


Github 的清点对象算法

6 回复 681 查看



(<https://www.shiyanlou.com/user/8490>) 实验楼管理员  (<https://www.shiyanlou.com/vip>)

2015-10-26 17:35


技术分享 (<https://www.shiyanlou.com/questions/?tag=技术分享>)

"清点对象" (counting objects)，Github需要实时计算出来，需要克隆的对象总数。



全部回答



实验楼管理员 (<https://www.shiyanlou.com/user/8490>)  (<https://www.shiyanlou.com/vip>)

(<https://www.shiyanlou.com/user/8490>) 使用Github的时候，你有没有见过下面的提示？

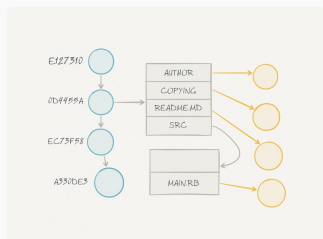
```
$ git clone https://github.com/torvalds/linux
Cloning into 'linux'...
remote: Counting objects: 4350078, done.
remote: Compressing objects: 100% (4677/4677), done.
Receiving objects: 4% (191786/4350078), 78.19 MiB | 8.70 MiB/s
```

这段提示说，远程代码库一共有4350078个对象需要克隆。

这就叫"清点对象" (counting objects)，Github需要实时计算出来，需要克隆的对象总数。

这个过程非常慢，根据Github的披露，像Linux kernel这样巨大的库，清点一次需要8分钟！也就是说，发出git clone命令后，会干等八分钟，然后才会开始真正的数据传输。这当然是无法忍受的。Github团队一直想解决这个问题。

后来，他们终于发现了一种新的算法，现在清点一次只要3毫秒！

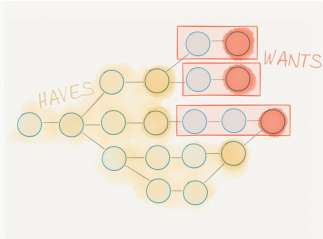


为了理解这个算法，你必须先知道，什么是Git的对象。简单说，对象就是文件，最重要的对象有三种。

- 快照对象 (Commit)
- 目录对象 (Directory)
- 文件对象 (File)

每次提交代码的时候，会生成一个commit对象，里面有对应的当前"目录对象"的名字。"目录对象"保存了代码根目录所含有的子目录和文件信息。每一个子目录就是另一个"目录对象"，每一个文件则是"文件对象"，里面是具体的文件内容。

所以，"清点对象"就是清点各种commit、目录、文件等。git clone和git fetch操作都需要清点对象，因为需要知道，到底下载哪些对象文件。




清点对象的原始算法如下。

- 列出本地所有分支最新的一个commit
- 列出远程所有分支最新的一个commit
- 两者进行比较，只要有不同，就意味着分支发生变动
- 每一个发生变动的commit，都清点其中具体变动的子目录和文件
- 追溯到当前commit的父节点，重复第四步，直至本地与远程的历史一致为止
- 加总所有需要变动的对象

上面的过程说明，“清点对象”是一个文件遍历算法，变动的对象会被一一清点到，这就意味着大量的文件读操作。对于大型代码库来说，这个过程非常慢。

2015-10-26 17:39

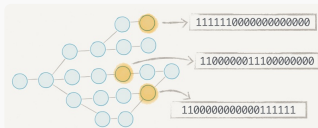


实验楼管理员 (<https://www.shiyanlou.com/user/8490>)  (<https://www.shiyanlou.com/vip>)

(<https://www.shiyanlou.com/user/8490>)

Github团队想到的新算法，是建立一个Bitmap索引，即为每一个commit生成一个二进制值。

打开本地Github仓库的.git/objects/pack/目录，你会看到一个索引文件和一个数据文件，它们就是Bitmap。简单说，这两个文件索引了当前代码库的所有对象，然后使用一个二进制值代表这些对象。有多少个对象，这个二进制值就有多少位。它的第n位，就代表数据文件里面的第n个对象。



每个commit都会有一个对应的二进制值，表示当前快照包含的所有对象。这些对象对应的二进制位都为1，其他二进制位都为0。

这样做的好处是，不用读取commit对象，只要读取这个二进制值，就会知道当前commit包含了哪些节点。更妙的是，两个二进制值只要做一次XOR运算，就会知道哪些位（即哪些对象）发生了变动。而且，因为新的对象总是添加到现有二进制位的后面，所以只要读取多出来的那些位，就知道当前commit比上一次commit多出了哪些对象。

这样一来，“清点对象”就变成了二进制值的比较运算，因此速度极快。进一步的介绍，请参看官方文档《Bitmap的解释》

(<https://github.com/gitster/git/commit/fff4275>), 《Bitmap的格式》

(<https://github.com/gitster/git/blob/master/Documentation/technical/bitmap-format.txt>)。

目前，Github的生产环境已经部署了这套算法，用户再也不用为了清点对象，而苦苦等待了。而且，Github团队还把它合并进了Git，这意味着，从此所有Git实现都可以使用Bitmap功能了，因此将来肯定还会有更多好玩的用法出现。

via:阮一峰的网络日志

链接: <http://www.ruanyifeng.com/blog/2015/09/git-bitmap.html> (<http://www.ruanyifeng.com/blog/2015/09/git-bitmap.html>)

2015-10-26 17:39



kaka826 (<https://www.shiyanlou.com/user/65418>)

(<https://www.shiyanlou.com/user/65418>)

编程珠玑里的bitmap

2015-10-26 17:54



大圣大圣你的金箍棒呢 (<https://www.shiyanlou.com/user/71302>)

(<https://www.shiyanlou.com/user/71302>)
!!! 想一想就觉得好爽啊。

2015-10-26 18:06




大圣大圣你的金箍棒呢 (<https://www.shiyanlou.com/user/71302>)

(<https://www.shiyanlou.com/user/71302>)
图缩得有点小。

2015-10-26 18:59



实验楼管理员 (<https://www.shiyanlou.com/user/8490>)  (<https://www.shiyanlou.com/vip>) 回复 大圣大圣你的金箍棒呢 (<https://www.shiyanlou.com/user/71302>)

(<https://www.shiyanlou.com/user/8490>)

点击图片即可放大图片了~

2015-10-27 09:42

登录 (<https://www.shiyanlou.com/login?next=/questions/2115>)后回答问题

我要提问

标签

- 课程相关 (<https://www.shiyanlou.com/questions/?tag=课程相关>)
- Linux (<https://www.shiyanlou.com/questions/?tag=Linux>)
- Python (<https://www.shiyanlou.com/questions/?tag=Python>)
- 实验环境 (<https://www.shiyanlou.com/questions/?tag=实验环境>)
- C/C++ (<https://www.shiyanlou.com/questions/?tag=C/C++>)
- 技术分享 (<https://www.shiyanlou.com/questions/?tag=技术分享>)
- 课程需求 (<https://www.shiyanlou.com/questions/?tag=课程需求>)
- 功能建议 (<https://www.shiyanlou.com/questions/?tag=功能建议>)
- Java (<https://www.shiyanlou.com/questions/?tag=Java>)
- 其他 (<https://www.shiyanlou.com/questions/?tag=其他>)
- Web (<https://www.shiyanlou.com/questions/?tag=Web>)
- Hadoop (<https://www.shiyanlou.com/questions/?tag=Hadoop>)
- SQL (<https://www.shiyanlou.com/questions/?tag=SQL>)
- NodeJS (<https://www.shiyanlou.com/questions/?tag=NodeJS>)
- PHP (<https://www.shiyanlou.com/questions/?tag=PHP>)
- Shell (<https://www.shiyanlou.com/questions/?tag=Shell>)
- 常见问题 (<https://www.shiyanlou.com/questions/?tag=常见问题>)
- Git (<https://www.shiyanlou.com/questions/?tag=Git>)
- HTML (<https://www.shiyanlou.com/questions/?tag=HTML>)
- 网络 (<https://www.shiyanlou.com/questions/?tag=网络>)
- 信息安全 (<https://www.shiyanlou.com/questions/?tag=信息安全>)
- HTML5 (<https://www.shiyanlou.com/questions/?tag=HTML5>)
- Android (<https://www.shiyanlou.com/questions/?tag=Android>)
- GO (<https://www.shiyanlou.com/questions/?tag=GO>)
- NoSQL (<https://www.shiyanlou.com/questions/?tag=NoSQL>)
- Ruby (<https://www.shiyanlou.com/questions/?tag=Ruby>)
- 训练营 (<https://www.shiyanlou.com/questions/?tag=训练营>)
- Perl (<https://www.shiyanlou.com/questions/?tag=Perl>)



实验楼客户端

即开即用



会员专属

(<https://www.shiyanlou.com/vip>)

相关问题

编写可靠shell脚本的八个建议 (<https://www.shiyanlou.com/questions/5726>)

Android Studio 高级配置 (<https://www.shiyanlou.com/questions/5694>)

Java 编程要点之 I/O 流详解 (<https://www.shiyanlou.com/questions/5660>)

能使用html/css解决的问题就不要使用JS (<https://www.shiyanlou.com/questions/5648>)

如何成为Python高手 (<https://www.shiyanlou.com/questions/5587>)

12个你未必知道的CSS小知识 (<https://www.shiyanlou.com/questions/5599>)

手把手教你搭建自己的 VPS 服务器 (<https://www.shiyanlou.com/questions/3056>)

对ASP.NET程序员都非常有用的85个工具 (<https://www.shiyanlou.com/questions/5516>)

TreeMap实现原理 (<https://www.shiyanlou.com/questions/5430>)

148个资源让你成为CSS专家 (<https://www.shiyanlou.com/questions/5386>)



动手做实验，轻松学IT。

(<http://weibo.com/shiyanlou2013>)

公司

关于我们 (<https://www.shiyanlou.com/aboutus>)

联系我们 (<https://www.shiyanlou.com/contact>)

加入我们 (<http://www.simplecloud.cn/jobs.html>)

技术博客 (<https://blog.shiyanlou.com/>)

合作

我要投稿 (<https://www.shiyanlou.com/contribute>)

教师合作 (<https://www.shiyanlou.com/labs>)

高校合作 (<https://www.shiyanlou.com/edu/>)

友情链接 (<https://www.shiyanlou.com/friends>)

服务

实战训练营 (<https://www.shiyanlou.com/bootcamp/>)

会员服务 (<https://www.shiyanlou.com/vip>)

学习路径

Python学习路径 (<https://www.shiyanlou.com/paths/python>)

Linux学习路径 (<https://www.shiyanlou.com/paths/linuxdev>)

[实验报告 \(https://www.shiyanlou.com/courses/reports\)](https://www.shiyanlou.com/courses/reports)

[常见问题 \(https://www.shiyanlou.com/questions/?tag=常见问题\)](https://www.shiyanlou.com/questions/?tag=常见问题)

[隐私条款 \(https://www.shiyanlou.com/privacy\)](https://www.shiyanlou.com/privacy)

[大数据学习路径 \(https://www.shiyanlou.com/paths/bigdata\)](https://www.shiyanlou.com/paths/bigdata)

[Java学习路径 \(https://www.shiyanlou.com/paths/java\)](https://www.shiyanlou.com/paths/java)

[PHP学习路径 \(https://www.shiyanlou.com/paths/php\)](https://www.shiyanlou.com/paths/php)

[全部 \(https://www.shiyanlou.com/paths/\)](https://www.shiyanlou.com/paths/)

Copyright @2013-2016 实验楼在线教育

蜀ICP备13019762号 (<http://www.miibeian.gov.cn/>) 站长统计 (http://www.cnzz.com/stat/website.php?web_id=5902315)