

**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 membership modeling is conducted by representing a document using the tree nodes or dishes that a document or a customer chooses according to the nIBP scenario. A tree stick-breaking process is implemented to select topic weights from a subtree for flexible topic modeling. Such an nIBP 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.