A Hierarchical Model of Web Summaries

Yves Petinot and **Kathleen McKeown** and **Kapil Thadani**
Department of Computer Science
Columbia University
New York, NY 10027
{ypetinot|kathy|kapil}@cs.columbia.edu

Abstract

We investigate the relevance of hierarchical topic models to represent the content of Web gists. We focus our attention on DMOZ, a popular Web directory, and propose two algorithms to infer such a model from its manually-curated hierarchy of categories. Our first approach, based on information-theoretic grounds, uses an algorithm similar to recursive feature selection. Our second approach is fully Bayesian and derived from the more general model, hierarchical LDA. We evaluate the performance of both models against a flat 1-gram baseline and show improvements in terms of perplexity over held-out data.

1 Introduction

The work presented in this paper is aimed at leveraging a manually created document ontology to model the content of an underlying document collection. While the primary usage of ontologies is as a means of organizing and navigating document collections, they can also help in inferring a significant amount of information about the documents attached to them, including path-level, statistical, representations of content, and fine-grained views on the level of specificity of the language used in those documents. Our study focuses on the ontology underlying DMOZ[1], a popular Web directory. We propose two methods for crystalizing a hierarchical topic model against its hierarchy and show that the resulting models outperform a flat unigram model in its predictive power over held-out data.

[1] http://www.dmoz.org

To construct our hierarchical topic models, we adopt the mixed membership formalism (Hofmann, 1999; Blei et al., 2010), where a document is represented as a mixture over a set of word multinomials. We consider the document hierarchy H (e.g. the DMOZ hierarchy) as a tree where internal nodes (category nodes) and leaf nodes (documents), as well as the edges connecting them, are known *a priori*. Each node N_i in H is mapped to a multinomial word distribution $Mult_{N_i}$, and each path c_d to a leaf node D is associated with a mixture over the multinonials $(Mult_{C_0} \dots Mult_{C_k}, Mult_D)$ appearing along this path. The mixture components are combined using a mixing proportion vector $(\theta_{C_0} \dots \theta_{C_k})$, so that the likelihood of string $\mathbf{w}$ being produced by path c_d is:

$$p(\mathbf{w}|\mathbf{c_d}) = \prod_{\mathbf{i=0}}^{|\mathbf{w}|} \sum_{\mathbf{j=0}}^{|\mathbf{c_d}|} \theta_{\mathbf{j}} \mathbf{p}(\mathbf{w_i}|\mathbf{c_{d,j}}) \quad (1)$$

where:

$$\sum_{j=0}^{|\mathbf{c_d}|} \theta_j = 1, \forall d \quad (2)$$

In the following, we propose two models that fit in this framework. We describe how they allow the derivation of both $p(w_i|c_{d,j})$ and θ and present early experimental results showing that explicit hierarchical information of content can indeed be used as a basis for content modeling purposes.

2 Related Work

While several efforts have focused on the DMOZ corpus, often as a reference for Web summarization

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 670–675,
Portland, Oregon, June 19-24, 2011.

tasks (Berger and Mittal, 2000; Delort et al., 2003) or Web clustering tasks (Ramage et al., 2009b), very little research has attempted to make use of its hierarchy as is. The work by Sun et al. (2005), where the DMOZ hierarchy is used as a basis for a hierarchical lexicon, is closest to ours although their contribution is not a full-fledged content model, but a selection of highly salient vocabulary for every category of the hierarchy. The problem considered in this paper is connected to the area of Topic Modeling (Blei and Lafferty, 2009) where the goal is to reduce the surface complexity of text documents by modeling them as mixtures over a finite set of topics[2]. While the inferred models are usually flat, in that no explicit relationship exists among topics, more complex, non-parametric, representations have been proposed to elicit the hierarchical structure of various datasets (Hofmann, 1999; Blei et al., 2010; Li et al., 2007). Our purpose here is more specialized and similar to that of Labeled LDA (Ramage et al., 2009a) or Fixed hLDA (Reisinger and Paşca, 2009) where the set of topics associated with a document is known *a priori*. In both cases, document labels are mapped to constraints on the set of topics on which the - otherwise unaltered - topic inference algorithm is to be applied. Lastly, while most recent developments have been based on unsupervised data, it is also worth mentioning earlier approaches like *Topic Signatures* (Lin and Hovy, 2000) where words (or phrases) characteristic of a topic are identified using a statistical test of dependence. Our first model extends this approach to the hierarchical setting, building actual topic models based on the selected vocabulary.

3 Information-Theoretic Approach

The assumption that topics are known *a-priori* allows us to extend the concept of *Topic Signatures* to a hierarchical setting. Lin and Hovy (2000) describe a *Topic Signature* as a list of words highly correlated with a target concept, and use a χ^2 estimator over labeled data to decide as to the allocation of a word to a topic. Here, the sub-categories of a node correspond to the topics. However, since the hierarchy is naturally organized in a generic-to-specific fashion, for each node we select words that have the least discriminative power between the node's children. The rationale is that, if a word can discriminate well between one child and all others, then it belongs in that child's node.

3.1 Word Assignment

The algorithm proceeds in two phases. In the first phase, the hierarchy tree is traversed in a bottom-up fashion to compile word frequency information under each node. In the second phase, the hierarchy is traversed top-down and, at each step, words get assigned to the current node based on whether they can discriminate between the current node's children. Once a word has been assigned on a given path, it can no longer be assigned to any other node on this path. Thus, within a path, a word always takes on the meaning of the one topic to which it has been assigned.

The *discriminative power* of a term with respect to node N is formalized based on one of the following measures:

Entropy of the *a posteriori* children category distribution for a given w.

$$Ent(w) = -\sum_{C \in Sub(N)} p(C|w) \log(p(C|w) \quad (3)$$

Cross-Entropy between the *a priori* children category distribution and the *a posteriori* children categories distribution conditioned on the appearance of w.

$$CrossEnt(w) = -\sum_{C \in Sub(N)} p(C) \log(p(C|w)) \quad (4)$$

χ^2 **score**, similar to Lin and Hovy (2000) but applied to classification tasks that can involve an arbitrary number of (sub-)categories. The number of degrees of freedom of the χ^2 distribution is a function of the number of children.

$$\chi^2(w) = \sum_{i \in \{w, \overline{w}\}} \sum_{C \in Sub(N)} \frac{(n_C(i) - p(C)p(i))^2}{p(C)p(i)} \quad (5)$$

To identify words exhibiting an unusually low discriminative power between the children categories, we assume a gaussian distribution of the score used and select those whose score is at least $\sigma = 2$ standard deviations away from the population mean[3].

[2] Here we use the term *topic* to describe a normalized distribution over a fixed vocabulary $\mathcal{V}$.

[3] Although this makes the decision process less arbitrary

Algorithm 1 Generative process for hLLDA

- For each topic $t \in H$
  - Draw $\beta_t = (\beta_{t,1}, \ldots, \beta_{t,V})^T \sim Dir(\cdot|\eta)$
- For each document, $d \in \{1, 2 \ldots K\}$
  - Draw a random path assignment $c_d \in H$
  - Draw a distribution over levels along c_d, $\theta_d \sim Dir(\cdot|\alpha)$
  - Draw a document length $n \sim \phi_H$
  - For each word $w_{d,i} \in \{w_{d,1}, w_{d,2}, \ldots w_{d,n}\}$,
    - Draw level $z_{d,i} \sim Mult(\theta_d)$
    - Draw word $w_{d,i} \sim Mult(\beta_{c_d}[z_{d,i}])$

3.2 Topic Definition & Mixing Proportions

Based on the final word assignments, we estimate the probability of word w_i in topic T_k, as:

$$P(w_i|T_k) = \frac{n_{C_k}(w_i)}{n_{C_k}} \tag{6}$$

with $n_{C_k}(w_i)$ the total number of occurrence of w_i in documents under C_k, and n_{C_k} the total number of words in documents under C_k.

Given the individual word assignments we evaluate the mixing proportions using corpus-level estimates, which are computed by averaging the mixing proportions of all the training documents.

4 Hierarchical Bayesian Approach

The previous approach, while attractive in its simplicity, makes a strong claim that a word can be emitted by at most one node on any given path. A more interesting model might stem from allowing soft word-topic assignments, where any topic on the document's path may emit any word in the vocabulary space.

We consider a modified version of hierarchical LDA (Blei et al., 2010), where the underlying tree structure is known *a priori* and does not have to be inferred from data. The generative story for this model, which we designate as hierarchical Labeled-LDA (hLLDA), is shown in Algorithm 1. Just as with Fixed Structure LDA[4] (Reisinger and Paşca, 2009), the topics used for inference are, for each document, those found on the path from the hierarchy root to the document itself. Once the target path $c_d \in H$ is known, the model reduces to LDA over the set of topics comprising c_d. Given that the joint distribution $p(\theta, z, w|c_d)$ is intractable (Blei et al., 2003), we use collapsed Gibbs-sampling (Griffiths and Steyvers, 2004) to obtain individual word-level assignments. The probability of assigning w_i, the i^{th} word in document d, to the j^{th} topic on path c_d, conditioned on all other word assignments, is given by:

$$p(z_i = j|\mathbf{z}_{-i}, \mathbf{w}, \mathbf{c}_d) \propto \frac{n^d_{-i,j} + \alpha}{|\mathbf{c}_d|(\alpha + 1)} \cdot \frac{n^{w_i}_{-i,j} + \eta}{V(\eta + 1)} \tag{7}$$

where $n^d_{-i,j}$ is the frequency of words from document d assigned to topic j, $n^{w_i}_{-i,j}$ is the frequency of word w_i in topic j, α and η are Dirichlet concentration parameters for the path-topic and topic-word multinomials respectively, and V is the vocabulary size. Equation 7 can be understood as defining the unormalized posterior word-level assignment distribution as the product of the current level mixing proportion θ_i and of the current estimate of the word-topic conditional probability $p(w_i|z_i)$. By repeatedly resampling from this distribution we obtain individual word assignments which in turn allow us to estimate the topic multinomials and the per-document mixing proportions. Specifically, the topic multinomials are estimated as:

$$\beta_{c_d[j],i} = p(w_i|z_{c_d[j]}) = \frac{n^{w_i}_{z_{c_d[j]}} + \eta}{\sum n^{\cdot}_{z_{c_d[j]}} + V\eta} \tag{8}$$

while the per-document mixing proportions $\theta_{\mathbf{d}}$ can be estimated as:

$$\theta_{d,j} \approx \frac{n^d_{\cdot,j} + \alpha}{n^d + |\mathbf{c_d}|\alpha}, \forall j \in 1, \ldots, \mathbf{c_d} \tag{9}$$

Although we experimented with hyper-parameter learning (Dirichlet concentration parameter η), doing so did not significantly impact the final model. The results we report are therefore based on standard values for the hyper-parameters ($\alpha = 1$ and $\eta = 0.1$).

5 Experimental Results

We compared the predictive power of our model to that of several language models. In every case, we

than with a hand-selected threshold, this raises the issue of identifying the true distribution for the estimator used.

[4]Our implementation of hLLDA was partially based on the UTML toolkit which is available at `https://github.com/joeraii/`

compute the perplexity of the model over the held-out data $\mathcal{W} = \{\mathbf{w_1} \dots \mathbf{w_n}\}$ given the model $\mathcal{M}$ and the observed (training) data, namely:

$$perpl_{\mathcal{M}}(\mathcal{W}) = \exp(-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\mathbf{w_i}|}\sum_{j=1}^{|\mathbf{w_i}|}\log p_{\mathcal{M}}(w_{i,j})) \tag{10}$$

5.1 Data Preprocessing

Our experiments focused on the English portion of the DMOZ dataset[5] (about 2.1 million entries). The raw dataset was randomized and divided according to a 98% training (31M words), 1% development (320k words), 1% testing (320k words) split. Gists were tokenized using simple tokenization rules, with no stemming, and were case-normalized. Akin to Berger and Mittal (2000) we mapped numerical tokens to the *NUM* placeholder and selected the $V = 65535$ most frequent words as our vocabulary. Any token outside of this set was mapped to the *OOV* token. We did not perform any stop-word filtering.

5.2 Reference Models

Our reference models consists of several n-gram ($n \in [1, 3]$) language models, none of which makes use of the hierarchical information available from the corpus. Under these models, the probability of a given string is given by:

$$p(\mathbf{w}) = \prod_{\mathbf{i=1}}^{|\mathbf{s}|}\mathbf{p}(\mathbf{w_i}|\mathbf{w_{i-1}}, \dots, \mathbf{w_{i-(n-1)}}) \tag{11}$$

We used the SRILM toolkit (Stolcke, 2002), enabling Kneser-Ney smoothing with default parameters.

Note that an interesting model to include here would have been one that jointly infers a hierarchy of topics as well as the topics that comprise it, much like the regular hierarchical LDA algorithm (Blei et al., 2010). While we did not perform this experiment as part of this work, this is definitely an avenue for future work. We are especially interested in seeing whether an automatically inferred hierarchy of topics would fundamentally differ from the manually-curated hierarchy used by DMOZ.

[5]We discarded the *Top/World* portion of the hierarchy.

5.3 Experimental Results

The perplexities obtained for the hierarchical and n-gram models are reported in Table 1.

	$\overline{\text{reg}}$	all
# documents	1153000	2083949
avg. gist length	15.47	15.36
1-gram	1644.10	1414.98
2-gram	352.10	287.09
3-gram	239.08	179.71
entropy	812.91	1037.70
cross-entropy	1167.07	1869.90
χ^2	1639.29	1693.76
hLLDA	941.16	983.77

Table 1: Perplexity of the hierarchical models and the reference n-gram models over the entire DMOZ dataset (all), and the non-Regional portion of the dataset ($\overline{\text{reg}}$).

When taken on the entire hierarchy (all), the performance of the Bayesian and entropy-based models significantly exceeds that of the 1-gram model (significant under paired t-test, both with p-value $< 2.2 \cdot 10^{-16}$) while remaining well below that of either the 2 or 3 gram models. This suggests that, although the hierarchy plays a key role in the appearance of content in DMOZ gists, word context is also a key factor that needs to be taken into account: the two families of models we propose are based on the bag-of-word assumption and, by design, assume that words are drawn *i.i.d.* from an underlying distribution. While it is not clear how one could extend the information-theoretic models to include such context, we are currently investigating enhancements to the hLLDA model along the lines of the approach proposed in Wallach (2006).

A second area of analysis is to compare the performance of the various models on the entire hierarchy versus on the non-Regional portion of the tree ($\overline{reg}$). We can see that the perplexity of the proposed models decreases while that of the flat n-grams models increase. Since the non-Regional portion of the DMOZ hierarchy is organized more consistently in a semantic fashion[6], we believe this reflects the ability of the hierarchical models to take advantage of

[6]The specificity of the Regional sub-tree has also been discussed by previous work (Ramage et al., 2009b), justifying a special treatment for that part of the DMOZ dataset.

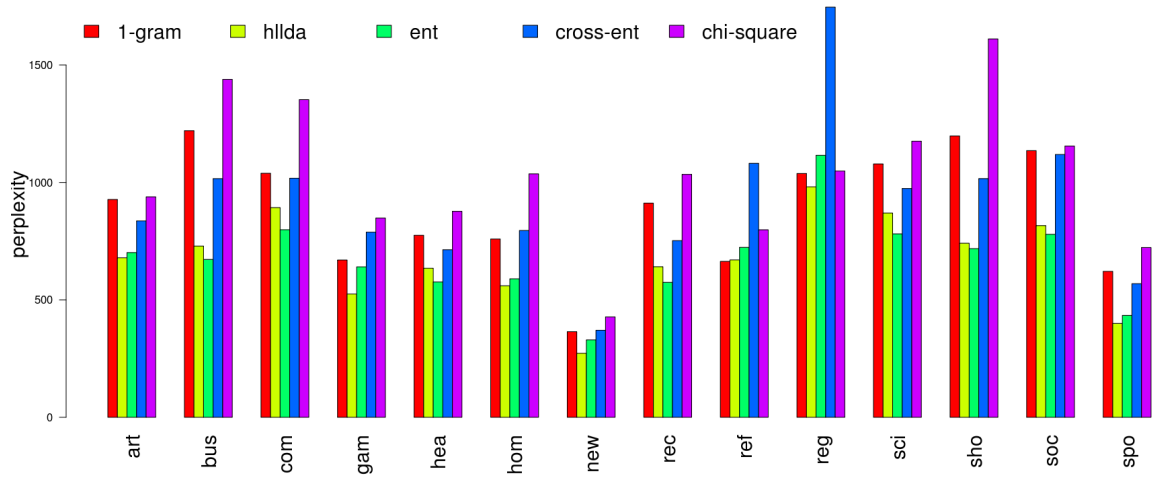

Figure 1: Perplexity of the proposed algorithms against the 1-gram baseline for each of the 14 top level DMOZ categories: Arts, Business, Computer, Games, Health, Home, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports.

the corpus structure to represent the content of the summaries. On the other hand, the Regional portion of the dataset seems to contribute a significant amount of noise to the hierarchy, leading to a loss in performance for those models.

We can observe that while hLLDA outperforms all information-theoretical models when applied to the entire DMOZ corpus, it falls behind the entropy-based model when restricted to the non-regional section of the corpus. Also if the reduction in perplexity remains limited for the entropy, χ^2 and hLLDA models, the cross-entropy based model incurs a more significant boost in performance when applied to the more semantically-organized portion of the corpus. The reason behind such disparity in behavior is not clear and we plan on investigating this issue as part of our future work.

Further analyzing the impact of the respective DMOZ sub-sections, we show in Figure 1 results for the hierarchical and 1-gram models when trained and tested over the 14 main sub-trees of the hierarchy. Our intuition is that differences in the organization of those sub-trees might affect the predictive power of the various models. Looking at sub-trees we can see that the trend is the same for most of them, with the best level of perplexity being achieved by the hierarchical Bayesian model, closely followed by the information-theoretical model using entropy as its selection criterion.

6 Conclusion

In this paper we have demonstrated the creation of a topic-model of Web summaries using the hierarchy of a popular Web directory. This hierarchy provides a backbone around which we crystalize hierarchical topic models. Individual topics exhibit increasing specificity as one goes down a path in the tree. While we focused on Web summaries, this model can be readily adapted to any Web-related content that can be seen as a mixture of the component topics appearing along a paths in the hierarchy. Such model can become a key resource for the fine-grained distinction between generic and specific elements of language in a large, heterogenous corpus.

Acknowledgments

This material is based on research supported in part by the U.S. National Science Foundation (NSF) under IIS-05-34871. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

A. Berger and V. Mittal. 2000. Ocelot: a system for summarizing web pages. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00)*, pages 144–151.

David M. Blei and J. Lafferty. 2009. Topic models. In *A. Srivastava and M. Sahami, editors, Text Mining: Theory and Applications. Taylor and Francis*.

David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *JMLR*, 3:993–1022.

David M. Blei, Thomas L. Griffiths, and Micheal I. Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. In *Journal of ACM*, volume 57.

Jean-Yves Delort, Bernadette Bouchon-Meunier, and Maria Rifqi. 2003. Enhanced web document summarization using hyperlinks. In *Hypertext 2003*, pages 208–215.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235.

Thomas Hofmann. 1999. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In *Proceedings of IJCAI'99*.

Wei Li, David Blei, and Andrew McCallum. 2007. Nonparametric bayes pachinko allocation. In *Proceedings of the Proceedings of the Twenty-Third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, pages 243–250, Corvallis, Oregon. AUAI Press.

C.-Y. Lin and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009a. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), Singapore*, pages 248–256.

Daniel Ramage, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. 2009b. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 54–63, New York, NY, USA. ACM.

Joseph Reisinger and Marius Paşca. 2009. Latent variable models of concept-attribute attachment. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 620–628, Morristown, NJ, USA. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing, vol. 2*, pages 901–904, September.

Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. 2005. Web-page summarization using clickthrough data. In *SIGIR 2005*, pages 194–201.

Hanna M. Wallach. 2006. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, U.S.*, pages 977–984.