Chapter 7

Conclusion

The subject of the thesis is universal lossless data compression algorithms. The background of this field of research was presented in Chapter 2. There was an analysis about a need for compression and situations in which universal lossless methods are useful. The algorithms used contemporarily were described in detail. A special, in-depth description of the Huffman coding, Arithmetic coding and LZW and Burrows Wheeler compression algorithms was provided as the thesis of the dissertation concerns this family of compression methods.

The importance of random access text data compression, which is one of the main core of this thesis is explained in the beginning of Chapter 3. The stages of Byte Pair Encoding Scheme [21] which is a universal random access text compression are presented in Section 3.4 with illustrations. We proposed significant improvements to the BPE and the coding and decoding algorithms of the improved algorithms are discussed in Chapter 3. The proposed algorithm BPE+ is discussed in Section 3.5. This algorithm uses the unused byte codes between 0 to 127 for replacement of most frequently appearing pairs of bytes. Another variation of BPE is proposed in Section 3.6 named as BPE++. In the process of replacement of frequent pairs, it was found that some of the byte codes totally vanished. BPE++ is finding such vanished byte
codes and using those codes for replacing the frequent pairs further. Byte Triplet Encoding (BTE) is another proposed method for random access text data compression, which replaces most frequently appearing triplets by the byte codes from 128 to 255. This is based on the fact that more than 80% of the English words are having length between 2 to 4 [39] and also most of the English words have prefixes and suffixes with length 3. Byte Quad Encoding (BQE) is another method proposed for random access text data compression which replaces the most frequently appearing quadruplets. This is also based on the observation made in the paper [39], and another fact that many English words have four lettered prefixes and suffixes.

BQE+ is another significant improvement to BQE that combines the concept of BPE++ with BQE. In Section 3.10, Byte Word Encoding method is proposed and discussed for random access text compression. There are three factors used for comparing the compression algorithms: compression ratio, compression time and decompression time. The compression ratio and compression time analysis of the proposed algorithms are presented in Section 3.11. The proposed algorithms provide significant improvement in the compression ratio. One of the main drawbacks of these kinds of random access text data compression algorithms is more compression time. The decompression is very fast in all proposed algorithms. These kinds of algorithms are best suited for an application which requires less decompression time without providing restriction to compression time. The detailed analysis of proposed algorithms is done and it is mathematically proved that the proposed algorithms perform more replacements than BPE.
The second core area of the thesis, text preprocessing is introduced and analyzed in Chapter 4. It is proved that by performing a lossless, reversible transformation to a source file prior to applying a compression algorithm, the data compression ratio may be improved [41, 42]. A few transformation methods are proposed in Chapter 5. Few techniques used to handle the most frequent characters space and eoln are presented in Section 5.1 and 5.3.1. In the proposed transformation an attempt is made to reduce possible byte values occurring after a byte by applying a reversible transformation. In Section 5.4, the proposed transformation is tested by the ten files of the Calgary Corpus[89]. The proposed methods improve BPC at an average of one bit per character. That is, the percentage of compression is improved by 12.5% at an average.

In the last chapter, the applications of data compression in the fields of networks, cryptography, and database management system are presented.

The results in this work can surely be a point of departure for further research. Finding the most frequently appearing pairs, triplets and quadruplets in a text file is a major task in the random access text compression algorithms. Most of the compression time is spent on searching. A new searching technique needs to be designed to find the most frequent appearing pairs, triplet and quadruplets within a least time. Transformation models become more successful, if the techniques are clear about the kind of input file to be transformed. The transformation techniques proposed in this thesis are applicable only to text files. There is a possibility of proposing a universal transformation that can be applied to all kinds of files to improve the compression ratio.