Long-xiang Wang, Xiao-she Dong, Xing-jun Zhang, Yin-feng Wang, Tao Ju, Guo-fu Feng, 2016. TextGen: a realistic text data content generation method for modern storage system benchmarks. *Frontiers of Information Technology & Electronic Engineering*, **17**(10):982-993.
http://dx.doi.org/10.1631/FITEE.1500332

TextGen: a realistic text data content generation method for modern storage system benchmarks

Key words: Benchmark, Storage system, Word-based compression

Corresponding author: Xing-jun Zhang
E-mail : xjzhang@mail.xjtu.edu.cn

ORCID: http://orcid.org/0000-0003-1434-7016

Motivation

- The data content significantly influences the compression performance and compression ratio, and affects the performance and space consumption of modern storage systems, which are called content-sensitive.
- To obtain accurate performance results, a benchmark needs to generate the realistic data content.
- Existing approach SDGen can only guarantee the storage systems benchmark result is accurate when byte-based Ziv-Lempel family compressors are enabled.
- However, SDGen cannot ensure that the compression ratio and compression performance results are accurate when word-based compressors are enabled.

Main idea

- To address the problem that the data content generated by existing methods is inaccurate at the word level, we present TextGen, a realistic text data content generation method for modern storage system benchmarks.
- Capturing the word-level properties that influence the compression performance and compression ratio of word-based compressors, and to use those characterizations to generate content.

Method

1. Segmenting the real-world text datasets into words, and count the frequency of each word to build a corpus, which has the form <word, frequency>..
2. Fitting the rank-frequency distribution to a lognormal distribution by maximum likelihood estimation (MLE).
3. Using the Monte Carlo method to generate dataset content.

Major results

- TextGen (lognormal distribution-based) has the highest throughput of dataset generation.

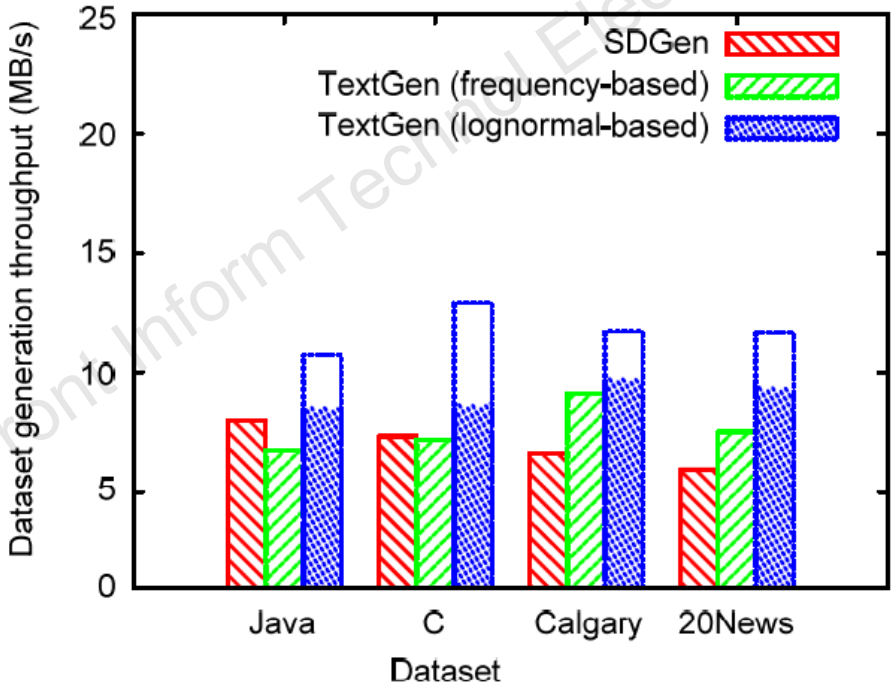

Fig. 10 Comparison of dataset generation throughput

Major results (Cont'd)

- The results show that the rate of deviation of TextGen is significantly higher than that of SDGen

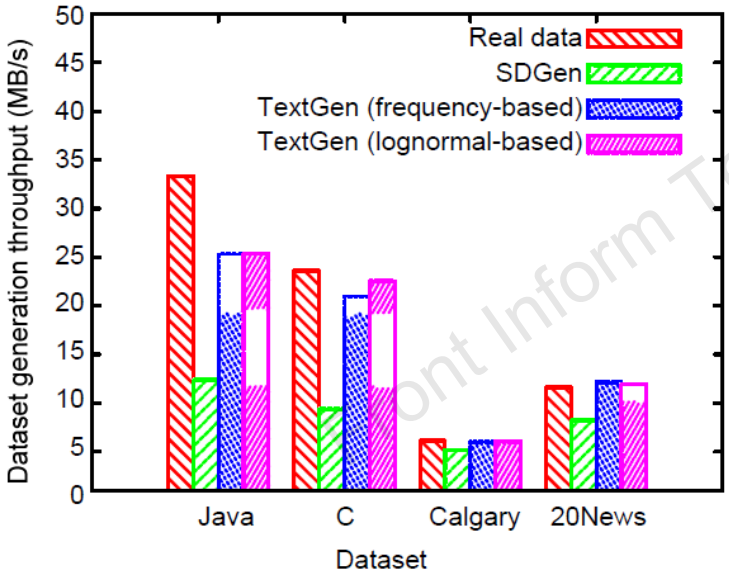

Fig. 13 Comparison of compression throughput by end-tagged dense code

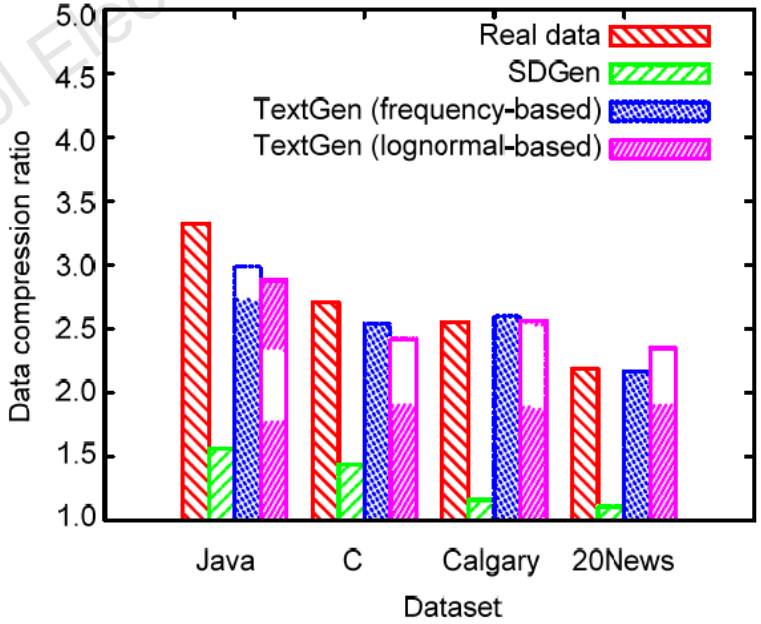

Fig. 15 Comparison of the compression ratio by end-tagged dense code

Conclusions

- We presented a lognormal-distribution-based text data content generation method for modern storage system benchmarks.
- To improve the performance of content generation, the lognormal distribution is fitted to the word-frequency distribution using maximum likelihood estimation (MLE).
- Experimental results show that the synthetic data perform accurately compared with the real data in the ETDC compressor test.