Combined Data Compression and Error Correction

A.Sasikumar and M.A.Ambroze

Fixed and Mobile Communications, University of Plymouth, Plymouth, UK
e-mail: M.Ambroze@plymouth.ac.uk

Abstract

The basic intention behind the project is to understand and familiarize with the concepts of data compression using LZW and encoding, error correction using Hamming codes. After achieving the desired results by creating algorithm for compression and error correction, gives a clear idea how it works on a text file. For a detailed analysis of compression, did study on comparing the rate of compression of WinZip and WinRAR.

Keywords

Data Compression – Lzw; Error Correction- Hamming Code

1 Introduction

In the field of information the quality of service of any data depends mainly on 1) the transferring data speed and 2) the rate of error (Shannon, 1949). The data transmission rate through a medium depends on a) the bandwidth of the channel and b) the size of the data. For the purpose of improving the rate of transmission, any of the above two parameters must be altered; and in most of the cases uses the second method decrease the size of information by compressing it. Due to compression there are mainly two advantages it reduces the time of transmission, and also shrink the storage space. And in the second part error, the chance of occurring error when the data is pass through a channel is more because of the effect of noise in that field. Due to those error factors the quality of the data will be compromised. So error correction is an important part in the field of data transfer. Hence in the field communication, both compression and error correction are two non-evitable factors. So the combination of both can able to transfer the data at high speed without so much of errors.

1.1 Aim and Objectives

The main aims of this project are; 1) to study the concepts of data compression and error correction in the field of communication. 2) Investigate the best combination of compression and error correction that gives lustiness to errors and attaining a diminution to the size of the data that to be transmitted.

The main objectives in the project are; a) Understanding the concept of compression, decompression, encoding and error correction b) Choosing proper method for data compression and error correction. c) Creating algorithm using

selected techniques. d) On the basis of those algorithms, find out the results, analyse it and draw graphs for it.

The programming part of the project is mainly split into three parts a) Data compression and decompression b) Encoding and error correction c) Combined data compression and error correction.

2 Background

2.1 Data Compression

The data compression technique starts in the year 1948 with Claude E.Shannon by his paper 'A Mathematical Theory of communication'. Shannon mainly explains about two types of compression lossless and lossy compression. Lossy compression is also known as rate distortion theory. Rate distortion theory and lossless data compression theory are jointly known as 'source coding theory' (Shannon, 1949). There are different methods are available for data compression and in that LZW was choose as the technique for this project. LZW is an advanced form of LZ78 created by Terry Welch in the year 1984. LZW is work on the basis of the dictionary. In LZW it creates its own dictionary on the basis of the input text, and these dictionary words were used for representing redundant data in the text. And at the decompression part it adds the redundant data on the basis of the dictionary that created at the time of compression (Welch, 1984).

2.2 Error Correction

Error correction is the technique that is used of correcting the errors that occurred while transferring the data through a channel. So these error correction coding provides a required level of exactness; as similar as that of the actual data. Accuracy of system can also attain by increasing the signal strength per unit of data, but the main advantage of using error correction to data is the accuracy can be achieved without any change in power (Prakash, 2006). There are different techniques available for error correction and hamming code is one among them. In the field of telecommunication hamming code is considered to be a linear error correcting code. The hamming codes were invented by Richard Hamming in the year 1950. These codes can able to detect two bits of errors and can correct single-bit error. A reliable form of communication can be achieved when hamming distance (hamming distance is defined as the bits position (error) difference between two files of same length) among the transmitted bit pattern and received bit is equal to or less than one, which means burst error is not occurred at the medium of transmission. But in demarcation, the simple parity code is not able to correct errors can only to detect the odd sequence of errors (Hamming, 1950).

3 WinSip WinRAR Comparison Study

The compression study then led me to study and compare different aspects compression tools like WinZip 12 and WinRAR 3.70.

Both WinZip and WinRAR are used for compression but there are lot of differences in their properties. And some of their properties are explained below.

The given below are some of the difference between WINZIP and WINRAR. When go through different aspects about WinZip and WinRAR, found that the compression ratio of WinRAR is slightly more than WinZip; so for proving that fact, compress different files using WinZip and WinRAR and find out their respective rate of compression. And those studies and its results are explained in this part.

3.1 Differences between WinZip and WinRAR

	WinZip	WinRAR
Multi OS Support	Only for Windows	Windows,Linux, Mac, DOS, OS/2
Compress Formats	ZIP	RAR, ZIP
Method Used	Deflate method(combination of LZ77 and Huffman coding)	LZ and Prediction by Partial Matching (PPM)
Dictionary Size	Default	Can Set Dictionary Size
Profile For Compression	Not Applicable	Create Our Own Profile For Compression
Extraction	ZIP,RAR,TAR,JAR,BZIP2,UUE,Z,CAB	RAR,ZIP,TAR,JAR.BZIP2,UUE,Z,CAB,ARJ,ACE,LZH,GZIP,7-ZIP,ISO.

Table 1: Difference between WinZip and WinRAR (WinRAR, 2009)

For this study; compressed around fifty to sixty different size files in default settings of both WinZip and WinRAR and compare their respective ratios, and found that the WinRAR compress a file more than that of WinZip. So for a deep analysis of the performance of WinZip and WinRAR, try to compress different format of file like .txt file, .jpg file, .bmp file, .wmv file and .mp3 files; and find out their respective rate of compression. And their respective graphs were plotted and those results are explained below.

3.2 Text File (.txt File)

In this section, using WinZip and WinRAR compress different text files of variable size, the below graph drawn on some of the files sizes and their compression ratios that did for the study, more than fifty files of different sizes were compressed and determine the compression ratios and compare their respective ratios of WinZip and WinRAR, and their result shows that the compression ratio was slightly more for WinRAR than WinZip and the ratio goes on increasing with the file size.

The following graph plot, file size on X axis and compression ratio on Y axis and the graph shows that the difference in compression ratio increases with the increase in file size. And also it shows; the files that were compressed more in WinRAR than WinZip, and their respective ratio was goes on increasing with the file size.

3.2.1 File Size vs Compression Ratio for Text Files

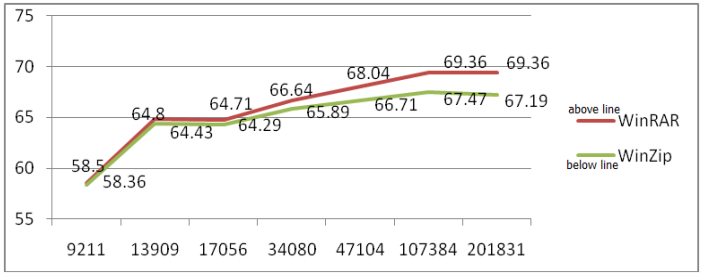

Figure 1: WinZip vs WinRAR for .txt files

3.2.2 Result

WinRAR compress text files more than that of WinZip.

3.3 Image (.JPG File)

In image compression, when compress different size of images; their result shows that the compression for images were more in WinZip than in WinRAR. But the difference is not as much as that in text files compression, even though their differences in size were almost as thousand bytes for both type of compression. The compression ratios for these .jpg images were very less, because these .jpg files were already compressed ones.

In the below graph, original file sizes in KBs on X axis and compressed file sizes in KBs on Y axis. And the graph shows that the WinZip has less number of bytes when compared to WinRAR, which means the compression is more in WinZip than in WinRAR.

3.3.1 Original File Size vs Compressed File Size (in KB)

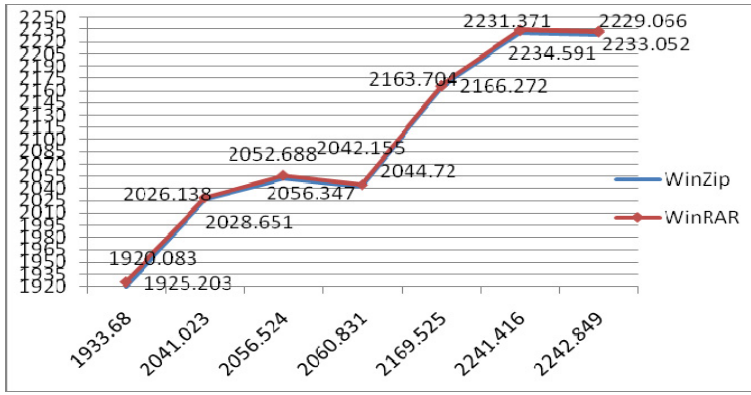

Figure 2: WinZip Vs WinRAR for .jpg files

3.3.2 Result

In compression for images (.jpg) WinZip compress more than WinRAR. In the above graph; it is not clearly shows that but we can found the truth by checking the file sizes of each compression. Even though it is only few bytes when compared to the original size, still more than thousand bytes difference were there between two.

3.4 Image Compression (.bmp File)

In WinZip, WinRAR compression study, next study was done on .bmp images, because in .jpg image files were already compressed ones so it was not clear about the idea of compression. For that different size .bmp images were downloaded from different website and did compression with both WinZip and WinRAR. In graph shown below, image sizes on X axis and compression ratios on Y axis, and when comparing their respective ratios shows that the compression ratio is more in WinRAR compared to WinZip.

3.4.1 WinZip vs WinRAR for .bmp IMAGES

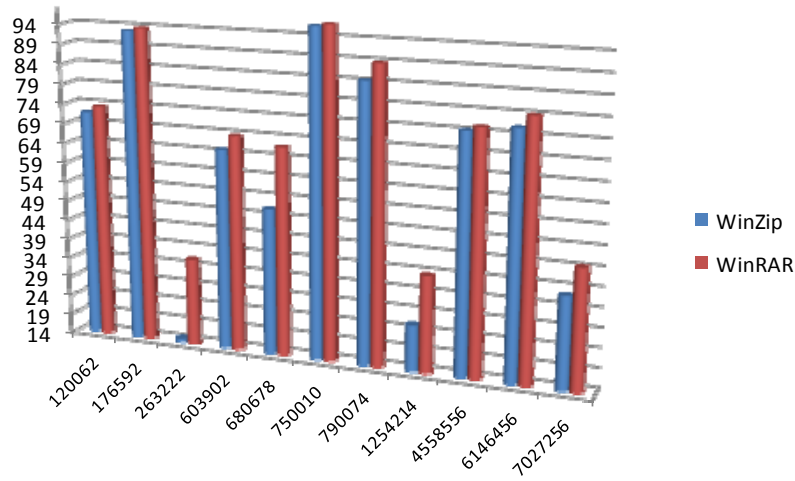

Figure 3: WinZip Vs WinRAR for .bmp files

3.4.2 Result

After comparing different .bmp images of variable size, the respective results shows compression rate is more for WinRAR than WinZip.

3.5 MP3 and Video

The next study of compression was on video and mp3 files, when compared different files of both of variable sizes; the compression ratios got in a varying mode, means in some mp3 and video files WinZip compress more than WinRAR and in some vice

versa. So it could not able to conclude which one is better for these mp3 and video files.

3.6 Conclusion FOR WinZip WinRAR Study

These studies were done on the basis of compression result of few files (around fifty files); but in some cases the result may vary because of the structure and content of the file. So the above results were written in a generalised way, so it means in some cases it may happen vice versa too.

4 Combined Compression and Error Correction

4.1 Result and Analysis

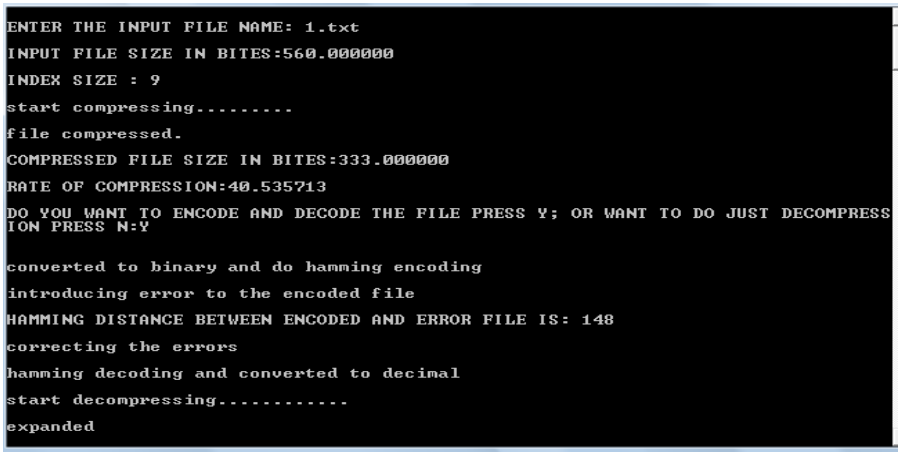

Figure 4: Result of the program

The main outputs of the program are 1) the size of the input file: which shows the number of bytes in the input for compression 2) the size of the output file: which shows the number of bytes in the compressed file 3) rate of compression: which indicate the rate at which it compress the input file 4) index size: it represent the bit size for compression 5) hamming distance: it shows the number of bit position difference between encoded file and error file (or number of errors) and these represent the main output of the program, and when compiling the program respective compressed, encode, error, correct, decode, decompress file were created which represent the output of each step. Here below shows different graph for the algorithm.

The below graph shows file size verses compression ratio of a normal text file. Graph shows that the compression ratio reach up to 25.6%. This means it compresses that particular file up to 25% of its original size.

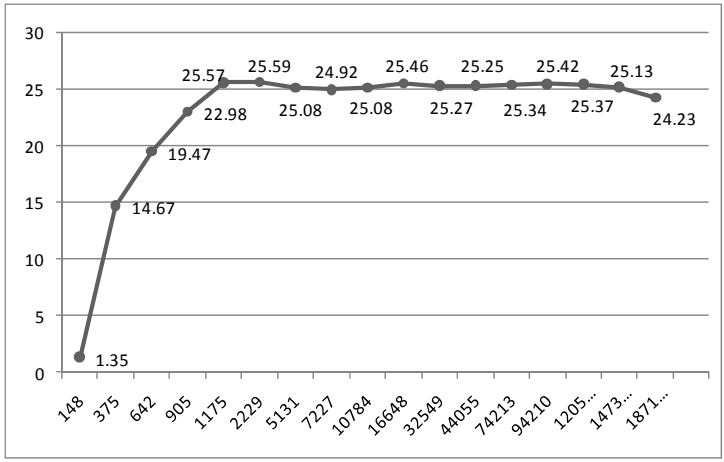

Figure 5: Graph for file size vs compression ratio

In the below graph; plotted with file size on X axis verses compression ratio for files of more repeated text on Y axis and the result shows the compression ratio reaches up to 85%. So by comparing the first graph and the below graph we can state that the rate compression of a file using LZW is depends on the content of the file (repeating sentences, words and symbols).

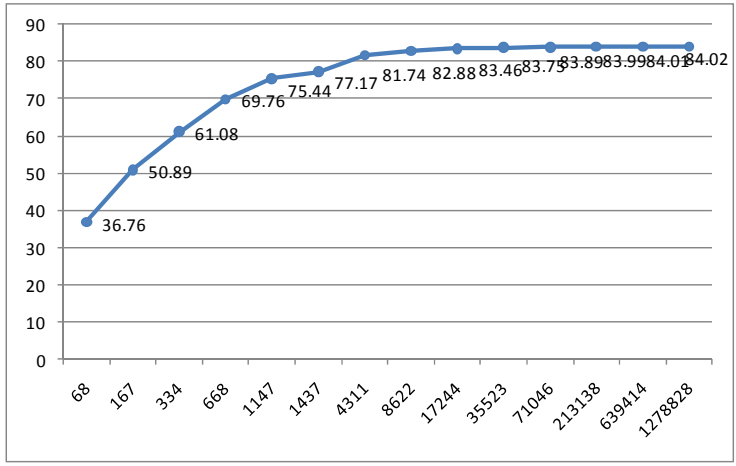

Figure 6: Graph for repeated text vs compression ratio

In next graph it shows about how compression rate changes with index size. Here graph was plotted compression rate verses size of the file. For a same file it compress with different index sizes and the result shows that for a small files when it compress the file with index size 9 shows maximum compression rate and then it decreases for the other index sizes. But considering a large file, the case was different the compression rate increases with increase in index size, i.e. minimum compression

rate when index size 9 and for every other index sizes it shows an increment in compression rate.

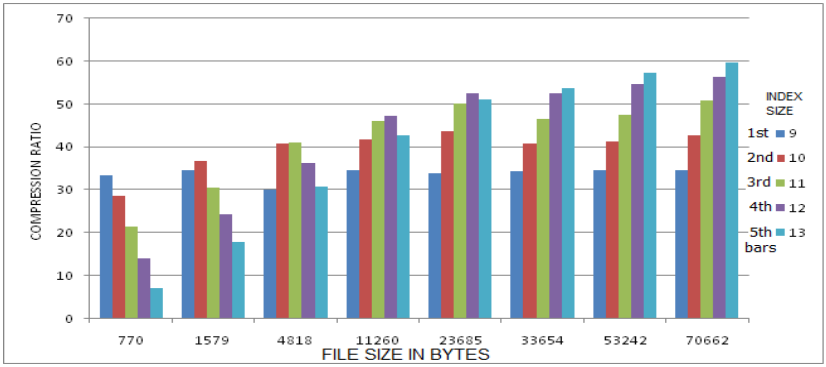

Figure 7: Graph for file size vs rate of compression for variable Index size

5 Conclusion and Future Work

The important aspects of this project is to familiarise with the concepts of data compression and error correction, after detailed study of two, give a clear cut idea about how it operates in the field of communication system. Since there are different techniques available for both the methods; so it is very difficult to say that which combinations produce the best result. Within a limited period of time; tries to find out the best and easy way to implement methods for compression and error correction technologies. Both LZW and Hamming code developed using C algorithm, and the results were achieved, but still the result not accomplished in a desired way, even though these result gave a good idea about how these technique works on a text file. The main concepts of LZW are; it compresses a text according to the content of the file and the rate of compression changes concordant with index size; these two factors were practically proved using with the help of algorithm and their respective graph were plotted, those graph were give a clear idea about these two concept of LZW compression. Another thing achieved by this project was about, how the encoding and error correction process done by using hamming code (11, 7). Next goal achieved was the understanding of effect of error in a compressed and encoded data, for those studies some files were compressed manually and program vice then introduce some errors to the file by deleting some bits from it and do the decompression and the effect of error for both compressed and encoded file are so desolating one. So error correcting is a significant factor in the field of communication.

So as mentioned early both compression and error correction are two non-evitable factors in the field of information theory. Proper compression technique helps the data to be transmitted in a faster manner and also reduce the storage space; the error correction technique assists the data to be retrieve accurately as possible as the original data that sent by the transmitter. So if both these technologies implement properly, i.e. the best combination of both give a great result to the field of communication. Compression and error correction technologies have a lot of

alteration over past few years, and still researches are going on these technologies. And it is very difficult to predict which combination of both gives a perfect result, because these technologies changes day by day. So we can hope in near future the best combination of both will find and help to transfer data more than twice or thrice faster and accurately as compared to today's technology.

6 References

Dipperstein, M. (2008), "Lempel-Ziv-Welch (LZW) Encoding Discussion and Implementation", http://michael.dipperstein.com/lzw/index.html#strings (Accessed 20 August 2009)

Hamming, R.W. (1950), "Error Detection and Error Correction Codes", Bell Systems Tech. Journal, Vol. 29, pp 147-160

Nelson, M. (1989), "LZW Data Compression", http://marknelson.us/1989/10/01/lzw-data-compression/ (Accessed 20 August 2009)

Prakash, A., Singh, A. K., and Tiwari, M. (2006), "Digital Principles and Switching Theory", New Age International (p) ltd, India, ISBN-10 : 812242306X

Shannon, C. E. (1949), "A mathematical theory of communications", University of Illinois Press.

Welch, T.A. (1984), "A Technique for High Performance Data Compression", IEEE Computer, Vol. 17, No. 6, pp. 8-19.

WinRAR Web Site (2009),"WinRAR 3.70 vs. WinZip 11.1 comparison", http://www.win-rar.com/winzipcomparison.html (Accessed 20 August 2009)