稿件编号：3085

Title：Higher-Order Smoothing: A Novel Semantic Smoothing Method for Text Classification

中文题目：高阶平滑：一种新的用于文本分类的语义平滑方法

Abstract: It is known that latent semantic indexing (LSI) takes advantage of implicit higher-order (or latent) structure in the association of terms and documents. Higher-order relations in LSI capture "latent semantics". These finding have inspired a novel Bayesian framework for classification named Higher-Order NaÄ³ve Bayes (HONB), which was introduced previously, that can explicitly make use of these higher-order relations. In this paper, we present a novel semantic smoothing method named Higher-Order Smoothing (HOS) for the Naive Bayes algorithm. HOS is built on a similar graph based data representation of the HONB which allows semantics in higher-order paths to be exploited. We take the concept one step further in HOS and exploit the relationships between instances of different classes. As a result, we move not only beyond instance boundaries, but also class boundaries to exploit the latent information in higher-order paths. This approach improves the parameter estimation when dealing with insufficient labeled data. Results of our extensive experiments demonstrate the value of HOS on several benchmark datasets.

中文摘要：潜在语义分析（LSI）利用词项和文本关联中的隐性高阶（潜在）结构。潜在语义分析中的这些高阶关系用于刻画“潜在语义”。这些发现启迪出了一种新的贝叶斯分类框架，也就是高阶朴素贝叶斯（HONB），它能够显性利用这些高阶关系进行分类。本文针对朴素贝叶斯算法提出了一种叫做高阶平滑（HOS）的新语义平滑方法。高阶平滑（HOS）利用高阶朴素贝叶斯（HONB）中的基于相似图的数据表示，寻求高阶路径所隐含的语义。我们利用高阶平滑（HOS）发现不同类别数据实例之间的关系。我们突破了不仅数据实例而且数据类别之间的界限，以寻求高阶路径上的语义信息。当仅有少量的标签数据时，本文所提出的方法能改进朴素贝叶斯中的参数估计。大量在标杆数据集的实验结果表明，高阶平滑（HOS）对于文本分类具有十分重要的价值。

Keywords: naive Bayes, semantic smoothing, higher-order naive Bayes, higher-order smoothing, text classification

中文关键词：朴素贝叶斯，语义平滑，高阶贝叶斯，高阶平滑，文本分类