Abstract

Model Selection is a task selecting set of potential models. This method is capable of establishing hidden semantic relations among the observed features, using a number of latent variables. In this paper, the selection of the correct number of latent variables is critical. In the most of the previous researches, the number of latent topics was selected based on the number
of invoked classes. This paper presents a method, based on backward elimination approach, which is capable of unsupervised order selection in PLSA. During the elimination process, proper selection of some latent variables which must be deleted is the most essential problem, and its relation to the final performance of the pruned model is straightforward. To treat this problem, we introduce a new combined pruning method which selects the best options for removal, has been used. The obtained results show that this algorithm leads to an optimized number of latent variables. In this paper, we propose a novel approach, namely DPMFS, to address this issue.

Reference

An Empirical Selection Method for Document Clustering

Index Terms

Computer Science
Information Retrieval

Key words

Document clustering          Model selection          EM
algorithm
Dirichlet Process Mixture Model
Feature Selection