**Web Search Engine: Index Quantization**

Haozhong Zheng

hz2675@nyu.edu

New York University

**ABSTRACT**

**1 INTRODUCTION**

**1.1 Background**

**1.1 Project Overview**

**2 RELATED WORKS**

**2.1 Search Engine**

**2.2 Data Compression**

**2.3 Quantization**

**3 METHODS**

**3.1 Linear Quantization**

**3.2 Logarithmic Quantization**

**3.3 Adaptive Float**

**4 IMPLEMENTATION**

**4.1 C++ Template Programming**

**4.2 Memory Counter**

**4.3 Bits Vector**

**4.4 Quantization**

**5 EXPERIMENT RESULTS**

**5.1**

**5.2**

**6 FUTURE WORKS**

**REFERENCES**

[1] Mallia, A., Siedlaczek, M., Mackenzie, J., & Suel, T. (2019). PISA: Performant indexes and search for academia. Proceedings of the Open-Source IR Replicability Challenge.

[2] Masui, K., Amiri, M., Connor, L., Deng, M., Fandino, M., Höfer, C., ... & Vanderlinde, K. (2015). A compression scheme for radio data in high performance computing. Astronomy and Computing, 12, 181-190.

[3] Lindstrom, P. (2014). Fixed-rate compressed floating-point arrays. IEEE transactions on visualization and computer graphics, 20(12), 2674-2683.

[4] Di, S., & Cappello, F. (2016, May). Fast error-bounded lossy HPC data compression with SZ. In 2016 ieee international parallel and distributed processing symposium (ipdps) (pp. 730-739). IEEE.

[5] Tao, D., Di, S., Chen, Z., & Cappello, F. (2017, May). Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization. In 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS) (pp. 1129-1139). IEEE.

[6] Liang, X., Di, S., Tao, D., Li, S., Li, S., Guo, H., ... & Cappello, F. (2018, December). Error-controlled lossy compression optimized for high compression ratios of scientific datasets. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 438-447). IEEE.

[7] Duwe, K., Lüttgau, J., Mania, G., Squar, J., Fuchs, A., Kuhn, M., ... & Ludwig, T. (2020). State of the Art and Future Trends in Data Reduction for High-Performance Computing. Supercomputing Frontiers and Innovations, 7(1), 4-36.

[8] Huang, C. M., & Harris, R. W. (1993). A comparison of several vector quantization codebook generation approaches. IEEE Transactions on Image Processing, 2(1), 108-112.

[9] Cai, J., Takemoto, M., & Nakajo, H. (2018, December). A deep look into logarithmic quantization of model parameters in neural networks. In Proceedings of the 10th International Conference on Advances in Information Technology (pp. 1-8).

[10] Tambe, T., Yang, E. Y., Wan, Z., Deng, Y., Reddi, V. J., Rush, A., ... & Wei, G. Y. (2019). Adaptivfloat: A floating-point based data type for resilient deep learning inference. arXiv preprint arXiv:1909.13271.