Generalized Component Analysis for Text with Heterogeneous Attributes

Xuerui Wang, Chris Pal, Andrew McCallum
Department of Computer Science
University of Massachusetts
140 Governors Drive
Amherst, MA 01003
xuerui@cs.umass.edu, pal@cs.umass.edu, mccallum@cs.umass.edu

ABSTRACT

We present a class of richly structured, undirected hidden variable models suitable for simultaneously modeling text along with other attributes encoded in different modalities. Our model generalizes techniques such as principal component analysis to heterogeneous data types. In contrast to other approaches, this framework allows modalities such as words, authors and timestamps to be captured in their natural, probabilistic encodings. A latent space representation for a previously unseen document can be obtained through a fast matrix multiplication using our method. We demonstrate the effectiveness of our framework on the task of author prediction from 13 years of the NIPS conference proceedings and for a recipient prediction task using a 10-month academic email archive of a researcher. Our approach should be more broadly applicable to many real-world applications where one wishes to efficiently make predictions for a large number of potential outputs using dimensionality reduction in a well defined probabilistic framework.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.2.8 [**Database Management**]: Database Applications—*data mining*

General Terms

Algorithms, Experimentation, Measurement, Performance

Keywords

Undirected Graphical Models, Topic Modeling, Text Mining, Author Prediction, Recipient Prediction, Multimodal Heterogeneous Data

1. INTRODUCTION

Many tasks in data mining involve the processing of high dimensional data with heterogeneous attributes. In practice, we often ignore the heterogeneity of attributes and assume that they come from the same source or distribution, and deal with the problem of data dimensionality by projecting the data into a lower dimensional representation. Principal component analysis (PCA) [14] is widely used in data mining and knowledge discovery to achieve dimensionality reduction from real valued input data. The singular value decomposition (SVD) of a centered data matrix can be used to obtain the *eigen decomposition* of the covariance matrix of data. The eigenvectors of the SVD are called the principal components [11] of the data.

Latent semantic analysis (LSA), proposed by Deerwester et al. [6], is a way to index documents through decomposing a term-document matrix using PCA. In this framework, a matrix consisting of integer word counts for each document is decomposed and the eigenvectors found tend to place higher weight on groups of words which correspond to the notion of semantic topics. Documents projected into the lower dimensional latent space can then be indexed more efficiently. While ad hoc methods can be used to augment a term document matrix with other information, the underlying assumptions of traditional PCA for modeling text are inappropriate.

In particular, recently Roweis and Ghahramani [21] outlined the connection between linear Gaussian latent variable models and a wide variety of methods including PCA. Probabilistic PCA is defined as a linear Gaussian latent variable model. The generative process defined by the model is that data in the reduced dimensional latent space are drawn from an isotropic Gaussian distribution. The observed data are then drawn from a Gaussian distribution with a linearly projected mean. Standard PCA can be easily derived as a special case of probabilistic PCA, where observed data are drawn deterministically from the linear projection of the lower dimensional data [26]. Other recent work has used this insight to obtain supervised forms of the probabilistic PCA techniques [32].

While vector space representations for text documents are widely used in a variety of fields and applications [23], from the probabilistic interpretation of PCA, it is clear that the model assumptions implicit within the original LSA approach for documents are inappropriate. For example, negative values in the term document matrix should not be possible within the framework of an underlying model. Furthermore, documents can consist of a rich variety of attributes such as information concerning authors, time stamps and various other relationships. We are interested in capturing

KDD'07, August 12–15, 2007, San Jose, California, USA.

this richness of information using an appropriate probabilistic model.

A number of probabilistically motivated methods have been proposed to obtain more realistic principal components for documents. One variation of LSA, the probabilistic LSA (PLSA) model was proposed by Hofmann [13] in which documents are represented with a more natural "bag of words" encoding in which each word arises from a hidden, discrete topic variable.

The general approach of PLSA has been extended to a method known as latent Dirichlet allocation (LDA) [3], which is now a state-of-the-art method for document topic modeling. LDA is a three-level hierarchical Bayesian model in which each item in a collection is modeled as a finite mixture over an underlying set of topics and each topic is modeled as an infinite mixture over an underlying set of topic probabilities. A survey of a number of similar techniques, also called discrete PCA, is given in [4]. These methods are all based on directed probabilistic graphical models where interactions between variables are encoded as conditional probabilities. However, for richly structured documents such modeling restrictions are limiting, as explained in Section 2.1.

To address existing modeling limitations we develop and present an approach here based on undirected probabilistic graphical models. Our model couples words encoded as draws of *discrete random variables* with a multidimensional *continuous latent variable* in a probabilistically principled framework. Extending this general approach we show how to model other attributes of documents such as authors and timestamps in a natural way. We then show how the well-known benefits of *supervised dimensionality reductions* are also easily obtained through a minor modification of our optimization procedure. Finally, we present qualitative results recovering topics within 13-year academic conference proceedings and a 10-month academic email archive. We then present quantitative results for authorship identification for research papers and recipient prediction for email messages.

2. GRAPHICAL MODELS AND TOPIC MODELING

It is common to describe and categorize probabilistic models as either directed or undirected graphical models (also known as, using Bayesian networks and Markov random fields). Models of words alone such as LDA [3, 9] are an example of a directed model. Other directed models have been proposed for heterogeneous information also associated with documents. For example, stochastic block structure models [20, 16] have been developed for relations between entities, the mixed membership technique [7] models words and research paper citations, words and authors are modeled by Steyvers et al. [24], senders and recipients in an email social network are modeled by McCallum et al. [17], words and relations such as voting patterns are modeled by Wang et al. [28], while words and their timestamps are modeled by Blei and Lafferty [2] and Wang and McCallum [27].

2.1 Directed Models

Directed graphical models can be described as 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 as a generative process and the direction of directed edges in the underlying graph can arguably be set either way. For example, when considering the authors and topics of documents, one can give reasonable arguments about either authors $\rightarrow$ topics or topics $\rightarrow$ authors. Particularly, when dealing with multiple modalities, the huge number of possible configurations of these directions between a large number of random variables have complicated the application of directed models to more complex multimodal, heterogeneous textual data.

Furthermore, in state-of-the-art hierarchical Bayesian models such as LDA, exact posterior inference over hidden topic variables and parameters is typically intractable and approximate inference techniques such as variational methods [15], Gibbs sampling [1] and expectation propagation [19] are employed to address these issues. As a result, the inference for obtaining a topic decomposition for a previously unseen document can be slow and troublesome.

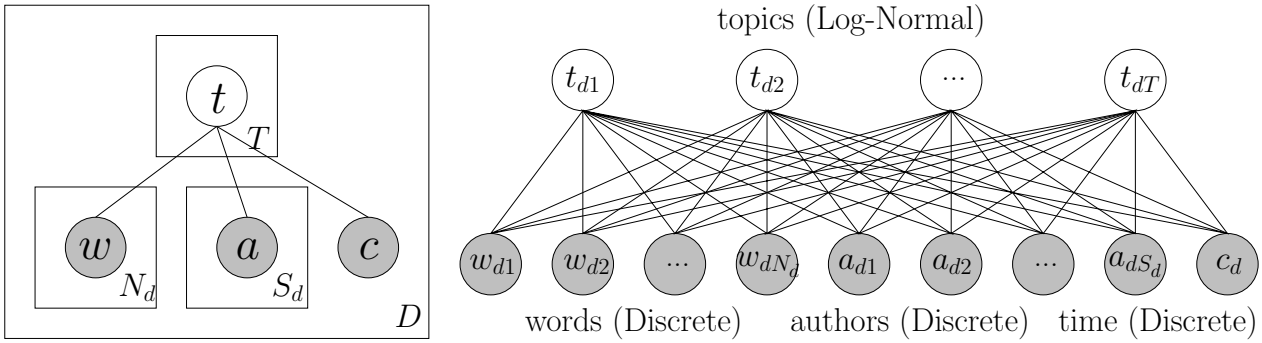
2.2 Undirected Models

Recently, a class of structured *undirected* latent variable models have gained attention for topic modeling – largely due to the fact that once model parameters have been optimized, inference of hidden topics for a new document has the complexity of a matrix multiplication, which is fast compared to hierarchical Bayesian models.

The exponential family harmonium (EFH) is one of the earlier pieces of work in this direction [29]. In Welling et al. [29], 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 *bucketed counts* of every word across the vocabulary of a text 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 provides the property that 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 leads to more difficult learning due to the intractable normalizing constant in these types of undirected models. Fortunately, the contrastive divergence [12] approach has been shown to be efficient for inference and effective for learning in these models. Further and more importantly, in many situations involving document processing, training can be done off-line, which gives us more freedom in learning.

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 [30] has been 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. This model, with some extensions, has been applied to video classification on a benchmark data set with good performance [31]. The rate adapting Poisson (RAP) model [8] is similar, but with Poisson distributions for words counts and Binomial (Bernoulli) distributions for hidden topics. The RAP model has been applied to document retrieval and object recognition to demonstrate its properties.



(a) Plate notation

(b) A corresponding restricted Boltzmann machine structure

Figure 1: Graphical representations for our models. Shaded random variables are *observed* word tokens, authors and timestamps of documents.

Most recently, a two-layer structured model has been shown to be very effective when it is applied to the Netflix movie ratings, a large but sparse tabular data set [22].

Undirected models of this structure have another important property that directed models lack: a more accurate characterization of rare words. As discussed in [30], 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 or limitation. In the harmonium-structured models, a word arises from a distribution influenced by all the topics. This different mechanism might play a crucial role in certain applications.

2.3 Our Approach

Textual documents such as research papers are very rich media that contains not only the text body considered by most of the topic models, but authors, citations, venues, and timestamps as well. Like the DWH model [30] for captioned images, we want to take advantage of the multimodal information from text documents.

In this paper, we propose a novel model, called generalized component analysis (GCA), based on the traditional 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, a random variable at 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.

To capture the rich structure of a document including attributes such as authors and timestamps we associate *different* coupling matrices for each of the different attributes. In general each attribute type is encoded as a different "bag of discrete attributes". Importantly, when conditioned on topics, draws from the bag are independent. However, when topics are unobserved, all draws are dependent. In our specific experiments here, we model words, authors and timestamps using this construction.

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 [29] a different connection matrix is needed for each word and word count level. As discussed in [29], various continuous exponential family distributions can be used to augment models for real valued attributes. The Poisson distributions adopted in [30] and [8] make it possible to use only one connection matrix, but when reconstructing the document counter vectors during contrastive divergence training (see Section 3), there is no guarantee that the reconstructed document has the same length of the original document. In such a case, at early stages of learning, the learning rate of the gradient update has to be carefully set to a small value as reported in [8] and this makes the model difficult to learn. Our model implicitly takes the document length as an input, and we find empirically that the learning process converges faster. Authors are also associated with a Discrete distribution in our setting. We now present the details of our model for generalized component analysis.

SYMBOL	DESCRIPTION
T	number of topics
D	number of documents
V	number of unique words
A	number of authors
C	number of discretized timestamps
N_d	number of word tokens in document d
S_d	number of authors in document d
M^w	$T \times (V-1)$ connection matrix for word
M^a	$T \times (A-1)$ connection matrix for author
M^c	$T \times (C-1)$ connection matrix for time
t_{di}	the i^{th} topic of document d
w_{dj}	the j^{th} word of document d
a_{dk}	the k^{th} author of document d
c_d	the (discretized) timestamp of document d

Table 1: Notation used in this paper

3. GENERALIZED COMPONENT ANALYSIS

In contrast to previous undirected topic models, in our new model, generalized component analysis (GCA), 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). The notation used in this paper is shown in Table 1. For clarity, we expand the model for document d 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:

1. Consider first, at hidden (topic) layer, a Log-normal distribution $p(t_{di}) = \text{Log-normal}(0, 1)$ for each latent topic in document d; and
2. at the observation layer,
   (a) a Discrete distribution $P(w_{dj}) = \text{Discrete}(\mathbf{0})$ for each word token in document d,
   (b) a Discrete distribution $P(a_{dk}) = \text{Discrete}(\mathbf{0})$ for each author of document d, and
   (c) a Discrete distribution $P(c_d) = \text{Discrete}(\mathbf{0})$ for the timestamp of document d;

where we 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 (for example, with words) with natural parameter θ_k ($k = 1, \cdots, V-1$) that can be transformed into the probability vector $\pi_k = e^{\theta_k} / \sum_{v=1}^{V} e^{\theta_v}$ (note here we set $\theta_V = 0$).

For simplicity, as shown in the above description, we do not use local potentials, but it is straightforward to define and learn these potentials as well, as demonstrated in previous harmonium-structured models [29, 30, 8]. Also, as in the DWH model, it is possible to mix together discrete and continuous distributions for different modalities, for example, we could utilize some continuous distribution such as Gaussian and Beta to model time without discretization as in [27].

Once we 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^w, M^a and M^c to obtain a joint probability distribution in exponential family form as follows:

$$P(\mathbf{t_d}, \mathbf{w_d}, \mathbf{a_d}, c_d) \propto \exp(\sum_{i=1}^{T}(-\log(t_{di}) - \frac{1}{2}\log^2(t_{di}) + (\sum_{j=1}^{N_d} M^w_{iw_{dj}} + \sum_{k=1}^{S_d} M^a_{ia_{dk}} + M^c_{ic_d})\log(t_{di}))) \quad (1)$$

where, for notational convenience, we set $M^w_{iV} = 0$, for $i = 1, \cdots, T$, $M^a_{iA} = 0$, for $i = 1, \cdots, T$, and $M^c_{iC} = 0$, 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,

$$\begin{aligned} & p(t_{di}|\mathbf{w_d}, \mathbf{a_d}, c_d) \\ = \ & \text{Log-normal}(\sum_{j=1}^{N_d} M^w_{iw_{dj}} + \sum_{k=1}^{S_d} M^a_{ia_{dk}} + M^c_{ic_d}, 1) \\ = \ & \text{Log-normal}(\sum_{j=1}^{V-1} M^w_{ij} m_{dj} + \sum_{k=1}^{S_d} M^a_{ia_{dk}} + M^c_{ic_d}, 1) \quad (2) \end{aligned}$$

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

$$P(a_{dk}|\mathbf{t_d}) = \text{Discrete}(\sum_{i=1}^{T} \log(t_{di}) M^a_{i\cdot}) \quad (4)$$

$$P(c_d|\mathbf{t_d}) = \text{Discrete}(\sum_{i=1}^{T} \log(t_{di}) M^c_{i\cdot}) \quad (5)$$

where m_{dj} is the count of word j in document d.

From the joint probability of all random variables (Eqn. 1), we can marginalize out the latent topic variables, and obtain the marginal likelihood of the observed document d. Note that there is no marginal independence between the observed variables although they are conditionally independent give the hidden topics.

$$\begin{aligned} & P(\mathbf{w_d}, \mathbf{a_d}, c_d) \\ \propto \ & \exp(\frac{1}{2}\sum_{i=1}^{T}(\sum_{j=1}^{V-1} M^w_{ij} m_{dj} + \sum_{k=1}^{S_d} M^a_{ia_{dk}} + M^c_{ic_d})^2) \end{aligned}$$

Our objective function, the marginal likelihood of the whole corpus, thus can be calculated (up to a normalizing constant) as

$$\begin{aligned} & \prod_{d=1}^{D} P(\mathbf{w_d}, \mathbf{a_d}, c_d) \\ \propto \ & \exp(\frac{1}{2}\sum_{d=1}^{D}\sum_{i=1}^{T}(\sum_{j=1}^{V-1} M^w_{ij} m_{dj} + \sum_{k=1}^{S_d} M^a_{ia_{dk}} + M^c_{ic_d})^2) \quad (6) \end{aligned}$$

3.1 Parameter Learning by Contrastive Divergence

Parameters of our model could be learned by gradient ascent on the marginal (log) likelihood in Eqn. 6. However, due to the intractability of the normalizing constant, it is difficult to calculate the gradient of the log-likelihood. We use contrastive divergence [12] which has been shown to greatly improve learning efficiency in harmonium architectures [29, 30, 8].

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 $\hat{m}_{dk}$, $d = 1, \ldots, D$, and $k = 1, \ldots, V-1$) from the truncated chain to approximate the model distribution.[1] In this way, the learning rule, by taking derivatives of the (unnomalized) log-likelihood objective function in Eqn. 6, can be written as the difference between the empirical average $\langle\cdot\rangle_{\tilde{p}}$ where $\tilde{p}$ denotes the empirical distribution determined by our observations, and the approximated (by contrastive divergence) model average $\langle\cdot\rangle_{p_{CD}}$ where p_{CD} denotes the model distribution approximated by the samples from the truncated Gibbs chain in our contrastive divergence learning.

$$\begin{aligned} \delta M^w_{ij} \propto & \sum_{d=1}^{D}(m_{dj}(\sum_{v=1}^{V-1} M^w_{iv} m_{dv} + \sum_{k=1}^{S_d} M^a_{ia_{dk}} + M^c_{ic_d}) \\ & - \hat{m}_{dj}(\sum_{v=1}^{V-1} M^w_{iv} \hat{m}_{dv} + \sum_{k=1}^{S_d} M^a_{i\hat{a}_{dk}} + M^c_{i\hat{c}_d})) - \frac{M^w_{ij}}{\sigma^2} \quad (7) \end{aligned}$$

[1] More details on contrastive divergence learning can be found in [12].

Algorithm 1 Learning via Contrastive Divergence

Input: document $\mathbf{w_d}, \mathbf{a_d}, c_d$ $(d = 1, \cdots, D)$, topic# T
Initialize connection matrix M^w, M^a, M^c randomly
repeat
  for $d = 1$ to D **do**
    for $i = 1$ to T **do**
      Draw t_{di}, according to $p(t_{di}|\mathbf{w_d}, \mathbf{a_d}, c_d)$ in Eqn. 2
    end for
    for $j = 1$ to N_d **do**
      Draw $\hat{w}_{dj}$, according to $P(w_{dj}|\mathbf{t_d})$ in Eqn. 3
    end for
    for $k = 1$ to S_d **do**
      Draw $\hat{a}_{dk}$, according to $P(a_{dk}|\mathbf{t_d})$ in Eqn. 4
    end for
    Draw $\hat{c}_d$, according to $P(c_d|\mathbf{t_d})$ in Eqn. 5
  end for
  for $i = 1$ to T **do**
    for $j = 1$ to $V - 1$ **do**
      Update M^w_{ij}, according to Eqn. 7
    end for
    for $k = 1$ to $A - 1$ **do**
      Update M^a_{ik}, according to Eqn. 8
    end for
    for $b = 1$ to $C - 1$ **do**
      Update M^c_{ib}, according to Eqn. 9
    end for
  end for
until M^w, M^a, M^c converge

Similarly, we can obtain the learning rules for the other two connection matrices,

$$\delta M^a_{ik} \propto \sum_{d=1}^{D}(I(k \in \mathbf{a_d})(\sum_{v=1}^{V-1} M^w_{iv} m_{dv} + \sum_{k=1}^{S_d} M^a_{ia_{dk}} + M^c_{ic_d}) - I(k \in \hat{\mathbf{a}}_\mathbf{d})(\sum_{v=1}^{V-1} M^w_{iv}\hat{m}_{dv} + \sum_{k=1}^{S_d} M^a_{i\hat{a}_{dk}} + M^c_{i\hat{c}_d})) - \frac{M^a_{ik}}{\sigma^2} \quad (8)$$

and,

$$\delta M^c_{ib} \propto \sum_{d=1}^{D}(I(b = c_d)(\sum_{v=1}^{V-1} M^w_{iv} m_{dv} + \sum_{k=1}^{S_d} M^a_{ia_{dk}} + M^c_{ic_d}) - I(b = \hat{c}_d)(\sum_{v=1}^{V-1} M^w_{iv}\hat{m}_{dv} + \sum_{k=1}^{S_d} M^a_{i\hat{a}_{dk}} + M^c_{i\hat{c}_d})) - \frac{M^c_{ib}}{\sigma^2} \quad (9)$$

where $I(q \in Q)$ and $I(a = b)$ are indicator functions, and the last terms in all formulae come from a Gaussian prior over parameters (with variance σ^2) which provides smoothing to help cope with sparsity in the training data [5]. This prior favors parameters that are closer to zero, and penalize (both positive and negative) large values of parameters. We summarize the above contrastive divergence learning procedures in Algorithm 1.

The introduction of this prior also helps alleviate the identifiability problem as reported in [29] and [8], that is, it makes the model more identifiable. Without further special handling of identifiability issues, we still get surprisingly accurate and interpretable results as shown in Section 5. Priors over weights can also 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 [25]. When given inappropriate number of topics, a model without a prior will try to duplicate some topic or create some random (but usually not trivial) topics. With priors, the spurious topics will gradually become trivial (near zero everywhere) since the priors push the weights toward zero where there is not enough data evidence supporting them.

3.2 Discriminative Learning

To explicitly emphasize that we want to infer one modality from other modalities, we can perform discriminative training by optimizing our model for a conditional likelihood (CL). Other alternatives include multi-conditional learning (MCL), a training criterion based on weighted combinations of multiple log conditional likelihoods [18].

The update rules can be simplified during discriminative learning since we do not need to reconstruct the modalities that we are not aiming to infer. In this section, we use an author prediction task as an example to illustrate how to do discriminative training in GCA, namely, we are interested in inferring authors from the given text and timestamp of a document. This task is fundamentally difficult considering there could often be hundreds of possible authors.

Using the discriminative learning criterion, we can obtain an alternative, simpler objective function $\prod_{d=1}^{D} P(\mathbf{a_d}|\mathbf{w_d}, c_d)$. Similar to Eqn. 7, 8, and 9 in regular training, we can arrive at the learning rules under discriminative learning as follows,

$$\begin{aligned}
\delta M^w_{ij} &\propto \sum_{d=1}^{D} m_{dj}(\sum_{k=1}^{S_d} M^a_{ia_{dk}} - \sum_{k=1}^{S_d} M^a_{i\hat{a}_{dk}}) - \frac{M^w_{ij}}{\sigma^2} \\
\delta M^a_{ik} &\propto \sum_{d=1}^{D}(I(k \in \mathbf{a_d})(\sum_{v=1}^{V-1} M^w_{iv} m_{dv} + \sum_{k=1}^{S_d} M^a_{ia_{dk}} + M^c_{ic_d}) \\
&- I(k \in \hat{\mathbf{a}}_\mathbf{d})(\sum_{v=1}^{V-1} M^w_{iv} m_{dv} + \sum_{k=1}^{S_d} M^a_{i\hat{a}_{dk}} + M^c_{ic_d})) - \frac{M^a_{ik}}{\sigma^2} \\
\delta M^c_{ib} &\propto \sum_{d=1}^{D} I(b = c_d)(\sum_{k=1}^{S_d} M^a_{ia_{dk}} - \sum_{k=1}^{S_d} M^a_{i\hat{a}_{dk}}) - \frac{M^c_{ib}}{\sigma^2}
\end{aligned}$$

In Section 5, we show the difference between the two training criteria, and empirically demonstrate that discriminative learning is significantly better for tasks such as author prediction for research papers and recipient prediction for email messages.

4. DATA SETS

We apply our models to two large text corpora, academic research papers and email messages of a researcher, and show the results in Section 5.

4.1 NIPS Data Set

The NIPS proceeding data set consists of the full text of the 13 years of proceedings from the Neural Information Processing Systems (NIPS) Conferences 1987 to 1999.[2] In addition to downcasing and removing stopwords and numbers, we also removed the words appearing less than five times in the corpus—many of them produced by OCR errors. Two-letter words (primarily coming from equations),

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

"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

Table 2: 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 10 positive words (left) and the top 10 negative ones (right) with the corresponding weights. Here, for displaying convenience, we have multiplied all the learned weights by a factor of 10. The learned topics are considered well known to exist within the NIPS community.

were removed, except for "ML", "AI", "KL", "BP", "EM" and "IR."

We also remove the authors who published fewer than 6 NIPS papers during 1987-1999, and only keep the papers co-authored by at least one of the remaining authors. Our data set contains 873 research papers, 125 authors, 13,576 unique words, and 1,173,343 word tokens in total. The timestamps we use are the publication years of the papers.

4.2 Academic Email Data Set

This data set consists of the last author's email archive of the ten months from January to October 2004 and here we only consider all the emails sent by McCallum to facilitate the recipient prediction task. In order to model only the new text entered by the author of each message, it is necessary to remove "quoted original messages" in replies. We eliminate this extraneous text by a simple heuristic: all text in a message below a "forwarded message" line or timestamp is removed. This heuristic does incorrectly delete text that are interspersed with quoted email text. Words are formed from sequences of alphabetic characters; stopwords are removed, and all text is downcased.

Similarly to the preprocessing steps used for the NIPS data set, we remove the recipients who got fewer than 6 emails from McCallum during that period and only keep the emails received by at least one of the remaining recipients. The data set contains 4,643 email messages, 190 recipients, 8,693 unique words, and 97,418 word tokens in total. Each document's timestamp is determined by the month the message was sent.

5. EXPERIMENTAL RESULTS

In this section, we first show several lists over words, authors and time for several learned topics, as anecdotal evidence, and then we compare our model with previous models in author prediction on the NIPS data set and recipient prediction on the Email data set.

5.1 Interpretable Topics for Conference Papers and Email Messages

We present the word list for a subset of topics learned within our weight matrices from the NIPS data set, first only using the text modality as shown in Table 2. Immediately, we can see that all the positive words provide a broad, 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.

In contrast, we believe that the topics which emerge when our model possesses author and time components tend to be more subtle and in some sense of higher fidelity. For example, Table 3 illustrates a VLSI topic and a Vision Science topic extracted from the NIPS data under the richer model. Interestingly, authors exhibit different co-occurrence patterns. For example, in our selection here, C. Koch is present in both the VLSI and Vision Science topic while T. Sejnowski is highly prominent only in the Vision Science topic and J. Platt is highly prominent only in the VLSI topic. A selection of (early) NIPS publications from these authors is given in Table 4 to further illustrate the effectiveness of our model and the relevance of these topics.

Table 5 depicts a selection of topics extracted from McCallum's email archive. The first topic concerns the writing of a paper with collaborators with usernames: *fuchun*, *wellner* and *mhay*. User *jensen* was also involved in earlier stages of the research and other people lower on the list were not involved with the paper but are collaborators and assistants. The second topic concerns the construction of a system for finding email contact information from the web. Users *culotta* and *ronb* were heavily involved in the system construction. User *pereira* is involved with the associated project called 'CALO.' Also interestingly, from the temporal modality, we can find this piece of work was primarily done in January and February, 2004, the annual spring paper submission season.

5.2 Author Prediction on NIPS Data Set

Author prediction for a document is fundamentally difficult: (1) in practice, the pool of potential authors for a given document could be very large; (2) the number of authors on a test document is often unknown. Obviously, accuracy across so many authors would not be informative. Here, we use the mean reciprocal rank (MRR) measure to evaluate the performance of models.

"VLSI"									
Words				**People**				**Year**	
analog	0.080	basis	-0.091	Platt, J	0.207	Sejnowski, T	-0.466	1991	0.138
pulse	0.068	representation	-0.089	Harris, J	0.205	Mel, B	-0.384	1997	0.122
chip	0.066	pca	-0.086	Alspector, J	0.153	Dayan, P	-0.349	1992	0.080
vlsi	0.066	representations	-0.084	Koch, C	0.153	Mozer, M	-0.346	1993	0.051
velocity	0.056	class	-0.084	Lazzaro, J	0.146	Zemel, R	-0.331	1998	0.050
synapse	0.049	structure	-0.081	Principe, J	0.137	Cottrell, G	-0.314	1989	0.032
circuit	0.047	face	-0.079	Mead, C	0.135	Tenenbaum, J	-0.304	1987	0.028
voltage	0.042	mixture	-0.078	Cauwenberghs, G	0.112	Wiles, J	-0.293	1996	0.021
trajectory	0.042	context	-0.075	Mjolsness, E	0.108	Pouget, A	-0.283	1999	0.000
circuits	0.041	zemel	-0.072	Maass, W	0.100	Ahmad, S	-0.267	1990	-0.013
"Vision Science"									
Words				**People**				**Year**	
orientation	0.083	decision	-0.095	Sejnowski, T	0.398	Bartlett, P	-0.278	1989	0.038
dominance	0.080	bounds	-0.089	Koch, C	0.375	Jaakkola, T	-0.240	1991	0.032
visual	0.080	risk	-0.088	Pouget, A	0.289	Shavlik, J	-0.239	1988	0.029
ocular	0.079	theorem	-0.087	Kawato, M	0.263	Tesauro, G	-0.207	1994	0.006
velocity	0.078	margin	-0.080	Obermayer, K	0.260	Thrun, S	-0.207	1996	0.005
stimuli	0.075	trees	-0.075	Nowlan, S	0.230	Bengio, Y	-0.199	1997	0.002
eye	0.075	boosting	-0.072	Dayan, P	0.219	Kowalczyk, A	-0.194	1999	0.000
cortex	0.070	policy	-0.070	Zemel, R	0.214	Cohn, D	-0.181	1992	-0.013
cortical	0.069	cost	-0.069	Lee, D	0.199	Baluja, S	-0.178	1990	-0.038
lgn	0.068	algorithms	-0.067	Li, Z	0.198	Lee, Y	-0.177	1987	-0.048

Table 3: Two topics (VLSI and Vision Science) 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 10 positive words/authors (left) and the top 10 negative ones (right) with the corresponding weights and the top 10 timestamps. We use our model to explicitly account for words, authors and time.

Terrence J. Sejnowski
— *Recurrent Eye Tracking Network Using a Distributed Representation of Image Motion.* NIPS 1991
— *Combining Visual and Acoustic Speech Signals with a Neural Network Improves Intelligibility.* NIPS 1989
John C. Platt
— *Analog Circuits for Constrained Optimization.* NIPS 1989
— *An Analog VLSI Chip for Radial Basis Functions.* NIPS 1992
Christof Koch
— *An Integrated Vision Sensor for the Computation of Optical Flow Singular Points.* NIPS 1998
— *Analog VLSI Circuits for Attention-Based, Visual Tracking.* NIPS 1996
— *An Analog VLSI Saccadic Eye Movement System.* NIPS 1993

Table 4: A selection of NIPS publications from several authors shown in Table 3.

In traditional information retrieval, given a query, we rank the documents in a corpus by some score, such as vector-space cosine similarity between document and query [23], and query likelihood [33] and take the top ones as the retrieved documents. Obviously, not all the retrieved documents are relevant to the given query. In our setting, we project a given test document and all of the training documents into latent space, and rank all the training documents according vector based cosine similarity with the test document. When the intersection of the author sets of the test document and a retrieved document is not empty, we output that the retrieved document as relevant. The reciprocal rank of a test document is the reciprocal of the rank at which the first relevant response was returned, or 0 if none of the responses contained a relevant answer. The score for a sequence of queries is the mean of the individual query's reciprocal ranks.

We randomly split the NIPS data set into training set (9/10, 786 documents) and test set (1/10, 87 documents). We compare our models (discriminatively trained and regularly trained) with the author-topic (AT) model [24] and singular value decomposition (SVD), all using 20 hidden topics.

We compare our results with GCA to those of the author-topic (AT) model — a Bayesian network, in which each author's interests are modeled with a *mixture* of topics [24]. In its generative process for each document d, a set of authors, a_d, is observed. To generate each word, an author x is chosen uniformly from this set, then a topic t is selected from a multinomial topic distribution θ_x that is specific to the author, and then a word w is generated from a topic-specific multinomial distribution ϕ_t over words. θ and ϕ are drawn from conjugate Dirichlet priors. The posterior estimates $\hat{\theta}$ and $\hat{\phi}$ of these two mixtures can be obtained

"Writing a Paper on Coreference"									
Words				**People**				**Month**	
paper	0.053	email	-0.019	fuchun	0.412	culotta	-0.073	5	0.470
model	0.047	people	-0.015	wellner	0.395	ronb	-0.065	2	0.335
section	0.037	find	-0.003	mhay	0.373	traustik	-0.030	9	0.208
inference	0.034	addrie	-0.001	jensen	0.192	viola	-0.021	6	0.157
results	0.030	clyde	-0.001	pereira	0.085	system	-0.021	3	0.106
models	0.028	calobase	-0.000	lafferty	0.073	lsaul	-0.017	8	0.031
work	0.028	sgml	-0.000	kate	0.060	souccar	-0.013	7	0.012
coreference	0.025	green	-0.000	mahadeva	0.050	tzhang	-0.013	10	0.000
ben	0.024	remotely	-0.000	casutton	0.045	jst@	-0.012	4	-0.135
text	0.023	emacs	-0.000	jean	0.040	szmiller	-0.012	1	-0.223
"Building a Contact Finding System"									
Words				**People**				**Month**	
aron	0.026	model	-0.006	culotta	0.357	fuchun	-0.168	1	0.334
data	0.021	research	-0.006	weili	0.116	wellner	-0.153	2	0.309
email	0.018	inference	-0.003	ronb	0.111	mhay	-0.131	9	0.112
people	0.015	spider	-0.003	pereira	0.105	saunders	-0.109	7	0.012
find	0.011	paper	-0.003	viola	0.097	mikem	-0.062	6	0.003
results	0.011	mccallum	-0.003	lafferty	0.086	jensen	-0.041	10	0.000
ron	0.010	fuchun	-0.002	traustik	0.074	adingle	-0.035	4	-0.060
calo	0.010	papers	-0.002	ghuang	0.046	slfeng	-0.015	5	-0.064
class	0.009	prototype	-0.002	hough	0.032	jean	-0.014	3	-0.093
training	0.009	mysql	-0.002	casutton	0.031	msindela	-0.014	8	-0.176

Table 5: A topic concerning the writing of a paper about coreference techniques (top) and a topic about building a system for finding email contacts (bottom). The "Title" above the word lists of each topic is our own summary of the topics. For each topic, we show the top 10 positive words/recipients (left) and the top 10 negative ones (right) with the corresponding weights and the top 10 timestamps. Topics were found within McCallum's email archive using our model for words, authors and timestamps encoded for each month.

Data Set	Discriminative Learning	Regular Learning	Author-Topic Model	SVD	Words Only
NIPS	0.8742	0.4590	0.5094	0.3791	0.2637
Email	0.6009	0.3060	0.3715	0.2706	0.2207

Table 6: The best mean reciprocal rank (MRR) score for GCA with discriminative learning, GCA with regular learning, the author-topic (AT) model and SVD. We also include the MRR score of GCA on the word part of the data set only to show the benefit of including additional information from non-word modalities. All models are trained with 20 hidden topics.

conveniently during the training stage by Gibbs sampling, variational methods, or expectation propagation.

To predict an author a of a given new document d, we can calculate the posterior probability of author a given $\mathbf{w}_d$ using Bayes rule, $P(a|\mathbf{w}_d) \propto P(\mathbf{w}_d|a)P(a)$. Here, for simplicity, we treat the timestamps of documents as additional words. The author(s) with highest posterior probabilities are our predictions. The prior of author a, $P(a)$, can be estimated by counting how many times he/she (co-)authored a paper in the training set, and the data likelihood of the words can be obtained by summing over all possible topic assignments of each token, as shown below,

$$P(\mathbf{w}_d|a) = \prod_{i=1}^{N_d} \left(\sum_{t=1}^{T} \hat{\theta}_{at} \hat{\psi}_{tw_{di}} \right).$$

We also compare our model versus SVD: we ignore the heterogeneity of the data, and aggregate the word counts, the authors, and the timestamps of the documents into a big matrix, conduct SVD analysis, and then find the lower dimensional representations of the documents.

The MRR scores are shown in Table 6. To demonstrate the advantage of incorporating information from multiple modalities, we also run the model on words only. As observed in Table 6, even with regular training, our model outperforms SVD and AT with uniform prior on authors (not shown in Table 6). With discriminative training, our model is significantly better than SVD and the author-topic model, achieving a MRR score more than twice as large as from SVD.

Note that, (1) author prediction on test documents for the author-topic model is relatively slow because we need sum over all possible topic assignments of each word token. On the other hand, both for GCA and SVD, this can be done by simple matrix multiplication; (2) training can be done offline, however we want to point out that discriminative learning is much more efficient in our setting than regular learning because we do not need to reconstruct the words and time during contrastive divergence learning.

We also show how the MRR scores change on the NIPS data set as the number of learning iterations increase in Fig-

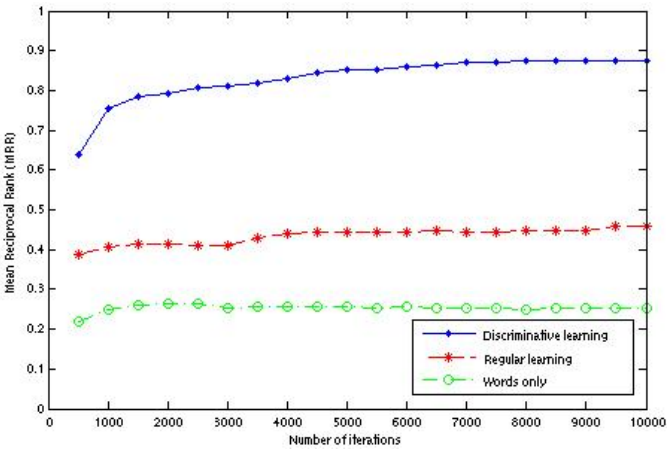


Figure 2: The mean reciprocal rank (MRR) scores vs. number of iterations for different models on the NIPS data set. All models are trained with 20 hidden topics.

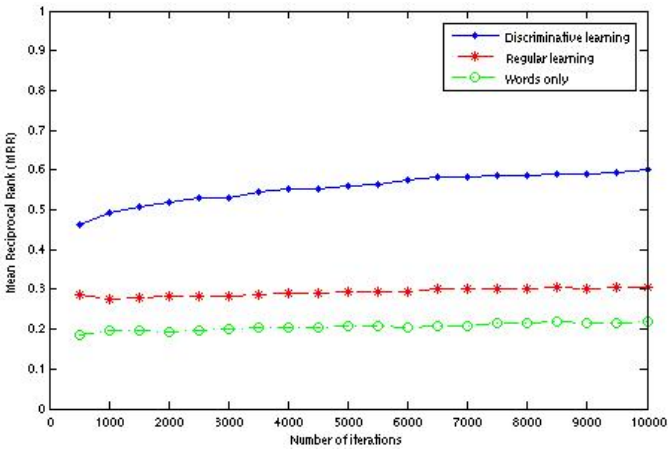


Figure 3: The mean reciprocal rank (MRR) scores vs. number of iterations for different models on the Email data set. All models are trained with 20 hidden topics.

ure 2. These curves also serve as additional evidence regarding whether the number of training iterations is sufficient.

5.3 Recipient Prediction on Email Data Set

Recipient prediction (also called CC prediction) has recently attracted significant interest. As an important office application, recipient prediction seems very similar to author prediction discussed in the previous section. We can easily adapt the same setting used for author prediction to do recipient prediction. We randomly split the Email data set into training set (9/10, 4,179 documents) and test set (1/10, 464 documents).

The MRR scores are reported in Table 6. Again, we can quickly see that the discriminatively trained GCA greatly outperforms other models. The MRR scores change on the Email data set as the number of learning iterations increase are shown in Figure 3 for different models.

Also, we can find that MRR scores are consistently worse than the ones on NIPS data set. Our conjecture is that the body message of an email is in general much shorter than a research paper; an email's body message can be as simple as one word. Additionally, the text in email body is composed by the sender, although it should reflect recipients' interests or expertise.

6. CONCLUSION AND DISCUSSION

We have proposed a new harmonium-structured undirected model for large text collections that simultaneously take into account information from multiple modalities. For the discrete attributes of documents such as words, unlike the previous models, the new model still allows the words to come from a discrete distribution in a 'bag-of-words' fashion. Thus, our model implicitly takes document length as input, which greatly increases the efficiency during the contrastive divergence learning.

We have shown interpretable topics over various document attributes (words, authors, time) on two large text collections, and demonstrate better mean reciprocal rank (MRR) performance, over other models, on author prediction task on the NIPS data set and recipient prediction task on the Email data set. Our models can be applied to tasks with similar objectives such as targeted advertising.

Our models with these hidden layer structures allow a great deal of flexibility to incorporate information from multiple modalities as demonstrated. 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 not comparable and weighting parameters are often needed as in [27]. We see great potential to combine a wide variety of information from other attributes 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 data with heterogeneous attributes.

7. 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-0326249, 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. The second author appreciates support by Microsoft Research under the Memex and eScience funding programs and support from Kodak Research. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

[2] D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA, 2006.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] W. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In M. Chickering and J. Halpern, editors, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 59–66, Banff, Alberta, Canada, 2004.

[5] S. F. Chen and R. Rosenfeld. A Gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University, CMU-CS-99-108, 1999.

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

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

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

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

[10] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, Vancouver, British Columbia, Canada, 2004.

[11] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[12] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[13] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, 1999.

[14] I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 2002.

[15] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pages 105–161, 1998.

[16] C. Kemp, T. L. Griffiths, and J. Tenenbaum. Discovering latent classes in relational data. Technical report, MIT CSAIL, 2004.

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

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

[19] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 2002.

[20] K. Nowicki and T. A. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455), 2001.

[21] S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.

[22] R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[23] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[24] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Technical report, University of California, Berkeley, Department of Statistics, 2004.

[26] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21(3):611–622, 1990.

[27] X. Wang and A. McCallum. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[28] X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and their attributes. In *Advances in Neural Information Processing Systems 18*, Vancouver, British Columbia, Canada, 2005.

[29] M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems 17*, Vancouver, British Columbia, Canada, 2004.

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

[31] J. Yang, Y. Liu, E. P. Xing, and A. Hauptmann. Harmonium-based models for semantic video representation and classification. In *Proceedings of the Seventh SIAM International Conference on Data Mining*, 2007.

[32] S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu. Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 464–473, New York, NY, USA, 2006. ACM Press.

[33] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information System*, 22(2):179–214, 2004.