**数据压缩博士申请课题**

**The Design, Implementation, and Deployment of a System to Transparently**

**Compress Hundreds of Petabytes of Image Files For a File-Storage Service**

**一、主要工作**

我们报告了我们使用另一种技术的经验:基于统计模型的特定于格式的透明文件压缩，该模型经过调优，可以在大型语料库上表现良好。

We report on our experience with a different technique: format-specific transparent file compression, based on a statistical model tuned to perform well on a large corpus

它用一种定制的统计模型取代了最低层的基线JPEG图像——无损霍夫曼编码——我们对该模型进行了调整，使其在存储在Dropbox中的大量JPEG图像语库中表现良好。

called Lepton, that replaces the lowest layer of baseline JPEG images— the lossless Huffman coding—with a custom statistical model that we tuned to perform well across a broad corpus of JPEG images stored in Dropbox.

**二、实现思路**

由于大数据通常存储在磁盘中，因此我们的想法是基于规则压缩在存储层对压缩后的数据进行随机更新。

Since large data are usually stored in disk, our idea is to develop random update over compressed data in storage layer based on rule compression.

在元素级别，我们引入了数据洞的概念，以允许在大数据块中进行更新。

in element level, we introduce the concept of data holes to allow updates in large data blocks.

在规则层，我们开发了哈希和计数数据结构，以有效地定位规则。

In rule level, we develop hashing and counting data structures for efficiently locating rules.

在DAG级别，我们限制了DAG的深度，以将规则拆分和合并的成本保持在一个小范围内。

In DAG level, we limit the depth of the DAG to retain the cost of rule split and merge within a small range.

1. **实现难点**

**Round-trip transparency.**

Lepton需要确定地恢复原始文件的精确字节

Lepton needs to deterministically recover the exact bytes of the original file

**Distribution across independent chunks.**

Lepton必须能够解压缩JPEG文件的任何子字符串，而不需要访问其他子字符串。

Lepton must be able to decompress any substring of a JPEG file, without access to other substrings

**Low latency and streaming.**

为了实现这一点，Lepton格式包括“霍夫曼切换词”，使解码器成为多线程的，并在请求后很快开始传输字节。

To achieve this, the Lepton format includes “Huffman handover words” that enable the decoder to be multithreaded and to start transmitting bytes soon after a request.

**Security.**

**Memory.**

1. **图片压缩的常见方法**

通用熵压缩。

Generic entropy compression.

有损图像压缩。

Lossy image compression

格式感知像素精确重压缩。

Format-aware pixel-exact recompression

支持格式、保存文件的再压缩。

**Format-aware, file-preserving recompression**

最后一组工具可以通过往返恢复原始文件的精确字节来重新压缩JPEG文件。

A final set of tools can re-compress a JPEG file with roundtrip recovery of the exact bytes of the original file.