COL761 (Session 2023-2024) Assignment 1

Team Name: only two

Team Members:

1)Varada Vinay Bhaskar 2020CS10405

2) Chinmayee Behera 2020CS10337

Contribution:

Varada Vinay Bhaskar – 50%

Chinmayee Behara – 50%

Reference:

https://github.com/VNSAditya02/COL761-Data-Mining/blob/main/HW1/fptree.cpp

Basically we are provided with a dataset of transactions we need to compress the given dataset and should be able to decompress it by using the map we used for compression.

Here we use the concept of frequent pattern tree to get the recurring patterns in the transactions dataset.

Once we find the recurring patterns with our required threshold we replace the recurring patterns with unique codes. As our data set contains non-negative positive integers we replace the recurring patterns with unique negative integers.

After replacing the frequent patterns with the unique codes assigned to them we reduce the size of the dataset and to regain the original data set from this compressed data set we maintain a map of frequent patterns and the unique negative codes assigned to them.

Steps of the Algorithm:

Step1: We use the FP-tree to find the frequent items in the dataset.

Step2:Next we use the Top frequent items that are above threshold and then we change the threshold for getting better compression.

Step3: Once we get the threshold we give the unique code to the frequent data items that are above our threshold and compress the data.

Step 4: We next repeat the above steps on the compressed data to compress it more. We do these iterations 14 times or till 58 mins if the iterations not get completed.

Decompression:

Once we have the compressed data we need to able to get the original data with zero loss so we already store the mapping while we are doing compression so we simply replace the frequent data items code with the original frequent item and we will be able get the original data.

Results for the provided Datasets:

For D_small.dat

```
intial and later count: 118252, 76837
Compression Ratio: 35.0227
intial and later count: 76837, 36153
Compression Ratio: 52.9484
intial and later count: 36153, 31477
Compression Ratio: 12.9339
intial and later count: 31477, 31377
Compression Ratio: 0.317692
intial and later count: 31377, 31337
Compression Ratio: 0.127482
intial and later count: 31337, 31331
Compression Ratio: 0.0191467
intial and later count: 31331, 31203
Compression Ratio: 0.408541
intial and later count: 31203, 31183
Compression Ratio: 0.0640964
intial and later count: 31183, 35636
Compression Ratio: -14.2802
intial and later count: 35636, 36301
Compression Ratio: -1.86609
intial and later count: 36301, 36335
Compression Ratio: -0.0936613
intial and later count: 36335, 36339
Compression Ratio: -0.0110087
intial and later count: 36339, 36339
Compression Ratio: 0
intial and later count: 36339, 36339
Compression Ratio: 0
Total Time Spent: 4216 milliseconds
```

Time=4.216 seconds

Compression percentage= 69%

For D_medium2.dat

```
intial and later count: 3960507, 3932156
Compression Ratio: 0.715843
intial and later count: 3932156, 3932156
Compression Ratio: 0
intial and later count: 3932156, 3723842
Compression Ratio: 5.2977
intial and later count: 3723842, 2964334
Compression Ratio: 20.3958
intial and later count: 2964334, 2891193
Compression Ratio: 2.46737
intial and later count: 2891193, 2847993
Compression Ratio: 1.49419
intial and later count: 2847993, 2818362
Compression Ratio: 1.04042
intial and later count: 2818362, 2796556
Compression Ratio: 0.773712
intial and later count: 2796556, 2782632
Compression Ratio: 0.497898
intial and later count: 2782632, 2772618
Compression Ratio: 0.359875
intial and later count: 2772618, 2765135
Compression Ratio: 0.269889
intial and later count: 2765135, 2758383
Compression Ratio: 0.244183
intial and later count: 2758383, 2753638
Compression Ratio: 0.172021
intial and later count: 2753638, 2749863
Compression Ratio: 0.137091
Total Time Spent: 185592 milliseconds
```

Time=185 secs

Compression Percentage = 30.52%

For D_medium.dat

```
intial and later count: 8019015, 6468540
Compression Ratio: 19.335
intial and later count: 6468540, 6461064
Compression Ratio: 0.115575
intial and later count: 6461064, 6454405
Compression Ratio: 0.103064
intial and later count: 6454405, 6448603
Compression Ratio: 0.0898921
intial and later count: 6448603, 6394309
Compression Ratio: 0.84195
intial and later count: 6394309, 6388269
Compression Ratio: 0.094459
intial and later count: 6388269, 6347214
Compression Ratio: 0.642662
intial and later count: 6347214, 6337727
Compression Ratio: 0.149467
intial and later count: 6337727, 6330779
Compression Ratio: 0.109629
intial and later count: 6330779, 6321977
Compression Ratio: 0.139035
intial and later count: 6321977, 6315381
Compression Ratio: 0.104334
intial and later count: 6315381, 6308814
Compression Ratio: 0.103984
intial and later count: 6308814, 6303338
Compression Ratio: 0.0867992
intial and later count: 6303338, 6188320
Compression Ratio: 1.82472
Total Time Spent: 664032 milliseconds
Total Integers count in the compressed file: 6188320
```

Time : 664 secs.

Compresses Percentage = 22.829%

For D_large.dat

Initial number of lines : 109360594

Compressed to : 9947439630

Time:1hr

Compresses ratio:99474396