超酷算法：喷泉码

2014/10/28 · IT技术 · 1 评论 · 喷泉码, 算法

是的，是更新这个十分罕见的超酷算法系列新的一集的时候了。如果你不熟悉这个系列，你可以看看之前的一些文章。

今天的主题是喷泉码，或者称为“无率码”。喷泉码是将一些数据，例如文件，转化为一个有效的任意数量的编码包的方法，这样只要你接收到稍大于信源数据包数量的编码包的子集，就可以恢复信源数据。换句话说，你创建了一个编码数据的“喷泉”，只要接收端接收到足够的“水滴”，就可以恢复文件，而不管它们接到哪一个遗漏了哪一个。

让喷泉码如此知名的原因是，它允许你在有损连接（比如说因特网）的情况下传输文件，而且传输过程不依赖于你是否知道丢包率，也不需要接收端反馈哪些数据包丢失了。可以看到在很多场景，从通过广播媒介传送一个静态文件，比如点播电视，到在多源并行下载中传播文件包，像BitTorrent那样，喷泉码都得到了很好的应用。

虽然从根本上喷泉码惊人地简单。它有许多种类，但是在本文中我们只介绍最简单的——LT码，或者Luby变换码。LT码生成编码包的步骤如下：

何选取最佳的d。
2. 从文件中随机选取 d 块，并把它们组合起来。这里我们可以用异或运算来组合这些块。
3. 传送合并的块，同时发送它由哪些块构成的信息。

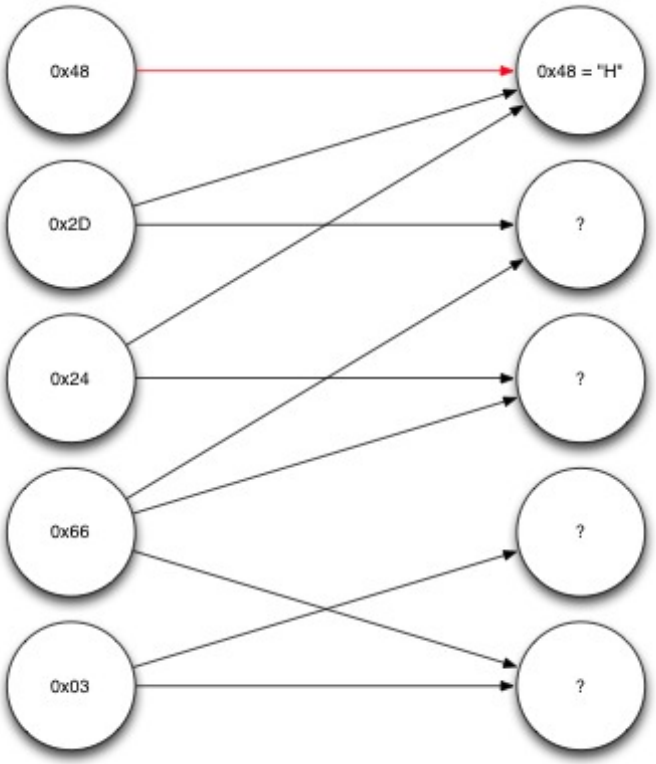
这些非常简单是不是？主要依赖于我们怎么选取块的数量并组合起来（叫做度分布），在接下来我们会简短的介绍一下。你可以从上面的描述中看到有些编码块最后只由单一源码块组成，而大部分将由多个源码块组成。

另外一个可能不是立刻显现的事情是，虽然我们确实不得不让接收端知道输出码块由哪些码块合并产生的，我们不需要详细地发送那个列表。如果发送端和接收端使用相同的伪随机数生成器（pseudo-random number generator，PRNG），我们可以用一个随机选择的种子来生成PRNG，并且用这个来选择度和该组源码块。然后我们只需要在发送编码块的同时发送种子，我们的接收端可以用相同的过程来重建我们使用过的源码块列表。

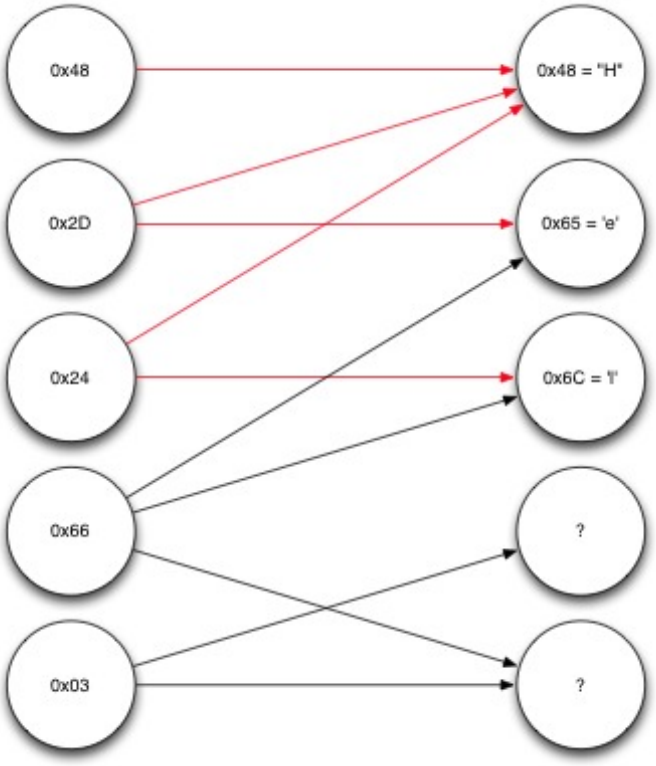
解码的过程有一点复杂，但是没有很复杂：

1. 重建用于生成编码块的源码块列表。
2. 对于列表中的每一个源码块，如果已经解码了，将它和编码块做异或运算，并且把它从源码块列表中移除。
3. 如果在列表剩下至少两个源码块，将编码块加入到一个等候区。
4. 如果在列表中只剩下一个源码块，我们已经成功的把另一个源码块解码了，那么把它加入到已解码文件中，迭代等候列表，重复以上过程直到有编码块包含它。

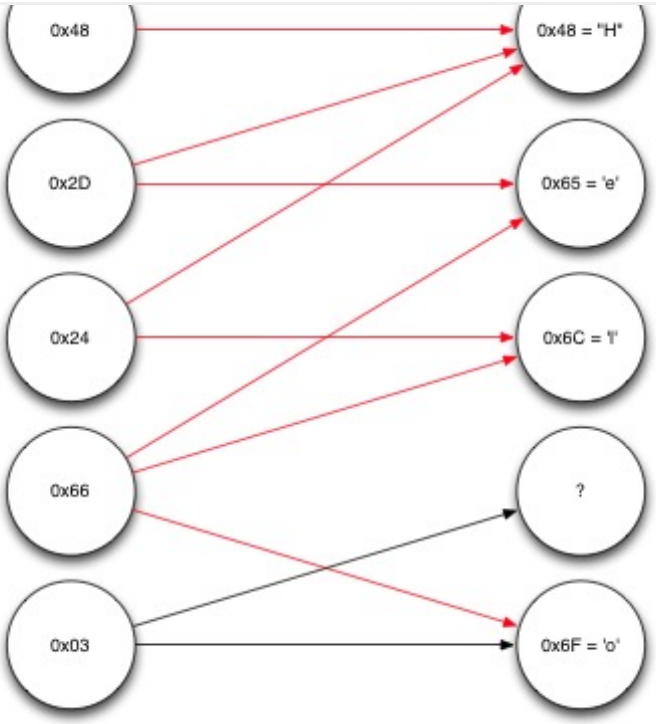
让我们通过一个译码实例来更清晰说明这个过程。假设我们收到5个编码块，每个长度是一个字节，并且我们知道每个源码块由哪些构成。我们可以用图来表示数据，如下所示：



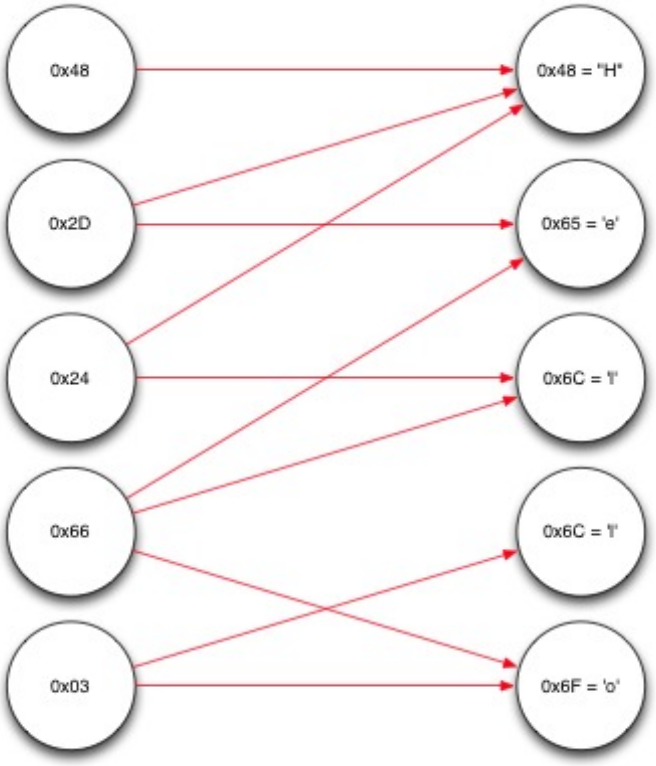
块（第一个源码块）构成，所以已经知道是哪个块。沿着指向第一个源码块箭头的反向，可以看到第二个和第三个编码块都只依赖于第一个源码块和另外一个源码块，由于我们知道第一个源码块，我们可以对它们做异或运算，如下图所示：



重复以上过程，我们可以看到我们现在有了足够的信息来解码第四个编码块，它依赖于第二个和第三个源码块，而这两个我们现在都知道了。对它们做异或运算，可以得到第五个也是最后一个源码块，如下所示：



最后，我们可以解码最后剩下的源码块，得到剩余的信息：



应该承认的是这是一个非常特殊的例子，这个例子刚好接收到我们在译码这个信息时需要的块，没有剩余的，并且是一个非常简单的顺序，但是这个例子很好的演示算法的原理。我确定你可以看到这个算法应用到大规模码块和大规模文件中会相当简单。

下，我们需要生成一些只包含一个源码块的编码块，然后可以开始译码了，大多数编码块依赖很少的其他编码块。这种理想的分布是存在的，叫做理想孤波分布。

不幸的是，理想孤波分布在实际情况中并没有这么理想，正如随机变量使得有些源码块不被任何编码块包含，或者当所有知道的块用完之后译码停止了。理想孤波分布的一个变形，叫做稳健孤波分布，在这方面进行了改进，用非常少的源码块生成更多的码块，也通过合并所有的或几乎所有的源码块生成一些码块，来帮助破译最后一些源码块。

简而言之，这就是喷泉码的，更确切的说是LT码的，工作原理。LT码是已知的喷泉码中效率最低的，但是最易解释的。如果你想进一步学习，我强烈推荐读这篇关于喷泉码的技术论文，也可以读Raptor码，Raptor码只比LT码增加了一点复杂度，但是在传输开销和计算上都显著的提高了它们的效率。

在我们总结之前有一个进一步的思考问题。对于系统来说喷泉码可能看起来很理想，比如说比特流，它允许种子生成和散布几乎无限制数量的码块，或多或少的消除了稀疏种子流“最后一块”的问题，而且确保两个随机选择的并行端几乎总有有用信息相互交换。但是它面临一个重大的问题：验证从并行端接收到的数据将会很难。

像比特流这样的协议使用安全散列函数，比如说SHA1，和一个可信任中心（最初的上传者），向所有的并行端发送一个权威散列表。每个并行然后可以验证他们下载的散列块的文件包，并且和权威散列进行对比。但是对于喷泉码，这个是很难的。根本没有方法在编码块上计算SHA1散列，更不要说单独块上的散列。我们不能相信我们的并行端计算的结果，因为它们可以对我们撒谎。我们可以等到我们得到全部文件，然后从无效码块列表出发，尝试推断什么样的编码块是无效的，但这是困难的也是不可靠的，而且信息来的时候可能已经为时已晚。一个可供选择的方法是让最初发布者公布一个公共密钥，并且标注所有的生成块。然后我们就可以验证编码块了，但代价是：现在只有最初发布者可以生成有效的编码块了，并且我们失去了最初使用喷泉码的很多好处。似乎我们被困住了。

还有另一种选择，而且已经证明是一个非常聪明的方案，叫做同态哈希，尽管它有自己的注意事项和缺点。我们将会在下一版的超酷算法中讨论。

关于作者：mathshelly

逆风的方向，更适合飞翔 （新浪微博：@mathshelly）

个人主页 · 我的文章 · 12

相关文章

- 漫画算法：什么是 B 树？ · 2
- 漫画算法：什么是跳跃表？ · 5
- 七大查找算法

可能感兴趣的话题

- 有同做 Android for ROS 的小伙伴么？欢迎交流
- 程序员清晰的职业规划会有多长？ · 4
- layout布局优化
- GreenDao多表联查
- PHP 的可能 · 1
- 怎么提高组织语言能力和表达能力？求指导下，主要在项目文档的撰写和开会对... · 1

登录后评论　新用户注册

直接登录

最新评论

疾风剑豪　06/24

恕我直言，写的让人看不懂

赞　回复

- 本周热门文章
- 本月热门文章
- 热门标签

1 不懂技术的管理者，给你们扫盲软件开...

2 10 个鲜为人知的 Linux 命令（5）

3 30 个实例详解 TOP 命令

4 10 个鲜为人知的 Linux 命令（3）

5 10 个鲜为人知的 Linux 命令（4）

6 分布式事务的一种实现方式--状态流转

7 QA 请勿忘初心

8 读懂 MySQL 执行计划

9 2017 最优秀的十大 Linux 服务器...

业界热点资讯 更多 »

F-35 战斗机的代码量达到 800 万行
1 天前 · 14

慕尼黑放弃 Linux，2020 年或将全面迁入 Windows
18 小时前 · 3

N 次跳票后，Fedora 27 正式版终于发布了
18 小时前 · 2

java

[最新的 Java SE 平台和 JDK 版本发布计划]
1 天前 · 3

[TIOBE 11 月编程语言排行榜，脚本语言怎么了？]
1 天前 · 4

精选工具资源 **更多资源 »**

[Whitewidow：SQL 漏洞自动扫描工具]
[数据库] · 2

[Caffe：一个深度学习框架]
[机器学习]

[静态代码分析工具清单：公司篇]
[静态代码分析]

[HotswapAgent：支持无限次重定义运行时类与资源]
[开发流程增强工具]

静态代码分析工具清单：开源篇（各语言）
静态代码分析

关于伯乐在线博客

在这个信息爆炸的时代，人们已然被大量、快速并且简短的信息所包围。然而，我们相信：过多“快餐”式的阅读只会令人“虚胖”，缺乏实质的内涵。伯乐在线内容团队正试图以我们微薄的力量，把优秀的原创文章和译文分享给读者，为“快餐”添加一些“营养”元素。

快速链接
网站使用指南 »
问题反馈与求助 »
加入我们 »
网站积分规则 »
网站声望规则 »

关注我们

新浪微博：@伯乐在线官方微博
RSS：订阅地址
推荐微信号

合作联系
Email：bd@Jobbole.com
QQ： 2302462408（加好友请注明来意）

更多频道

小组 – 好的话题、有启发的回复、值得信赖的圈子
头条 – 分享和发现有价值的内容与观点
相亲 – 为IT单身男女服务的征婚传播平台
资源 – 优秀的工具资源导航
翻译 – 翻译传播优秀的外文文章
文章 – 国内外的精选文章
设计 – UI,网页，交互和用户体验
iOS – 专注iOS技术分享
安卓 – 专注Android技术分享
前端 – JavaScript, HTML5, CSS
Java – 专注Java技术分享
Python – 专注Python技术分享