**Bayesian Nonparametric Learning for Hierarchical and Sparse Topics**

**Abstract:**

This paper presents the Bayesian nonparametric (BNP) learning for hierarchical and sparse topics from natural language. Traditionally, the Indian buffet process provides the BNP prior on a binary matrix for an infinite latent feature model consisting of a flat layer of topics. The nested model paves an avenue to construct a tree model instead of a flat-layer model. This paper presents the nested Indian buffet process (nIBP) to achieve the sparsity and flexibility in topic model where the model complexity and topic hierarchy are learned from the groups of words. The mixed membershipmodeling is conducted by representing a document using the tree nodes or dishes that a document or a customer chooses according to the nIBP scenario. Atree stick-breaking process is implemented to select topic weights from a subtree for flexible topic modeling. Such annIBP relaxes the constraint of adopting a single tree path in the nested Chinese restaurant process (nCRP) and, therefore, improves the variety of topic representation for heterogeneous documents. A Gibbs sampling procedure is developed to infer the nIBP topic model. Compared to the nested hierarchical Dirichlet process (nhDP), the compactness of the estimated topics in a tree using nIBP is improved. Experimental results show that the proposed nIBP reduces the error rate of nCRP and nhDP by 18% and 8%on Reuters task for document classification, respectively.

**Architecture:**

****

**Introduction:**

Unsupervised learning has a broad goal of extracting salient features and discovering structural information from the collected data. It is challenging to extract reliable features and their latent structure from abundant heterogeneous documents which are prone to be redundant, noisy, ambiguous, mismatched and ill-posed. We aim to construct a flexible latent variable model to meet the heterogeneous conditions and annotate the observed documents for prediction of future documents. In the past decade, the unsupervised learning via probabilistic topic model has been successfully developed for document categorization, collaborative filtering, document summarization and other natural language systems. The latent features or semantic topics are learned from a bag of words which are semantically similar across different documents. Conventionally, the parametric topic model based on latentDirichlet allocation (LDA) was constructed for modeling a set of D documents w = {wd} = {wdi}. Each vocabulary word wdi = v is represented by an associated parameter θdi which is driven by a topic label zdi = k using the multinomial parameters β = {βvk } = {p(wdi = v|zdi = k)}. The multinomial parameters πd = {πdk } = {p(zdi = k)} ofK topics in document d are represented by a Dirichlet distribution with parameters α = {αk }. Graphical representation of LDA is depicted. LDA is known as a finite-dimensional mixture representation for documents which assumes that

1) number of topics K is fixed,

2) topics k within or across documents d are independent,

3) all topics are used to represent a target document even though some of them are irrelevant to the document.

Bayesian nonparametrics (BNPs) b

ased on the hierarchical model, the nested model, and the sparse model are developed to provide nonparametric priors for topic model which relax these three assumptions.

**Existing system:**

 In general, the hierarchical model, nested model and sparse model provide the functions of mixed membership, tree model and feature selection for data representation, respectively. These functions offer the fundamentals to develop a flexible topic model to reflect heterogeneous documents, the ambiguous or out-of-domain contents, and the multi-level or multi-document aspects, etc. Table I shows three categories of BNPs under topic and non-topic models. The approximate inference algorithms using Markov chain Monte Carlo and variational inference were developed to implement these models. This paper presents a hybrid model which simultaneously considers three functions for document representation. The nested Indian buffet process (nIBP) is proposed to build a random and infinite-dimensional tree model from multiple documents. The sparsity of the estimated hierarchical topics is imposed. The documents are modeled according to the shared topics in a tree with multiple levels of abstraction. This nIBP provides a general strategy to build a structural feature model instead of non-structural model using IBP. The selection of features for each document is incrementally driven by a beta process. We propose the nIBP compound hDP (also denoted by nIBP-hDP or nIBP topic model). This process is different from hBPwhich carries out the Bernoulli process without tree model and topic information.We build a hierarchical topic model.

**Proposed system:**

we evaluate the relation between sparsity and complexity when applying nhDP and nIBP topic models. The sparsity (described in Section III-B) is calculated and averaged over different documents. This measure is seen as the proportion of tree nodes which are not selected in representation of individual document. The model complexity described in Section V-A is controlled by the number of topics and the accumulated topic complexity. This relation by using AP and WSJ datasets where the results from five folds of training data are shown. The subtree selection based on TSBP1 is performed.We find that nIBP consistently attains higher sparsity and lower complexity than nhDP over two datasets. dataset has lower complexity and higher sparsity when compared with AP dataset. This result reflects the condition that the domains or topics of WSJ are more specialized or focused than those of AP. The number of topics which are inferred from six WSJ documents with different numbers of words in the documents (Nd = 105, 210, 290, 425, 504, 670). We find that the number of topics sampled for representation of a document is increased by the length of document when applying nhDP and nIBP using TSBP1 and TSBP2 for subtree selection. The nCRP using SBP selects a single path so that only three topics are selected for each document. This is an evidence that TSBPs in nhDP and nIBP catch more topic variations than SBP in nCRP for document modeling.

**Modules:**

* **Hierarchical Dirichlet Process:**
* **Tree Stick-Breaking Process:**
* **Evaluation for Topic Sparsity and Complexity:**

**Hierarchical Dirichlet Process:**

The hDP in (1) deals with the representation of multiple documents where each document wd is associated with a mixture model from a local DP which determines how much a member from a shared set of mixture components or topic unigram models contributes to that document. The words in different documents are represented by a global mixture model which is drawn from a global DP. Bayesian hierarchy is formed by treating the global measure G0 as the base measure of a DP to draw the document specific measure Gd.

**Tree Stick-Breaking Process:**

In general, the hierarchical model, nested model and sparse model provide the functions of mixed membership, tree model and feature selection for data representation, respectively. These functions offer the fundamentals to develop a flexible topic model to reflect heterogeneous documents, the ambiguous or out-of-domain contents, and the multi-level or multi-document aspects, etc. Table I shows three categories of BNPs under topic and non-topic models.

**Evaluation for Topic Sparsity and Complexity:**

 These two documents share the same path with green nodes which are viewed as the correlated topics between two documents. We can see that d673 addresses the price of gold, energy, coffee, etc., while d1012 mentions the war, weapon, oil, etc. These two documents collaboratively talk about Gulf region, oil, Iraq, etc, as seen in green nodes. In addition, the documents in AP are found to be more heterogeneous than those in WSJ so that more paths are drawn for AP documents. But, the count of the most frequent topic word in AP is less than that in WSJ because the data diversity in AP is larger than that of WSJ. This result is consistent with the scope of target domains in two corpora.

**Algorithm:**

**A Gibbs sampling algorithm:**

Gibbs sampling is one MCMC. technique suitable for the task. The idea in Gibbs sampling is to generate posterior samples. by sweeping through each variable (or block of variables) to sample from its conditional. distribution with the remaining variables fixed to their current values.



**Future work:**

 We evaluate different topic models for document representation and classification by using four public-domain datasets: AP: This is a collection of 1341 randomly selected broadcast news documents with 219 720 words from the Associated Press newswire (AP) 1988-1990 dataset. The vocabulary contains 5183 words.We used a Bernoulli variable for a document which was associated with each topic to determine whether the document contributed to that topic or not. Conditioned on this variable, each topic was represented by amultinomial distribution over a small subset of documents. The topics were drawn layer by layer in a nested fashion. These topics covering different branches were comparable with the dishes in different restaurants.We aim to construct a flexible latent variable model to meet the heterogeneous conditions and annotate the observed documents for prediction of future documents. In the past decade, the unsupervised learning via probabilistic topic model has been successfully developed for document categorization, collaborative filtering, document summarization and other natural language systems.

**Conclusion:**

We have surveyed the BNP learning based on hierarchical model, nested model and sparse model. A new BNP learning was presented to construct a structural focused topic model where the topics along tree paths are flexibly selected for representation of a document in heterogeneous conditions. We used a Bernoulli variable for a document which was associated with each topic to determine whether the document contributed to that topic or not. Conditioned on this variable, each topic was represented by amultinomial distribution over a small subset of documents. The topics were drawn layer by layer in a nested fashion. These topics covering different branches were comparable with the dishes in different restaurants. Finding the topics and their proportions for individual document was implemented through a TSBP which captured a variety of latent features. A Gibbs sampling procedure was implemented to carry out the nIBP. The posterior probabilities were derived to draw the documentbased subtree paths and the word-based topics. The experiments on AP, WSJ, Reuters and NIPS datasets showed the merits of sparsity control and tree model for BNP text representation according to the metrics of perplexity and topic coherence. The merit was also illustrated for document classification by using Reuters dataset. But, the computation demand was substantially increased.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL