Abstract: Dynamic Chunking Algorithm Exploiting File Similarity Information

Young Chan Moon[1], Ider Lkhagvasuren[1], Ho Min Jung[1], Wan Yeon Lee[2], Chuck Yoo[3] and Young Woong Ko[1]

[1]*Dept. of Computer Engineering, Hallym University, Chuncheon, Korea*
{ycmoon, ider, chorogyi, yuko}@hallym.ac.kr

[2]*Dept. of Computer Science Dongduk Womens University Seoul, Korea*
wanlee@dongduk.ac.k

[3]*Dept. of Computer Science and Engineering, Korea University, Seoul, Korea*
hxy@os.korea.ac.kr

Abstract

Data deduplication is widely used in storage systems to prevent duplicated data blocks. In this paper, we suggest a dynamic chunking approach using fixed-length chunking and file similarity technique. The fixed-length chunking struggles with "boundary shift problem" and shows poor performance when handling duplicated data files. The key idea of this work is to utilize duplicated data information in the file similarity information. We can easily find several duplicated point by comparing hash key value and file offset within file similarity information. We consider these duplicated points as a hint for starting position of chunking.

With this approach, we can significantly improve the performance of data deduplication system using fixed-length chunking. In experiment result, the proposed dynamic chunking results in significant performance improvement for deduplication processing capability and shows fast processing time comparable to that of fixed length chunking.

Acknowledgement

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2009-0076520)