Submitted by:

Goutham,

Premchand.

**Assignment 2:**

**Compression Techniques:**

The techniques used for representing the information in a compact form are known as compression techniques. It is required because more and more of the information that we generate and use is in digital form required to represent multimedia data can be huge. There exists variety of techniques for data compression. All are based on different ideas and are suitable for different types of data. They compress data by removing redundancies.

The major reasons behind using compression techniques is we can reduce storage requirement, processing time, transmission duration. It allows more efficient way of saving memory, transmission bandwidth and transmission duration.

**Types of compression techniques:**

**Lossless Compression Technique:** In this technique, no data is lost and the exact replica of the original file can be retrieved by decrypting the encrypted file. One of the lossless type compression is text compression. In this type of compression generally the encrypted file is used for storing or transmitting data. For general purpose use we need to decrypt the file.

**Lossy Compression Technique:** This technique is generally used for image, audio, video. Here, the compression process ignores some less important data and the exact replica of the original file can’t be retrieved from the compressed file. To decompress the compressed data we can get a closer approximation of the original file.

**Comparitive Analysis of Compression Techniques:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Compression**  **Technique** | Huffman | Shannon Fano | LZW | Run Length Encoding | Arithmetic Coding | Bit Reduction Algorithm | J-bit encoding | Transform coding | Discrete Cosine Transform | Discrete Wavelet Transform |
| **Type** | Lossless | Lossless | Lossless | Lossless | Lossless | Lossless | Lossless | Lossy | Lossy | Lossy |
| **Functionality** | uses variable length code for encoding a source symbol which is derived based on the estimated probability of occurrence for each possible value of the source symbol | The optimal code length for a symbol is –log b P, where b is the number of symbols used to make output codes and P is the probability of the input symbol. Initially a frequency table is generated and then a particular procedure is followed to produce the code table from frequency | A dynamic dictionary is created depending upon the presence of substring chosen from the original file. Then the substring is matched with the Dictionary | It is a simple and popular data compression algorithm. It is based on the idea of replacing a long sequence of the same symbol by a shorter sequence. | It is a method of replacing each bit with a codeword. Arithmetic coding replaces a string of input data with a single floating point number as an output | The idea is that this program reduces the standard 7-bit encoding to some application specific 5-bit encoding system and then pack into a byte array. | The main idea of this algorithm is to split the input data into two data where the first data will contain original non-zero byte and the second data will contain bit value explaining position of non-zero and zero bytes | It is lossy technique which results in a lower quality copy of the original input. The knowledge of the application is used to choose information to discard, thereby lowering its band width. | A discrete cosine transform (DCT) expresses a finite sequence of data points in the terms of the sum of cosine functions oscillating at different frequencies. | It is an implementation of the wavelet transform using a discrete set of the values scales and translations obeying some defined rules. |
| **Operation** | A table is created incorporating the no of occurrences of an individual symbol. This table is known as frequency table and is arranged in a certain order. Then a tree is generated from that table, in this tree high frequency symbols are assigned codes which have fewer bits, and less frequent symbols are assigned codes with many bits. In this way the code table is generated | Each part of the table will be divided into two segments. The algorithm has to ensure that either the upper and the lower part of the segment have nearly the same sum of frequencies. This procedure will be repeated until only single symbols are left. The original data can be coded with an average length of 2.26 bit. Linear coding of 5 symbols would require 3 bit per symbol. But, before generating a Shannon-Fano code tree the table must be known or it must be derived from preceding data. | If the string is found then a reference of the dictionary is mentioned in the encoded file, if the string is not found then a new dictionary entry is made with a new reference. In all algorithms the encoded file contains the code table/ Dictionary and the encoded text; the encoder matches the codes with the directory and retrieves the original text iteratively | Consider a character run of 20 'B' characters which normally would require 20 bytes to store.  The considered sequence is “BBBBBBBBBBBBBBBBBBBB”.  This sequence can now be compressed to the form “20B”.  With RLE, this would only require two bytes to store, the count (20) is stored as the first byte and the symbol (B) as the second byte | It treats the whole string data as one unit. A message is represented by a half-open interval [x, y) where x and y are real numbers between 0 and 1. Initially, the interval is [0, 1). When the message becomes longer, the length of the interval shortens and the number of bits needed to represent the interval increases | Using this algorithm, about 256 characters per message which is typically 160 characters per message through the same 7-bit GSM network could be send. This program reduces the standard 7-bit encoding to some application specific 5-bit encoding system and then pack into a byte array. | It manipulates each bit of data inside file to minimize the size without losing any data after decoding. This basic algorithm is intended to be combined with other data compression algorithms in order to optimize the compression ratio. The performance of this algorithm is measured by comparing combination of different data compression algorithms | In transform coding , knowledge of the application is used to choose information to discard, thereby lowering its band width. The remaining information can be compressed via number of methods, when the output is decoded, the result may not be identical to the original input, but is expected to be close enough for the purpose of the application. | DCTs are used to convert data in the summation of series of cosine waves oscillating at different frequencies. It expresses a finite sequence of data points in the terms of the sum of cosine functions oscillating at different frequencies. There are very similar to Fourier transforms but DCT involves uses of cosine functions are much more efficient as fewer function are needed to approximate a signal | This transform decomposes the signal into mutually orthogonal set of wavelets which is the main differences from the continuous wavelet transform or its implementation from the discrete time series sometimes called Discrete –time continuous wavelet transform. It is an implementation of the wavelet transform using a discrete set of the values scales and translations. |
| **End Result** | The two rarest symbols are connected first, followed by other 2 symbols. The new parent nodes have certain frequencies respectively and are brought together in the next step. The resulting node and the remaining symbols are subordinated to the root node that is created in a final step. | Generation of Shannon Fano code tree which represents the compressed data | Since it is a dictionary coder, a dynamic dictionary is created depending upon the presence of substring chosen from the original file | It requires only a small amount of hardware and software resources. It is a data compression algorithm that is supported by most bitmap file formats, such as TIFF, BMP, and PCX | Arithmetic coding replaces a string of input data with a single floating point number as an output. | This method will reduce the size of a string considerably when the string is lengthy and the compression ratio is not affected by the content of the string | Both of the data can be separately compressed with other data compression algorithm to achieve maximum compression ratio. | A compressed data is obtained at the output with loss in some amount of original data. | Signal approximations at the output which are done efficiently by making use of the cosine functions. | A mutually orthogonal set of wavelets is obtained at the output after decomposition of a signal. |
| **Application** | Used in lossless data compression schemes. | It is a variable length code for encoding a source symbol | Used in data compression, storing or retrieving files | Used in the scenarios where a larger sequence is to be replaced by a shorter sequence | The main purpose of this technique is to give an interval to each potential bit data. It treats the whole string data as one unit. | This algorithm was originally implemented for use in an SMS application | Most required in the applications where there is a high demand of high compression ratios. | It is a type of data compression for natural data like audio signals or images. | It is a lossy compression technique which is widely used in area of image and audio compression. | Discrete Wavelet Transform is applied to the image block generated by the pre-processor |