

文件加密压缩解压器

王述熠 2016013273

李 帅 2016013270

苏宇荣 2015080045

开发环境

- 平台：Windows 10
- IDE：Visual Studio 2017
- 汇编器：MASM32

实现原理

1. 压缩

压缩算法使用了静态哈夫曼编码算法。静态哈夫曼编码算法需要对待压缩的原始数据进行两遍扫描。

第一遍，逐字节地将待压缩的文件数据读取出来，将每字节转义解释为256种字符的编码，并统计原始数据中各转义字符出现的频率，利用得到的频率值构造优先级队列（小顶堆），随后创建哈夫曼树，得到每种字符对应的压缩编码，最后将上述有关信息保存起来，便于第二遍进行编码映射时候使用。

第二遍，逐字节地将待压缩的文件数据读取出来，根据前面得到的哈夫曼树对每个读取出来的字节进行编码映射，并将编码信息逐比特地写入到内存缓存中。如果文件最后一个的字节编码不足以填满缓存的最后一个字节，则对缓存内最后一个字节内未填满的比特写0。

随后，将编码解码所需要使用到的文件长度信息、字符频率信息、下面即将提到的加密信息以及内存缓存好的压缩编码信息等，一同写入到文件中。文件默认以".tql"作为后缀。

2. 加密

我们对文件的加密涉及两个部分：

1. 在压缩文件中储存加密信息用于解压时比对：我们对密码的处理采用md5算法。若用户不选择加密，则内部将密码串视为""。随后，计算密码串的md5值，并将md5值按特定方式存入文件（下述）。
2. 在压缩文件中只显式储存转义字符的权重，不直接储存编码/解码对应的二进制信息。由于哈夫曼树的构造不具有唯一性，这意味着，破解者若是只获得了权重信息而没有获取哈夫曼树的构建算法过程，恢复原始数据仍然将比较困难。

3. 压缩文件结构

- 第1字节 - 第4字节：压缩前大小
- 第5字节 - 第8字节：压缩后大小
- 随后16字节：密码md5值
- 随后256 * 4字节：转义字符权重
- 最后部分：压缩编码

4. 解压

压缩是解压的逆过程。分为如几个部分进行：

- 检查名字是否对应压缩文件
- 检查用户输入密码的md5信息是否与压缩文件中储存的密码md5信息匹配
- 利用压缩文件中的转义字符权重重新构建哈夫曼树，获得原始编码信息并构建解码映射
- 根据解码映射，恢复原始文件

难点

- 实现哈夫曼树、哈夫曼森林的非线性结构
- 基于小顶堆实现了优先级队列，提升了建树过程的时间效率
- 压缩文件结构的设计
- 加密逻辑的设计与具体实现
- 在汇编环境下对文件进行比特级别的操作，设计众多位运算

创新点

使用了小顶堆实现优先级队列，提升了建树过程的时间效率，对比一般使用列表的实现更加高效。

运行

双击tql.exe直接运行，或在cmd中打开运行。根据提示操作。

下面为示例。

1. 压缩

```
[ Usage ]:
- compress : compress the file into a .tql file.
- decompress : decompress a .tql file.
- quit : quit the program.
[ Input Command ]: compress
[ Input Filename ]: chinese.dic
[ Input Password ]: 233333
[ INFO ]: Compressing ...
[ INFO ]: File [ chinese.dic ] was compressed into [ chinese.dic.tql ] successfully...
[ Input Command ]: quit
[ INFO ]: The program is going to exit, press 'Enter' to continue. _
```

2. 解压

```
[ Usage ]:
- compress : compress the file into a .tql file.
- decompress : decompress a .tql file.
- quit : quit the program.
[ Input Command ]: decompress
[ Input Filename ]: chinese.dic.tql
[ Input Password ]: 233333
[ INFO ]: Decompressing ...
[ INFO ]: File [ chinese.dic.tql ] was decompressed into [ chinese.dic ] successfully...
[ Input Command ]: quit
[ INFO ]: The program is going to exit, press 'Enter' to continue.
```

压缩效果

本工具理论上可以对任何格式的文件进行压缩，但受限于哈夫曼压缩算法的限制，实际压缩中，表现较好的为文本文件的格式。

exe文件夹中提供了几份示例文件，下面展现工具的压缩空间效率。

文件名	压缩前大小	压缩后大小	压缩率
chinese.dic	2171KB	1852KB	85.31%
山河入梦.txt	486KB	429KB	88.27%
test.bmp	1801KB	1873KB	104.00%
Homework_4.pdf	123KB	117KB	95.12%

小组分工

姓名	分工
王述熠	mapper / md5 / main 相关过程
李帅	huffman_buffer 相关过程
苏宇荣	huffman_node / huffman_forest / priority_queue 相关过程