##  加入文档相似属性的LDA主题模型在新闻分类语料库上的应用

1、问题描述

主题建模已被广泛的应用于文本挖掘方面。例如LDA技术，在文本主题挖掘学习方面已有成功应用。我们在传统LDA模型基础之上，加入文档相关相似属性构建新的主题模型，并在真实的新闻分类语料库上应用实验，已验证新模型对比传统LDA模型的优越性。

2、模型设计

首先，我们定义了一种新的元数据类型，相关相似度。这种相关相似性可以来自于一个特定的应用程序中人为的判定或从其它数据中自动导出。

我们假设给出了一组文档，记为 D。我们进一步假设有距离函数 dist(di,dj),di,dj 来自于 D，来计算相关相似性，我们假设给出一个集合 T 三联体如下：



受最大边缘方法的启发，我们重写约束条件如下：



从上面的定义可知，我们的目标函数为最小化下面的损失函数：



其次，我们设计出一种模型，结合了LDA以及上述定义的相关相似度约束。并提出有效的算法已适用新的主题模型。

只考虑标准的 LDE 模型，我们的目标是寻找最大化下面所示的目标函数时的最优参数：



为了使这个目标函数与前面介绍的限制条件相联系，我们简单的把这两项相加：



为了把这个目标函数它转化为一个非限制性优化问题，我们首先定义如下的转化函数：

 

之后我们改变Dirichlet prior 为Gaussian prior ，我们把目标函数变化为下面的形式：

 

我们采用Gibbs-EM 方法优化前边提到的目标函数（8），回想一下，我们已经定义隐藏变量ž表示话题任务。 利用隐藏变量，目标函数可以如下来优化。在第E步，我们在tth 迭代修改参数λ(t) 并获得假定的隐变量分配p(z|w, λ(t), β) 。在第M步，我们解决了如下的优化问题：

 

其中，Eq[f] 是f关于分布q的预期值，并且



利用Gibbs-EM，代替评估精确的条件分布p(z|w, (t), ) ，我们利用Gibbs-EM抽样粗略估计。

3、数据采集

我们使用两个广泛使用的文本语料库，20 新闻组和 TDT2。20个新闻文本语料库是一个收集约20000新闻组文件，分区均匀在20个不同的新闻组。我们使用这个版本的数据集预处理，在文件分为训练集和测试集。在训练集的文档中，我们随机选择了100个文件，从每个类别和删除停止的话和非常短的文件（文件少于3个字），从而留下我们与14538个不同的单词的1997个文件。

TDT2 语料库包含可分为 96 类的 11,201 文件。只有最大的 20 类别被保留，和那些出现在多个类别的文件被删除。我们也随机抽取 100 个文档从每个类别作为 20 新闻组数据集的策略。最后，还有与 12,166 不同的单词左边的 1,998 文件。

4、实验验证

在我们的实验中，我们执行200次Gibbs-EM。每次运行，我们跑了 100 次迭代的吉布斯采样和另一个 10 迭代梯度下降。我们设置狄利克雷先验 = 0.1，高斯先验模型的方差 = 1。我们还固定的主题数是 20 （每个数据集类别的数目相同）。请注意，我们不调整此参数，因为我们的目标不是来找主题的最优数量。

为自动获取模型的三重态约束，我们在训练文档集根据他们的真实类别标签取样一组三重态。具体来说，通过随机抽样两个文件在同一类别和一个文件从另一个不同的类别，我们生成一个三重态实例采用随机抽样。我们的实验中，我们从每个数据集的训练集生成 100 K、 50 K 和 10 K 三重态实例。试验运行在硬件环境为4芯和4GB内存的机器上。

5、下一步任务

1.基于模型设计的理论分析，完成模型的详细设计与实现

2.对数据集进行主题分析

3.基于实验结果对比传统基于LDA主题模型，进行新模型的优越性分析