哈尔滨工业大学

数据科学与大数据技术专业

实验报告

课程名称：计算机算法基础

实验项目：贪心算法

实验题目：哈夫曼编码与译码方法

实验日期：2020.12

班级：1804105

学号：1180400510

姓名：石瑞河

|  |  |  |
| --- | --- | --- |
| 设计成绩 | 报告成绩 | 指导老师 |
|  |  | 张岩 |

**一、实验目的**

利用贪心算法实现一个哈夫曼编码与解码系统

**二、实验要求及实验环境**

**实验要求：**

1. 从文件中读入任意一篇英文文本文件，分别统计英文文本文件中各字符（包括标点符号和空格）的使用频率；

2. 根据已统计的字符使用频率构造哈夫曼编码树，并给出每个字符的哈夫曼编码（字符集的哈夫曼编码表）；

3. 将文本文件利用哈夫曼树进行编码，存储成压缩文件（哈夫曼编码文件）；

4. 计算哈夫曼编码文件的压缩率；

5. 将哈夫曼编码文件译码为文本文件，并与原文件进行比较。

**实验环境：**

硬件环境



软件环境：VSCode； Mingw-w64

**三、设计的基本思想和算法的原理**

哈夫曼编码是一种以哈夫曼树（最优二叉树，带权路径长度最小的二叉树）

为基础的基于统计学的变长编码方式。其基本思想是：将使用次数多的代码转换

成长度较短的编码，而使用次数少的采用较长的编码，并且保持编码的唯一可解

性。在计算机信息处理中，经常应用于数据压缩。是一种一致性编码法（又称"

熵编码法"），用于数据的无损耗压缩。本实验要求利用贪心算法实现一个完整

的哈夫曼编码与译码系统。

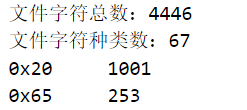
哈夫曼编码基于频率编码，将出现频率高的字符编码为短码，出现频率低的字符编码为长码，最后实现平均编码长度较定长编码小的效果。

哈夫曼树的构建可以通过贪心实现，维护一个堆，每次取出频率最小的两个结点合成一个新结点，把原结点作为新结点的子树，再把这个新结点放回堆中，直到堆中只剩下一个结点（或者堆为空）

也可以通过自上而下的划分构建哈夫曼树，将字符按出现频率排序，将序列按频率划分为最相近的两部分，在左右两个子序列中重复次过程，当子序列中只有一种字符时为其编码。

**四、实验结果与分析**

文件有4446个字符，有67种不同的字符，huffman树生成了133个结点

原始文件4446字节

编码转换后的01串为21616字节，即应该为21616/8 = 2702字节

考虑到哈夫曼编码随文件不同而不同，有必要将编码树保存在文件中才能进行传输和解压缩，需要设计一种协议。用1个字节表示哈夫曼树的序号编码位数x，1个字节表示字符字节数y，用x个字节表示哈夫曼树的结点数，2x+y个字节编码哈夫曼树的结点（包括左儿子序号、右儿子序号、编码的字符）。本题中，取x=y=1,哈夫曼树信息需要 1+1+133\*3=401个字节

故压缩率为



可见，哈夫曼编码确实可以对文件起压缩效果，对本次实验所用文件，体积可减小30%。

查看decode\_result.txt,与原文件haff\_no\_zh\_cn.cpp相同，字节数均为4446字节，内容也没有差别

**五、经验体会与不足**

哈夫曼编码确实能够起到不错的压缩效果，但是对于每一个的文件，它的哈夫曼编码的结果都不一样，如果需要网络传输，必须要定义一套存储编解码信息的协议，会带来额外的成本，字符种类较多时成本会急剧上升，且必须依赖于一套已有的通用编码。

**六、附录：源代码（带注释）**

见附件

源代码：Huffman.cpp

输入文本：Haff\_no\_zh\_cn.cpp (源代码的一个中间版本去掉中文注释后得到)

程序输出：result.txt 输出的编码信息等

encode\_result.txt 编码结果

decode\_result.txt 解码结果