ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 2, February 2013

A Survey on the different text data compression techniques

Rajinder Kaur
Student (M.Tech-CSE)

Mrs. Monica Goyal
Assistant Professor (CSE)

Guru Kashi University, Talwandi Sabo (Bathinda)

Abstract: **-** Compression is useful because it helps us to reduce the resources usage, such as data storage space or transmission capacity. Compression methods have a long list. In this paper, we shall discuss only the lossless compression techniques and not the lossy techniques as related to our work. In this, reviews of different basic lossless data compression methods are considered. The methods such as Shannon-Fano Coding, Huffman coding, Run Length Encoding and Arithmetic coding are considered. Lempel Ziv scheme is also considered which a dictionary based technique. A conclusion is derived on the basis of these methods.

Keywords- Data compression, Lossless compression, Huffman Coding

1. Introduction:-

Compression is the conversion of data in such a format that requires few bits usually formed to store and transmit the data easily and efficiently. Compression is done to reduce amount of data and needed to reproduce that data. And the compression is done either to reduce the volume of information in case of text, fax and images or reduce bandwidth in case of speech, audio and video.

We show the compression method with the help of following figure 1:

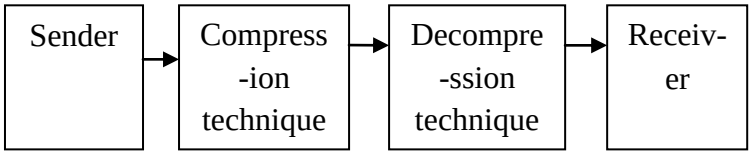

Fig 1: diagrammatic representation of Compression

A data compression is that it converts a string of characters in another representation into a new string which have the same data in small length as much as possible. Data compression may be viewed as the study of information theory in which the main objective for the efficient coding and to minimize the speed of transmission bandwidth. The main purposes of this paper to shows the variety of various lossless compression techniques and their comparative study.

1.2 Classification of compression methods

We have two types of compression methods:

Lossless compression: - It is used to reduce the amount of source information to be transmitted in such a way that when compressed information is decompressed, there is not any loss of information.

Lossy compression: - The aim of lossy compression is normally not to reproduce a exact copy of the information after

decompression. In this case some information is lost after decompression.

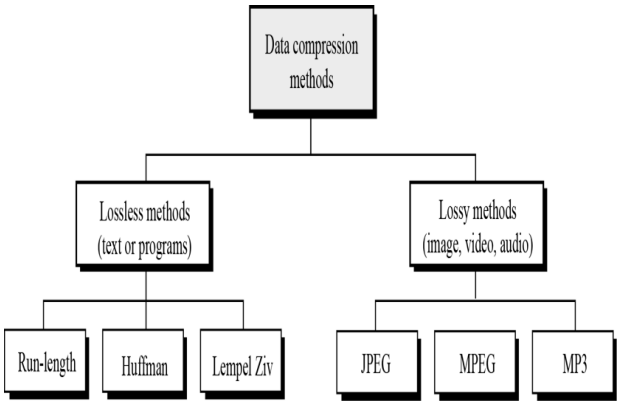

Fig: - Tree representation of compression methods

But in this paper we concentrate only on the lossless compression methods which are used on the text data formatting and define them with the help of some algorithms. The question of the better methods from, "lossless" or "lossy" is that they have its own uses with lossless techniques better in text formats and lossy technique better in images, audio and video. There are some lossless compression techniques which we are now explain in our paper are Run Length coding, Arithmetic Coding, Huffman coding, Shannon Fano and LZW family methods and compare their performance.

2. Literature Review:-

S. Shanmugasundaram and R. Lourdusamy, "A Comparative Study of Text Compression Algorithms" There are lot of data compression algorithms which are available to compress files of different formats. This paper provides a survey of different basic lossless data compression algorithms. Experimental results and comparisons of the lossless compression algorithms using Statistical compression techniques and Dictionary based compression techniques were performed on text data [1].

Md. RubaiyatHasan, "Data Compression using Huffman based LZW Encoding Technique" A method and system for transmitting a digital image (i.e., an array of pixels) from a digital data source to a digital data receiver. More the size of the data be smaller, it provides better transmission speed and saves time. In this communication we always want to transmit data efficiently and noise free [2].

S. Kaur and V.S.Verma,"Design and Implementation of LZW Data Compression Algorithm "In this paper, LZW data compression algorithm is implemented by finite state machine, thus the text data can be effectively compressed [3].

U.Khuranaand, A.Koul, "Text Compression And Superfast Searching" A new compression technique that uses referencing through two-byte numbers (indices) for the purpose of encoding has been presented. The technique is efficient in providing high compression ratios and faster search through the text [4].

3. Lossless Compression Methods:-

3.1Repetitive Sequence Suppression or Run Length Encoding: -The first step in this technique is read file then it scans the file and find the repeating string of characters [6].when repeating characters found it will store those characters with the help of escape character followed by that character and count the binary number of items it is repeated. This method is useful for image having solid black pixels. This algorithm is also effective for repeating of characters. But it is not effective if data file has less repeating of characters.

We can compress the run-length symbols using Huffman coding, arithmetic coding, or dictionary based methods.

3.1.1 Huffman Coding:- The Huffman coding algorithm is named after its inventor, David Huffman, who developed the method as a student in a class on information theory at MIT in 1950[1].

Huffman Coding Algorithm— it is a bottom-up approach

1. Initialization: Put the old nodes in a list sorted according to their frequency counts.

2. Repeat the following steps until the sorted list has only one node left:

(1) From the list pick two nodes with the lowest frequency counts. Form a Huffman sub tree that has these two nodes as child nodes and create a parent node.

(2) Assign the sum of the children's frequency to the parent node and insert it into the list such that the order is maintained.

(3) Delete the children from the sorted list.

3. Assign a 0 and 1codeword to the two branches of the tree on the path from the root.

After the Huffman tree, the method creates a prefix code for each node from the alphabet by traversing the tree from the root to the node. It creates 0 for left node and 1 for a right node.

3.1.2 Shannon Fano Coding technique:-

It is used to encode messages depending upon their probabilities [1]. The method is defined as given below:-

1. For a given list of symbol create a probability table.
2. Sorting the table based on the probability and places the most frequent symbol at the top of a list.
3. The table is divided into equally two halves upper and lower which having a same probability as much as possible.
4. The upper half of the list defined with '0' digit and the lower half with a '1'.
5. Repeat the steps 3 and 4 for each of the two halves then further divide the groups and adding bits to the codes and stop the process when each symbol has a corresponding leaf on the tree.

3.1.3 Arithmetic coding technique:-Arithmetic coding is change the method of replacing each bit with a codeword. So it replaces a string of input data with a single floating point number as a output. The main purpose of this technique is to given an interval to each potential bit data.

• Arithmetic coding is a more modern coding method that usually than Huffman coding.

• Huffman coding assigns a codeword to each symbol which has an integral bit length. Arithmetic coding can treat the whole string data as one unit.

• A message is represented by a half-open interval [*a*, *b*) where *a* and *b* are real numbers between 0 and 1. Initially, the interval is [0, 1). When the message becomes longer, the length of the interval shorts and the number of bits needed to represent the interval increases.

3.1.4 LZW (Lempel-Ziv Welch) compression method:-

LZW is the most popular method. This technique has been applied for data compression [6]. The main steps for this technique are given below:-

- Firstly it will read the file and given a code to each character.
- If the same characters are found in a file then it will not assign the new code and then use the existing code from a dictionary.
- The process is continuous until the characters in a file are null.

4. Conclusion

There we talked about a need of data compression, and situations in which these lossless methods are useful. The algorithms used for lossless compression are described in brief. A special, Run-length coding, statistical encoding and dictionary based algorithm like LZW, are provided to the concerns of this family compression methods.

In the Statistical compression techniques, Arithmetic coding technique performs with an improvement over Huffman coding, over Shannon-Fano coding and over Run Length Encoding technique. Another area of research would be to implement the compression scheme so that searching is faster.

5. References

[1] S. Shanmugasundaram and R. Lourdusamy, "A Comparative Study of Text Compression Algorithms" International Journal of Wisdom Based Computing, Vol. 1 (3), December 2011

[2]V.K.Govindan and B.S. Shajeemohan,"IDBE - An Intelligent Dictionary Based Encoding Algorithm for Text Data Compression for High Speed Data Transmission Over Internet **"**

[3]P.G.Howard and J.C.Vitter, Fellow IEEE," Arithmetic Coding For Data Compression".

[4] S. Kaur and V.S.Verma," Design and Implementation of LZW Data Compression Algorithm", International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.4, July 2012

[5] U .Khuranaand, A.Koul, **"**Text Compression and Superfast Searching", Thapar Institute of Engineering and Technology, Patiala, Punjab, India-147004

[6] Y. M. Kamir, M. Deris. M. Sufian, and A. A.F. Amri,"Study of Efficiency and Capability LZW++Technique in Data Compression", World Academy of Science, Engineering and Technology 35 2009

6. Bibliographies:-

Rajinder Kaur received the B.Tech degree in Computer Science Engineering from the Yadwindra College of engineering, Punjabi University Patiala, Punjab in 2011, and pursuing M.Tech degree in Guru Kashi University, Talwandi Sabo, Punjab. Currently, she is doing her thesis work on a new data compression technique. Her topic of interest is data compression. Email rajechatha@gmail.com

Monica Goyal received his B-Tech degree in computer science & Engineering from Guru Teg Bhadhur Khalsa Institute of Engineering & Technology College, of Chappianwali, Malout, under PTU, Punjab in 2007, and M-Tech. degree in Computer Science Engineering from Yadwindra college of Engineering and Technology from Punjabi University, Patiala, Punjab. Her research area is in digital image processing. At present, she is engaged in Guru Kashi University, Talwandi Sabo, Punjab as an Assistant Professor in Computer Science Engineering & Information Technology department.Email:monikagoyal84@gmail.com