School of Computer Science
University of Petroleum & Energy Studies, Dehradun

Synopsis Report (2018-19)

PROJECT TITLE

Lossless Data Compression Algorithms and their Comparison.

ABSTRACT

In today's world, With the advent of the Internet and mobile devices with limited resources and with the growing requirements of information storage and data transfer, Cloud Computing has become an important aspect, but cloud computing also require physical infrastructure, somewhere down the lane. This exponential sub purge of data leads to high demand for data processing that leads to a high computational requirement which is usually not available at the user's end. Compression reduces the redundancy in data representation thus increasing effective data density. [1] Data Compression is a technique which is used to decrease the size of data. This is very useful when some huge files have to be transferred over networks or being stored on a data storage device and the size is more than the capacity of the data storage or would consume so much bandwidth for transmission in a network. With the limited physical infrastructure for storage, data compression has gained even more importance these days. There are number of data compression algorithms, which are dedicated to compressing different data formats. Even for a single data type, there are number of different compression algorithms, which use different approaches. In this project, we will examine lossless data compression algorithms like Huffman encoding algorithm, Lempel-Ziv-Welch algorithm, and Shannon-Fano algorithm and comparing their performance. [2]

Keywords: Cloud Computing, Data Compression, Huffman encoding algorithm, Lempel-Ziv-Welch algorithm, Shannon-Fano algorithm.

INTRODUCTION

Compression is the art of representing the information in a compact form rather than its original or in uncompressed form [3]. In other words, using the data compression, the size of a particular file can be reduced. This is very useful when processing, storing or transferring a huge file, which needs lots of resources. If the algorithms used to encrypt works properly, there should be a significant difference between the original file and the compressed file. When data compression is used in a data transmission application, speed is the primary goal. Speed of transmission depends upon the number of bits sent, the time required for the encoder to generate the coded message and the time required for the decoder to recover the original message. In a data storage application, the degree of compression is the primary concern. There are two types of data compressions ie. Lossless data compression and Lossy data compression.

Table 1.0 :Comparison between Lossy and Lossless Data Compression Technique.

Lossless data compression	Lossy data compression
In Lossless data compression algorithms, the original data can be recovered from compressed data after applying decompression algorithm	In Lossy compression algorithms it permanently reduces the original data by eliminating certain information, especially redundant information, after decompressing the file only the part of original data is recovered.
Lossless compression is generally used for text data or spreadsheet files, where even a very small amount of data loss can be detected by users.	Lossy compression is generally used for video and sound, where a certain amount of information loss will not be detected by most users.
No loss in information so compression rate is small.	In return for accepting this distortion in reconstructed data we obtain high compression rate.
Less data can be accommodated in channel.	More data can be accommodated in channel.
E.g(*i*)*Fax Machine*,(*ii*)*Radiological Imaging*	*E.g*.(*i*)*Telephone System*,(*ii*)*Video CD*

Various lossless data compression algorithms have been proposed and used. Some of the main techniques in use are the Huffman Coding, Run Length Encoding, Arithmetic Encoding and Dictionary Based Encoding [4]. We will examine the performance of the Huffman Encoding Algorithm, Shannon Fano Algorithm, and Lempel Zev Welch

Algorithm. In particular, performance of these algorithms in compressing text data will be evaluated and compared.

Shannon-Fano Algorithm:

In Shannon–Fano Algorithm, the symbols are arranged in order from most probable to least probable, and then divided into two sets whose total probabilities are as close as possible to being equal. All symbols then have the first digits of their codes assigned; symbols in the first set receive "0" and symbols in the second set receive "1". As long as any sets with more than one member remain, the same process is repeated on those sets, to determine successive digits of their codes. When a set has been reduced to one symbol, of course, this means the symbol's code is complete and will not form the prefix of any other symbol's code. The algorithm works, and it produces fairly efficient variable-length encodings; when the two smaller sets produced by a partitioning are in fact of equal probability, the one bit of information used to distinguish them is used most efficiently. Unfortunately, Shannon–Fano does not always produce optimal prefix codes. [5]

Huffman Encoding:

The Huffman encoding works on variable length encoding rather than fixed length encoding. frequency of every character is calculated, and the lengths of the assigned codes are based on the frequencies of corresponding characters. The most frequent character gets the smallest code and the least frequent character gets the largest code. The variable-length codes assigned to input characters are Prefix Codes, means the codes (bit sequences) are assigned in such a way that the code assigned to one character is not prefix of code assigned to any other character. This is how Huffman Coding makes sure that there is no ambiguity when decoding the generated bit stream for this task a binary tree is created using the symbols as leaves according to their probabilities and paths of those are taken as the code words.

Following are the steps of algorithm-

1. Count the frequency of each character in a text file to be encoded.
2. Create a node of each different character and store them in the queue in ascending order of their frequency.
3. Build a tree by removing the first two elements of the queue and create a new node by joining those two nodes, keeping the first node on left and second on right and then the

weight of new node will be the sum of those two nodes then add the new node formed in the queue.

4. Complete building the tree.
5. Assign 0 to left edge and 1 to right edge at every level.
6. Now we can write the Huffman code for each character using the tree and combination of 0's and 1's.

The Lempel-Ziv Welch Algorithm:

It is a dictionary-based compression algorithm As in Dictionary the set of all possible word of a language is stored similarly in LZW algorithm a dictionary is used to store or index the previously used string patterns In the compression process those index value is used instead of repeating. The dictionary is created dynamically in the compression process and no need to transfer it with the encoded message for decompressing. In the decompression process, the same dictionary is created dynamically. Therefore, this algorithm is an adaptive compression algorithm.

Performance of compression algorithm [6] is based on space efficiency and the time complexity

The compression behaviour of algorithm is dependent on redundancy of symbol in source file therefore it is difficult to measure the performance of compression algorithm. There are some following measurements used to evaluate the performances of compression algorithms.

Compression Ratio –Compression Ratio is the ratio between the size of the compressed file and the size of the source file

Compression ratio= (size of compressed file)/ (size of source file)

Compression Factor-Compression factor is inverse of Compression ratio It tells how much time our file has been compressed

Compression Factor= (size of source file)/ (size of compressed file)

Saving Percentage-It tales the shrinkage of the source file in percentage

Saving Percentage= (size before compression-size after compression)/ (size before compression) %

Time complexity –The time complexity is measured by the number of clocks used to encode or decode the source code. The algorithm that uses less clock cycle to encode or decode is considered more efficient.

Code Efficiency Average code length is the average number of bits required to represent a single code word. If the source and the
lengths of the code words are known; the average code length can be calculated using the following equation.

$$\bar{l} = \sum_{j=1}^{n} p_j l_j$$

, where p_j is the occurrence probability of j th symbol of the source message, l_j is the length of the particular code word for that symbol and $L = \{l_1, l_2, ... l_n\}$.

In this project we will be comparing all the three algorithms on the basis of these factors mentioned above.

LITERATURE REVIEW

Data Compression is the way that you can use the space on cloud i.e. Server in an optimal way. In this project we will Lossless Data Compression algorithms which can reconstruct the original message exactly from the compressed message Here is the conclusion of some of the reference paper that we review to make our project better and to know more technologies that we can use in our system.

- In the paper [6] by S.R. KODITUWAKKU, Department of Statistics & Computer Science, University of Peradeniya, Sri Lanka, U. S. AMARASINGHE, Postgraduate Institute of Science, University of Peradeniya, Sri Lanka; Among the available lossless compression algorithms they considered the Run Length Encoding Algorithm, Huffman Encoding, The Shannon Fano Algorithm,

Arithmetic Encoding, The Lempel Ziv Welch Algorithm for study. They carried out an experimental comparison of a number of different lossless compression algorithms for text data. On the basis of compression times, decompression times and saving percentages of all the algorithms, they found that the Shannon Fano algorithm can be considered as the most efficient algorithm among the selected ones. The values which they calculated are in the acceptable range and it also shows better results for the larger files.

- In the paper [7] by Laxmi Shaw, Student Member, IEEE, Daleef Rahman, and Aurobinda Routray, Senior Member, IEEE; The authors have examined the different lossless compression methods for single and multichannel EEG signals, and their performance with respect to their relative Compression ratios has been analysed. They evaluate their proposed algorithms, analysis of which showed that a very high CR (Compression Ratio) in different publicly available database. They also analysed that among the existing methods for the single-channel EEG compression scheme, the linear prediction followed by the context-based error modelling showed the best results. The increase in CR by applying the context-based error modelling is high for the first-order predictor, whereas the increase is small for higher-order predictors. The MVAR model and the bivariate autoregression model were examined for the multichannel EEG compression. The results show that these proposed methods in combination outperform the existing MVAR and the bivariate autoregression model.

- In the paper [8] by Mohammad Hosseini, Network Systems Lab; In this research paper author introduced two types of compression, lossless and lossy compression, and some major concepts, algorithms and approaches in data compression and discussed their different applications and the way they work. They also evaluated two of the most important compression algorithms based on simulation results. Then as his next contribution, he thoroughly discussed two major everyday applications regarding data compression; JPEG as an example for image compression and MPEG as an example of video compression in our everyday life. At the end of this survey he discussed major issues in leveraging data compression algorithms and the state-of-the art research works done regarding energy saving in topworld-discussed area in networking which is Wireless Sensor Networks.

PROBLEM STATEMENT

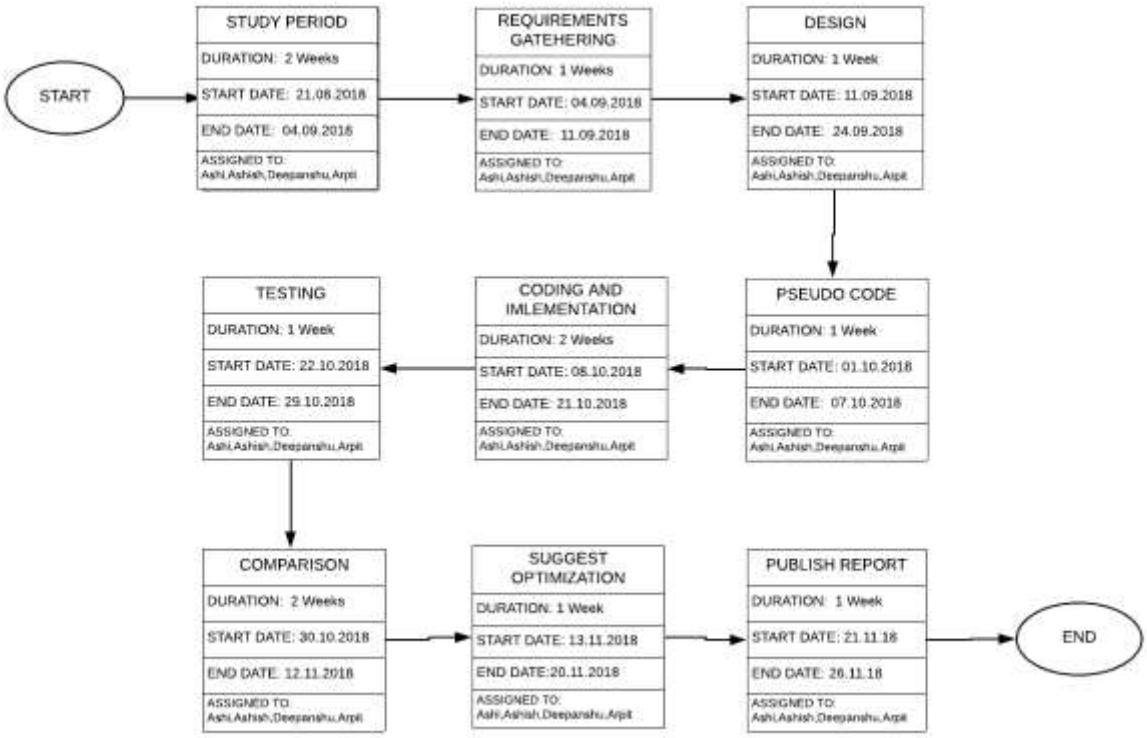
Storage on the cloud is a limited resource. Even though more storage space can be purchased, it seems better to utilize the given space to the fullest. The solution to this problem is data compression. Compress data to save space and then store it on the cloud, also in doing so, we save the data transmission cost over the network and make our cloud storage even more efficient.

SYSTEM REQUIREMENTS

- **Hardware Interface:**
  - 64 bits processor architecture supported by windows.
  - Minimum RAM requirement for proper functioning is 8 GB.
  - Required input as well as output devices.

- **Software Interface:**
  - C Compiler (GCC).
  - AWS CLOUD SERVICES.
  - Socket Programming Library in C.

SCHEDULE (PERT CHART)



REFERENCES

[1] Monika Soni , Dr Neeraj Shukla “Data Compression Techniques in Cloud Computing”

[2] Mohammad Hosseini “A Survey of Data Compression Algorithms and their Applications”

[3] Pu, I.M., 2006, Fundamental Data Compression, Elsevier, Britain.

[4] Kesheng, W., J. Otoo and S. Arie, 2006. Optimizing bitmap indices with efficient compression, ACM Trans. Database Systems, 31: 1-38.

[5] https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=15&ved=2ahUKEwjW8piM_q_dAhVHyrwKHXE2DHoQFjAOegQIABAC&url=http%3A%2F%2Fecehithaldia.in%2Fteaching_material%2FShanon-Fano1586521731.pdf&usg=AOvVaw0MHM4foSS-sDhzqyRAVfaE

[6] S.R. Kodituwakku ,U. S. Amarasinghe “Comparision of Lossless data compression algorithms for text data”

[7] Highly Efficient Compression Algorithms for Multichannel EEG,Laxmi Shaw , Student Member, IEEE , Daleef Rahman, and Aurobinda Routray, Senior Member, IEEE

[8] A Survey of Data Compression Algorithms and their Applications ,Mohammad Hosseini

[9] Network Systems Lab, School of Computing Science, Simon Fraser University, BC, Canada,Email: mohammad hosseini@sfu.ca