Data Compression through Duplicate Elimination and Tagging

Jeffrey Thomas, Computer Systems Lab 2008-2009, Period 2

Abstract

Data compression is a valuable tool to save memory and time when sending data between computers. I believe to have created a new method to compress data based on simple probability.

Introduction

In order for a data compression algorithm to be valid for use, it must reach its optimal case a majority of time it is run on different data files. If the algorithm is a success, it has the potential to be exceptionally useful in the world of computing, as it can perform on any type of file, and should be able to optimally perform every time it runs.

Methodology

The algorithm works by finding patterns of bits where the first half is identical to the second half. The second half is deleted, and a marker noting its former position is placed at the beginning of the stream of bits. The process is repeated for the rest of the stream. For any group of 2^n bits, ¼ will be duplicates. With enough data to be effectively infinite compared to the group size, ¼ of the groups should be compressed on any run. The algorithm was tested using sixty-four different inner variables, with each combination tested twenty times. The final values were the average of the results.

```
        01100110
        10001000
```

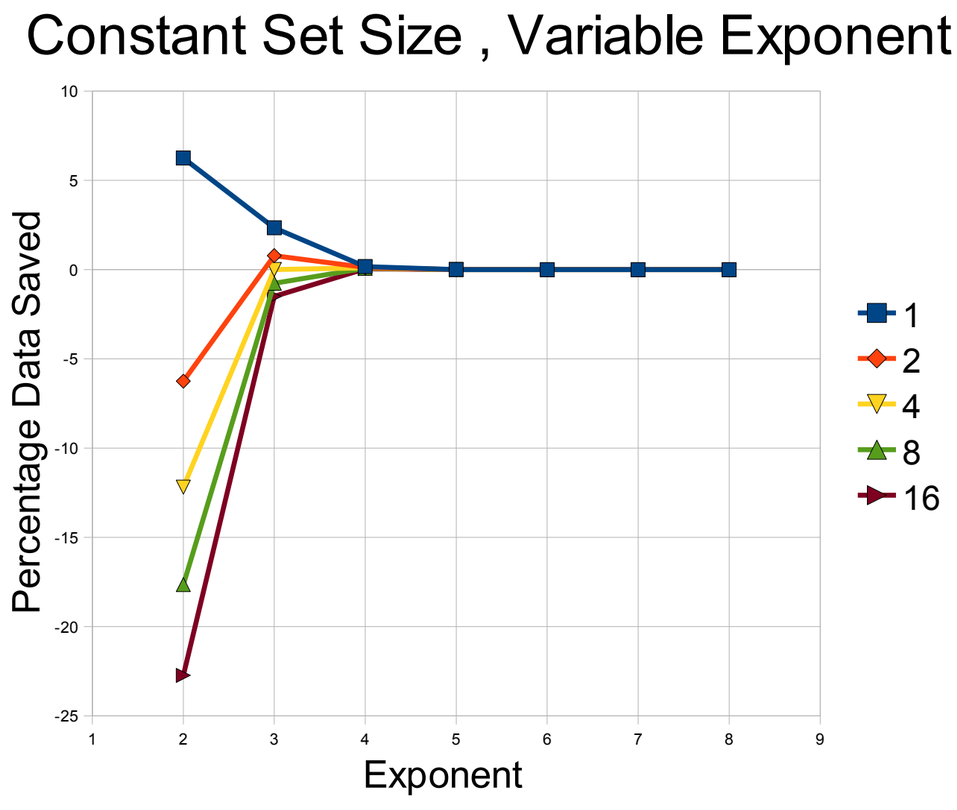
Fig 1. Duplicate bit patterns

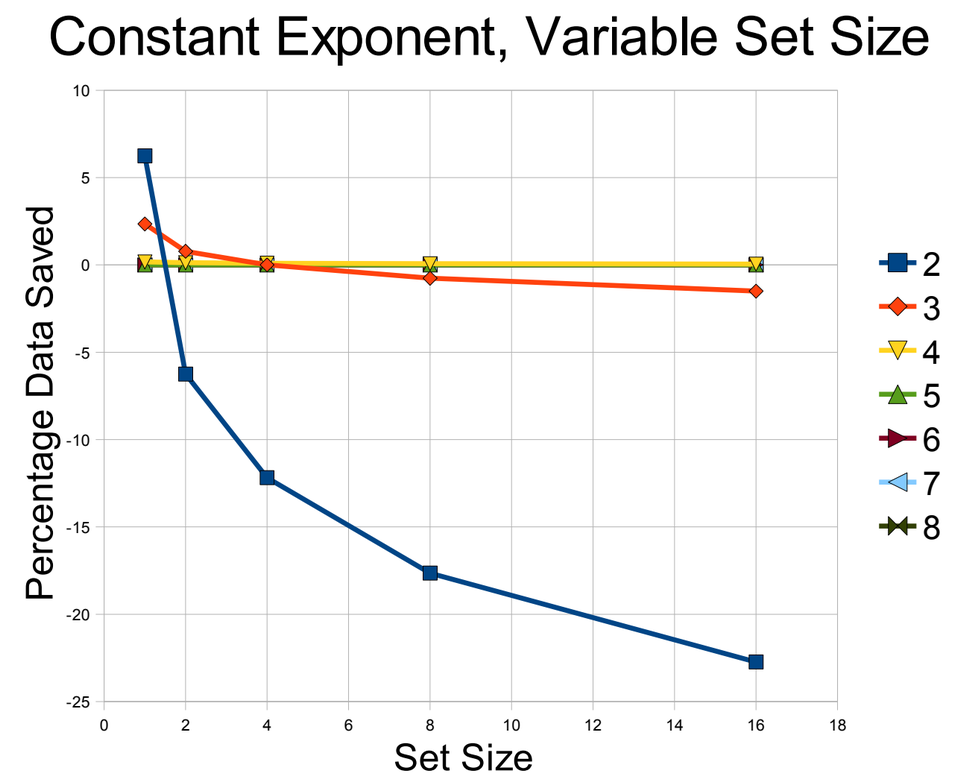
```
        00101100
        11100111
```

Fig 2. Not Duplicate patterns

Results

As seen below, increasing either variable leads to a decrease in overall compression.





Conclusion

As tested, the algorithm is not viable for widespread use. However, further testing with larger test files could yeild different results, as this project was limited by memory usage. A different setup allowing more memory usage could use larger test files, and perhaps yeild different results.