**创新性探索课题实践报告**

# 选题: 利用概率统计方法实现英文文本压缩

## 小组成员

|  |  |
| --- | --- |
| **姓 名** | **学 号** |
| **张超** | **22336290** |
|  |  |

## 摘要（200字以内）

本研究旨在利用哈夫曼编码技术对英文文本进行压缩。传统的数据压缩方法在处理文本数据时存在一定的局限性，而哈夫曼编码作为一种无损压缩方法，具有更好的性能和广泛的应用前景。

研究方法主要围绕哈夫曼编码的原理展开，通过构建频率统计表，对英文文本中的字符进行编码，并生成相应的哈夫曼树，将文本中出现频率较高的字符用较短的编码表示，从而实现整体上的数据压缩。研究动机则源于对数据传输效率的追求，希望通过提高数据压缩率，减少存储和传输开销，以应对日益增长的数据量和对高效通信的需求。

实验结果显示，通过哈夫曼编码对英文文本进行压缩，显著减小原始数据的体积，达到较高的压缩比。有助于在有限的存储空间内存储更多的数据。

## 引言

在信息时代，数据的存储和传输成为一项关键挑战。文本数据是信息传递的基础，因此提高文本数据的压缩效率对于节省存储空间、降低传输成本具有重要意义。哈夫曼编码作为一种经典的无损压缩方法，被广泛应用于文本压缩领域。本研究旨在深入探讨利用哈夫曼编码对英文文本进行压缩的方法，以提高数据处理效率。

目前，文本压缩领域已有多种方法和算法，包括传统的基于字典的方法和近期基于深度学习的技术。虽然这些方法在一定程度上取得了成功，但在处理英文文本时，仍存在一些挑战，如压缩率不高、算法复杂等。哈夫曼编码作为一种经典的无损压缩方法，具有较好的性能，但在当前文本压缩研究中并未充分发挥其潜力。因此，本研究旨在通过深入研究哈夫曼编码在英文文本压缩中的应用，探索其在提高压缩率和减小数据体积方面的优势。

创新之处：

与传统的哈夫曼编码相比，我在此基础上进行了一些创新。传统的哈夫曼编码技术需要将哈夫曼编码写入文本文件之中，以便于后续解压缩。但是这会增加压缩后文本文件的体积，所以我对哈夫曼编码技术进行了一些调整，以避免将哈夫曼编码写入压缩后的文本文件。具体内容在下面的创新点板块阐述。

## 方法介绍

哈夫曼编码是一种无损数据压缩算法，其主要目标是通过赋予出现频率较高的符号短的编码，而给予出现频率较低的符号长的编码，从而实现对数据的高效压缩，主要步骤为：

**1.频率统计**： 哈夫曼编码的第一步是统计待编码数据中每个符号（通常是字符或字节）的出现频率。这些频率信息用于构建编码树。

**2.构建哈夫曼树**： 使用频率信息构建哈夫曼树，一种特殊的二叉树。构建哈夫曼树的基本思想是将出现频率较低的符号作为叶子节点，然后通过合并这些节点生成新的父节点，直到构建整个树。构建过程中，频率越高的符号离根节点越近。

**3.分配编码**： 从哈夫曼树的根节点开始，沿着路径到达每个叶子节点，分配0和1的编码，形成哈夫曼编码表。由于频率较高的符号路径较短，它们的编码也较短，从而实现了压缩。

**4.生成编码数据**： 使用生成的哈夫曼编码表，将原始数据中的每个符号替换为对应的哈夫曼编码，生成压缩后的数据。在解压时，可以使用相同的哈夫曼编码表将编码的数据还原为原始数据。

## 创新点（如有）

传统的哈夫曼编码压缩技术需要将哈夫曼编码放进压缩后的文本文件之中，因为对于每个文本文件的哈夫曼编码不同，只有这样才能进行后续解压缩，但是这占用了一部分不必要的文件体积，甚至有可能超过压缩后文本本身的体积。在学习到“抽样分布”这一章节时，我学习到可以用大量的样本去估计总体的分布。将这个知识运用到这个项目中来，我想到可以统计大量样本的文本文件的字符出现频率，得到一个基于大量样本的哈夫曼编码，以此来估计我们所要压缩的文本文件的哈夫曼编码。这样我们就只需要得到这个基于大量样本的哈夫曼编码，就可以对任意的文本文件进行压缩，同时不需要把哈夫曼编码写进压缩后的文本文件。当数据量较少时，哈夫曼编码写入压缩文件将占很大的空间，此时采用上述方式将能有效提高压缩比。

但是这个方法也有不足之处。采用这种由大量样本估计出来的哈夫曼编码，由于估计的哈夫曼编码较真实的哈夫曼的不准确性，文本文件的压缩比将会大大降低，甚至可能压缩文件的体积大于源文件。并且当数据量较大时，此时采用传统的哈夫曼编码压缩技术压缩比会较高，因为此时哈夫曼编码的数据存储空间和文本的数据存储空间比起来微乎其微，但它的哈夫曼编码是真实的，所以压缩比较高。

综上，数据较小时适合我创新的压缩方式，数据量较大时适合传统的哈夫曼压缩方式，我结合两者的优势，将两种算法一起实现，用户可以根据压缩文件大小选择要使用哪种压缩方式。

## 实验结果

我一共使用了八组数据进行测试，在这八组数据中有有四组数据量较小，有四组数据量较大。

我分别使用两种压缩方式进行压缩，使用传统的哈夫曼压缩得到的平均压缩比为22.23%，而使用创新哈夫曼压缩得到的平均压缩率为15.58%。

但是如果结合两种压缩方式，即前四个样本用创新哈夫曼压缩，后四个样本用传统哈夫曼压缩，所得平均压缩率为27.21%。

由此可见将两种方式结合使用可以达到更高的压缩率。

## 程序说明

打开电脑终端，进入到该程序所在文件目录下。

使用传统的哈夫曼压缩方式（压缩文本的数据量较大时，如在10kb以上），请在命令行中输入：

**huffman zip inputfilename outputfilename**

如图：



使用创新的哈夫曼压缩方式（压缩文本的数据量较小时，如在10kb以下），请在命令行中输入：

**huffman minidata\_zip infilename outfilename**

如图：



对传统的哈夫曼压缩的文件进行解压缩，请在命令行中输入：

**huffman unzip infilename outfilename**

如图：



对创新的哈夫曼压缩的文件进行解压缩，请在命令行中输入：

**huffman minidata\_unzip infilename outfilename**

如图：



如需对代码重新编译，请在命令行输入（确保电脑已安装g++）：

**g++ -o huffman \*.cpp**

## 参考文献

1.Huffman, D. A. (1952). "A Method for the Construction of Minimum-Redundancy Codes." Proceedings of the IRE, 40(9), 1098-1101.

## 具体分工说明

张超（22336290）1：本次项目由我独立完成。我独立完成了查找资料，确定实现文本压缩的算法，并最终确定为哈夫曼编码。我用c++语言写出哈夫曼编码文本压缩的算法，并用python爬虫在对应网站上爬取测试数据，并在这个过程中不断改正bug，同时我也发现了传统的哈夫曼编码不足之处，所以在算法的实现中加入了一些新的想法，使文本的压缩率更高。

通过这个项目，我深刻理解了哈夫曼编码的原理和应用，锻炼了问题解决和编码能力。同时，我学到了思考和创新的重要性，明白了如何将理论知识转化为实际应用。这次经历让我更深入地理解了数据压缩领域，为我的职业发展积累了宝贵经验。