doi:10.1631/FITEE.1500332

题目：TextGen：用于新型存储系统基准测试的真实文本数据集生成方法

概要：新型存储系统通过内置数据压缩功能提高性能，并节省存储空间。因此，数据内容会显著影响存储系统基准测试结果。由于真实数据集规模庞大，难以复制到目标测试系统，并且大多数数据集由于隐私性无法进行共享。因此，基准测试程序需要人工生成测试数据集。为了保证测试结果的准确性，需要根据影响存储系统性能的真实数据集特征信息生成数据。现有方法 SDGen 在字节级别上分析真实数据集内容分布特征，并以此生成数据集，因此能够保证内置字节级压缩算法的存储系统测试结果准确。但是 SDGen 并未分析真实数据集的词级别内容分布特征，因此不能保证内置词级别压缩算法的存储系统测试结果准确，本文提出了一种基于 Lognormal 概率分布模型的文本数据集生成方法 TextGen。该方法根据真实数据集的词切分结果建立语料库，分析语料库中词的分布特征，利用最大似然估计得到词分布的 Lognormal 模型参数，根据模型采用蒙特卡洛方法生成数据内容。该方法生成数据集所消耗的时间只与生成数据集规模相关，具有线性的时间复杂度 $O(n)$。本文收集了四种数据集验证方法有效性，并通过一种典型的词级别压缩算法——ETDC（End-Tagged Dense Code）进行测试。实验结果表明：相比 SDGen，TextGen 生成文本数据集性能更高，并且，生成数据集用于压缩测试后与真实数据集的压缩速率、压缩率相似程度更高。

关键词：基准测试；存储系统；基于词的压缩算法