Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations

Wei Li WEILI@CS.UMASS.EDU
Andrew McCallum MCCALLUM@CS.UMASS.EDU
University of Massachusetts, Dept. of Computer Science

Abstract

Latent Dirichlet allocation (LDA) and other related topic models are increasingly popular tools for summarization and manifold discovery in discrete data. However, LDA does not capture correlations between topics. In this paper, we introduce the *pachinko allocation* model (PAM), which captures arbitrary, nested, and possibly sparse correlations between topics using a directed acyclic graph (DAG). The leaves of the DAG represent individual words in the vocabulary, while each interior node represents a correlation among its children, which may be words or other interior nodes (topics). PAM provides a flexible alternative to recent work by Blei and Lafferty (2006), which captures correlations only between *pairs* of topics. Using text data from newsgroups, historic NIPS proceedings and other research paper corpora, we show improved performance of PAM in document classification, likelihood of held-out data, the ability to support finer-grained topics, and topical keyword coherence.

1. Introduction

Statistical topic models have been successfully used to analyze large amounts of textual information in many tasks, including language modeling, document classification, information retrieval, document summarization and data mining. Given a collection of textual documents, parameter estimation in these models discovers a low-dimensional set of multinomial word distributions called "topics". Mixtures of these topics give high likelihood to the training data, and the highest probability words in each topic provide keywords that briefly summarize the themes in the text collection. In addition to textual data (including news articles, research papers and email), topic models have also been applied to images, biological findings and other non-textual multi-dimensional discrete data.

Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a widely-used topic model, often applied to textual data, and the basis for many variants. LDA represents each document as a mixture of topics, where each topic is a multinomial distribution over words in a vocabulary. To generate a document, LDA first samples a per-document multinomial distribution over topics from a Dirichlet distribution. Then it repeatedly samples a topic from this multinomial and samples a word from the topic.

The topics discovered by LDA capture correlations among words, but LDA does not explicitly model correlations among topics. This limitation arises because the topic proportions in each document are sampled from a single Dirichlet distribution. As a result, LDA has difficulty modeling data in which some topics cooccur more frequently than others. However, topic correlations are common in real-world text data, and ignoring these correlations limits LDA's ability to predict new data with high likelihood. Ignoring topic correlations also hampers LDA's ability to discover a large number of fine-grained, tightly-coherent topics. Because LDA can combine arbitrary sets of topics, LDA is reluctant to form highly specific topics, for which some combinations would be "nonsensical".

Motivated by the desire to build accurate models that can discover large numbers of fine-grained topics, we are interested in topic models that capture topic correlations.

Teh et al. (2005) propose hierarchical Dirichlet pro-

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

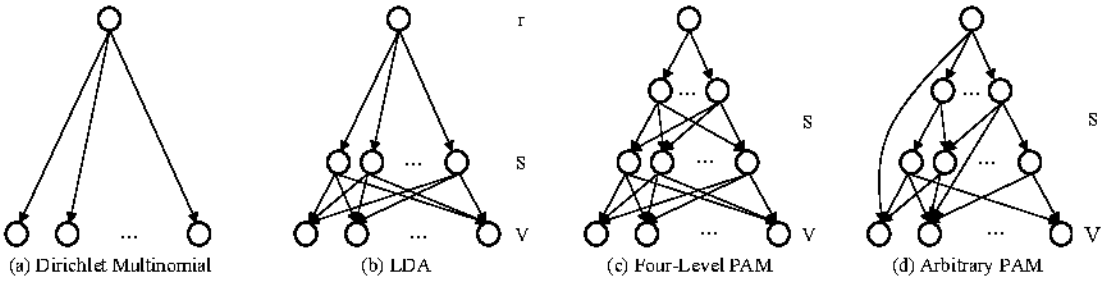

Figure 1. Model structures for four topic models (a) Dirichlet Multinomial: For each document, a multinomial distribution over words is sampled from a single Dirichlet. (b) LDA: This model samples a multinomial over topics for each document, and then generates words from the topics. (c) Four-Level PAM: A four-level hierarchy consisting of a root, a set of super-topics, a set of sub-topics and a word vocabulary. Both the root and the super-topics are associated with Dirichlet distributions, from which we sample multinomials over their children for each document. (d) PAM: An arbitrary DAG structure to encode the topic correlations. Each interior node is considered a topic and associated with a Dirichlet distribution.

cesses (HDP) to model groups of data that have a pre-defined hierarchical structure. Each pre-defined group is associated with a Dirichlet process, whose base measure is sampled from a higher-level Dirichlet process. HDP can capture topic correlations defined by this nested data structure, however, it does not automatically discover such correlations from unstructured data. A simple version of HDP does not use a hierarchy over pre-defined groups of data, but can be viewed as an extension to LDA that integrates over (or alternatively selects) the appropriate number of topics.

An alternative model that not only represents topic correlations, but also learns them, is the correlated topic model (CTM) (Blei & Lafferty, 2006). It is similar to LDA, except that rather than drawing topic mixture proportions from a Dirichlet, it does so from a logistic normal distribution, whose parameters include a covariance matrix in which each entry specifies the correlation between a pair of topics. Thus in CTM topics are not independent, however note that only pairwise correlations are modeled, and the number of parameters in the covariance matrix grows as the square of the number of topics.

In this paper, we introduce the *pachinko allocation* model (PAM), which uses a directed acyclic graph (DAG) structure to represent and learn arbitrary-arity, nested, and possibly sparse topic correlations. In PAM, the concept of topics are extended to be distributions not only over words, but also over other topics. The model structure consists of an arbitrary DAG, in which each leaf node is associated with a word in the vocabulary, and each non-leaf "interior" node corresponds to a topic, having a distribution over its children. An interior node whose children are all leaves would correspond to a traditional LDA topic. But some interior nodes may also have children that are other topics, thus representing a mixture over topics. With many such nodes, PAM therefore captures not only correlations among words (as in LDA), but also correlations among topics themselves.

For example, consider a document collection that discusses four topics: *cooking*, *health*, *insurance* and *drugs*. The *cooking* topic co-occurs often with *health* , while *health*, *insurance* and *drugs* are often discussed together. A DAG can describe this kind of correlation. Four nodes for the four topics form one level that is directly connected to the words. There are two additional nodes at a higher level, where one is the parent of *cooking* and *health*, and the other is the parent of *health*, *insurance* and *drugs*.

In PAM each interior node's distribution over its children could be parameterized arbitrarily. In the remainder of this paper, however, as in LDA, we use a Dirichlet, parameterized by a vector with the same dimension as the number of children. Thus, here, a PAM model consists of a DAG, with each interior node containing a Dirichlet distribution over its children. To generate a document from this model, we first sample a multinomial from each Dirichlet. Then, to generate each word of the document, we begin at the root of the DAG, sampling one of its children according to its multinomial, and so on sampling children down the DAG until we reach a leaf, which yields a word. The model is named for pachinko machines—a game popular in Japan, in which metal balls bounce down around a complex collection of pins until they land in various bins at the bottom.

Note that the DAG structure in PAM is extremely flexible. It could be a simple tree (hierarchy), or an arbitrary DAG, with cross-connected edges, and edges

skipping levels. The nodes can be fully or sparsely connected. The structure could be fixed beforehand or learned from the data. It is easy to see that LDA can be viewed as a special case of PAM: the DAG corresponding to LDA is a three-level hierarchy consisting of one root at the top, a set of topics in the middle and a word vocabulary at the bottom. The root is fully connected to all the topics, and each topic is fully connected to all the words. (LDA represents topics as multinomial distributions over words, which can be seen as Dirichlet distributions with variance 0.)

In this paper we present experimental results demonstrating PAM's improved performance over LDA in three different tasks, including topical word coherence assessed by human judges, likelihood on held-out test data, and document classification accuracy. We also show a favorable comparison versus CTM and HDP on held-out data likelihood.

2. The Model

In this section, we detail the pachinko allocation model (PAM), and describe its generative process, inference algorithm and parameter estimation method. We begin with a brief review of latent Dirichlet allocation.

Latent Dirichlet allocation (LDA) (Blei et al., 2003) can be understood as a special-case of PAM with a three-level hierarchy. Beginning from the bottom, it includes: $V = \{x_1, x_2, ..., x_v\}$, a vocabulary over words; $T = \{t_1, t_2, ..., t_s\}$, a set of topics; and r, the root, which is the parent of all topic nodes and is associated with a Dirichlet distribution $g(\alpha)$. Each topic is represented as a multinomial distribution over words and sampled from a given Dirichlet distribution $g(\beta)$. The model structure of LDA is shown in Figure 1(b).

To generate a document, LDA samples a multinomial distribution over topics from $g(\alpha)$, then repeatedly samples a topic from this multinomial, and a word from the topic.

Now we introduce notation for the pachinko allocation model. PAM connects words in V and topics in T with an arbitrary DAG, where topic nodes occupy the interior levels and the leaves are words. Two possible model structures are shown in Figure 1(c) and (d). Each topic t_i is associated with a Dirichlet distribution $g_i(\alpha_i)$, where α_i is a vector with the same dimension as the number of children in t_i. In general, g_i is not restricted to be Dirichlet. It could be any distribution over discrete children, such as logistic normal. But in this paper, we focus only on Dirichlet and derive the inference algorithm under this assumption.

To generate a document d, we follow a two-step process:

1. Sample $\theta^{(d)}_{t_1}$, $\theta^{(d)}_{t_2}$, ..., $\theta^{(d)}_{t_s}$ from $g_1(\alpha_1)$, $g_2(\alpha_2)$, ..., $g_s(\alpha_s)$, where $\theta^{(d)}_{t_i}$ is a multinomial distribution of topic t_i over its children.
2. For each word w in the document,
   - Sample a topic path $\mathbf{z}_w$ of length L_w: $< z_{w1}, z_{w2}, ..., z_{wL_w} >$. z_{w1} is always the root and z_{w2} through z_{wL_w} are topic nodes in T. z_{wi} is a child of $z_{w(i-1)}$ and it is sampled according to the multinomial distribution $\theta^{(d)}_{z_{w(i-1)}}$.
   - Sample word w from $\theta^{(d)}_{z_{wL_w}}$.

Following this process, the joint probability of generating a document d, the topic assignments $\mathbf{z}^{(d)}$ and the multinomial distributions $\theta^{(d)}$ is

$$P(d, \mathbf{z}^{(d)}, \theta^{(d)}|\alpha) = \prod_{i=1}^{s} P(\theta^{(d)}_{t_i}|\alpha_i) \times \prod_{w} (\prod_{i=2}^{L_w} P(z_{wi}|\theta^{(d)}_{z_{w(i-1)}}) P(w|\theta^{(d)}_{z_{wL_w}}))$$

Integrating out $\theta^{(d)}$ and summing over $\mathbf{z}^{(d)}$, we calculate the marginal probability of a document as:

$$P(d|\alpha) = \int \prod_{i=1}^{s} P(\theta^{(d)}_{t_i}|\alpha_i) \times \prod_{w} \sum_{\mathbf{z}_w} (\prod_{i=2}^{L_w} P(z_{wi}|\theta^{(d)}_{z_{w(i-1)}}) P(w|\theta^{(d)}_{z_{wL_w}})) \mathrm{d}\theta^{(d)}$$

Finally, the probability of generating a whole corpus is the product of the probability for every document:

$$P(\mathbf{D}|\alpha) = \prod_{d} P(d|\alpha)$$

2.1. Four-Level PAM

While PAM allows arbitrary DAGs to model the topic correlations, in this paper, we focus on one special structure in our experiments. It is a four-level hierarchy consisting of one root topic, s_1 topics at the second level, s_2 topics at the third level and words at the bottom. We call the topics at the second level super-topics and the ones at the third level sub-topics. The root is connected to all super-topics, super-topics are fully connected to sub-topics and sub-topics are fully connected to words (Figure 1(c)). We also make a simplification similar to LDA: the multinomial distributions for sub-topics are sampled once for the whole corpus, from a single Dirichlet distribution $g(\beta)$. The multinomials for the root and super-topics are still sampled individually for each document. As we can see, both the

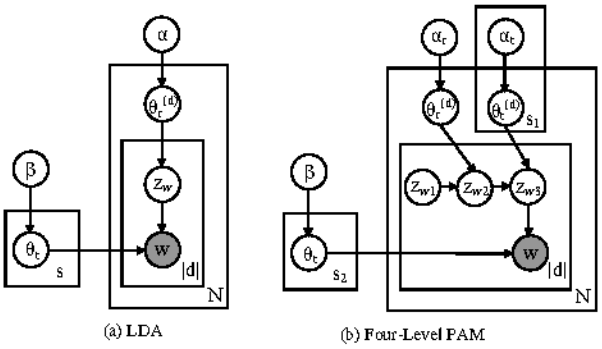

Figure 2. Graphical models for (a) LDA and (b) four-level PAM

model structure and generative process for this special setting are similar to LDA. The major difference is that it has one additional layer of super-topics modeled with Dirichlet distributions, which is the key component capturing topic correlations here. We present the corresponding graphical models for LDA and PAM in Figure 2.

2.2. Inference and Parameter Estimation

The hidden variables in PAM include the sampled multinomial distributions Θ and topic assignments $\mathbf{z}$. Furthermore, we need to learn the parameters in the Dirichlet distributions $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_s\}$. We could apply the Expectation-Maximization (EM) algorithm for inference, which is often used to estimate parameters for models involving hidden variables. However, EM has been shown to perform poorly for topic models due to many local maxima.

Instead, we apply Gibbs Sampling to perform inference and parameter learning. For an arbitrary DAG, we need to sample a topic path for each word given other variable assignments enumerating all possible paths and calculating their conditional probabilities. In our special four-level PAM structure, each path contains the root, a super-topic and a sub-topic. Since the root is fixed, we only need to jointly sample the super-topic and sub-topic assignments for each word, based on their conditional probability given observations and other assignments, integrating out the multinomial distributions Θ; (thus the time for each sample is in the number of possible paths). The following equation shows the joint probability of a super-topic and a sub-topic. For word w in document d, we have:

$$P(z_{w2} = t_k, z_{w3} = t_p | \mathbf{D}, \mathbf{z}_{-w}, \alpha, \beta) \propto \frac{n_{1k}^{(d)} + \alpha_{1k}}{n_1^{(d)} + \sum_{k'} \alpha_{1k'}} \times \frac{n_{kp}^{(d)} + \alpha_{kp}}{n_k^{(d)} + \sum_{p'} \alpha_{kp'}} \times \frac{n_{pw} + \beta_w}{n_p + \sum_m \beta_m}.$$

Here we assume that the root topic is t_1. z_{w2} and z_{w3} correspond to super-topic and sub-topic assignments respectively. $\mathbf{z}_{-w}$ is the topic assignments for all other words. Excluding the current token, $n_x^{(d)}$ is the number of occurrences of topic t_x in document d; $n_{xy}^{(d)}$ is the number of times topic t_y is sampled from its parent t_x in document d; n_x is the number of occurrences of sub-topic t_x in the whole corpus and n_{xw} is the number of occurrences of word w in sub-topic t_x. Furthermore, α_{xy} is the yth component in α_x and β_w is the component for word w in β.

Note that in the Gibbs sampling equation, we assume that the Dirichlet parameters α are given. While LDA can produce reasonable results with a simple uniform Dirichlet, we have to learn these parameters for the super-topics in PAM since they capture different correlations among sub-topics. As for the root, we assume a fixed Dirichlet parameter. To learn α, we could use maximum likelihood or maximum a posteriori estimation. However, since there are no closed-form solutions for these methods and we wish to avoid iterative methods for the sake of simplicity and speed, we approximate it by moment matching. In each iteration of Gibbs sampling, we update

$$\begin{aligned} mean_{xy} &= \frac{1}{N} \times \sum_d \frac{n_{xy}^{(d)}}{n_x^{(d)}}; \\ var_{xy} &= \frac{1}{N} \times \sum_d (\frac{n_{xy}^{(d)}}{n_x^{(d)}} - mean_{xy})^2; \\ m_{xy} &= \frac{mean_{xy} \times (1 - mean_{xy})}{var_{xy}} - 1; \\ \alpha_{xy} &\propto mean_{xy}; \\ \sum_y \alpha_{xy} &= \frac{1}{5} \times exp(\frac{\sum_y log(m_{xy})}{s_2 - 1}). \end{aligned}$$

For each super-topic x and sub-topic y, we first calculate the sample mean $mean_{xy}$ and sample variance var_{xy}. $n_{xy}^{(d)}$ and $n_x^{(d)}$ are the same as defined above. Then we estimate α_{xy}, the yth component in α_x from sample mean and variance. N is the number of documents and s_2 is the number of sub-topics.

Smoothing is important when we estimate the Dirichlet parameters with moment matching. From the equations above, we can see that when one sub-topic y does not get sampled from super-topic x in one iteration, α_{xy} will become 0. Furthermore from the Gibbs sampling equation, we know that this sub-topic will never have the chance to be sampled again by this super-topic. We introduce a prior in the calculation of sample means so that $mean_{xy}$ will not be 0 even if $n_{xy}^{(d)}$ is 0 for every document d.

3. Experimental Results

In this section, we present example topics that PAM discovers from real-world text data and evaluate against LDA using three measures: topic clarity by human judgement, likelihood of held-out test data, and document classification accuracy. We also compare held-out data likelihood with CTM and HDP.

In the experiments we discuss below, we use a fixed four-level hierarchical structure for PAM, which includes a root, a set of super-topics, a set of sub-topics and a word vocabulary. For the root, we always assume a fixed Dirichlet distribution with parameter 0.01. We can change this parameter to adjust the variance in the sampled multinomial distributions. We choose a small value so that the variance is high and each document contains only a small number of super-topics, which tends to make the super-topics more interpretable. We treat the sub-topics in the same way as LDA and assume they are sampled once for the whole corpus from a given Dirichlet with parameter 0.01. So the only parameters we need to learn are the Dirichlet parameters for the super-topics, and multinomial parameters for the sub-topics.

In Gibbs sampling for both PAM and LDA we use 2000 burn-in iterations, and then draw a total of 10 samples in the following 1000 iterations. The total training time for the NIPS dataset (as described in Section 3.2) is approximately 20 hours on a 2.4 GHz Opteron machine with 2GB memory.

3.1. Topic Examples

Our first test dataset comes from Rexa, a search engine over research papers (http://Rexa.info). We randomly choose a subset of abstracts from its large collection. In this dataset, there are 4000 documents, 278438 word tokens and 25597 unique words. Figure 3 shows a subset of super-topics in the data, and how they capture correlations among sub-topics. For each super-topic x, we rank the sub-topics $\{y\}$ based on the learned Dirichlet parameter α_{xy}. In Figure 3, each circle corresponds to one super-topic and links to a set of sub-topics as shown in the boxes, which are selected from its top 10 list. The numbers on the edges are the corresponding α values. As we can see, all the super-topics share the same sub-topic in the middle, which is a subset of stopwords in this corpus. Some super-topics also share the same content sub-topics. For example, the topics about *scheduling* and *tasks* co-occur with the topic about *agents* and also the topic about *distributed systems*. Another example is *information retrieval*. It is discussed along with both the *data mining* topic and the *web, network* topic.

Table 1. Example topic pairs in human judgement.

PAM	LDA	PAM	LDA
control	control	motion	image
systems	systems	image	motion
robot	based	detection	images
adaptive	adaptive	images	multiple
environment	direct	scene	local
goal	con	vision	generated
state	controller	texture	noisy
controller	change	segmentation	optical
5 votes	0 vote	4 votes	1 vote
PAM	LDA	PAM	LDA
signals	signal	algorithm	algorithm
source	signals	learning	algorithms
separation	single	algorithms	gradient
eeg	time	gradient	convergence
sources	low	convergence	stochastic
blind	source	function	line
single	temporal	stochastic	descent
event	processing	weight	converge
4 votes	1 vote	1 vote	4 votes

Table 2. Human judgement results. For all the categories, 5 votes, $\geq$ 4 votes and $\geq$ 3 votes, PAM has more topics judged better than LDA.

	LDA	PAM
5 votes	0	5
$\geq$ 4 votes	3	8
$\geq$ 3 votes	9	16

3.2. Human Judgement

We provided each of five human evaluators a set of topic pairs, one each from PAM and LDA, anonymized and in random order. Evaluators were asked to choose which one has stronger sense of semantic coherence and specificity.

These topics were generated using the NIPS abstract dataset (NIPS00-12), which includes 1647 documents, a vocabulary of 11708 words and 114142 word tokens. We use 100 topics for LDA, and 50 super-topics and 100 sub-topics for PAM. The topic pairs are created based on similarity. For each sub-topic in PAM, we find its most similar topic in LDA and present them as a pair. We also find the most similar sub-topic in PAM for each LDA topic. Similarity is measured by the KL-divergence between topic distributions over words. After removing redundant pairs and dissimilar pairs that share less than 5 out of their top 20 words,, we provide the evaluators with a total of 25 pairs. We present four example topic pairs in Table 1. There are 5 PAM topics that every evaluator agrees to be the better ones in their pairs, while LDA has none. And out of 25 pairs, 19 topics from PAM are chosen by the majority ($\geq$ 3 votes). We show the full evaluation results in Table 2.

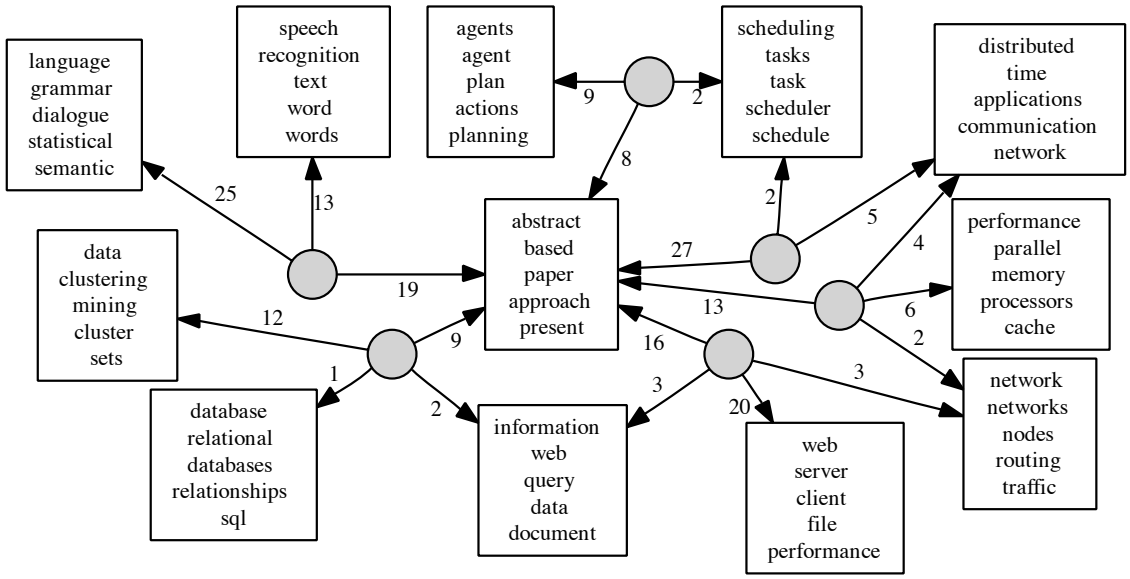

Figure 3. Topic correlation in PAM. Each circle corresponds to a super-topic each box corresponds to a sub-topic. One super-topic can connect to several sub-topics and capture their correlation. The numbers on the edges are the corresponding α values for the (super-topic, sub-topic) pair.

3.3. Likelihood Comparison

In addition to human evaluation of topics, we also provide quantitative measurements to compare PAM with LDA, CTM and HDP. In this experiment, we use the same NIPS dataset and split it into two subsets with 75% and 25% of the data respectively. Then we learn the models from the larger set and calculate likelihood for the smaller set. We use 50 super-topics for PAM, and the number of sub-topics varies from 20 to 180.

To calculate the likelihood of heldout data, we must integrate out the sampled multinomials and sum over all possible topic assignments. This problem has no closed-form solution. Previous work that uses Gibbs sampling for inference approximates the likelihood of a document d by the harmonic mean of a set of conditional probabilities $P(d|\mathbf{z}^{(d)})$, where the samples are generated using Gibbs sampling (Griffiths & Steyvers, 2004). However, this approach has been shown to be unstable because the inverse likelihood does not have finite variance (Chib, 1995) and has been widely criticized (e.g. (Newton & Raftery, 1994) discussion).

In our experiments, we employ a more robust alternative in the family of non-parametric likelihood estimates—specifically an approach based on empirical likelihood (EL), *e.g.* (Diggle & Gratton, 1984). In these methods one samples data from the model, and calculates the empirical distribution from the samples. In cases where the samples are sparse, a kernel may be employed. We first randomly generate 1000 documents from the trained model, based on its own generative process. Then from each sample we estimate a multinomial distribution (directly from the sub-topic mixture). The probability of a test document is then calculated as its average probability from each multinomial, just as in a simple mixture model. Unlike in Gibbs sampling, the samples are unconditionally generated; therefore, they are not restricted to the topic co-occurrences observed in the held-out data, as they are in the harmonic mean method.

We show the log-likelihood on the test data in Figure 4, averaging over all the samples in 10 different Gibbs sampling. Compared to LDA, PAM always produces higher likelihood for different numbers of sub-topics. The advantage is especially obvious for large numbers of topics. LDA performance peaks at 40 topics and decreases as the number of topics increases. On the other hand, PAM supports larger numbers of topics and has its best performance at 160 sub-topics. When the number of topics is small, CTM exhibits better performance than both LDA and PAM. However, as we use more and more topics, its likelihood starts to decrease. The peak value for CTM is at 60 topics and it is slightly worse than the best performance of PAM. We also apply HDP to this dataset. Since there is no pre-defined data structure, HDP does not model any topic correlations but automatically learns the number of topics. Therefore, the result of HDP does not change with the number of topics and it is similar to the best result of LDA.

Table 3. Document classification accuracies (%).

class	# docs	LDA	PAM
graphics	243	83.95	86.83
os	239	81.59	84.10
pc	245	83.67	88.16
mac	239	86.61	89.54
windows.x	243	88.07	92.20
total	1209	84.70	87.34

We also present the likelihood for different numbers of training documents in Figure 5. The results are all based on 160 topics except for HDP. As we can see, the performance of CTM is noticeably worse than the other three when there is limited amount of training data. One possible reason is that CTM has a large number of parameters to learn especially when the number of topics is large.

3.4. Document Classification

Another evaluation comparing PAM with LDA is document classification. We conduct a 5-way classification on the comp subset of the 20 newsgroup dataset. This contains 4836 documents with a vocabulary size of 35567 words. Each class of documents is divided into 75% training and 25% test data. We train a model for each class and calculate the likelihood for the test data. A test document is considered correctly classified if its corresponding model produces the highest likelihood. We present the classification accuracy for both PAM and LDA in Table 3. According to the sign test, the improvement of PAM over LDA is statistically significant with a p-value < 0.05.

4. Related Work

Previous work in document summarization has explored topic hierarchies built with a probabilistic language model (Lawrie et al., 2001).The dependence between a topic and its children is captured by relative entropy and encoded in a graph of conditional probabilities. Unlike PAM, which simultaneously learns all the topic correlations at different levels, this model incrementally builds the hierarchy by identifying topic terms for individual levels using a greedy approximation to the Dominating Set Problem.

Hierarchical LDA (hLDA) (Blei et al., 2004) is a variation of LDA that assumes a hierarchical structure among topics. Topics at higher levels are more general, such as stopwords, while the more specific words are organized into topics at lower levels. To generate a document, it samples a topic path from the hierarchy and then samples every word from those topics. Thus hLDA can well explain a document that discusses a

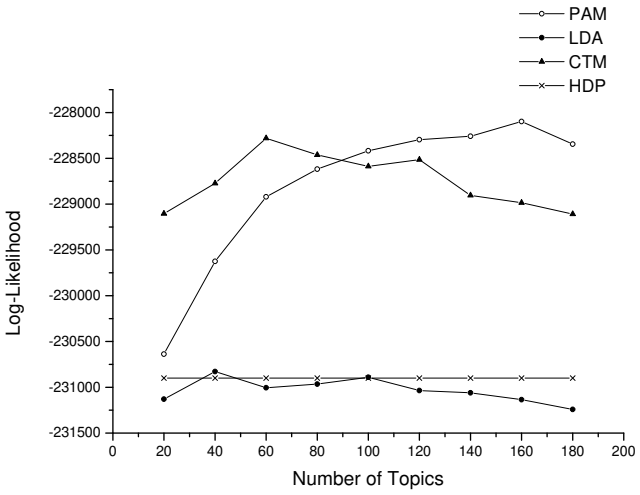

Figure 4. Likelihood comparison with different numbers of topics: the results are averages over all samples in 10 different Gibbs sampling and the maximum standard error is 113.75.

mixture of *computer science*, *artificial intelligence* and *robotics*. However, for example, the document cannot cover both *robotics* and *natural language processing* under the more general topic *artificial intelligence*. This is because a document is sampled from only one topic path in the hierarchy. Compared to hLDA, PAM provides more flexibility because it samples a topic path for each word instead of each document. Note that it is possible to create a DAG structure in PAM that would capture hierarchically nested word distributions, and obtain the advantages of both models.

Another model that captures the correlations among topics is the correlated topic model introduced by Blei and Lafferty (2006). The basic idea is to replace the Dirichlet distribution in LDA with a logistic normal distribution. Under a single Dirichlet, the topic mixture components for every document are sampled almost independently from each other. Instead, a logistic normal distribution can capture the pairwise correlations between topics based on a covariance matrix. Although CTM and PAM are both trying to model topic correlations directly, PAM takes a more flexible approach that can capture n-ary and nested correlations. In fact, CTM is very similar to a special-case structure of PAM, where we create one super-topic for every pair of sub-topics. Not only is CTM limited to pairwise correlations, it must also estimate parameters for each possible pair in the covariance matrix, which grows as the square of the number of topics. In contrast, with PAM we do not need to model every pair of topics but only sparse mixtures of correlations, as determined by the number of super-topics.

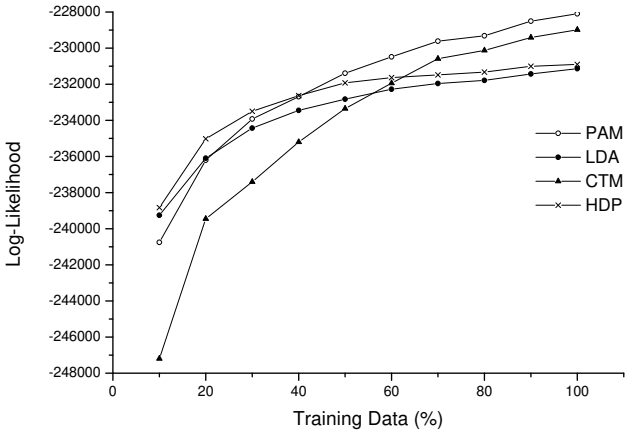

Figure 5. Likelihood comparison with different amounts of training data: the results are averages over all samples in 10 different Gibbs sampling and the maximum standard error is 171.72.

In this paper, we have described PAM operating on fixed DAG structures. It would be interesting to explore methods for learning the number of (super- and sub-) topics and their nested connectivity. This extension is closely related to hierarchical Dirichlet processes (HDP) (Teh et al., 2005), where the per-document mixture proportions over topics are generated from Dirichlet processes with an infinite number of mixture components. These models have been used to estimate the number of topics in LDA. In addition, when the data is pre-organized into nested groups, HDP can capture different topic correlations within these groups by using a nested hierarchy of Dirichlet processes. In future work, we plan to use Dirichlet processes to learn the numbers of topics at different levels, as well as their connectivity. Note that unlike HDP, PAM does not rely on pre-defined hierarchical data, but automatically discovers topic structures.

5. Conclusion

In this paper, we have presented pachinko allocation, a mixture model that uses a DAG structure to capture arbitrary topic correlations. Each leaf in the DAG is associated with a word in the vocabulary, and each interior node corresponds to a topic that models the correlation among its children, where topics can be not only parents of words, but also other topics. The DAG structure is completely general, and some topic models like LDA can be represented as special cases of PAM. Compared to other approaches that capture topic correlations such as hierarchical LDA and correlated topic model, PAM provides more expressive power to support complicated topic structures and adopts more realistic assumptions for generating documents.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval 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, and under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor. We also thank Charles Sutton and Matthew Beal for helpful discussions, David Blei and Yee Whye Teh for advice about a Dirichlet process version, Sam Roweis for discussions about alternate structure-learning methods, and Michael Jordan for help naming the model.

References

Blei, D., Griffiths, T., Jordan, M., & Tenenbaum, J. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems 16*.

Blei, D., & Lafferty, J. (2006). Correlated topic models. In *Advances in neural information processing systems 18*.

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

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*.

Diggle, P., & Gratton, R. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society*.

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

Lawrie, D., Croft, W., & Rosenberg, A. (2001). Finding topic words for hierarchical summarization. *Proceedings of SIGIR'01* (pp. 349–357).

Newton, M., & Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society*.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2005). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*.