基于GZIP文件压缩

[本节目标]

- **1. 什么是文件压缩**
- **2. 为什么需要压缩**
- **3. 压缩的分类**
- **4. 压缩的原理**
- **5. GZIP的压缩原理简介**

1. 什么是文件压缩

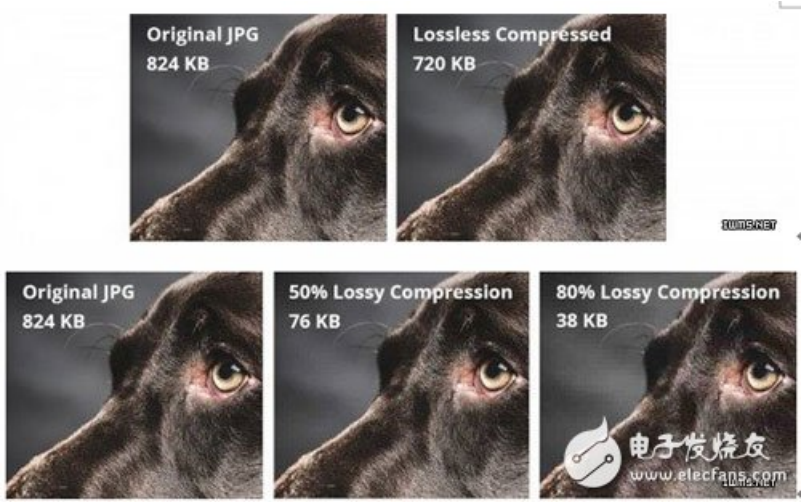
> 文件压缩是指**在不丢失有用信息的前提下，缩减数据量以减少存储空间**，提高其传输、存储和处理效率，或**按照一定的算法对文件中数据进行重新组织，减少数据的冗余和存储的空间**的一种技术方法。

2. 为什么需要压缩

> 1. 紧缩数据存储容量，减少存储空间
> 2. 可以提高数据传输的速度，减少带宽占用量，提高通讯效率
> 3. 对数据的一种加密保护，增强数据在传输过程中的安全性

3. 压缩的分类

> - **有损压缩**
>
> 有损压缩是利用了人类对图像或声波中的某些频率成分不敏感的特性，**允许压缩过程中损失一定的信息；虽然不能完全恢复原始数据，但是所损失的部分对理解原始图像的影响缩小，却换来了大得多的压缩比**，即指使用压缩后的数据进行重构，重构后的数据与原来的数据有所不同，但不影响人对原始资料表达的信息造成误解。
> - **无损压缩**
>
> 对文件中数据按照特定的编码格式进行重新组织，**压缩后的压缩文件可以被还原成与源文件完全相同的格式，不会影响文件内容**，对于数码图像而言，不会使图像细节有任何损失。



4. 压缩本质　本质：想办法让文件变小-->要能够还原

压缩的目的是让文件变小，减少文件所占的存储空间。那怎么才能让文件变小呢？方式比较多，比如：

1. **专有名词采用的固定短语**

   比如：陕西科技大学，简称陕科大，就可以提到压缩的目的，但只能针对于大家所熟知的专有名词。

2. **缩短文件中重复的数据**

   1. 比如文件中存放数据为：**mnoabczxyuvwabc123456abczxydefgh**

      对文件中重复数据使用(距离，长度)对进行替换，压缩之后的结果为：

      mnoabczxyuvw(9,3)123456(18, 6)defgh

3. **给文件中每个字节找一个更短的编码**

   比如文件中存放数据为：ABBBCCCCCDDDDDDD

字符	静态等长编码	动态不等长编码
A	00	100
B	01	101
C	10	11
D	11	0

采用静态等长编码压缩： 00010101 10101010 10000000 00000000

采用动态不等长编码压缩：10010110 11011111 11111100 00000000

源文件16个字节，压缩完成之后占4个字节，可以起到压缩的目的。

5. GZIP压缩原理简介

GZIP压缩总共经历了两个阶段：

1. 利用变形的LZ77压缩算法，先快速的从重复语句层面进行快速压缩
2. 利用huffman编码方式再从字节上面进行压缩

将两种压缩算法结合起来，以达到高效快速的压缩，具体实现，请参考下文。