Chunk Content is not Enough: Chunk-Context Aware Resemblance Detection for Deduplication Delta Compression

Xuming Ye [1,†], Xiaoye Xue [1,†], Wenlong Tian [1,2,*], Ruixuan Li [3], Weijun Xiao [4], Zhiyong Xu[5], and Yaping Wan [1,2]

[1] University of South China, China
[2]Hunan provincial base for scientific and technological innovation cooperation China
[3]Huazhong University of Science and Technology, China
[4]Virginia Commonwealth University, USA
[5]Suffolk University, USA

In this paper, we propose a novel chunk-context-aware resemblance detection algorithm called CARD. By introducing machine learning into deduplication, the chunk feature will embed the chunk-context information after the N-sub-chunk shingles based initial feature extraction and BP-Neural network training. In the predicting process, each chunk's initial feature corresponds to a chunk-context feature. Finally, the cloud calculates the different part among resemblance chunks based on these feature by delta encoding. Only the different part is stored. The basic workflow corresponds to Figure 1. For more detailed illustrations, please see our full paper here[1].

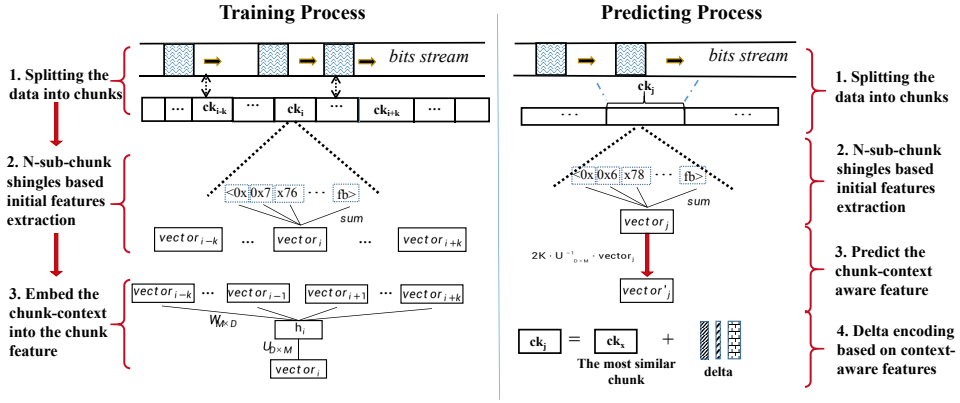

Figure 1: The CARD Workflow

Acknowledgements

This work is supported by the National Natural Science Foundation of China under grants U1836204, Teaching Reform Foundation under grant 2019YB-XJG30, Natural Science Foundation of Hunan Province of China under grant 2021JJ40468, National Training Program of Innovation and Entrepreneurship for Undergraduates under grants 202110555052, S202110555215 and the Ministry of Education Humanities and Foundation on Humanities and Social Sciences under grant 20YJC880027.

† These authors contributed to the work equally and should be regarded as co-first authors
∗ Corresponding Author (Email: wenlongtian@usc.edu.cn)

[1]https://arxiv.org/abs/2106.01273