Mixtures of Hierarchical Topics with Pachinko Allocation

David Mimno MIMNO@CS.UMASS.EDU
Wei Li WEILI@CS.UMASS.EDU
Andrew McCallum MCCALLUM@CS.UMASS.EDU
Department of Computer Science, University of Massachusetts, Amherst

Abstract

The four-level *pachinko allocation model* (PAM) (Li & McCallum, 2006) represents correlations among topics using a DAG structure. It does not, however, represent a nested hierarchy of topics, with some topical word distributions representing the vocabulary that is shared among several more specific topics. This paper presents *hierarchical PAM*—an enhancement that explicitly represents a topic hierarchy. This model can be seen as combining the advantages of hLDA's topical hierarchy representation with PAM's ability to mix multiple leaves of the topic hierarchy. Experimental results show improvements in likelihood of held-out documents, as well as mutual information between automatically-discovered topics and human-generated categories such as journals.

1. Introduction

Topic models are an important tool because of their ability to identify latent semantic components in unlabeled text data. Recently, attention has focused on models that are able not only to identify topics but also to discover the organization and cooccurrences of the topics themselves.

In this paper, we focus on discovering topics organized into hierarchies. A hierarchical topical structure is intuitively appealing. Some language is shared by large numbers of documents, while other language may be restricted to a specific subset of a corpus. Within these subsets, there may be further divisions, each with its own characteristic words. We believe that a topic model that takes such structure into account will have two primary advantages over a "flat" topic model. First, explicitly modeling the hierarchical cooccurrence patterns of topics should allow us to learn better, more predictive models. For example, knowing that hockey and baseball are both contained in a more general class "sports" should help to predict what words will be contained in previously unseen documents. Second, a hierarchical topic model should be able to describe the organization of a corpus more accurately than a topic model that does not represent such structure.

A natural representation for a hierarchical topic model is to organize topics into a tree. This approach is taken in the *hierarchical LDA* model of Blei et al. (2004). In hLDA, each document is assigned to a path through the topic tree, and each word in a given document is assigned to a topic at one of the levels of that path. A tree structured hierarchical topic model has several limitations. First, it is critically important to identify the correct tree. In order to learn the tree structure, the hLDA model uses a non-parametric nested Chinese restaurant process (NCRP) to provide a prior on tree structures. Second, it is not unusual for documents that are in clearly distinct subsets of a corpus to share a topic. For example, various topics in a professional sports sub-hierarchy and various topics in a computer games sub-hierarchy would both use similar words describing "games," "players," and "points." The only way for sports and computer gaming to share this language would be for both sub-hierarchies to descend from a common parent, which may not be the most appropriate organization for the corpus.

Another approach to representing the organization of topics is the *pachinko allocation model* (PAM) (Li & McCallum, 2006). PAM is a family of generative models in which words are generated by a directed acyclic graph (DAG) consisting of distributions over words and distributions over other nodes. A simple example of the PAM framework, *four-level PAM*, is described in Li and McCallum (2006). There is a single node

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

at the top of the DAG that defines a distribution over nodes in the second level, which we refer to as super-topics. Each node in the second level defines a distribution over all nodes in the third level, or sub-topics. Each sub-topic maps to a single distribution over the vocabulary. Only the sub-topics, therefore, actually produce words. The super-topics represent clusters of topics that frequently cooccur.

In this paper, we develop a different member of the PAM family and apply it to the task of hierarchical topic modeling. This model, *hierarchical PAM* (hPAM), includes multinomials over the vocabulary at each internal node in the DAG. This model addresses the problems outlined above: we no longer have to commit to a single hierarchy, so getting the tree structure exactly right is not as important as in hLDA. Furthermore, "methodological" topics such as one referring to "points" and "players" can be shared between segments of the corpus.

Computer Science provides a good example of the benefits of the hPAM model. Consider three subfields of Computer Science: Natural Language Processing, Machine Learning, and Computer Vision. All three can be considered part of Artificial Intelligence. Vision and NLP both use ML extensively, but all three subfields also appear independently. In order to represent ML as a single topic in a tree-structured model, NLP and Vision must both be children of ML; otherwise words about Machine Learning must be spread between an NLP topic, a Vision topic, and an ML-only topic. In contrast, hPAM allows higher-level topics to share lower-level topics. For this work we use a fixed number of topics, although it is possible to use non-parametric priors over the number of topics.

We evaluate hPAM, hLDA and LDA based on the criteria mentioned earlier. We measure the ability of a topic model to predict unseen documents based on the empirical likelihood of held-out data given simulations drawn from the generative process of each model. We measure the ability of a model to describe the hierarchical structure of a corpus by calculating the mutual information between topics and human-generated categories such as journals. We find a 1.1% increase in empirical log likelihood for hPAM over hLDA and a five-fold increase in super-topic/journal mutual information.

2. Models

2.1. LDA

The standard LDA topic model represents each document as a mixture of topics. Details of this model are discussed in Blei et al. (2003). Documents in LDA are linked only through a single non-informative Dirichlet prior. The model therefore makes no attempt to account for the distribution of topic mixtures.

2.2. hLDA

The hLDA model, which is described in Blei et al. (2004), represents the distribution of topics within documents by organizing the topics into a tree. Each document is generated by the topics along a single path of this tree. When learning the model from data, the sampler alternates between choosing a new path through the tree for each document and assigning each word in each document to a topic along the chosen path.

In hLDA, the quality of the distribution of topic mixtures depends on the quality of the topic tree. The structure of the tree is learned along with the topics themselves using a nested Chinese restaurant process (NCRP). The NCRP prior requires two parameters: the number of levels in the tree and a parameter γ. At each node, a document sampling a path chooses either one of the existing children of that node, with probability proportional to the number of other documents assigned to that child, or to a new child node, with probability proportional to γ. The value of γ can therefore be thought of as the number of "imaginary" documents in an as-yet-unsampled path.

2.3. PAM

Pachinko allocation models documents as a mixture of distributions over a single set of topics, using a directed acyclic graph to represent topic cooccurrences. Each node in the graph is a Dirichlet distribution. At the top level there is a single node. Besides the bottom level, each node represents a distribution over nodes in the next lower level. The distributions at the bottom level represent distributions over words in the vocabulary. In the simplest version, there is a single layer of Dirichlet distributions between the root node and the word distributions at the bottom level. These nodes can be thought of as "templates"—common cooccurrence patterns among topics. PAM does not represent word distributions as parents of other distributions, but it does exhibit several hierarchy-related phenomena. Specifically, trained PAM models often include what appears to be a "background" topic: a single topic with high probability in all super-topic nodes. Earlier work with four-level PAM suggests that it reaches its optimal performance at numbers of topics much larger than previously published topic models (Li & McCallum, 2006).

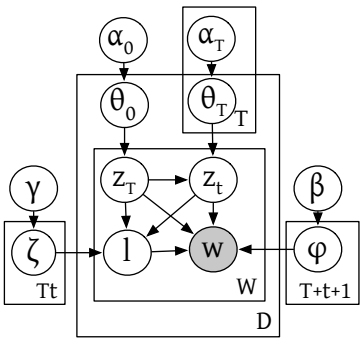


Figure 1. The graphical model representation of the hPAM model 1 generative process. Each document draws a multinomial over super-topics θ_0 and, for each super-topic, a multinomial over sub-topics θ_T. Each path through the DAG, defined by a super-topic/sub-topic pair Tt, has a multinomial ζ_{Tt} over which level of the path outputs a given word.

2.4. Hierarchical PAM

In this paper, we extend the basic PAM structure to represent hierarchical topics. We define hierarchical PAM (hPAM) as a PAM model in which every node, not just the nodes at the lowest level, is associated with a distribution over the vocabulary. This is an extremely flexible framework for hierarchical topic modeling. We present results for two variations of hPAM, but there are many other possibilities for hPAM topic models.

In the first variation, hPAM model 1, for each path through the DAG there is a distribution ζ_{Tt} on the levels of that path. These distributions are shared by all documents. The generative model is as follows.

1. For each document d, sample a distribution θ_0 over super-topics and a distribution θ_T over sub-topics for each super-topic.
2. For each word w,
   (a) Sample a super-topic z_T from θ_0.
   (b) Sample a sub-topic z_t from θ_{z_T}.
   (c) Sample a level ℓ from $\zeta_{z_T z_t}$.
   (d) Sample a word from ϕ_0 if $\ell = 1$, ϕ_{z_T} if $\ell = 2$, or ϕ_{z_t} if $\ell = 3$.

The second variation, hPAM model 2, is similar to model 1, but does not include the distributions over path levels. Instead, the Dirichlet distribution of each internal node has one extra dimension. For example, in a model with 10 super-topics and 20 sub-topics, the root node has an 11-dimensional distribution and each super-topic has a 21-dimensional distribution. This extra "exit" dimension corresponds to the event that a word is emitted directly by the internal node, without ever reaching the bottom of the DAG. The generative model is as follows.

1. For each document d, sample a distribution θ_0 over super-topics and a distribution θ_T over sub-topics for each super-topic.
2. For each word w,
   (a) Sample a super-topic z_T from θ_0. If $z_T = 0$, sample a word from ϕ_0.
   (b) Otherwise, sample a sub-topic z_t from θ_{z_T}. If $z_t = 0$, sample a word from ϕ_{z_T}.
   (c) Otherwise, sample a word from ϕ_{z_t}.

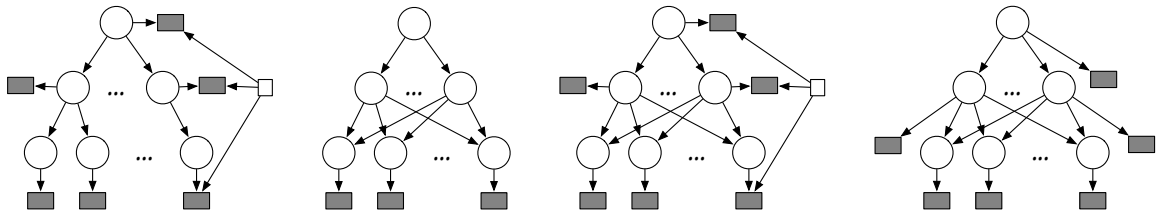
We train both models by Gibbs sampling. The multinomials θ_0, θ_T, and ζ_{Tt} can be integrated out analytically, resulting in standard Dirichlet-multinomial distributions. For each word, we sample a super-topic T, a sub-topic t and a level ℓ. The sampling distribution in hPAM model 1 for a word given the remaining topics is

$$p(z_{Ti}, z_{ti}, \ell_i | \mathbf{w}, \mathbf{z_{T\backslash i}}, \mathbf{z_{t\backslash i}}, \ell_{\backslash i}, \alpha, \beta, \gamma) \quad \propto \quad \frac{\alpha_T + N_d^T}{\sum_{T'} \alpha_{T'} + N_d} \frac{\alpha_{Tt} + N_d^{Tt}}{\sum_{t'} \alpha_{Tt'} + N_d^T} \times \frac{\gamma + N_{Tt}^{\ell}}{3\gamma + N_{Tt}} \frac{\beta_w + N_{Tt\ell}^{w}}{\sum_{w'} \beta_{w'} + N_{Tt\ell}}. \tag{1}$$

The distribution for hPAM model 2 is similar but with different values for N_d^{Tt} and without the γ term.

Learning the hPAM model with Gibbs sampling might involve substantially more computation than hLDA. Each word samples both a path through a DAG and a level, as opposed to hLDA where each document picks a single path through a tree and each word must only sample a level on that path. In fact however, it is possible to take advantage of the structure of the likelihood function of hPAM to make the sampling process more efficient. The predictive distribution in Equation 1 can be divided into two parts: the final term, which depends on the actual word in question and the first three terms, which only depend on the path through the DAG. As each word is reassigned, the factors for paths that do not involve either the old super-topic or the new super-topic do not change, and therefore do not need to be recomputed.

A more substantial increase in performance can be gained by not evaluating the complete sampling distribution, which contains $|T||t||\ell|$ elements. Rather,



(a) hLDA (b) PAM (c) hPAM 1 (d) hPAM 2

Figure 2. Generative structures in hLDA, PAM, and hPAM models 1 and 2. hLDA and hPAM include multinomials over words (represented by gray boxes) at each node, with a separate distribution over levels for each path (represented by white squares). hLDA is a tree structure: only one topic at each level is connected to a given topic at the next lower level. PAM and hPAM are directed acyclic graphs, so each node at a given level has a distribution over all nodes on the next lower level.

we marginalize over paths and levels to get a distribution over output topics, which has the same number of elements as there are nodes in the DAG: $|T| + |t| + 1$. As a result, we can sample an output topic, and then separately sample a path to that topic, that is, either a single super-topic (if we chose a distribution connected to a sub-topic), a single sub-topic (if we chose a distribution connected to a super-topic) or a super-topic and then a sub-topic (if we chose the root node's distribution). For the special case where we choose the root distribution, we also maintain a distribution over super-topics given the root topic, marginalized over sub-topics. In all cases, avoiding the calculation of a full super-topics-by-sub-topics matrix substantially improves performance.

Unlike hLDA, hPAM does not learn a tree structure of topics. Instead, it represents the hierarchical structure of topics through the Dirichlet-multinomial parameters of the internal node distributions. Training those parameters is therefore a critical part of the hPAM system. We train both the root node's distribution over super-topics and the super-topic distributions over sub-topics using Gibbs EM, as described by Wallach (2006). We allow the model to run for a number of burn-in iterations and then begin periodically taking samples of the number of words assigned to each super-topic in each document. Then we estimate the parameters using the fixed-point iteration method presented by Minka (2000). We then repeat this process.

We have observed consistent patterns in the learned Dirichlet-multinomial parameters. Parameters at the top level tend to be quite small, generally around 0.05. This corresponds to the fact that although it is possible for a document to contain words drawn from different top-level branches, documents tend to consist of predominantly one broad topic. Sub-topic distributions are mostly centered on a small number of topics, corresponding to super-topics having more predictable distributions over sub-topics.

2.5. Further hPAM models

The PAM framework is extremely flexible. As a result, it is easy to imagine other variations that would also capture hierarchical topic structures. One might consider DAGs that are not fully connected from one level to the next. In this variation, a super-topic might have its own "private" sub-topics, along with some number of shared "public" sub-topics. One could construct DAGs in which internal nodes have a set of word multinomials. For example, in hPAM model 2, it is possible to add or subtract word distributions at each internal node by increasing or decreasing the dimension of that node's Dirichlet distribution. Other DAG structures could be constructed so that every internal node is connected to all descendant nodes, so that a word could go from the root node directly to a sub-topic or through a super-topic. Another variation of hPAM model 2 might include a separate beta-binomial distribution over the decision to output a word or choose a sub-topic for each internal node.

It is also possible to construct models within the PAM family using non-parametric priors. Li et al. (2007) use hierarchical Dirichlet processes (HDP). We assume a hierarchical DAG structure for PAM and model each topic with a Dirichlet process. The Dirichlet processes at the same level are further organized into one individual HDP, which is used to estimate the number of topics at that level. As in the generative process for PAM, each word is associated with a topic path. The topic assignments provide a hierarchical grouping of data: a topic sampled from an upper-level HDP is used by lower-level HDPs to sample more topics.

3. Evaluation

We compare the performance of PAM, hLDA, and both hPAM models 1 and 2. Where possible, we also compare the performance of a flat LDA model as a baseline. In both hPAM and hLDA we have restricted consideration to three-level hierarchies of topics, although neither model is fundamentally restricted in the number of topic levels.

We evaluate each model on a corpus consisting of 5000 abstracts from 11 journals in the Medline database. Hyperparameter settings are important in all of these models. We have observed, for example, that in hLDA the parameter specifying a symmetric Dirichlet over words in topics (η in Blei et al. (2004)) has a much larger impact on the learned tree structure than the γ NCRP prior. The published value $\eta = 0.1$ is an order of magnitude greater than similar parameters in the topic modeling literature, for example $\beta = 0.01$ in Rosen-Zvi et al. (2004). We found that using the parameter 0.01 in hLDA caused the number of topics, particularly sub-topics, to become very large. hLDA with $\eta = 0.01$ finds around 15 super-topics and more than 300 subtopics, while with $\eta = 0.1$ it converges consistently to approximately 10 super-topics and 100 sub-topics. hLDA seems to require a stronger symmetric prior distribution over words in order for existing paths to be able to "compete" with empty new paths. In addition, this value affects empirical likelihood: an LDA model trained with $\beta = 0.1$ showed a 2.5% improvement over another model trained on the same corpus with $\beta = 0.01$. We use the standard parameter for PAM and LDA. For hPAM, model 1 works best with $\beta = 0.01$ and model 2 works best with 0.1. The parameter for levels of a path (α in hLDA and γ in hPAM) is set to 10.

For each Gibbs sampling run, we initialize each model randomly. For hLDA, we initialize the tree by adding documents successively, generating new paths based on the NCRP prior. We run all models for 5000 iterations.

3.1. Empirical Likelihood

A good hierarchical topic model should be able to generalize to unseen data. We follow Li and McCallum (2006) in using the empirical likelihood method of Diggle and Gratton (1984). For each generative model we draw 1000 samples using the model's generative process. In the case of hPAM, this involves sampling multinomials from the learned Dirichlet parameters over super-topics and over sub-topics for each super-topic. We then use the resulting weights on paths and the estimated ζ and ϕ distributions to construct a single multinomial over the vocabulary for each sample. These samples then represent an estimate of the model's distribution over word cooccurrences. We can easily calculate the log likelihood of each of the held-out testing documents under each sampled multinomial, providing an estimate of the log likelihood of the document under the model's distribution. Further discussion of the advantages of this method and the disadvantages of inverse likelihood calculations are in Li and McCallum (2006).

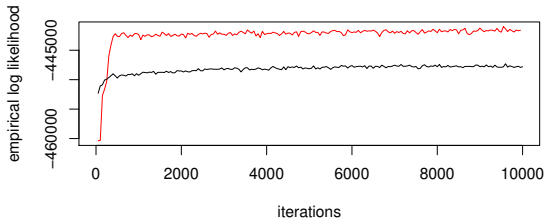


Figure 3. Empirical likelihood for hLDA (bottom line) and hPAM model 2 (top line) over 10,000 iterations. Both models converge quickly to their final range of values.

We show results for the empirical likelihood for five models, hLDA, PAM, hPAM models 1 and 2, and flat LDA. For all models except hLDA, we vary the number of topics. Results are averaged over five-fold cross validation. The hLDA model does not require a fixed number of topics, so we report a single average. We found that the number of topics selected is consistent across several runs and several folds of cross-validation. For example, the number of super-topics varied between 7 and 13, while the total number of leaf topics was between 85 and 106.

hPAM produces better empirical likelihood than PAM, hLDA and LDA. Results are shown for all five models in Figure 4 and for both hPAM models and hLDA in Tables 1 and 2. We would prefer a model that is good at predicting unseen documents with the maximum number of topics for finer granularity and better interpretability. LDA drops sharply above 20 topics. PAM achieves higher likelihood with more topics, peaking around 40 sub-topics. hPAM is more stable at larger numbers of topics, showing little decline above 60 sub-topics, and performs better than hLDA at most configurations of topics. Both PAM and hPAM model 1 perform better with more super-topics, while hPAM model 2 is relatively insensitive to the number of super-topics.

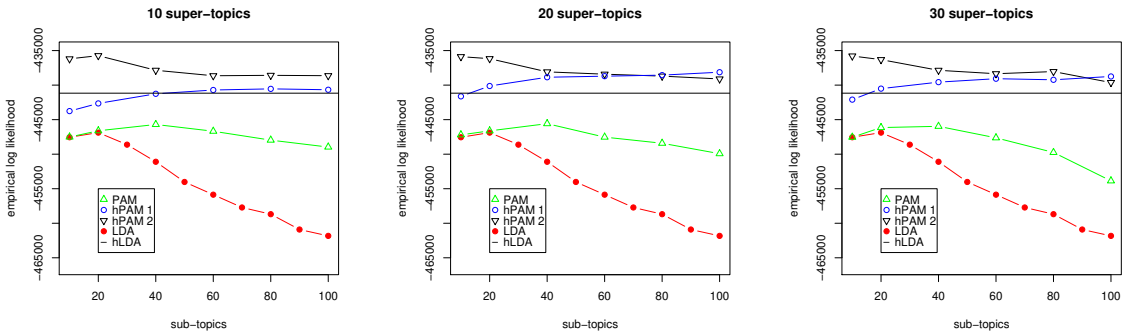


Figure 4. Empirical likelihood for PAM, hPAM models 1 and 2, hLDA and LDA on Medline data for various numbers of topics. Results for PAM and hPAM models 1 and 2 are shown for 10, 20, and 30 super-topics. LDA is shown for comparison. The horizontal line represents hLDA. Results are averaged over five-fold cross validation. The hPAM models provide the best performance.

Table 1. Model log likelihood, empirical log likelihood and topic/journal mutual information results for hPAM model 1 on 11 journals from Medline, averaged over five-fold cross validation. For the same corpus, hLDA averages an empirical log likelihood of -441167 with topic/journal mutual information of 0.26. $|T|$ and $|t|$ are the number of super-topics and sub-topics, respectively.

$\|T\|$	$\|t\|$	Mod. LL	Emp. LL	MI
10	10	-3964968	-436196	1.03
10	20	-4864974	-442650	1.23
10	40	-5094752	-441258	1.26
10	60	-5314754	-440708	1.28
10	80	-5492167	-440543	1.27
10	100	-5633893	-440672	1.25
20	10	-3939949	-435881	1.14
20	20	-4937777	-440128	1.33
20	40	-5221963	-438860	1.35
20	60	-5462150	-438707	1.36
20	80	-5683108	-438556	1.33
20	100	-5820878	-438132	1.37
30	10	-3921605	-435784	1.35
30	20	-4949808	-440510	1.35
30	40	-5284247	-439570	1.37
30	60	-5543105	-439072	1.36
30	80	-5719039	-439229	1.35
30	100	-5854591	-438748	1.36

Table 2. Model log likelihood, empirical log likelihood and topic/journal mutual information results for hPAM model 2. Model log likelihood and empirical log likelihood are better than model 1, but this variation of hPAM is less able to predict journals based on top-level topics.

$\|T\|$	$\|t\|$	Mod. LL	Emp. LL	MI
10	10	-3964968	-436196	1.03
10	20	-3923563	-435736	0.66
10	40	-3873560	-437853	0.31
10	60	-3854149	-438642	0.31
10	80	-3856613	-438586	0.24
10	100	-3859365	-438635	0.40
20	10	-3939949	-435881	1.14
20	20	-3907674	-436172	0.75
20	40	-3864218	-438082	0.19
20	60	-3843792	-438417	0.21
20	80	-3852512	-438698	0.22
20	100	-3861470	-439109	0.36
30	10	-3921605	-435784	1.35
30	20	-3886158	-436311	0.76
30	40	-3854115	-437855	0.29
30	60	-3839399	-438353	0.18
30	80	-3848539	-438038	0.16
30	100	-3864588	-438752	0.20

3.2. Predicting document labels using top-level branches

In order to evaluate the ability of each model to discover the hierarchical structure of a corpus, we focus on the top level branches of each hierarchy. A model that effectively identifies hierarchical structure should at least be able to divide a collection into its primary topical components.

We have deliberately constructed a corpus of biomedical journals with distinct topical divisions, including journals on neurology, hematology, and virology. We measure the mutual information of the journal, which is not taken into account during sampling, with the top level branches of the hLDA tree and the hPAM DAG. In other words, if a word is assigned to a given super-topic, how well can the model predict which journal it is in, and vice-versa?

In the hPAM models, every word is assigned to a particular super-topic, so we construct the joint probability of super-topics and document labels (ie journal). In hLDA, an entire document is assigned to a path, so we count all the words in a document.

A comparison of empirical likelihood and topic/journal mutual information is shown in Figure 5. We find that hPAM model 1 is consistently better able to predict which journal a word belongs to given its topic assignment than all other models. Journal/super-topic mutual information is between 1.35 and 1.54, where for hLDA it is 0.26. hPAM model 2 is closer to hLDA, but is still substantially better at its best configuration of topics. The hLDA model depends heavily on its ability to find a good tree structure of topics. hPAM models are more flexible by representing documents as mixtures of top-level branches.

Interestingly, the non-hierarchical LDA model performs similarly to hPAM with respect to mutual information. The best results for LDA are comparable to results for hPAM with small numbers of super-topics. PAM performs extremely poorly at this task, with super-topic/journal mutual information close to statistical independence.

3.3. Quality of topics

Qualitatively, the hPAM model exhibits some of the desirable properties that we identified earlier. Table 4 shows the top four sub-topics for each super-topic learned by hPAM model 2 with 10 super-topics and 20 sub-topics. The model identifies common, relatively uninformative words, which are assigned to the root topic. The super-topics are readily interpretable. Super-topic distributions over sub-topics tend to be very non-uniform, making them sparse and interpretable.

Some sub-topics are very specific to a single super-topic, such as *leukemia* while others are shared by several super-topics, such as those describing the mode of the study (*patients* vs. *rats*) and those describing a scientific focus (*gene, neurons,* and *virus, hiv*). Finally certain methodological topics are prominent in most super-topics, such as study results (*results*) and quantitative evaluation (*levels*). These topics are examples of the benefit of using a DAG structure rather than a tree: coherent clusters of methodological words can be shared between topics without being an ancestor.

We also show topics for the 20 Newsgroups corpus. In this example, three largely unrelated super-topics (Christianity, cryptography and Mideast politics) all share a sub-topic with words about discussion and argument.

4. Discussion

One of the most obvious advantages of an NCRP-based model like hLDA over the hPAM model we describe

Table 3. A topical hierarchy learned with hPAM model 2, with 10 super-topics and 20 sub-topics. The top four sub-topics are shown for each super-topic. Most documents use methodological topics such as *results, study* and *levels, increased.* Shared sub-topics distinguish types of studies (*patients* vs. *rats*) and the focus of work (*neurons, genes,* and *proteins*). Some sub-topics are predominant in only one super-topic, such as *leukemia*.

virus infection cells infected cell viral gene replication
- rna replication virus dna viral
  - results study specific studies role
  - protein proteins binding virus domain
  - gene genes expression sequence protein
- spinal nerve pain cord rats
  - rats receptor induced kg administration
  - ca neurons glutamate receptor hippocampal
  - neurons nucleus expression cells fos
- cardiac heart ventricular myocardial left
  - patients risk years clinical ci
  - disease risk ad women subjects
  - levels increased significantly compared
- cells cd cell marrow specific
  - results study specific studies role
  - leukemia cell expression aml myeloid
  - patients therapy treatment disease dose

Table 4. Sample topics from the 20 Newsgroups corpus by hPAM model 1. The topic *agree, reason, matter* contains words about argument and discussion. This topic is commonly used by several top-level topics (religion, Mideast politics, cryptography), but comprises a sufficiently coherent linguistic cluster that it is not absorbed into the root topic.

writes article don time apr
- god jesus christ people christian
  - faith wrong read spiritual passage
  - agree reason matter statement means
  - history support community house involved
- key government encryption president clipper
  - agree reason matter statement means
  - power arms president home vote
  - history support community house involved
- israel jews israeli jewish arab
  - history support community house involved
  - side left happened committee region
  - agree reason matter statement means
- turkish armenian armenians people turkey
  - side left happened committee region
  - history support community house involved
  - hundred clothes tyre bosnians origin
- file ftp windows window image
  - bit fax manager lib uk
  - site dec sources key public
  - release size function appreciated box

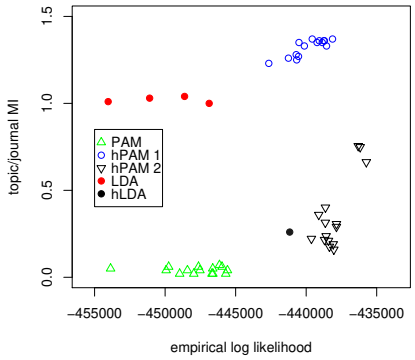


Figure 5. A comparison of the empirical log likelihood and mutual information on Medline data. Points for models other than hLDA represent different topic configurations. hPAM model 1 has the highest mutual information. hPAM model 2 has high likelihood but lower mutual information. Non-hierarchical PAM super-topics have almost no statistical dependence on journal.

here is that the number of topics does not need to be specified in advance. Although the non-parametric prior is clearly helpful, the difference is ultimately not as significant as it seems. Although some topic configurations are better than others, the range of "good" numbers of topics is quite broad.

hPAM is substantially faster than hLDA, making some amount of searching in the parameter space practical. Our implementation of hLDA is between 3–10 times slower than hPAM on a per-iteration basis. The improvement in efficiency comes from several sources. The ability to use static data structures in hPAM reduces bookkeeping overhead and pointer arithmetic. The predictive distribution of hPAM factors in a way that reduces sampling time to a sum of the number of super- and sub-topics rather than a product.

The hPAM model is reasonably robust to bad parameterizations. When hPAM is given too many topics, it has the option to simply not assign any words to a given sub-topic. Since each super-topic tends to prefer specific sub-topics, if the number of sub-topics is much larger than the number of super-topics, many sub-topics will effectively have no "parents". We have observed that when the number of topics increases, words are primarily assigned to super-topics, with only a few sub topics receiving large numbers of tokens. In this case, hPAM effectively reverts to a flat LDA model. We can thus be confident that although a badly parameterized model may not produce a very good topical hierarchy, it will not utterly fail. The distribution of words to sub-topics might also be used as a criterion for model selecting.

Finally, there is no reason that a Pachinko Allocation model could not take advantage of non-parametric priors over the number of topics. As discussed previously, it is possible to specify a prior over the DAG structure of the model using hierarchical Dirichlet processes. The results presented here suggest, however, that the non-parametric nature of hLDA does not provide much additional modeling power.

hPAM combines the hierarchical nature of hLDA with the topic mixing abilities of PAM. The resulting model is effective at discovering mixtures of topic hierarchies.

Acknowledgements

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, in part by NSF grant # CNS-0551597, 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 author(s) and do not necessarily reflect those of the sponsor.

References

Blei, D., Griffiths, T., Jordan, M., & Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. *NIPS*.

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

Diggle, P. J., & Gratton, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society B*, *46*, 193–227.

Li, W., Blei, D., & McCallum, A. (2007). Nonparametric Bayes pachinko allocation.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *ICML*.

Minka, T. (2000). Estimating a Dirichlet distribution.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *UAI*.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. *ICML*.