doi:10.1631/FITEE.1601688

题目：词加权有监督主题模型：多标签文本分类

概要：有监督主题模型已成功应用于多标签文本分类任务。代表性模型包括有监督隐含狄利克雷分配模型（labeled latent Dirichlet allocation，L-LDA）和判别隐含狄利克雷分配模型（dependency-LDA）。这些已有模型忽略单词类别频率信息，即训练集中单词出现的类别数量，对分类任务的影响。对此引入类别频率信息，提出一个类别频率词权重方法（class frequency weight, CF-weight）。CF-weight 方法基于如下假设：具有较高（或较低）类别频率的单词在分类问题中具有较低（或较高）判别力。将 CF-weight 方法应用于 L-LDA 和 dependency-LDA 模型。实验结果表明，相比传统有监督主题模型，基于 CF-weight 的模型在多标签分类性能上具有优势。

关键词：有监督主题模型；多标签分类；类别频率；有监督隐含狄利克雷分配模型；判别隐含狄利克雷分配模型