A Classification of Lossless and Lossy Data Compression Schemes

Lee Chin Kho, S. S. Ngu, A. Joseph, D. A. A. Mat, K. Kuryati

Abstract*: Data compression is a promising scheme to increase memory system capacity, performance and energy advantages. The compression performance could affect the overall network performance when compression scheme is implemented in a communication field. Many data compression schemes have been introduced. Most of other researchers choose very limited parameters to analyze the performance of the selected data compression scheme. This paper classifies the major data compression schemes according to nine different perspectives, such as homogeneity, purpose, accuracy, structuring of the data, repetition distance, structure sharing, number of passes, sampling frequency, and sample size ratio. Various data compression schemes are examined and classified according to the parameters mentioned above. The classification will provide researchers with the in-depth insight on the potential role of compression schemes in memory components and network performance of future extreme-scale systems.

Keywords* : *Data Compression, Lossless, Homogeneity, Accuracy.

I. INTRODUCTION

Data compression is the process of converting data into another format that requires less storage (more efficient) than the original format with some satisfactory accuracy. It is considered as one of the information encoding techniques. Sometimes it is also referred to as bit-rate reduction when applied in networking. In 1848, Morse Code was introduced, which is considered to be the first modern data compression [1-3].

Data compression theory is an extension of the basic information theory, same as any other encoding. Data compression is mainly focused on statistical inference information theory. Compression can be either lossless or lossy. Compressed data from lossless compression must be decompressed to exactly its original value. Compression belongs to algorithmic information theory category for lossless compression and rate-distortion information theory for lossy compression. Both areas were established by Claude Shannon, in the late 1940s and early 1950s as the base for all communication, signaling and data handling [4, 5]. Shannon Fano (SF) coding was the first compression scheme built based on information theory.

When a data unit or sequence of units from the source is com-pressed, the resulting compressed representation will be referred to as the representing code within this study. The representing code should achieve the desired target, either faster transmission, lesser storage or energy consumption, and the degree of compression based on unit by unit individually, which could vary and rarely considered by itself. The aggregate overall compression degree is more significant for the whole data together with the overall performance especially when compressing data from different sources.

Designers of data compression schemes have to handle tight trade-off between the conflicting targets. Those targets are the degree of compression, and the computational resources required (time, temporary storage and energy). Lossy compression faces an additional target that is the amount of distortion introduced, which is highly dependent on the degree of compression. The suitable position in the trade-off limited space is usually decided during design or implementation to target the application of the compression scheme according to a specific situation [3, 6]. When data compression is applied in computer networks, compression is traditionally activated manually by the user at the sender end devices before transferring. When the receiver end devices receive the compressed data, the user again activates decompression manually. Sometimes data compression function is embedded in computer applications or lower layers of end devices. Either way, compression has been largely restricted to end-to-end use. The compression performance could affect the overall network performance.

The objective of this study is to classify the major data compression schemes based on different perspectives. There are 11 common compression schemes such as Joint Photographic Experts Group (JPEG), Run Length Encoding (RLE), Huffman, Arithmetic, Lempel Ziv (LZ), Van Jacobson's Header (VJHC), IP Header Compression (IPHC), Compression Real-time Transport Protocol (CRTP), RObust Header Compression (ROHC), Adaptive Compression-based Technique (ACT), and Lightweight DeCompression (LDC) are considered here. The rest of this review is organized as follows. Section 2 discusses the related work and motivation. Section 3 explains the perspectives that use to classifying of compression schemes and Section 4 discusses the listed compression schemes. The classification results are illustrated in Section 5 and conclusion in Section 6.

Revised Manuscript Received on January 5, 2020

Lee Chin Kho, Department of Electrical Engineering, University Of Adelaide, Adelaide, Australia

Sze Song Ngu, Department of Electrical Engineering, University Of Adelaide, Adelaide, Australia

Annie Joseph, Senior Lecturer, Department of Electronic, University Malaysia

Dayang Azra Binti Awang Mat, Senior Lecturer in Faculty of Engineering, University Malaysia

Kuryati Kipli, Department of Electrical and Electronics, Faculty of Engineering, UNIMAS