1. **Wuthikrai Tharachatr,** Export Logistics Specialist at a chemical company in Thailand.
2. **Asst. Prof. Dr. Thotsapon Sortrakul,** Director of Master of Science Program, Assumption University.

Efficient Bitstream Compression using Historical-Based Grouping and Size-Based Coding Rearrangement

Abstract— **Data storage becomes one of the important factors for most of large enterprises in increasing the cost, chip area, and dissipation. Large enterprises usually create redundancy of repeated digital messages through communication and documentary. In order to minimize these kinds of redundant bitstream, this paper presents a novel algorithm for lossless data compression by developing Historical-Based grouping and Sized-Based Coding Rearrangement approaches to reduce the code size used in compress redundant data flows. The Historical-Based grouping technique improves efficiency of data compression by implementing knowledge-based dictionary that is adaptive and can analyze the existing text coding with the incoming text by considering frequency of same text that has been compressed. In this case, if a compressed text besides that particular compressed text is the same, then this algorithm will group them into a "phase" by using just a particular compression code or symbol to identify that phase. This technique can also efficiently reduce compressed data size by extending a particular compression text length from "word" to "phase", from "phase" to "sentence", from "sentence" to "paragraph", indefinitely depended on repetition of the same binary patterns and amount of available memory allocated to store it. This technique is also able to group various groups of duplicated texts into a single code if those text groups are located in the same text area, same length of space between groups of texts, and set special symbols to link among each group of texts if space lengths are not the same. The objective of this technique is to analyze the historical coded data to minimize the coded size either to group or not to group the same binary patterns based on their frequencies. However, coding of the same binary patterns by using only the knowledge-based dictionary approaches might not be able to efficiently minimize the compressed patterns since frequencies of specific patterns are different. The data that being encoded should be frequency-based analyzed and grouped as long as possible by using Historical-Based Grouping approach, then the compressed output should be rearrange based on their coding sizes and text lengths in which the result of total bit size should be concerned and minimized. Normally, longest or mostly-use coded text should be matched with the shortest code or symbol to ensure that every codes are efficiently utilized and total memory space requirement for that particular file is minimized.**

I. Introduction

Due to dramatic growth of data storage and transfer for most enterprises nowadays, it is necessary to have larger memories to store numbers of binary generated from everyday communication and documentary. Especially, for large or multinational enterprises where various parts of digital communication and documentary are duplicated and redundant, many data compression techniques are used to reduce the storage requirement by compressing the data binaries. The techniques offer an approach to reduce communication and save storage costs by using available bandwidth and memory space efficiently.

In term of analyzing the efficiency of data compression, compression ratio is widely accepted in measuring the efficiency, which is defined as:

$$Compression\ Ratio = \frac{Compression\ Size}{Original\ Size}$$

Data compression can be categorized into two main techniques consisting of Lossless Data Compression and Lossy Data Compression. Lossless Data Compression can decode the data exactly what it encoded. It is normally used for all kinds of text, scientific and statistical databases, and so on to ensure that the decoded output will be exactly identical to the original data. For Lossy Data Compression, the compressed data will be slightly distorted in which it is mostly used for digital sound or image compression where the distortion is