Journal of Machine Learning Research 12 (2011) 2749-2775 Submitted 10/09; Revised 6/11; Published 10/11

Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes

Elias Zavitsanos[*] IZAVITS@IIT.DEMOKRITOS.GR
Georgios Paliouras PALIOURG@IIT.DEMOKRITOS.GR
Institute of Informatics and Telecommunications
NCSR "Demokritos"
Patriarhou Gregoriou and Neapoleos Street
15310 Athens, Greece

George A. Vouros GEORGEV@AEGEAN.GR
Department of Information and Communication Systems Engineering
University of the Aegean
Karlovassi, 83200 Samos Island, Greece

Editor: David Blei

Abstract

This paper presents hHDP, a hierarchical algorithm for representing a document collection as a hierarchy of latent topics, based on Dirichlet process priors. The hierarchical nature of the algorithm refers to the Bayesian hierarchy that it comprises, as well as to the hierarchy of the latent topics. hHDP relies on nonparametric Bayesian priors and it is able to infer a hierarchy of topics, without making any assumption about the depth of the learned hierarchy and the branching factor at each level. We evaluate the proposed method on real-world data sets in document modeling, as well as in ontology learning, and provide qualitative and quantitative evaluation results, showing that the model is robust, it models accurately the training data set and is able to generalize on held-out data.

Keywords: hierarchical Dirichlet processes, probabilistic topic models, topic distributions, ontology learning from text, topic hierarchy

1. Introduction

In this paper we address the problem of modeling the content of a given document collection as a hierarchy of latent topics given no prior knowledge. These topics represent and capture facets of content meaning, by means of multinomial probability distributions over the words of the term space of the documents. The assignment of documents to latent topics without any preclassification is a powerful text mining technique, useful among others for ontology learning from text and document indexing.

In the context of this modeling problem, probabilistic topic models (PTMs) have attracted much attention. While techniques for terminology extraction and concept identification from text rely on the identification of representative terms using various frequency measures, such as the TF/IDF (Salton and McGill, 1986) or C/NC value (Frantzi et al., 2000), PTMs aim to discover topics that are soft clusters of words, by transforming the original term space into a latent one of meaningful features (topics).

∗. Also in the Department of Information and Communication Systems Engineering, University of the Aegean, Greece.

Much work on PTMs focuses on a flat clustering of the term space into topics, while the creation of a hierarchical structure of topics without user involvement or pre-defined parameters still remains a challenging task. The goal of discovering a topic hierarchy that comprises levels of topic abstractions is different from conventional hierarchical clustering. The internal nodes of this type of hierarchy reflect the topics, which correspond to the shared terminology or vocabulary between documents. In contrast, hierarchical clustering usually groups data points, for instance documents, resulting in internal nodes that constitute cluster summaries. Conventional techniques such as agglomerative clustering, allow objects to be grouped together based on a similarity measure, but the hierarchy is generally the result of hard clustering. This form of clustering limits the applicability of the techniques, since a document is assigned to only one topic and may not be retrieved upon a search on a related topic.

Furthermore, conventional clustering models texts based explicitly on syntax. It tends to cluster words that appear in similar local contexts. On the other hand, topic models attempt to capture through syntax, latent semantics. They cluster words that appear in a similar global context, in the sense that they try to generalize beyond their place of appearance in a text collection, in order to reflect their intended meaning.

The role of hierarchical topic models regarding text modeling and natural language processing (NLP) is very important. The hierarchical modeling of topics allows the construction of more accurate and predictive models than the ones constructed by flat models. Models of the former type are more probable to predict unseen documents, than the latter. In most text collections, such as web pages, a hierarchical model, for instance a web directory, is able to describe the structure and organization of the document collection more accurately than flat models. This, however, ultimately depends on the nature of the data set and the true generative process of the documents themselves. Assuming that the higher levels of the hierarchy capture generic topics of a particular domain, while lower-level ones focus on particular aspects of that domain, it is expected that a hierarchical probabilistic topic model would be able to "explain" or could have generated the data set. In other words, the likelihood of such a model given the data set would probably be higher than the likelihood of other flat models (Mimno et al., 2007; Li et al., 2007; Li and McCallum, 2006).

Despite recent activity in the field of HPTMs, determining the hierarchical model that best fits a given data set, in terms of the structure and size of the learned hierarchy, still remains a challenging task and an open issue. In this paper, we propose a method that deals with some of the limitations of the current models, regarding the representation of input data as latent topics. In particular, we aim to infer a hierarchy of topics and subtopics, such that each topic is more general than its subtopics, in the sense that if a document can be indexed by any of the subtopics it should also be indexed by the topic itself. Moreover, we demand to infer the hierarchy without making any assumption either about the number of topics at any level of the hierarchy, or about the height of the hierarchy. The proposed method, given a collection of text documents, produces a hierarchical representation in the form of a topic hierarchy, adopting a nonparametric Bayesian approach. The resulting hierarchy specifies each topic as a multinomial probability distribution over the vocabulary of the documents. Moreover, internal nodes are also represented as multinomial probability distributions over the subtopics of the hierarchy. In addition to the basic model, we also present a variant that produces a topic hierarchy, by modeling the vocabulary only at the leaf level and considering topics in the inner levels to be multinomial distributions over subtopics. Although the evaluation of such models is also an open issue, we demonstrate the effectiveness of the model in different tasks through an extensive evaluation, providing qualitative and quantitative results.

In what follows, we start by a quick review of the family of probabilistic topic models and hierarchical models (Section 2). Section 3 presents the proposed method, namely topic hierarchies of hierarchical Dirichlet processes (hHDP), along with its variant. Section 4 provides an extensive evaluation of hHDP, including comparisons to other models and applications to different tasks, while Section 5 summarizes the paper and presents future directions.

2. Hierarchical Probabilistic Topic Models

Probabilistic topic models (PTMs) (Griffiths and Steyvers, 2002) are generative models for documents. Documents are assumed to be mixtures of topics and topics are probability distributions over the words of some vocabulary. The vocabulary may comprise all the words that appear in the documents or a part of them, for example excluding the stop-words. PTMs are based on the De Finetti theorem (Finetti, 1931), which states that an exchangeable sequence of random variables is a mixture of independent and identically distributed random variables. In the case of text data, PTMs treat documents as "bag-of-words." The words in the documents are infinitely exchangeable without loss of meaning, and thus, the joint probability underlying the data is invariant to permutation. Based on this assumption of exchangeability, the meaning of documents does not depend on the specific sequence of the words, that is, the syntax, but rather on their "ability" to express specific topics either in isolation or in mixture. Given the latent variables, (the topics), the words are assumed to be conditionally independent and identically distributed in the texts.

Figure 1 represents the underlying idea of the generative nature of PTMs. Topics, represented as clouds, are probability distributions over words (puzzle pieces) of a predefined vocabulary. According to the mixture weights that reflect the probability of a topic to participate in a document, words are sampled from the corresponding topics, in order for documents to be generated.

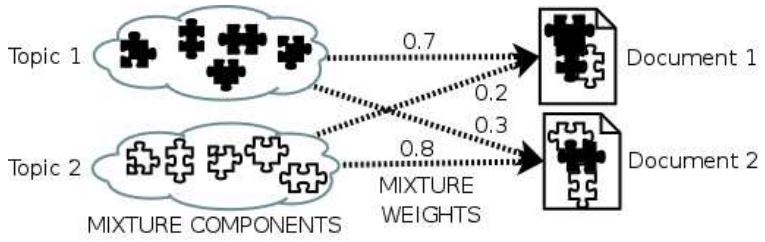

Figure 1: The generative nature of PTMs: Documents are mixtures of topics. Topics are probability distributions over words (puzzle pieces). The probability of participation of a topic in a document is defined by the mixture weights. Inspired by Steyvers and Griffiths (2007).

In the rest of the paper, we will refer to the document collection as D, consisting of d_1, d_2, ..., d_N documents. The set of the latent topics will be defined as T, consisting of t_1, t_2, ..., t_K topics. We will refer to the distribution of topics as θ_K, indicating the dimensionality K of the distribution, and finally, ϕ_V will stand for the distribution of the words of the vocabulary V.

Following the principles of PTMs, the generative model of probabilistic latent semantic analysis (PLSA) (Hofmann, 2001) specifies a simple generative rule for the words in a document d_i, according to which, each word of a training document d_i comes from a randomly chosen topic t_i. The topics are drawn from a document-specific distribution over topics θ_K, and there exists one such

distribution for each d_i. Hence, the set of the training documents D defines an empirical distribution over topics. In PLSA, the observed variable d_i is actually an index into the training set D, and thus, there is no natural way for the model to handle previously unseen documents, except through marginalization (Blei et al., 2003).

The model of PLSA has been extended by latent Dirichlet allocation (LDA) (Blei et al., 2003). The generative model of LDA, being a probabilistic model of a corpus, represents each d_i as random mixture over latent topics T. The mixture indicator is selected once per term, rather than once per document as in PLSA. The estimated topics are represented as multinomial probability distributions over the terms of the documents, while each d_i is represented as a Dirichlet random variable θ, the dimensionality of which, is predefined and equal to the number of estimated latent topics. In contrast to PLSA, LDA states that each word of both the observed and unseen documents is generated by a randomly chosen topic, which is drawn from a distribution with a randomly chosen parameter. This parameter is sampled once per document from a smooth distribution over topics.

A question that usually arises when using models like LDA is how many topics the estimated model should have, given the document collection. The problem is harder when multiple parameters are shared among documents, as in LDA. The problem is addressed by sharing a discrete base distribution among documents. A hierarchical Dirichlet process (HDP) creates such a discrete base distribution for the document Dirichlet processes (DPs) by sampling from another DP. In such a Bayesian hierarchy, the root DP uses the Dirichlet distribution of the topics as a base distribution and each document samples from it.

Although LDA is a true generative probabilistic model for documents and HDP is a convenient mechanism for inferring the number of topics, relations of any type or correlations between the estimated topics are not taken into account. In fact, a flat and soft clustering of the term space of the documents into topics is provided. Thus, there is a need for hierarchical models that are able to capture relations between the latent topics in order to represent common shared structure, as explained in Section 1.

A method for producing a tree-like structure of latent topics is presented in Gaussier et al. (2002), as an extension of the PLSA model. According to hierarchical probabilistic latent semantic analysis (HPLSA), the data set D is assumed to have been generated by a hierarchical model. For each d_i, a document class is picked from a predefined number of classes, with some probability. Then, a d_i is chosen based on the conditional probability of a document given the class. Again, given the class, a topic t_i is sampled for that d_i. Finally, a word is generated given the sampled topic t_i. A class here represents a group of documents sharing some common thematic feature. According to this model, documents and words are conditionally independent given the class. In a typical hierarchy, documents are assigned to classes at the leaves of the hierarchy, while words are sampled from topics which occupy non-leaf nodes of the hierarchy. The number of classes actually defines the number of leaves of the hierarchy. The model extends PLSA in the sense that if one topic per class is sampled, then the result is the flat clustering of PLSA. If on the other hand, a single topic is sampled for more than one class, then it is placed on a higher level and represents shared knowledge between these classes. However, the model inherits known problems of PLSA, such as the large number of parameters that need to be estimated, which grow linearly with the size of the corpus, a problem that LDA seems to deal with, since the latter treats the distribution θ_K as a hidden random variable, rather than a large set of individual parameters which are explicitly linked to the training set.

Another approach to capturing relations between topics is the correlated topic models (CTM) (Blei and Lafferty, 2006), an extension of LDA. The generative process of this model is identical to that of LDA, with the exception that the topic proportions are drawn from a logistic normal distribution, rather than a Dirichlet as in the case of LDA. The parameters of this distribution include a covariance matrix, the entries of which specify the correlations between pairs of topics. Correlations are introduced by topics that appear in the same context, in the sense that they appear together in documents (or parts of documents). The advantage of this model is that the covariance matrix may include positive covariance between two topics that co-occur frequently and negative between two topics that co-occur rarely, while with the Dirichlet approach, we actually express the expectation of each topic to occur, according to the weights of the mixture proportions, and how much we expect any given document to follow these proportions. In CTM only pairwise correlations between topics are modeled. Hence, the number of parameters grows as the square of the number of topics.

The Pachinko allocation model (PAM) (Li and McCallum, 2006) deals with some of the problems of CTM. PAM uses a directed acyclic graph (DAG) structure to represent and learn arbitrary, nested and possibly sparse topic correlations. PAM connects the words of the vocabulary V and topics T on a DAG, where topics occupy the interior nodes and the leaves are words. Each topic t_i is associated with a Dirichlet distribution of dimension equal to the number of children of that topic. The four-level PAM, which is presented in Li and McCallum (2006), is able to model a text collection through a three-level hierarchy of topics with arbitrary connections between them. However, PAM is unable to represent word distributions as parents of other word distributions and also requires the length of the path from the root node to the leaves to be predefined.

The hierarchical latent Dirichlet allocation (hLDA) model (Blei et al., 2004) was the first attempt to represent the distribution of topics as a tree-structure by providing at the same time uncertainty over the branching factor at each level of the tree. In hLDA, each document is modeled as a mixture of L topics defined by θ_L proportions along a path from the root topic to a leaf. Therefore, each document d_i is generated by the topics along a single path of this tree. Hence, each d_i is about a specific topic (a leaf topic) and its abstractions along the path to the root. Multiple inheritance, in the sense of assigning more than one topic to a super-topic, is not modeled. When estimating the model from data, for each d_i, the sampler chooses an existing or a new path through the tree and assigns each word to a topic along the chosen path. Thus, both internal and leaf topics generate words for new documents. In order to learn the structure of the tree, a nested Chinese restaurant process (nCRP) is used as a prior distribution. Assuming that the depth (L) of the hierarchy is provided a priori, the nCRP prior actually controls the branching factor at each level of the hierarchy. It expresses the uncertainty about possible L-level trees and thus, the problem of modeling the corpus is reduced to finding a good, in the sense of maximum likelihood, L-level tree among them.

Aiming to support multiple inheritance between topics, and extending PAM to express word distributions as parents of other word distributions, the work in Mimno et al. (2007) presents the hierarchical Pachinko allocation model (HPAM), in which every node is associated with a distribution over the vocabulary of the text collection. There are actually two variants of the model. In the first variant, each path through the DAG is associated with a multinomial distribution on the levels of the path, which is shared by all documents. In the second one, this distribution does not exist, but the Dirichlet distribution of each internal node has one extra "exit" dimension, which corresponds to the event that a word is produced directly by the internal node, without reaching the leaf topics of the DAG. The three-level model that is presented in Mimno et al. (2007) comprises a root topic, a level of super-topics and a level of sub-topics and it uses $(T+1)$ Dirichlet distributions to model

the text collection. One distribution incorporates a hyper-parameter α_0 and serves as a prior over the super-topics. The remaining T distributions incorporate a hyper-parameter α_T, which serves as a prior over their sub-topics. The difference between the priors α_0 and α_T is that they produce different distributions θ_0 and θ_T over super-topics and subtopics respectively.

While the models belonging in the PAM family provide a powerful means to describe inter-topic correlations, they have the same practical difficulty as many other topic models in determining the number of topics at the internal levels. For this purpose, a non-parametric Bayes version of PAM has been presented in Li et al. (2007). This model is actually a combination of the hLDA model, in the sense of determining the number of topics T at the internal levels, and of the four-level PAM (Li and McCallum, 2006). Each topic t_i is modeled by a Dirichlet process and the Dirichlet processes at each level are further organized into a hierarchical Dirichlet process (HDP), which is used to estimate the number of topics at this level. Apart from this, the model follows the basic PAM principles. During the generation of a document, after sampling the multinomial distributions over topics from the corresponding HDPs, a topic path is sampled repeatedly according to the multinomials for each word in the document d_i. The resulting hierarchy is limited to three levels and comprises the root topic, the next level of super-topics and the final level of sub-topics, which are the ones that are able to generate words.

Representing all topics as multinomial distributions over words is more appealing, than representing only the leaf topics. For this purpose, the work in Zavitsanos et al. (2008) uses the LDA model iteratively to produce layers of topics and then establishes hierarchical relations between them, based on conditional independences, given candidate parent topics. The branching factor at each level is decided by the number of discovered relations, since topics that are not connected to others are disregarded. The issue of the depth of the hierarchy is addressed in that work by measuring the similarity of the newly generated topics to the existing ones. However, the number of the generated topics at each level is predefined.

In summary, some topic models support a latent hierarchy of topics, but allow the generation of words only at the leaf level. Others are able to generate words at each level, but depend on a predefined depth of the hierarchy. In particular, hLDA is able to infer the branching factor at each level, but still requires the depth of the hierarchy to be known a priori. In addition, in contrast to the simple LDA, in the case of hLDA, documents can only access the topics that lie across a single path in the learned tree. Hence, LDA, which places no such restrictions in the mixture of topics for each document, can be significantly more flexible than hLDA. The models belonging in the PAM family seem to be able to address these issues, especially the non-parametric Bayes version of PAM (Li et al., 2007) that exploits some of the advantages of hLDA. However, the fact that the resulting hierarchy comprises three levels and produces words only at the leaves is limiting. It seems possible to extend the hierarchy to more levels, but this would require the depth to be known a priori and would impose an increase on the number of parameters to be estimated. Finally, parameters such as the number of topics or the number of levels need to be estimated using cross-validation, which is not efficient even for non-hierarchical topic models like LDA. Table 1 summarizes the properties of the aforementioned models.

The evaluation of topic models is also an open issue. The majority of the work reviewed in this section assesses the inferred hierarchy on the basis of how "meaningful" the latent topics are to humans. In this spirit, new evaluation measures (Chang et al., 2009) have been proposed that try to capture aspects of how humans evaluate topic models and especially the inferred hierarchy. Thus,

Model	Topic hierarchy	Infer number of topics	Infer number of levels	Multiple inheritance	Generate words at all nodes
PLSA	×	×	×	×	✓
LDA	×	×	×	×	✓
HDP	×	✓	×	×	✓
HPLSA	✓	×	×	×	✓
CTM	×	×	×	×	✓
PAM	✓	×	×	✓	×
hLDA	✓	✓	×	×	✓
HPAM	✓	×	×	✓	✓
NPPAM	✓	✓	×	✓	×

Table 1: Comparison of topic models. The first column is the acronym of the model. The second column shows whether the model is able to organize the topics hierarchically. The third and fourth columns depict the ability of the model to infer the number of topics and levels respectively. The last two columns indicate whether the model's topics share subtopics, and whether the model produces words at all nodes.

the emphasis is on how topic models infer the latent structure of the input documents, rather than on how well they generate documents. Based on this observation, we propose an algorithm that:

- Determines the depth of the learned hierarchy.
- Infers the number of topics at each level of the hierarchy.
- Allows sharing of topics among different documents.
- Allows topics to share subtopics.
- Allows a topic at any level of the hierarchy to be specified as a distribution over terms.
- Has a non-parametric Bayesian nature and thus exhibits all the advantages of such techniques.

In addition, we present a variant that models only the leaf levels as probability distributions over words and results in a hierarchical topic clustering of the text collection. The basis for the methods proposed in this paper is the model of a hierarchical Dirichlet process (HDP).

3. Topic Hierarchies of Hierarchical Dirichlet Processes (hHDP)

In this section we present the hHDP method in two variants. The first variant results in a hierarchy whose internal nodes are represented as probability distributions over topics and over words. Thus it performs a hierarchical vocabulary clustering (hvc). The second variant provides a hierarchical topic clustering (htc) of the corpus, where only leaf nodes are represented as distributions over words. We will refer to the first variant as hvHDP, and to the second as htHDP. We divide the section into two subsections, providing insights about the proposed method and information regarding the sampling scheme.

3.1 Stacking HDPs

Starting with the criteria that we posed at the end of Section 2, we want to be able to infer the number of topics at each level. For this purpose we use the mixture model of hierarchical Dirichlet processes (HDP) (Teh et al., 2006), which is illustrated in Figure 2.

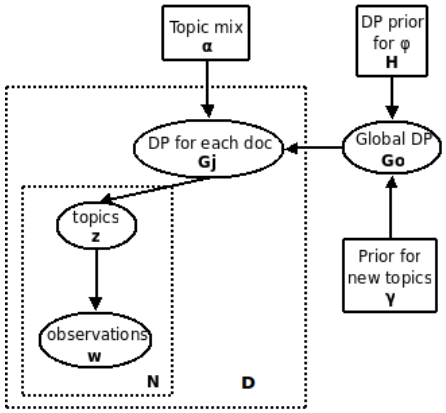

Figure 2: The HDP mixture model. Assuming a text collection of M documents, each of length N, there is a DP G_j for each document to draw word distributions. There is a global, higher-level DP (G_0) that maintains the global distribution of word distributions.

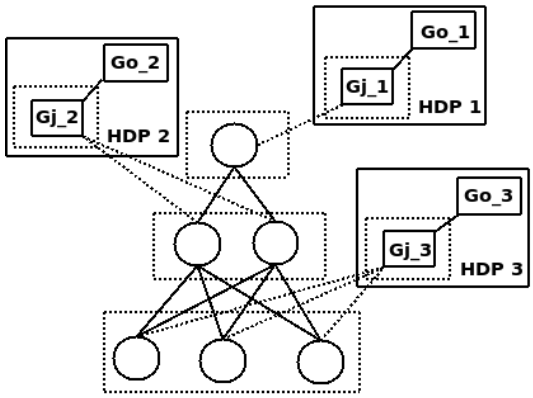

Figure 3: The association of the HDPs with the topic hierarchy. There is an HDP associated with each level. There are as many DPs (G_j) as the documents at each level, connected to all topics of the level. Each level also comprises a global DP (G_0) that is connected to all the G_j in this level.

In the proposed method (Figure 3), at each level of the hierarchy, there is a DP (G_j) for each document and a "global" DP (G_0) over all the DPs at that level. Therefore, each level of the topic hierarchy is associated with a HDP. An important characteristic of this approach is that the number of topics of each level is automatically inferred, due to the non-parametric Bayesian nature of

the HDP. In addition, it allows the topics at each level to be shared among the documents of the collection. Figure 3 depicts the DPs associated with different levels of the topic hierarchy.

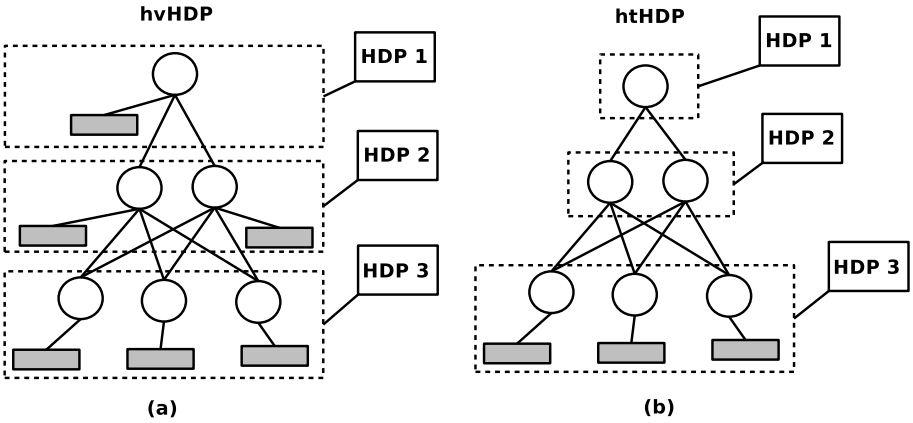

Figure 4: (a) hvHDP. (b) htHDP. Topics are represented as circles, while word distributions as gray boxes. hvHDP consists of topics that are both distributions over subtopics and over words. htHDP represents only leaf topics as distributions over words.

Therefore, at each level, a HDP is assumed, according to Figure 4, which is modeled as shown in Figure 3. The HDP at each level is used to express uncertainty about the possible number of mixture components, that is, the latent topics.

Among the models mentioned in Section 2, hPAM and hLDA are the closest "relatives" of hvHDP in terms of the representation of the corpus through an inferred hierarchy. They both have internal nodes containing words. However, in hLDA a topic is not allowed to have more than one parent, while in hPAM and hHDP this is allowed. On the other hand, while hPAM needs the number of internal topics to be fixed a priori, hLDA and hHDP are able to infer the number of topics at each level of the hierarchy, due to their non-parametric Bayesian nature. Moreover, while the model of hLDA requires that each document is made of topics across a specific path of the hierarchy, hPAM and hHDP provide much more flexibility, since topics can be shared among super-topics. Overall, hHDP combines the strengths of hPAM and hLDA, extending also the non-parametric approach to include the estimation of the depth of the learned hierarchy, which is further explained in the following paragraphs.

The PAM and the non-parametric PAM models are similar to the second version of hHDP (htHDP). The topics of the PAM models generate words at the leaf level and the models are based on a fixed three-level hierarchy. The simple PAM model needs the number of internal topics to be known a priori, while its non-parametric version uses the CRP to decide the number of super-topics and sub-topics. The obvious advantage of htHDP is its full non-parametric nature that does not impose restrictions on the depth and the branching factor at each level of the hierarchy.

3.2 Estimation of the Hierarchy

Regarding the estimation of the latent structure, exact inference of the hierarchy given a document collection is intractable. For this purpose we use Gibbs sampling, which climbs stochastically the

posterior distribution surface to find an area of high posterior probability and explores its curvature (Andrieu et al., 2003). Although the method of Gibbs sampling lacks theoretical guarantees, it has been proven to be appropriate for this type of data analysis and for inferring latent variables given the distribution of the observations and the corresponding priors. More information about sampling methods in machine learning can be found in Andrieu et al. (2003).

The sampling scheme of hHDP estimates both the number of topics at each level and the number of levels of the learned hierarchy. As shown in Figure 5, starting at the leaf level, we use HDP to infer the number of leaf topics as if no hierarchy is to be built. We then build the hierarchy bottom-up until reaching a level with a single node (the root topic). Each level is modeled as a HDP, estimating the appropriate number of topics.

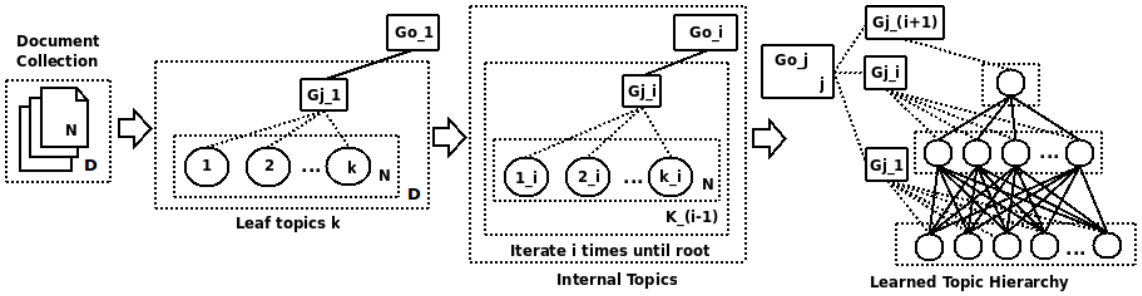

Figure 5: Bottom-up probabilistic estimation of the topic hierarchy: Starting with a corpus of M documents, the leaf topics are inferred first. The word distributions for each leaf topic make up the observations ("documents") for the estimation of the next level up. The procedure is repeated until the root topic is inferred.

Figure 5 presents the steps of the sampling scheme. We start with the text collection, which provides the observations, that is, the words, for the estimation. The words constitute the term space. At the first step that infers the leaf level, in a Chinese Restaurant Franchise analogy, we assume that the documents correspond to restaurants and the words to customers. The next steps differ for the two variants of hHDP.

In hvHDP, where topics are both distributions over subtopics and over words, the inference of the non-leaf levels treats topics, instead of documents, as restaurants. Thus, each inferred leaf topic maintains a distribution over the term space as its representation. Based on this distribution, it is treated as an observation for the inference of the next level up. Having inferred the topics at the leaf level, we know the mixture proportions that the documents of the collection follow. Similarly, each inferred topic maintains a distribution over the term space and a distribution over the subtopics below it, following the corresponding proportions inferred for this topic. Therefore each internal topic maintains a distribution over words and a distribution over subtopics. This procedure is repeated until we infer a single topic, which serves as the root of the hierarchy. In other words, at the leaf level we allocate documents to leaf topics, while at the intermediate levels we allocate topics to super-topics. The sampling scheme that we propose for hvHDP is described in Algorithm 1.

Therefore, the main contribution of this sampling scheme is the estimation of the non-leaf topics from "artificial" documents that correspond to estimated topics of lower levels. This procedure

Data: Term - Document matrix of frequencies
Result: Estimated topic hierarchy
set M=number of documents
set V=vocabulary size
estimate leaf topics K
set $T = K$
while $\mid T \mid > 1$ **do**
  // **transform document space**
  set $M = K$
  set input=MxV matrix of frequencies
  estimate topics K of next level up
  set $T = K$
end

Algorithm 1: Estimation of the topic hierarchy for the hierarchical vocabulary clustering hHDP method (hvHDP).

supports the non-parametric inference of the depth of the hierarchy. Together with the use of the HDP for the estimation of the number of topics at each level, it makes the estimation of the topic hierarchy completely non-parametric.

Regarding the second variant of the model (htHDP), where the internal topics are distributions only over subtopics and not words, the inference procedure differs in the modeling of non-leaf topics. Leaf topics serve now as customers, changing the term space, maintaining at the same time the restaurant space, which consists of the original documents. As observations for the inference of the next level up, we use the distributions of topics at the lower level over the original documents. Therefore, while in the first variant of hHDP, we had a topic - term matrix of frequencies as input for the estimation of an intermediate level of the hierarchy, in htHDP, we have a document - topic matrix of frequencies for the sampling procedure. The hierarchy estimated by htHDP is expected to be shallower than that inferred by hvHDP. This is because the term space is reduced when moving a level up. The procedure is repeated until we infer a single topic, which serves as the root topic. The proposed sampling scheme is described in Algorithm 2.

The last step in Figure 5 shows the overall model that is estimated. A topic hierarchy is derived from the corpus and a non-parametric Bayesian hierarchy is used at each level of the topic hierarchy. The first hHDP variant satisfies the criteria that we set in Section 2: internal topics are represented as distributions over words and over subtopics, topics can share subtopics at the lower level in the hierarchy, and topics across any level of the hierarchy are shared among documents. The degree of sharing topics across documents is expressed through the inferred parameters of the model, and this sharing of topics reflects the sharing of common terminology between documents. The non-parametric nature of this process is due to HDP that models each level of the hierarchy.

3.3 Level-wise Estimation

In hHDP, the estimation of each level is performed through posterior sampling of a HDP. At each level we integrate out all the probability measures G_i, the base measures G_0 and the tables. The metaphor of the "Chinese restaurant franchise" (CRF) is ofter used to illustrate the sampling scheme of the HDP. According to that metaphor, there are D restaurants and each one has an infinite number

Data: Term - Document matrix of frequencies
Result: Estimated coarse topic hierarchy
set M=number of documents
set V=vocabulary size
estimate leaf topics K
set $T = K$
while $\mid T \mid > 1$ **do**
  // **transform term space**
  set $V = K$
  set input=MxV matrix of frequencies
  estimate topics K of level up
  set $T = K$
end

Algorithm 2: Estimation of the topic hierarchy for hierarchical topic clustering version of hHDP (htHDP).

of tables. On each table the restaurant serves one of the infinitely many dishes that other restaurants may serve as well. A customer enters the restaurant. The customer not only chooses a table (which corresponds to topic sampling from Gj appearing in G_0), but also chooses whether she may have a dish popular among several restaurants (topic sharing among documents).

Based on the CRF metaphor, the collapsed sampling scheme includes only the sampling of the dishes, and the calculation of the number of tables that serve a specific dish in each restaurant. Thus, the sampling of an existing topic z at a specific level, given a word w_{ji} and the previous state of the Markov chain $z_{\neg ji}$ uses Equation (1), or equivalently Equation (2). On the other hand, the sampling of a new topic z_{new}, given a word w_{ji} and the previous state of the Markov chain $z_{\neg ji}$ uses Equation (3), or equivalently Equation (4).

$$p(z_{ji} = z \mid w_{ji}, z_{\neg ji}) \propto \frac{n_{j\cdot z} + \frac{\alpha t_z}{t_\cdot + \gamma}}{n_{j\cdot\cdot} + \alpha} \cdot \phi_z(w_{ji}) \tag{1}$$

$$p(z_{ji} = z \mid w_{ji}, z_{\neg ji}) \propto \frac{n_{j\cdot z} + \frac{\alpha t_z}{t_\cdot + \gamma}}{n_{j\cdot\cdot} + \alpha} \cdot \frac{n_{\cdot iz} + H}{n_{\cdot\cdot z} + VH} \tag{2}$$

$$p(z_{ji} = z_{new} \mid w_{ji}, z_{\neg ji}) \propto \frac{\alpha\gamma}{(n_{j\cdot\cdot} + \alpha)(t_\cdot + \gamma)} \cdot \phi_z(w_{ji}) \tag{3}$$

$$p(z_{ji} = z_{new} \mid w_{ji}, z_{\neg ji)} \propto \frac{\alpha\gamma}{(n_{j\cdot\cdot} + \alpha)(t_\cdot + \gamma)} \cdot \frac{1}{V} \tag{4}$$

In Equations (1) to (4), besides the hyper-parameters α and γ, $n_{j\cdot z}$ is the number of words in document j that are associated to topic z, $n_{j\cdot\cdot}$ is the number of words in document j, t_z is the number of tables that serve the dish z, and $t_\cdot$ is the total number of tables. The factor $n_{j\cdot z}$ emulates the draw of an existing dish of restaurant G_j, while the factor $\frac{\alpha t_z}{t_\cdot + \gamma}$ emulates the draw of a dish from the base restaurant G_0 that maintains all the dishes. The factor $\frac{\alpha\gamma}{t_\cdot + \gamma}$ emulates the draw of a new dish from

the global DP with hyper-parameter H. Finally, $\phi_z(w_{ji})$ stands for the word distribution $p(w \mid z)$. In addition, $n_{\cdot iz}$ is the number of occurrences of word i in topic z, $n_{\cdot\cdot z}$ is the total number of words assigned to topic z, and finally, H and V are the prior DP hyper-parameter for word distributions and the total number of words respectively.

Following the sampling of topic indicators, we calculate the number of tables that serve a specific dish at each restaurant, since we need that parameter for the sampling of topics. That is, we calculate the factor t_z, which influences the likelihood of a new table in document j via the factor $\frac{\alpha t_z}{t_\cdot + \gamma}$. We estimate this number by simulating a DP with hyper-parameter α, since we are interested in each document that is associated to a probability measure G_j, and parameter α provides control over the topic mixture. Algorithm 3 describes this process.

Data: $n_{j\cdot z}$, hyper-parameter α
Result: Number of tables in document j serving topic z
// **if no words exist then no tables are needed if** $n_{j\cdot z} = 0$ **then**
| return 0
end
// **if only one word exists, one table is needed**
if $n_{j\cdot z} = 1$ **then**
| return 1
end
// **if more words exist, simulate the DP**
set $t_z = 1$
for *all words w in* $[1, n_{j\cdot z}]$ **do**
| draw *rand* from Random
| set $DP_{table} = \alpha/(w+\alpha)$
| **if** $rand < DP_{table}$ **then**
| | set $t_z = m_t + 1$
| **end**
end

Algorithm 3: Estimation of the number of tables that serve a specific dish (topic) in each restaurant. The parameters $t_z, n_{j\cdot z}$ are the ones used in Equations (1) to (4).

According to Algorithm 3, the estimation of the number of tables is performed for each restaurant, for the customers that have been assigned to new tables, not present in the previous sampling iteration. The factor t_z can only change when a word is assigned to a new topic. Due to the "rich gets richer" property of the DP, some tables become unoccupied. Then, the probability that this table will be occupied again in the future is zero, since this is proportional to $n_{j\cdot z}$, which will be zero. Therefore, when estimating a new level bottom-up, the number of tables tends to decrease. In addition, in hvHDP, at each level of the hierarchy we transform the inferred topics to documents. This introduces a bound on the number of tables, since we decrease the restaurant space, which in turn bounds the number of sharing components, that is, the topics. The same holds for htHDP, where the term space is dramatically reduced at each level, placing in this way a stronger bound on the number of sharing components. For this reason, the second variant of hHDP converges faster to a single topic, producing smaller hierarchies.

More formally, and according to Teh and Jordan (2010), $t_z \in O(\alpha log\frac{n_{j..}}{\alpha})$. Since G_0 is itself a draw from a DP, we have that $K \in O(\gamma log\sum_j \frac{t_z}{\gamma}) = O(\gamma log(\frac{\alpha}{\gamma}\sum_j log\frac{n_{j..}}{\alpha}))$. Assuming J groups, each of average size N, we have that $K \in O(\gamma log\frac{\alpha}{\gamma}Jlog\frac{N}{\alpha}) = O(\gamma log\frac{\alpha}{\gamma} + \gamma logJ + \gamma loglog\frac{N}{\alpha})$. Thus, the number of topics scales doubly logarithmically in the size of each group and logarithmically in the number of groups. In summary, the HDP expresses a prior belief that the number of topics grows very slowly in N.

4. Evaluation and Empirical Results

In this section, we present experiments using real data sets in order to demonstrate and evaluate the proposed method. We perform experiments on two different tasks, in order to obtain a good overview of the performance of the model. The goal is to measure how well the estimated hierarchy fits a heldout data set of a specific domain, given a training data set of that domain and to what extent the proposed method can be used for knowledge representation and help bootstrap an ontology learning method. In particular, we divide the section into two subsections. The first one (Section 4.1) concerns document modeling and provides qualitative and quantitative results, while Section 4.2 applies the model to the task of ontology learning.

4.1 Document Modeling

Given a document collection, the task is to retrieve the latent hierarchy of topics that represents and fits well, in terms of perplexity, to the data set. We fit hHDP and compare it with LDA and hLDA on various data sets using held-out documents.

In particular, we use 10-fold cross validation and report perplexity figures for each method. Perplexity is commonly used to evaluate language models, but it has also been used to evaluate probability models in general (Blei et al., 2003; Teh et al., 2006). Better models that avoid overfitting tend to assign high probabilities to the test events. Such models have lower perplexity as they are less surprised by the test sample. In other words, they can predict well held-out data that are drawn from a similar distribution as the training data. Hence, in our evaluation scenario, a lower perplexity score indicates better generalization performance. Equation (5) defines the perplexity on a test set D consisting of words $w_1, w_2, ..., w_N$.

$$Perplexity(D) = exp\{-\sum_{i=1}^{N}\frac{1}{N}\log p(w_i)\} \tag{5}$$

As an example of the results obtained by hvHDP, Figure 6 presents part of the latent structure that was discovered from the NIPS data set. The NIPS data set is a benchmark corpus that has been used in related work (Blei et al., 2004). It contains abstracts of the corresponding conferences from 1987 to 1999. Specifically, the data set comprises 1732 documents and no pre-processing took place before the learning of the hierarchy, resulting in an unrestricted vocabulary of 46873 terms. The model ran for 1000 iterations of the Gibbs sampler with fixed hyper-parameters. In particular, the Dirichlet process priors H and γ were set to 0.5 and 1.0 respectively, while the parameter α of the topic mixture was set to 10.0. The values selected for the hyper-parameters are similar to the values selected for related tasks in the literature (Mimno et al., 2007).

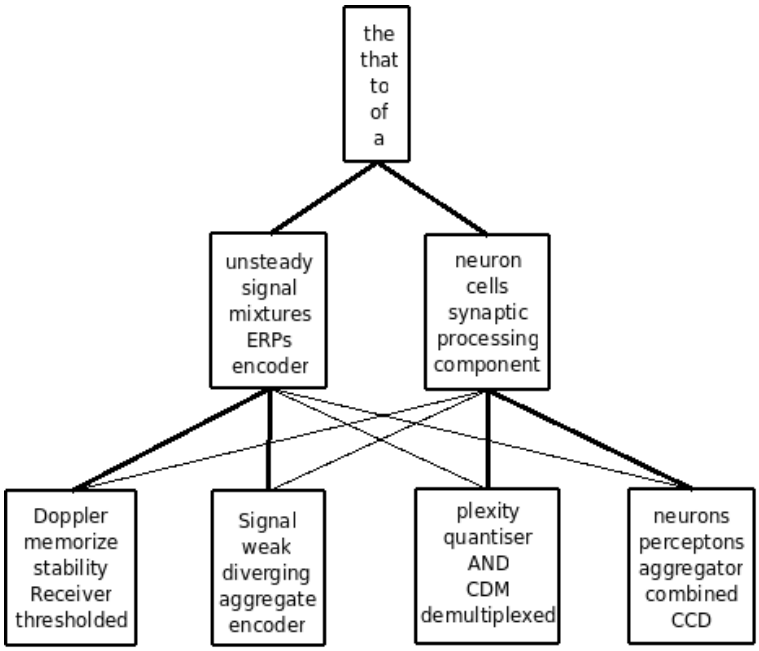

Figure 6: Part of the hierarchy estimated from the NIPS data set. The learned hierarchy contains 54 topics, inferred by the hHDP model without any user-specified parameters. Thick lines represent edges of high probability, while thinner ones stand for edges of lower probability.

As shown in Figure 6, the model discovered interesting topics from the field of the conference. Stop words are first grouped together at the root node representing a very general "topic" that connects equiprobably the two topics of the conference, signal processing and neural networks. Taking into account the context of the NIPS conferences, we believe that we have discovered a rather realistic hierarchical structure of 54 topics that fits well the field in question.

Similarly, Figure 7 illustrates part of the hierarchy that was produced by mixing two corpora together and running hvHDP on the resulting data set. In this experiment we wanted to investigate how the mixing of documents of different domains affects the resulting hierarchy, and in particular to see whether we can identify a sub-hierarchy of one domain inside the complete hierarchy that was learned. For this reason, we used 100 documents from the tourism domain and 1000 documents from the domain of molecular biology, resulting in a total of 1100 documents.

In Figure 7 only edges of high probability are shown for clarity reasons. The two separate sub-hierarchies, corresponding to the different domains are evident. The sub-hierarchy that corresponds to the tourism data set (inside the circle in the figure) is much smaller than that of the domain of molecular biology.

In order to obtain a quantitative evaluation of the method on document modeling, we used five different data sets. We also fitted the models of hLDA and LDA to the same data sets, as well as two other baseline models that we have implemented. The first, based on a uniform model (UM), is not trained and generates words following a uniform distribution, irrespective of the data set.

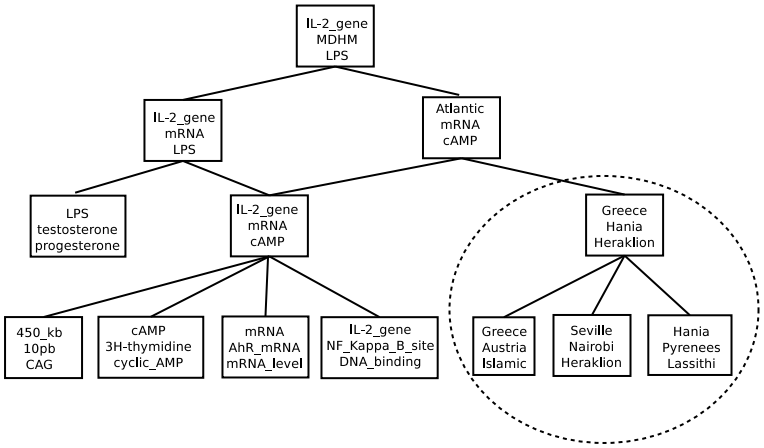

Figure 7: Part of the hierarchy estimated from a data set containing 1000 articles regarding molecular biology and 100 regarding tourism information.

The second that we call memory model (MM), memorizes the given data set and generates words according to the multinomial probability distribution of each document of the data set.

The different evaluation data sets that we used are the following: (a) the Genia data set,[1] from the domain of molecular biology, (b) the Seafood corpus,[2] comprising texts relative to seafood enterprises, (c) the Lonely Planet corpus,[3] consisting of texts from the tourism domain, (d) the Elegance corpus,[4] comprising nematode biology abstracts, and finally, (e) the NIPS data set[5] that includes abstracts from the corresponding conferences between the years 1987 and 1999. Table 2 summarizes basic statistics of the five data sets.

Data Set	#Docs	TermSpace	Domain
Genia	2000	16410	Molecular biology
Seafood	156	13031	Seafood enterprises
Lonely Planet	300	3485	Tourism
Elegance	7300	35890	Nematode biology
NIPS	1732	46873	NIPS conferences

Table 2: Data Sets

In the specific experimental setup we used the same hyper-parameters for all data sets. As mentioned above, for hHDP, $H = 0.5$, $\gamma = 1.0$ and $\alpha = 10.0$. In the case of hLDA, $\eta = 0.5$ and $\gamma = 1.0$, and we varied the number of levels, while in the case of LDA, we varied the number of topics from 10 to 120. Figure 8 illustrates the behavior of the models in the different data sets

1. The GENIA project can be found at `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi`.
2. The Seafood corpus can be found at `http://users.iit.demokritos.gr/~izavits/datasets/Seafood_corpus.zip`.
3. The Lonely Planet travel advise and information can be found at `http://www.lonelyplanet.com/`.
4. The Elegance corpus can be found at `http://elegans.swmed.edu/wli/cgcbib`.
5. The NIPS data set can be found at `http://books.nips.cc`.

for different numbers of LDA topics. Specifically, the figures plot the perplexity of the various models against the number of the discovered topics. In the case of hHDP the number of topics is inferred automatically and cannot vary with the LDA or the hLDA parameters. The hLDA model is parameterized by the number of levels. However, when changing the number of levels, the number of topics also changes. The model itself decides the branching factor at each level, and thus the total number of topics changes. A first observation in the results that we obtained is that in all cases, the simple UM results in very high perplexity values, between 2500 and 45000 that we do not depict in Figure 8 for reasons of readability of the graphs. Moreover, the MM performs worse in general than the rest of the models.

In order to interpret the different results obtained in the five different data sets, we measured the heterogeneity between the training and the held-out data in each case. More specifically, we measured the difference in the distribution of words between training and held-out data, using the mean total variational distance (TVD) (Gibbs and Su, 2002), according to Equation (6). The higher the TVD, the bigger the difference between the training and the held-out set. Table 3 presents the results of this measure in terms of the mean TVD in a 10-fold cross measurement. Based on these figures, the Genia data set seems to be the most homogeneous, while NIPS is the least.

$$TVD = \frac{1}{2}\sum_{i} \mid p(i) - q(i) \mid . \tag{6}$$

Data Set	Mean TVD
Genia	$1.2 * 10^{-5}$
Seafood	$3.5 * 10^{-5}$
Lonely Planet	$2.2 * 10^{-5}$
Elegance	$3.8 * 10^{-5}$
NIPS	$5.2 * 10^{-5}$

Table 3: Mean Total Variational Distance between the training and the held-out parts of the data sets.

Additionally, in order to validate the graphs of Figure 8, we measured the significance of the results, using the Wilcoxon signed-rank test. This test is suitable for this kind of experiment, since it is non-parametric and does not assume that the samples follow a specific distribution. In particular, we performed the test for the mean perplexity values, for each value of the number of topics. According to the test, the perplexity of a model is significantly lower than that of another model, if the output probability of the test is below 0.05, which is a threshold that is commonly used in statistical analysis.

In all data sets, the most interesting comparison is that between hHDP and hLDA. Thus, Table 4 depicts the ranges of topics for which the proposed model performs significantly better than the one of hLDA and the one of LDA. These ranges are also marked on the x axis of Figure 8 in all diagrams.

Examining the results on the Genia data set (Figure 8a), the lowest perplexity is achieved by hvHDP, while hLDA approaches the same perplexity for a number of topics around 60. LDA and htHDP obtain higher perplexity.

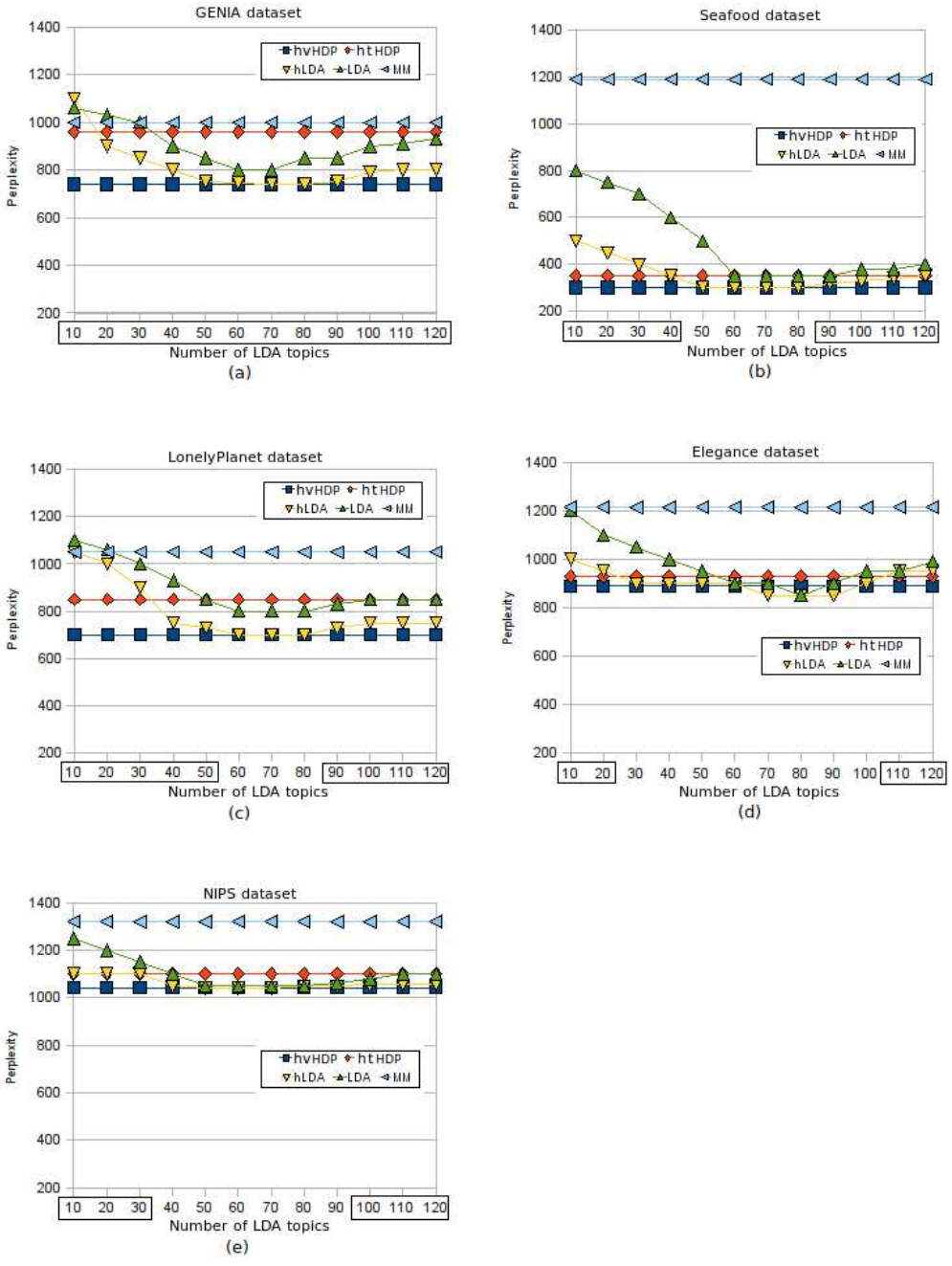

Figure 8: The behavior of the models on the five different data sets in terms of perplexity. The models: hvHDP, htHDP, hierarchical Latent Dirichlet Allocation (hLDA), Latent Dirichlet Allocation (LDA), and Memory Model (MM). Diagrams (a)-(e) illustrate the perplexity of the models for the Genia, Seafood, LonelyPlanet, Elegance and NIPS data sets respectively. Topic ranges where statistically significant improvement over existing models is achieved are marked on the *x* axis.

Comparison	Genia	Seafood	LP	Elegance	NIPS
hvHDP	$1-120$	$1-40$	$1-50$	$1-20$	$1-30$
hLDA		$90-120$	$90-120$	$110-120$	$100-120$
hvHDP	$1-120$	$1-120$	$1-120$	$1-50$	$1-120$
LDA				$100-120$	

Table 4: Significant differences between hvHDP and hLDA and between hvHDP and LDA in all corpora. Each cell presents topic ranges for which hvHDP performs significantly better than hLDA or LDA.

The comparison between hLDA and hvHDP showed that for all the cases in this data set, hvHDP obtains significantly lower perplexity values (Table 4).

Regarding the Seafood data set (Figure 8b), hLDA and LDA catch up with hvHDP after 40 and 60 topics respectively. htHDP also achieves good performance in this case. Regarding the statistical significance of the differences, Table 4 validates that hvHDP performs better than hLDA for a range of topics between $1-40$ and between $90-120$.

In the LonelyPlanet data set (Figure 8c), only hLDA manages to approach the good performance of hvHDP for a number of topics between $60-80$ (Table 4). The LDA and htHDP models perform worse. The htHDP is again much worse than the first variant of hHDP.

Concerning the Elegance data set (Figure 8d), all models, besides MM, achieve similar performance within a specific range of topics (50 to 120). Furthermore, this is the only data set where hLDA and LDA are observed to achieve better results than hvHDP, though not statistically significant and for a very small range of topics (around 80).

Finally, in the NIPS data set, (Figure 8e), hLDA and LDA manage to equal hvHDP for a certain range of topics and present a better performance than htHDP in a large range of topics. For this data set, the statistical test showed that hvHDP is better than hLDA in the range $10-30$ and $100-120$ topics and better than LDA in the whole range of topics, although for a certain range both models achieve similar perplexity values. On the other hand, the second version of hHDP outperforms only LDA between 10 and 20 topics.

The results illustrate clearly the suitability of hHDP for document modeling tasks. It discovers a hierarchy that fits well the given data sets, without overfitting them, thus achieving low values of perplexity. The competing models of hLDA and LDA manage only at their best to reach the performance of the proposed model. Furthermore, the performance of these models seems to be very sensitive to the chosen number of topics (number of levels in the case of hLDA). This observation makes the non-parametric modeling of hHDP particularly important. Comparing hLDA to the simple LDA, it is also quite clear that the hierarchical modeling of topics adds significant value to the model.

Regarding the naive UM and MM models, these are only used as baselines and they perform poorly. The experiments show that an overfitted model, such as MM, has low predictive performance outside the training set. On the other hand, a uniform model is not able to predict at all the test set, achieving the worst results.

A final important observation that is not evident in the numeric results, is that for a large number of topics, hLDA tends to construct a single path, rather than a hierarchy. Perhaps this can be attributed to the difficulty of identifying sufficiently different topics at various levels of abstraction,

when requesting a large depth for the hierarchy. By assigning all topics to a single branch, the model becomes equivalent to LDA. When this happens, the perplexity value of the two models is also very similar. The Wilcoxon statistical tests have indicated that in the Elegance data set and for a number of topics around 80, hLDA does not perform significantly better than LDA, while in the NIPS data set, the same situation holds for a number of topics between 50 and 100.

Perplexity has been criticized, since it is mainly used for the evaluation of language models. In addition, recent advances in topic modeling evaluation suggest the use of unbiased assessment of topic models. For this reason, we decided to conduct an additional experiment, measuring the log-likelihood of these models using the left-to-right sequential sampler (Buntine, 2009). This sampler improves on the algorithm proposed in Wallach et al. (2009), by providing unbiased estimates of the model likelihood for sufficiently large sample sizes. Since hvHDP, hLDA and LDA achieve the best results in terms of perplexity, we compare these models. Having the models trained on a portion (90%) of the data sets, we calculate the log-likelihood of the models on the remaining 10% that constitute the held-out data, using 10-fold cross validation. Figure 9 presents the results of this experiment, in terms of the mean log-likelihood.

The main result shown in Figure 9 is the same as in Figure 8. hHDP outperforms the other methods with statistical significance in most cases. The other methods, especially hLDA, approach the performance of hHDP if the right number of topics is chosen somehow. Therefore, the experiment has confirmed the value of estimating the number of topics and the depth of the hierarchy in a completely non-parametric way.

As an additional experiment on the task of document modeling, we assessed the ability of the method to estimate a known hierarchy, which is used to generate a set of documents. In particular, based on the hierarchy inferred for the Seafood data set, we generated a set of documents with the same average length as the original data set. Thus, we started at the root node of the hierarchy, and traversed it stochastically, based on the parameters of the model, which are the probabilities of each subtopic. When reaching a leaf topic we chose a word to be generated according to the probability distribution of that topic. In this manner, we generated a total of 156 documents, as many as the original data set, exhibiting similar word distributions. Then, we ran hvHDP on this "artificial" data set, estimating a latent hierarchy, which we compared manually against the one used to create the data set. From this comparison we concluded that all the topics of the estimated hierarchy have been correctly inferred. However, the estimated hierarchy comprises fewer topics, a fact that in terms of quantitative results implies a drop in recall.

4.2 Ontology Learning

The aim of this experiment was to validate the suitability of the proposed method on the task of ontology learning. The vocabulary clustering version of hHDP (hvHDP) estimates topics that are defined as distributions over words. It is, therefore, of particular interest to investigate how close these distributions are to a gold-standard hierarchy, given the corresponding data set. Such an experiment would highlight the potential of the method in other domains, such as automated ontology construction, and would provide qualitative and quantitative results regarding the performance of the method. In this experiment, we also compare hvHDP with hLDA.

Ontology learning (Gomez-Perez and Manzano-Macho, 2003; Maedche and Staab, 2003) refers to the set of methods and techniques used for either building an ontology from scratch, enriching, or adapting an existing ontology in an automated fashion, using various sources of information.

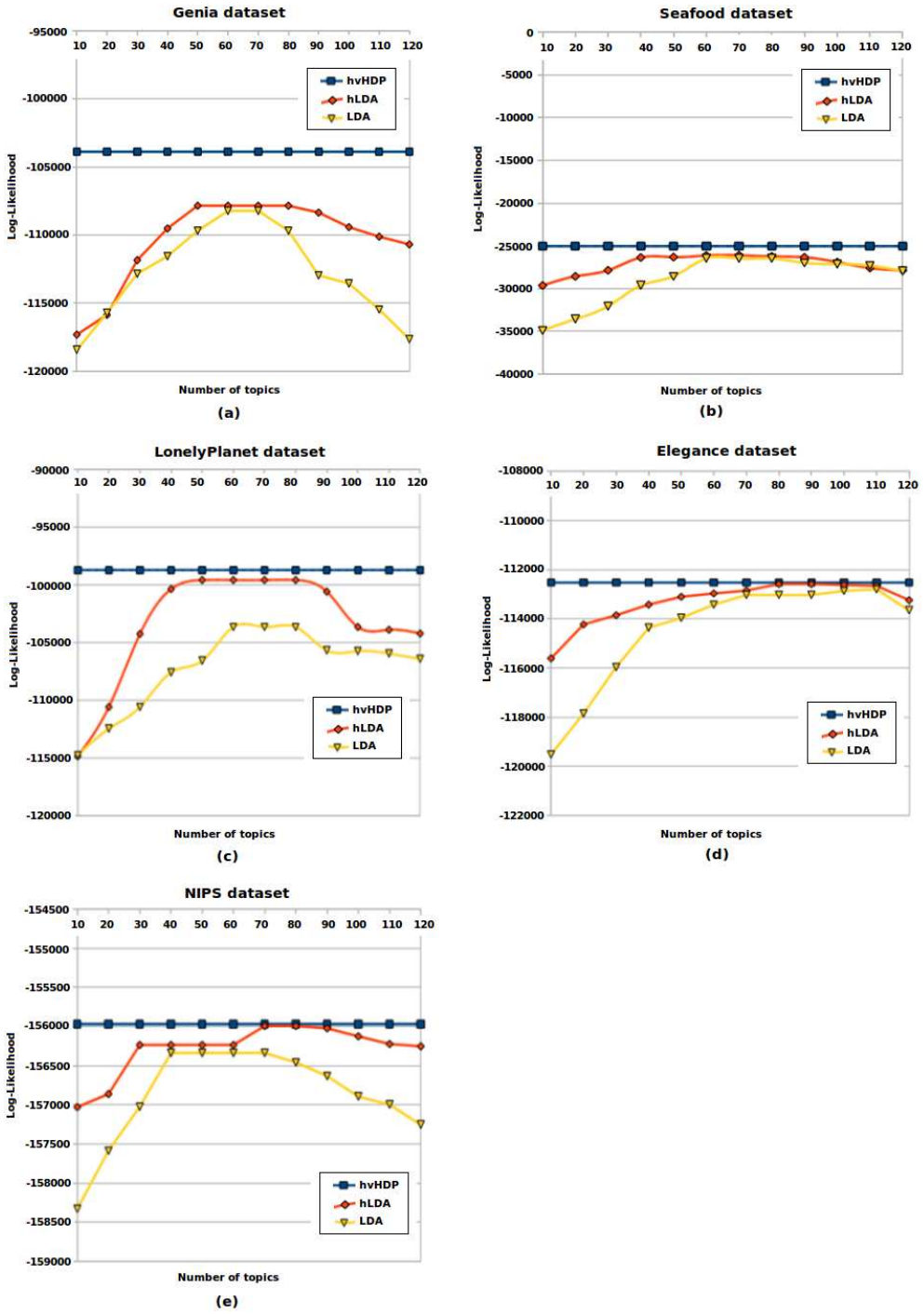

Figure 9: The behavior of the models (hvHDP, hierarchical Latent Dirichlet Allocation (hLDA), and Latent Dirichlet Allocation (LDA)) on the five different data sets in terms of log-likelihood. Diagrams (a)-(e) illustrate the perplexity of the models for the Genia, Seafood, LonelyPlanet, Elegance and NIPS data sets respectively.

This task is usually decomposed into three steps: (a) identification of topics, (b) building of the hierarchical backbone, and (c) enriching with further semantic relations. Regarding the sources of information, we focus here on text collections.

Both hvHDP and hLDA can be used to perform the first two steps of the ontology learning process, that is, identification of concepts and hierarchy construction, given the data set. Thus, we merge the aforementioned two steps into one, and we assume that the estimated latent topics correspond to ontology concepts. Therefore, in this task, we construct a topic ontology from scratch that comprises only hierarchical relations, given a collection of text documents and we compare it to a given gold standard ontology.

For this purpose, we use the Genia and the Lonely Planet data sets and the corresponding ontologies, which serve as gold standards for evaluation. The Genia ontology comprises 43 concepts that are connected by 41 subsumption relations, which is the only type of relation among the concepts. The Lonely Planet ontology contains 60 concepts and 60 subsumption relations among them. For our experiments, the only pre-processing applied to the corpus was to remove stop-words and words appearing fewer than 10 times.

The estimation of the hierarchy was achieved through 1000 iterations of the Gibbs sampler with fixed hyper-parameters $H = 0.5$ and $\gamma = 1.0$ for the Dirichlet priors and $\alpha = 10.0$ for the topic mixture. The evaluation was performed using the method proposed in Zavitsanos et al. (2010). This method is suitable for the evaluation of learned ontologies, since it represents the concepts of the gold ontology as multinomial probability distributions over the term space of the documents and provides measures in the closed interval of [0,1] to assess the quality of the learned structure.

In particular, the evaluation method first transforms the concepts of the gold ontology into probability distributions over the terms of the data set, taking into account the context of each ontology concept. In a second step, the gold ontology is matched to the learned hierarchy, based on how "close" the gold concepts and the learned topics are. The final evaluation is based on the measures of P and R that evaluate the learned hierarchy in the spirit of precision and recall respectively, as well as F that is a combined measure of P and R. The corresponding formulae are given in Equations (7), (8) and (9).

$$P = \frac{1}{M}\sum_{i=1}^{M}(1 - SD_i)PCP_i \tag{7}$$

$$R = \frac{1}{M}\sum_{i=1}^{M}(1 - SD_i)PCR_i \tag{8}$$

$$F = \frac{(\beta^2 + 1)P * R}{(\beta^2 R) + P} \tag{9}$$

In Equations (7) - (9), M is the number of matchings between learned topics and gold concepts and SD is a distance measure between concepts, ranging in $[0,1]$. Specifically, the total variational distance (TVD) (Gibbs and Su, 2002) of Equation (10) was used to assess the similarity between topics and gold concepts.

$$TVD = \frac{1}{2}\sum_{i} \mid P(i) - Q(i) \mid \tag{10}$$

In Equation (10), $P(\cdot)$ and $Q(\cdot)$ are multinomial probability distributions over words that represent a gold concept and a learned topic. The estimated topics are already represented as multinomial probability distributions over the term space of the data set, while the concepts of the gold ontology are also transformed into multinomial probability distributions over the same term space. Thus, the comparison between topics and gold concepts becomes straightforward.

The matching scheme compares the distributional representations of topics and gold concepts and finds the best matching in the sense that the most similar word distributions among the two hierarchies will be matched. More details about how the matching is performed can be found in Zavitsanos et al. (2010). The *PCP* and *PCR* (*probabilistic cotopy precision and recall*) factors in Equations (7) and (8) respectively, are influenced by the notion of semantic cotopy (Maedche and Staab, 2002). The cotopy set of a concept C is the set of all its direct and indirect super and subconcepts, including also the concept C itself. Thus, for a matching i, of a topic T in the learned ontology and a concept C in the gold ontology, PCP_i is defined as the number of topics in the cotopy set of T matched to concepts in the cotopy set of C, divided by the number of topics participating in the cotopy set of T. For the same matching i, PCR_i is defined as the number of topics in the cotopy set of T matched to concepts in the cotopy set of C, divided by the number of topics participating in the cotopy set of C.

Values of the P, R and F measures close to 1 indicate that the resulting hierarchy is close to the gold ontology, while values close to 0 indicate the opposite. Finally, we set $\beta = 1$ in Equation (9), hence using the harmonic mean of P and R.

Figure 10 depicts a part of the gold ontology on the left and a part of the estimated hierarchy on the right. The labels on the latent topics of the learned hierarchy correspond to the best TVD match of each topic with a gold concept. As it is shown in the figure, hHDP estimated a hierarchy very close to the gold standard. Thin edges between topics represent relations of low probability, while thicker edges carry higher probability.

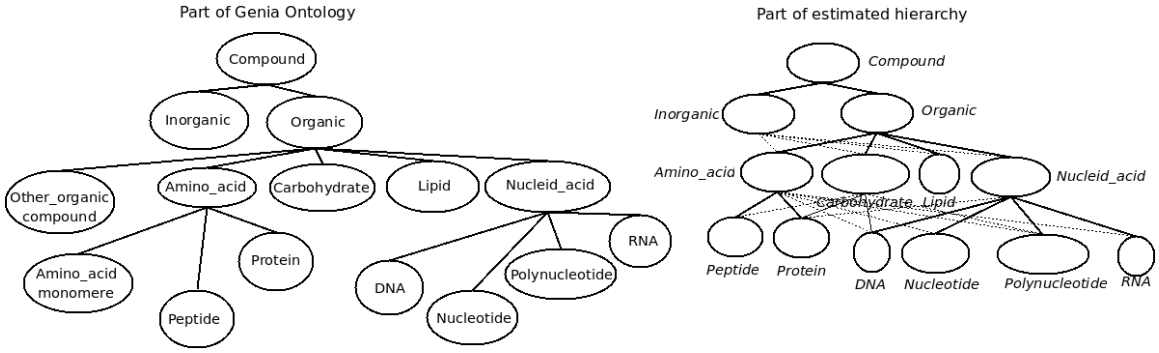

Figure 10: Part of the Genia ontology on the left and part of the estimated hierarchy on the right. The labels on the topics of the learned hierarchy correspond to the best match of each topic to a gold concept, according to TVD.

Regarding the estimated hierarchy, it comprises 38 topics in total, while the gold ontology comprises 43. Recall from Section 3 that the method estimates a probability distribution for each topic over all topics of the next level. Hence, we expect to learn a hierarchy comprising more relations than the gold ontology. However, relations with low probability, as the ones depicted with thin lines

in Figure 10, can be ignored. In addition, the way the hierarchy is estimated, through Gibbs sampling, infers the probability distributions, based on the assignments of words to topics and topics to subtopics. Through sampling, it is possible for fragments of documents not to be allocated to every estimated topic, and for subtopics not to be allocated to every super-topic. This leads to some zero values in the probability distributions of topics. Therefore, there exist cases where the probability of an edge in the resulting hierarchy may be zero. This fact provides extra flexibility to the method, since it permits the construction of unbalanced hierarchies and prunes edges that are definitely not necessary.

In the general case though, the learned hierarchy is expected to have more edges than the gold ontology has. Therefore, pruning mechanisms may be of particular importance for the task of ontology learning.

In the case of the Lonely Planet data set, hvHDP estimated a smaller hierarchy than the gold standard, achieving lower quantitative results in terms of P, R and F. The difficulty in estimating a hierarchy of similar size to the gold standard is due to the nature of the data set and the gold ontology. In particular, half of the gold concepts had only one instance and in general, most of the concepts were insufficiently instantiated in the data set.

Regarding hLDA, in the case of the Genia data set, the best quantitative results were obtained for an estimated hierarchy of depth equal to 6. In this case, hLDA performed similarly to hHDP in terms of P, R and F. However, in the case of the Lonely Planet data set the performance of the model was poor. In particular, the best quantitative results were obtained for an estimated hierarchy of 3 levels. However, these results are much lower than that of hvHDP for the same data set.

Table 5 presents the quantitative results of the experiments, in terms of P, R and F for both hHDP and LDA. For the proposed method, two cases are foreseen. The first case concerns the evaluation of the learned hierarchy as is, without any post-processing. The performance of hHDP is low, because the evaluation method is rather strict. The evaluation method does not take into account the probabilities on the edges connecting a topic to all its sub-topics, but rather assumes that all edges are of equal importance and penalizes the learned hierarchy for its high connectivity.

Therefore, through this first evaluation, we conclude that the original, highly connected hierarchy may not be usable as is. For this reason, we include another set of evaluation results in Table 5 that we call "pruned." This is actually the same method without the low probability relations between the topics. In particular, we keep relations with probability higher than 0.1. The pruned hierarchy is significantly closer to the gold standard than the unpruned one.

	Genia			LonelyPlanet		
Method	P	R	F	P	R	F
hHDP	0.65	0.60	0.624	0.22	0.15	0.17
hHDP-pruned	0.88	0.80	0.838	0.35	0.23	0.27
hLDA	0.62	0.55	0.58	0.07	0.01	0.017

Table 5: Quantitative results for the task of Ontology Learning.

In summary, we conclude that hvHDP can be applied to the task of ontology learning with promising results. Its ability to identify topics and at the same time build the taxonomic backbone can facilitate the learning of ontologies in a purely statistical way, providing a powerful tool that is independent of the language and the domain of the corpus. The proposed method discovered correctly the majority of the identifiable gold concepts in the experiment and constructed a hierarchy

that is very close to the gold standard. Furthermore, it constructed the taxonomy and inferred the correct depth without any user parameters (except the pruning threshold) in a statistical way and without any prior knowledge.

5. Conclusions

We have introduced hHDP, a flexible hierarchical probabilistic algorithm, suitable for learning hierarchies from discrete data. hHDP uses the "bag-of-words" representation of documents. The method is based on Dirichlet process priors that are able to express uncertainty about the number of topics at each level of the hierarchy. We have also presented a bottom-up non-parametric discovery method for the latent hierarchy, given a collection of documents. Since exact inference is known to be intractable in such non-parametric methods, approximate inference was performed, using the Gibbs sampling method, which provided accurate estimates.

An important contribution of this paper is the inference of the correct number of topics at each level of the hierarchy, as well as the depth of the hierarchy. Its Bayesian non-parametric nature requires no user parameters regarding the structure of the latent hierarchy. The Dirichlet process priors, as well as the bottom-up procedure for the estimation of the hierarchy, provide a flexible search in the space of different possible structures, choosing the one that maximizes the likelihood of the hierarchy for the given data set. Moreover, hHDP does not impose restrictions and constraints on the usage of topics, allowing multiple inheritance between topics of different layers and modeling the internal nodes as distributions of both subtopics and words.

We provided extensive experimental results for the proposed method in two different evaluation scenarios: (a) document modeling in five real data sets, comparing against state-of-the-art methods on the basis of perplexity, and (b) applying the method to an ontology learning task, comparing the learned hierarchy against a gold standard. The evaluation showed that hHDP is sufficiently robust and flexible. The proposed method discovered meaningful hierarchies and fitted well the given data sets. Finally, we have concluded that such methods are suitable for the task of ontology learning, since they are able to discover topics and arrange them hierarchically, in a way that is independent of the language and the domain of the data set, and without requiring any prior knowledge of the domain.

The very promising results that we obtained in this work, encouraged us to study and improve hHDP further. One possible improvement is the use of Pitman-Yor processes, which are generalizations of Dirichlet processes and produce power-law distributions. Natural language text is known to follow such distributions and therefore we may be able to model documents more accurately. In addition, we intend to apply the method to different tasks, including the learning of folksonomies from user-generated tags. Also, due to its statistical nature, it would be interesting to evaluate hHDP on different types of data sets, including images, time series and events. Finally, another future direction is to bootstrap hHDP from an existing ontology and infer the remaining parameters using the corresponding data set.

Acknowledgments

We would like to acknowledge support for this work from the research and development project ONTOSUM,[6] funded by the Greek General Secretariat for Research and Technology and the research and development project SYNC3.[7] We would also like to thank David Mimno and Andrew McCallum for helping us with the implementation of hLDA. Finally, we would like to thank Anna Sachinopoulou for providing us the Seafood corpus during her visit at our laboratory in NCSR "Demokritos."

References

C. Andrieu, N.D. Freitas, A. Doucet, and M.I. Jordan. An introduction to MCMC for machine learning. *Machine Learning Journal*, 50:5–43, 2003.

D.M. Blei and J.D. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems 18*, 2006.

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

D.M. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*, 2004.

W. Buntine. Estimating likelihoods for topic models. In *Asian Conference in Machine Learning*, 2009.

J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D.M. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*, 2009.

B. De Finetti. Funzione caratteristica di un fenomeno aleatorio. In *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale*, pages 251–299. 1931.

K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

E. Gaussier, C. Goutte, K. Popat, and F. Chen. A hierarchical model for clustering and categorizing documents. In *Advances in Information Retrieval - Proceedings of the 24th BCS-IRSG European Colloquium on IR Research*, pages 229–247, 2002.

A.L. Gibbs and F.E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

A. Gomez-Perez and D. Manzano-Macho. A survey of ontology learning methods and techniques. Technical Report Deliverable 1.5, 2003. Ontology-based Information Exchange for Knowledge Management and Electronic Commerce.

6. For ONTOSUM see also `http://www.ontosum.org/`.
7. For SYNC3 see also `http://www.sync3.eu/`.

T. Griffiths and M. Steyvers. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 381–386, 2002.

T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 41:177–196, 2001.

W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 577–584, 2006.

W. Li, D. Blei, and A. McCallum. Nonparametric Bayes pachinko allocation. In *Uncertainty in Artificial Intelligence*, 2007.

A. Maedche and S. Staab. Measuring similarity between ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW)*, pages 251–263, 2002.

A. Maedche and S. Staab. Ontology learning. In *Handbook on Ontologies in Information Systems*. Springer, 2003.

D. Mimno, W. Li, and A. McCallum. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th International Conference on Machine Learning*, pages 633–640, 2007.

G. Salton and M.H. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986.

M. Steyvers and T. Griffiths. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*. Hillsdale, NJ: Erlbaum, 2007.

Y.W. Teh and M.I. Jordan. Hierarchical bayesian nonparametric models with applications. In *Bayesian Nonparametrics: Principles and Practice. Cambridge, UK: Cambridge University Press*. 2010.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006.

H.M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *International Conference on Machine Learning*, 2009.

E. Zavitsanos, S. Petridis, G. Paliouras, and G.A. Vouros. Determining automatically the size of learned ontologies. In *18th European Conference on Artificial Intelligence, ECAI*, 2008.

E. Zavitsanos, G. Paliouras, and G.A. Vouros. Gold standard evaluation of ontology learning methods through ontology transformation and alignment. *IEEE Transactions on Knowledge and Data Engineering. IEEE Computer Society Digital Library. IEEE Computer Society, http://doi.ieeecomputersociety.org/10.1109/TKDE.2010.195*, 2010.