowicki & Snijders, 2001; Kemp et al., 2004), of words and research paper citations (Erosheva et al., 2004), of word sequences with Markov dependencies (Griffiths et al., 2004; Wallach, 2006; Wang et al., 2005a), of words and their authors (Steyvers et al., 2004), of words in a social network of senders and recipients (McCallum et al., 2005), of words and relations (such as voting patterns) (Wang et al., 2005b), as well as words and their timestamps (Blei & Lafferty, 2006; Wang & McCallum, 2006).

Directed graphical models can be described as a generative processes and thus enjoy modeling and computational benefits conferred from conditional independencies such as simple sampling procedures. However, in many applications, the dependency between two random variables in directed models can be difficult to describe and specify and the direction of directed edges in the underlying graph can arguably be set either way. Importantly, posterior inference over hidden topic variables and parameters for directed models with structures similar to LDA is typically intractable and approximate inferences techniques such as variational methods(Jordan et al., 1998), Gibbs sampling(Andrieu et al., 2003) and expectation propagation (Minka & Lafferty, 2002) are employed to address these issues.

Recently, a class of structured undirected topic models has begun to draw increased interest — largely due to the fact that inference of hidden topics can be fast compared to directed, LDA inspired models. The Exponential family harmonium (EFH) is one of the earlier works in this direction (Welling et al., 2004). In



(a) Plate notation (b) A corresponding restricted Boltzmann machine structure

Figure 1. Graphical representations for our models. Shaded random variables are *observed* word tokens.

(Welling et al., 2004) a specific model for latent semantic indexing of documents is also outlined in which a consistent conditional Gaussian distribution for hidden (topic) variables is coupled with a corresponding Bernoulli or Discrete distribution for *discretized counts* of words across the vocabulary of a document collection.

The two-layer structures in EFHs have an important property: the random variables at the two layers are conditionally independent given each other, which means the mapping from one layer to the other layer can be done by a simple matrix multiplication (and possibly some trivial follow-up transformations).

However, there is no free lunch. The faster inference corresponds to more difficult learning due to the intractable normalizing constant in these types of undirected models. Fortunately, the contrastive divergence (Hinton, 2002) approach has been shown to be efficient for inference and effective for learning in these models. Further, in many situations involving document processing learning can be done off-line. We will discuss this in more detail in Section 2.

Based on the two-layer factorization structure of an EFH, there are several other undirected topic models that have been recently proposed for various tasks. For example, a dual-wing harmonium (DWH) model (Xing et al., 2005) has been proposed and applied to captioned images. In this model hidden topics are conditional Gaussian given words and word counts are distributed according to a Poisson distribution and Gaussians for color histograms. The rate adapting Poisson (RAP) models (Gehler et al., 2006) are similar, but with Poisson for words counts and Binomial (Bernoulli) for hidden topics. Applications to document retrieval and object recognition demonstrate its effectiveness.

Undirected models of this structure have another important property that directed models lack: a more accurate characterization of rare words. As discussed in Xing et al. (2005), in directed models such as latent Dirichlet allocation, a word is always generated from a single topic. When its count is low, this behavior becomes a very strong assumption/limitation. In the harmonium-structured models, a word is always from a distribution influenced by all the topics. This different mechanism might play a crucial role in certain applications.

In this paper, we propose a novel model based on similar two layer factorization structure but with dramatically different semantics. At the hidden layer, previous models assume either Gaussian distributions or Binomial (Bernoulli) distributions. In our model, conditioned on observations the hidden layer follows a log-normal distribution and takes advantage of both continuity and positivity. We believe that in this setting more interpretable results arise.

Furthermore, we associate a Discrete distribution for the identity of each observed word, thus *each word token* is drawn in a replicated fashion akin to traditional 'bag-of-words' models. Note here that all the word tokens share a common connection matrix between word layer and topic layer. By contrast, in (Welling et al., 2004) a different connection matrix is needed for each word and word count level. The Poisson distributions adopted in Xing et al. (2005) and Gehler et al. (2006) make it possible to use only one connection matrix, but when reconstructing the document counter vectors, there is no guarantee that the reconstructed document has the same length of the original document. In such a case, at early stage of learning, the learning rate of the gradient update has to be carefully set to a small value as reported in Gehler et al. (2006) and make the model difficult to learn in long run.

More importantly, to the best of our knowledge, none of the previous undirected models is able to interpret the learned topics, and our model is the first evidence that undirected models can also give clearly interpretable topics, which are useful for model diagnosis,

SYMBOL	DESCRIPTION
T	number of topics
D	number of documents
V	number of unique words
N_d	number of word tokens in document d
M	$T \times (V-1)$ connection matrix
t_{di}	the i^{th} topic of document d
w_{dj}	the j^{th} word of document d

Table 1. Notation used in this paper

interpretation, summarization and data mining.

2. Our Model

In contrast to previous undirected topic models, in our new model, words are encoded as individual observations instead of word counts. Because of the conditional independencies between two layers, we can describe the model in plate notation, shown in Figure 1(a). Notations used in this paper are shown in Table 1. We expand the model for document d for clarity as shown in Figure 1(b) into a restricted Boltzmann machine or exponential family harmonium structure.

Following a common approach for describing a general exponential family two layer architecture, we specify our model as follows: Consider first a Log-normal distribution $p(t_{di}) = \text{Log-normal}(0, 1)$ at hidden (topic) layer and Discrete distribution $P(w_{dj}) = \text{Discrete}(\mathbf{0})$ at the observation (words) layer, where use the notation Log-normal(μ, σ^2) for a Log-normal distribution with parameters μ and σ^2 — the mean and variance of the variable's logarithm, and Discrete(θ) is a Discrete distribution with natural parameter θ_k ($k = 1, \cdots, V-1$) that can be transformed to the probability vector $\pi_k = e^{\theta_k} / \sum_{v=1}^{V} e^{\theta_v}$ (note here we set $\theta_V = 0$). For simplicity, we do not use local potentials, but it is straightforward to define and learn these potentials as well.

Once we have defined the form we wish the observed and hidden layers to take, we couple the random variables within the two layers by the connection matrix M to get a joint probability distribution in exponential family form as follows:

$$P(\mathbf{w_d}, \mathbf{t_d}) \propto \exp(\sum_{i=1}^{T}(-\log(t_{di}) - \frac{1}{2}\log^2(t_{di})) + \sum_{i=1}^{T}\sum_{j=1}^{N_d} M_{iw_{dj}} \log(t_{di})) \quad (1)$$

where, for notation convenience, we set $M_{iV} = 0$, for $i = 1, \cdots, T$.

Consequently, it is easy to verify the conditional distributions still remain in the same exponential family but with shifted parameters,

$$p(t_{di}|\mathbf{w_d}) = \text{Log-normal}(\sum_{j=1}^{N_d} M_{iw_{dj}}, 1) = \text{Log-normal}(\sum_{k=1}^{V-1} M_{ik} c_{dk}, 1) \quad (2)$$

$$P(w_{dj}|\mathbf{t_d}) = \text{Discrete}(\sum_{i=1}^{T} \log(t_{di}) M_{i\cdot}) \quad (3)$$

where c_{dk} is the count of word k in document d.

From the joint probability of all random variables (Eqn. 1), we can marginalize out the latent topic variables, and get the marginal likelihood of the observed document d,

$$P(\mathbf{w_d}) \propto \exp(\frac{1}{2}\sum_{i=1}^{T}(\sum_{k=1}^{V-1} M_{ik} c_{dk})^2)$$

The marginal likelihood of the whole corpus (our objective function) thus can be calculated as

$$\prod_{d=1}^{D} P(\mathbf{w_d}) \propto \exp(\frac{1}{2}\sum_{d=1}^{D}\sum_{i=1}^{T}(\sum_{k=1}^{V-1} M_{ik} c_{dk})^2) \quad (4)$$

Note here, we can only compute the marginal likelihood up to a normalizing constant.

2.1. Parameter Learning by Contrastive Divergence

Parameters of our model can be learned by gradient ascent on the marginal (log) likelihood in Eqn. 4. However, due to the intractability of the normalizing constant, it is difficult to calculate the gradient of the log-likelihood. We use contrastive divergence (Hinton, 2002) which has been shown to greatly improve learning efficiency in harmonium architectures (Welling et al., 2004; Xing et al., 2005; Gehler et al., 2006). The main idea of contrastive divergence is that we can truncate a Gibbs sampler with only one (or a few) iterations, and use the distribution of the samples (say, $\hat{\mathbf{w}}_\mathbf{d}$ or equivalently $\hat{c}_{dk}$, $d = 1, \ldots, D$, and $k = 1, \ldots, V-1$) from the truncated chain to approximate the model distribution. In this way, the learning rule can be written as the difference between the empirical average and the approximated (by contrastive divergence) model average,

$$\delta M_{ik}$$

Algorithm 1 Learning via Contrastive Divergence

Input: document $\mathbf{w_d}$ $(d = 1, \cdots, D)$, topic# T
Initialize connection matrix M randomly
repeat
  for $d = 1$ to D **do**
    for $i = 1$ to T **do**
      Draw t_{di}, according to Eqn. 2
    end for
    for $k = 1$ to N_d **do**
      Draw $\hat{w}_{dk}$, according to Eqn. 3
    end for
  end for
  for $i = 1$ to T **do**
    for $j = 1$ to $W - 1$ **do**
      Update M_{ij}, according to Eqn. 5
    end for
  end for
until M converges

$$\begin{aligned} \propto & \quad \frac{\partial \log \prod_{d=1}^{D} P(\mathbf{w_d})}{\partial M_{ik}} - \frac{\partial \log \prod_{d=1}^{D} P(\hat{\mathbf{w}}_\mathbf{d})}{\partial M_{ik}} - \frac{M_{ik}}{\sigma^2} \\ \propto & \quad \sum_{d=1}^{D} (c_{dk} \sum_{v=1}^{V-1} M_{iv} c_{dv} - \hat{c}_{dk} \sum_{v=1}^{V-1} M_{iv} \hat{c}_{dv}) - \frac{M_{ik}}{\sigma^2} \end{aligned} \tag{5}$$

where the last term comes from a Gaussian prior over parameters (with variance σ^2) that provides smoothing to help cope with sparsity in the training data (Chen & Rosenfeld, 1999). This prior favors parameters that are closer to zero, and penalize (positive and negative) large values. We summarize the learning procedures in Algorithm 1.

The introduction of this prior also helps alleviate the identifiability problem as reported in Welling et al. (2004) and Gehler et al. (2006), that is, makes the model more identifiable. Without further special handling of identifiability issues, we still get surprising good results as shown in Section 4. Priors over weights can influence the effectiveness of dimensionality reduction. A corpus usually has an intrinsic number of topics that is unknown, and in general, we either try many settings and select the best, or use nonparametric methods to estimate this number(Teh et al., 2004). When given inappropriate number of topics, a model with prior will try duplicate some topic or create some random (but not trivial) topics. With priors, the spurious topics will gradually become trivial (near zero everywhere) since the priors push the weights toward zero where no enough data evidence support them.

2.2. Multi-Conditional Learning

To explicitly emphasize that we want to capture co-occurrence patterns, another way is to maximize the conditional probability of $\prod_{j=1}^{N_d} P(w_{dj}|\mathbf{w}_\mathbf{d}^{-j})$ using multi-conditional learning principle, where $\mathbf{w}_\mathbf{d}^{-j}$ is all the observed words in document d excluding the j^{th} word.

Multi-conditional learning (MCL) is a training criterion based on a product of multiple conditional likelihoods(McCallum et al., 2006). When combining the traditional conditional probability of label given input with a generative probability of input given label the later acts as a surprisingly effective regularizer. When applied to models with latent variables, MCL combines the structure-discovery capabilities of generative topic models, with the accuracy and robustness of discriminative classifiers, such as logistic regression and conditional random fields. Results on several standard text data sets have been shown significant reductions in classification error due to MCL regularization, and substantial gains in precision and recall due to the latent structure discovered under MCL(McCallum et al., 2006).

Note that our configuration of Discrete distribution with 'bag-of-words' assumption makes it possible to take advantage of MCL rather easily, and it is not straightforward to apply MCL to other models with similar structures, such as the dual-wing harmonium model(Xing et al., 2005) and the rate adapting Poisson model(Gehler et al., 2006).

Using the multi-conditional learning creterion, we can get an alternative objective function as

$$\begin{aligned} & \prod_{d=1}^{D} \prod_{j=1}^{N_d} P(w_{dj}|\mathbf{w}_\mathbf{d}^{-j}) \\ \propto & \quad \exp(\frac{1}{2} \sum_{d=1}^{D} \sum_{i=1}^{T} \sum_{v=1}^{V-1} ((2 \sum_{k=1}^{V-1} M_{ik} c_{dk} - M_{iv}) M_{iv} c_{dv}) \end{aligned}$$

Similar to Eqn. 5, we can have a learning rule under MCL, which is surprisingly simple due to our 'bag-of-word' setting here,

$$\begin{aligned} \delta M_{ik} \quad \propto \quad & \sum_{d=1}^{D} (2 c_{dk} \sum_{v=1}^{V-1} M_{iv} c_{dv} - 2 \hat{c}_{dk} \sum_{v=1}^{V-1} M_{iv} \hat{c}_{dv} \\ & + M_{ik} (\hat{c}_{dk} - c_{dk})) - \frac{M_{ik}}{\sigma^2} \end{aligned}$$

In Section 4, we show the different between the two training criteria, and empirically demonstrated that MCL helps discover more distinct topics than simple maximum likelihood.

3. Data Sets

We apply our models to two large well-known text corpora and show the results in Section 4.

3.1. NIPS Data Set

The NIPS proceeding data set consists of the full text of the 13 years of proceedings from 1987 to 1999 Neural Information Processing Systems (NIPS) Conferences.[1] All the text is downcased, stopwords removed, but not stemmed. The dataset contains 1,740 research papers, 13,649 unique words, and 2,301,375 word tokens in total.

3.2. 20 Newsgroups Data Set

The 20 newsgroups data set we use only keeps the 10,000 words with highest average mutual information with the class label.[2] All the text is downcased, stopwords removed and stemmed with a Porter stemmer.The data set contains 18,796 documents, 10,000 unique words, 1,848,207 word tokens in total. Each document has one of the 20 newsgroup names as its label.

4. Experimental Results

In this section, we first show several word lists for several learned topics as anecdotal evidence, and then we compare our model with previous models in information retrieval experiments on the newsgroups data set.

4.1. Interpretable Topics

We show the word list for a subset of topics learned within our weight matrices from the NIPS data set in Table 2. Immediately, we can see that all the positive words provide a vivid summary of topics well known to exist within the NIPS community: Biological Neuroscience, Reinforcement Learning and Probabilistic Methods. Other topics not shown exhibit words characteristic of topics such as Computational Neuroscience. Interestingly, the negatively weighted words are also common words in other topics, and serve to separate this topic from others possibly confused with it. A similar subset of clean topics emerges from the 20 newsgroups data set, with clear Religion, Image and General Computer topics emerging as illustrated in Table 4.

We also calculated the average cosine similarity between topics learned by maximum likelihood and MCL, and found that the MCL criterion does help discover more distinct topics (average cosine similarity: 0.2281) than maximum likelihood (average cosine similarity: 0.3201). We also observe a subtler distinction between topics found using this method. For example, Table 3 illustrates Pattern Recognition, 'Neural Networks' and 'Classification and Regression' topics. We found the MCL optimization was better at separating a Classification and Regression topic from a Probabilistic Methods topic. The topics are equally good with some interesting differences, such as commonly co-occurring words having lower weights. As explained in Section 2, we also find several trivial low-weight topics thanks to the prior we adopted, when we increase the number of topics.

4.2. Information Retrieval

In information retrieval, given a query, we rank the documents in corpus by some score, such as vectorspace based cosine similarity between document and query, and query likelihood (Zhai & Lafferty, 2004) and take the top ones as the retrieval documents. Obviously, not all the retrieved documents are relevant to the given query, precision and recall are the most common measure for retrieval performance. Precision can be understood as the ratio of retrieved and relevant documents to all retrieved documents, that is,

$$\mathrm{P} = \frac{|\{\text{relevant document}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall, on the other hand, can be thought as the ratio of retrieved and relevant documents to all relevant documents in the corpus, that is,

$$\mathrm{R} = \frac{|\{\text{relevant document}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Using our model as a retrieval system, we can rank documents in a corpus by the (cosine) similarity between the latent (topic) representation of the documents and a given query. We use a small version of the 20 newsgroup data: only the 100 words with highest average mutual information with the class label are kept and we remove the documents do not contain these 100 words. We randomly split the data set into training set (9/10, 16,218 documents) and test set (1/10, 1,802 documents). If a retrieved document has the same label as the test query document, they are relevant.

We use a 20-topic run (3,000 iterations) on the training set to learn the parameters and calculate the average precision and recall across all the test documents

[1] http://www.cs.toronto.edu/∼roweis/data.html

[2] http://www.kyb.tuebingen.mpg.de/bs/people/pgehler/rap/#datasets

Biological Neuroscience				Reinforcement Learning				Probabilistic Methods			
cells	.439	training	-.556	learning	.318	image	-.536	data	.364	state	-.512
cell	.361	networks	-.500	policy	.266	data	-.444	model	.307	time	-.454
firing	.360	error	-.472	reinforcement	.252	images	-.431	mixture	.271	neuron	-.449
cortex	.357	network	-.470	control	.239	recognition	-.345	gaussian	.260	neural	-.429
cortical	.355	speech	-.465	state	.234	feature	-.315	likelihood	.225	system	-.422
stimulus	.327	neural	-.461	action	.233	object	-.271	image	.221	control	-.405
spike	.314	classifier	-.436	actions	.158	visual	-.270	distribution	.217	neurons	-.373
synaptic	.310	class	-.412	weight	.153	features	-.263	bayesian	.213	analog	-.363
synapses	.275	word	-.410	states	.151	gaussian	-.241	images	.204	network	-.359
motion	.268	state	-.407	controller	.150	classification	-.233	em	.189	circuit	-.335
orientation	.262	recognition	-.406	optimal	.125	mixture	-.227	density	.183	action	-.334
excitatory	.255	classifiers	-.386	weights	.121	models	-.217	models	.182	synaptic	-.317
visual	.253	classification	-.370	error	.117	model	-.211	posterior	.163	chip	-.316
inhibitory	.243	set	-.359	time	.115	likelihood	-.190	prior	.148	networks	-.287
response	.243	hmm	-.354	neuron	.105	set	-.189	regression	.146	states	-.285
stimuli	.240	algorithm	-.344	sutton	.102	orientation	-.184	kernel	.144	memory	-.279
spatial	.238	hidden	-.342	gradient	.101	classifier	-.180	log	.135	recurrent	-.263
direction	.233	test	-.337	recurrent	.101	face	-.179	classification	.134	current	-.263
membrane	.231	mixture	-.334	agent	.096	class	-.171	class	.133	policy	-.259
eye	.229	data	-.333	learn	.096	test	-.169	parameters	.124	reinforcement	-.256

Table 2. The three topics from a 20-topic run of our model on 13 years of NIPS research papers. The **Title** above the word lists of each topic is our own summary of the topics. For each topic, we show the top 20 positive words (left) and the top 20 negative ones (right) with the corresponding weights. Here, for displaying convenience, we have multiplied all the learned weights by a factor of 10.

Pattern Recognition				'Neural Networks'				Classification and Regression			
recognition	.734	policy	-.574	input	.900	itly	-.106	functions	.419	units	-.075
image	.687	weight	-.537	output	.820	construc	-.105	class	.403	visual	-.069
images	.663	action	-.504	hidden	.617	nash	-.100	classifier	.391	motion	-.068
object	.577	reinforcement	-.454	model	.593	ination	-.099	regression	.368	unit	-.057
speech	.547	learning	-.428	state	.550	probabilit	-.098	classifiers	.361	task	-.055
visual	.489	convergence	-.420	speech	.544	rival	-.097	bounds	.350	direction	-.054
word	.483	optimal	-.414	training	.540	aleksander	-.097	gaussian	.330	learning	-.052
features	.465	actions	-.400	models	.538	laxation	-.097	loss	.323	eye	-.050
feature	.419	error	-.395	weights	.529	arthur	-.096	theorem	.322	object	-.049
objects	.384	neuron	-.344	error	.506	duplicating	-.096	density	.317	motor	-.048
face	.379	controller	-.341	patterns	.504	cedures	-.096	approximation	.315	cortex	-.048
hmm	.287	gradient	-.338	inputs	.502	affirmative	-.095	bound	.315	action	-.047
classification	.284	theorem	-.335	unit	.500	hindered	-.095	matrix	.314	velocity	-.047
segmentation	.272	reward	-.330	weight	.500	allan	-.094	classification	.312	position	-.044
system	.262	sutton	-.328	net	.489	glasgow	-.094	vector	.312	activity	-.044
context	.261	finite	-.324	architecture	.488	delaying	-.094	kernel	.302	control	-.042
frame	.258	function	-.308	word	.483	mutations	-.094	distribution	.298	cortical	-.042
classifier	.257	stochastic	-.304	systems	.482	mech	-.094	log	.290	reinforcement	-.041
orientation	.255	control	-.296	control	.474	concomitantly	-.093	data	.277	head	-.041
vision	.245	time	-.291	learning	.471	mother	-.093	neural	.270	cells	-.040

Table 3. The three topics from a 20-topic, MCL run of our model on 13 years of NIPS research papers. The **Title** above the word lists of each topic is our own summary of the topics. For each topic, we show the top 20 positive words (left) and the top 20 negative ones (right) with the corresponding weights. Here, for displaying convenience, we have multiplied all the learned weights by a factor of 10.

at different recall levels, and plot the Precision-Recall curve in Figure 2. We also compare our model with the (a) RAP model (Gehler et al., 2006) also with 20 topics but with 30,000 updates and (b) the TF-IDF representation, with cosine similarity, where

$$\text{TF-IDF}_{dw} = \frac{c_{dw}}{N_d} \log \frac{D}{|\{\text{documents containing } w\}|}$$

As shown in Figure 2, we can see that at low recall – where we are primarily interested – the precision of our model is superior to both RAP and TF-IDF. Note that, due to the small vocabulary size, the precisions are relatively low.

5. Conclusion and Discussion

We have proposed a new harmonium-structured undirected model for large text collections. Unlike the previous models, the new model still allows the words to come from a discrete distribution in a 'bag-of-words' fashion. In contrast to the directed topic models such as Latent Dirichlet Allocation, a word is always drawn from a distribution taking into account all possible topics, instead of a topic-specific distribution. We show interpretable word lists for topics, and demonstrate better information retrieval performance.

It is well known that the precision of dimensionality

Religon				Images				Computer			
god	.0107	max	-.0256	jpeg	.0125	god	-.0051	window	.0130	god	-.0141
peopl	.0074	giz	-.0183	gif	.0077	wire	-.0033	graphic	.0115	jpeg	-.0073
lord	.0053	bhj	-.0179	imag	.0064	lord	-.0026	pub	.0110	jehovah	-.0073
jehovah	.0052	output	-.0171	color	.0053	law	-.0024	server	.0109	lord	-.0064
armenian	.0051	qax	-.0156	qualiti	.0045	presid	-.0024	ftp	.0108	jesu	-.0053
jesu	.0046	entri	-.0144	format	.0042	jesu	-.0024	system	.0100	peopl	-.0043
presid	.0045	bxn	-.0130	viewer	.0042	entri	-.0023	mail	.0096	christ	-.0041
don	.0036	file	-.0119	compress	.0037	jehovah	-.0021	data	.0090	father	-.0033
christian	.0036	program	-.0115	convert	.0035	state	-.0021	user	.0084	christian	-.0030
live	.0033	window	-.0090	displai	.0033	christian	-.0020	comput	.0081	don	-.0029
christ	.0032	nrhj	-.0085	pixel	.0032	year	-.0020	anonym	.0080	armenian	-.0026
govern	.0031	line	-.0084	quantiz	.0029	question	-.0019	softwar	.0078	son	-.0022
dai	.0031	biz	-.0077	free	.0028	live	-.0019	widget	.0078	mormon	-.0018
didn	.0030	printf	-.0072	graphic	.0028	ground	-.0018	applic	.0077	bibl	-.0014
father	.0030	check	-.0070	bit	.0027	christ	-.0018	list	.0076	output	-.0014
state	.0028	jpeg	-.0068	version	.0027	armenian	-.0018	includ	.0074	vers	-.0014
thing	.0027	char	-.0067	zip	.0025	dai	-.0018	support	.0072	gif	-.0013
law	.0026	stream	-.0066	softwar	.0024	hous	-.0016	run	.0072	didn	-.0013
made	.0024	section	-.0066	quicktim	.0024	person	-.0016	version	.0071	sin	-.0013
fact	.0022	info	-.0064	mirror	.0024	time	-.0016	motif	.0070	live	-.0012

Table 4. The three topics from a 50-topic run of our model on the 20 newsgroup data set. The **Title** above the word lists of each topic is our own summary of the topics. For each topic, we show the top 20 positive words (left) and the top 20 negative ones (right) with the corresponding weights. Note here, all words are stemmed.

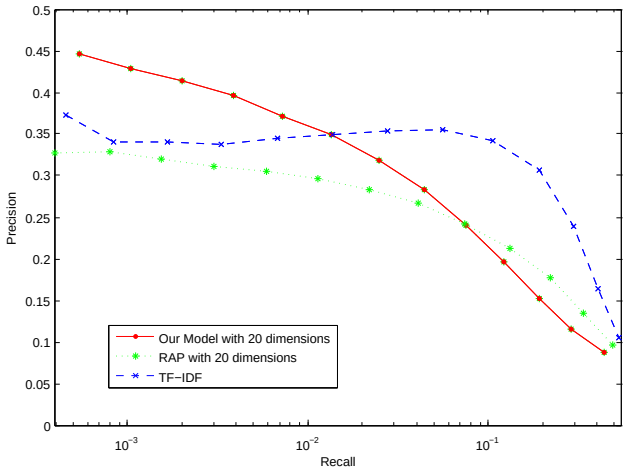


Figure 2. Precision-Recall plot on the 20 newsgroups data set of 100 words vocabulary for our model, RAP and TF-IDF.

reduction based models tends to increase with a larger input vocabulary. While many real world tasks do allow off-line computation to be performed, the optimization time for large vocabulary experiments can be challenging, taking over half a day for model optimization. However, we have ongoing experiments with larger vocabulary sizes underway.

Undirected models with these hidden layer structures allow a great deal of flexibility to incorporate information from multiple modalities as demonstrated in Xing et al. (2005). In directed models, typically when a new source of information is introduced, dependencies with other variables are carefully hand specified, and in many cases, dependencies are too complicated to be explicitly expressed. Furthermore, likelihoods from different modalities are often are not comparable and weighting parameters are often needed as in Wang and McCallum (2006). We see great potential to combine a wide variety of information from the text document (such as words, authors, timestamp, venue, citations, etc.), and robustly create extremely rich models that could have been particularly hard to devise in a directed model. We believe the model presented in this paper and other similar ones will play an important role in modeling multi-modal heterogeneous data.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0427594, in part by a grant from the Eastman Kodak Company, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. (2003). An introduction to mcmc for machine learning. *Machine Learning*, *50*, 5–43.

Blei, D., & Lafferty, J. (2006). Dynamic topic models.

Proceedings of the 23rd International Conference on Machine Learning.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chen, S. F., & Rosenfeld, R. (1999). *A gaussian prior for smoothing maximum entropy models* (Technical Report). Carnegie Mellon University, CMU-CS-99-108.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *101(Suppl. 1)*.

Gehler, P., Holub, A., & Welling, M. (2006). The rate adapting Poisson model for information retrieval and object recognition. *Proceedings of the 23rd International Conference on Machine Learning*.

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

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2004). Integrating topics and syntax. *Advances in Neural Information Processing Systems 17*.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*, 1771–1800.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1998). An introduction to variational methods for graphical models. *Proceedings of the NATO Advanced Study Institute on Learning in graphical models* (pp. 105–161).

Kemp, C., Griffiths, T. L., & Tenenbaum, J. (2004). *Discovering latent classes in relational data* (Technical Report). MIT CSAIL.

McCallum, A., Corrada-Emanuel, A., & Wang, X. (2005). Topic and role discovery in social networks. *Proceedings of the 18th International Joint Conference on Artificial Intelligence*.

McCallum, A., Pal, C., Druck, G., & Wang, X. (2006). Multi-conditional learning: Generative/discriminative training for clustering and classification. *Proceedings of the 21st National Conference on Artificial Intelligence*.

Minka, T., & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*.

Nowicki, K., & Snijders, T. A. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, Washington.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2004). *Hierarchical Dirichlet processes* (Technical Report). University of California, Berkeley, Department of Statistics.

Wallach, H. (2006). Topic modeling: beyond bag-of-words. *Proceedings of the 23rd International Conference on Machine Learning*.

Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Wang, X., McCallum, A., & Wei, X. (2005a). *Topical n-grams: Phrase and topic discovery, with an application to information retrieval* (Technical Report). University of Massachusetts, Amherst.

Wang, X., Mohanty, N., & McCallum, A. (2005b). Group and topic discovery from relations and their attributes. *Advances in Neural Information Processing Systems 18*.

Welling, M., Rosen-Zvi, M., & Hinton, G. (2004). Exponential family harmoniums with an application to information retrieval. *Advances in Neural Information Processing Systems 17*.

Xing, E., Yan, R., & Hauptmann, A. G. (2005). Mining associated text and images with dual-wing harmoniums. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information System*, *22*, 179–214.