从 GITHUB 到长尾理论

Github 社区的文化创意

2016-6-18

南京大学软件学院
141250019 崔浩

引言

通过 Github 网站提供的 API（应用程序编程接口），我们从 Github 上获取了约 30 万条用户信息，并存储在云端服务器数据库中。我们以此为基础进行数据分析，探究该社区用户数据分布特征，尤其是其中出现的长尾现象，以此发掘该社区近年来发展壮大背后的文化创意，以及分析该社区未来的发展趋势。

关键词

Github， 数据挖掘， 长尾理论， 帕累托分布， 文化创意

一、 Github：从开源代码库到开源协作社区

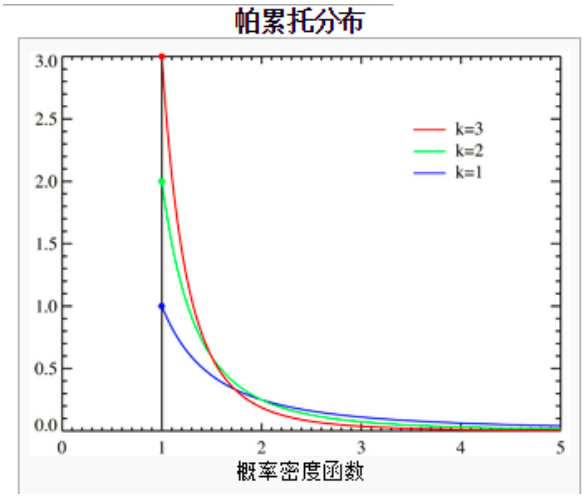
什么是 Github？根据维基百科的解释，“GitHub 是一个利用 Git 进行版本控制、专门用于存放软件代码与内容的共享虚拟主机服务”。用通俗的语言讲，Github 是一个免费的仓库，在这样的仓库里，你可以创建自己的小仓库，并存放属于自己的代码。除此之外，用户还可以邀请其他人和自己一起完成项目，在完成项目的过程中，Github 会协助进行版本控制，防止多人同时修改同一份文件而引起不必要的文件冲突。

这样一个纯技术的社区，却在 2007 年 10 月 1 日创建之后逐渐演化。截止到 2015 年，GitHub 已经有超过九百万注册用户和 2110 万代码库。（数据来源：维基百科）比这些数据更有价值的是，Github 已经逐渐演化为软件从业者交流技术，相互学习，寻找合作者或是发现开源项目的社交平台。Github 的核心功能依旧十分优秀，但越来越多的软件从业者开始活跃在这样的平台下，关注自己感兴趣的用户，订阅感兴趣的开源项目，并获得这些项目和用户的最新动态。

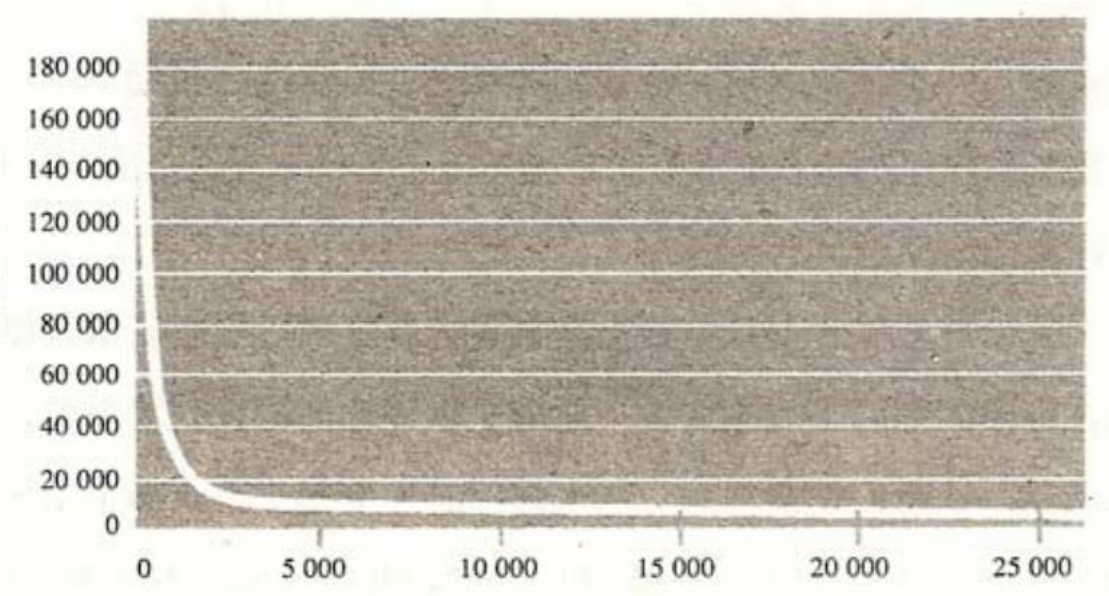
在 Github 发展演变的背后，有着怎样的特征？

二、 长尾理论与帕累托法则

帕累托法则，也称为二八定律或 80/20 法则，指的是 80%的结果取决于 20%的原因。帕累托分布是从大量真实世界的现象中发现的幂定律分布，分布图如下：



帕累托法则，往往引导人们专注热门，追逐最热潮流。大量的商家会关注为自己带来80%利益的 20%的群体，我们更关注吸引了 80%票房的 20%的电影，追捧销售量占据市场80%的专辑，尽管他们只占有专辑总量的 20%……而长尾理论却将关注点转向经常被忽略的80%，长尾理论指出，我们的文化和经济重心正在加速转移，从需求头部的少数大热门转向需求尾部曲线的大量利基产品和市场。长尾理论的条件是：转变需求的成本与原先几乎不变甚至更少，满足尾部需求的途径多样便捷。

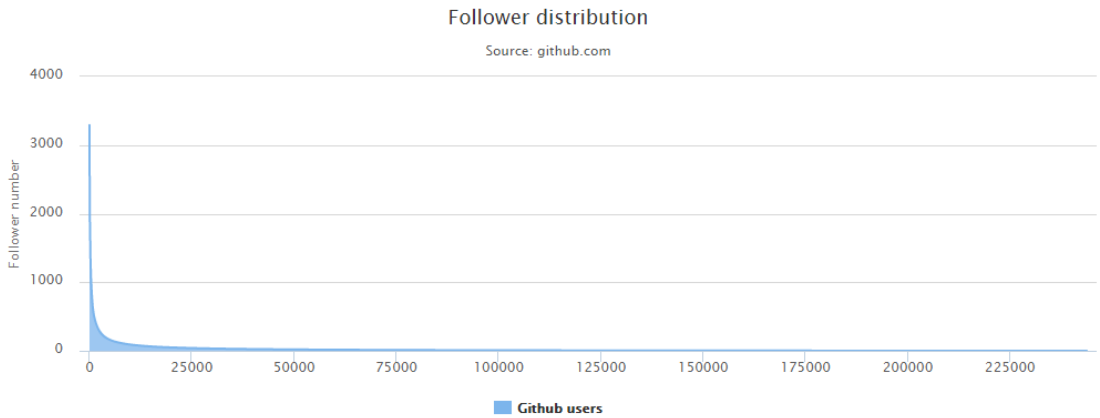


在很多情况下，部分分布会呈现上图所示的状态，呈现一条很长的尾巴，这条长尾或许比想象更长。若能有效开发长尾，利基产品可以创造一个客观的大市场。长尾的动力，一方面是 X 轴的延长，即内容越多，长尾越长，另一方面是获得利基产品的途径越多，产品越扁平化。

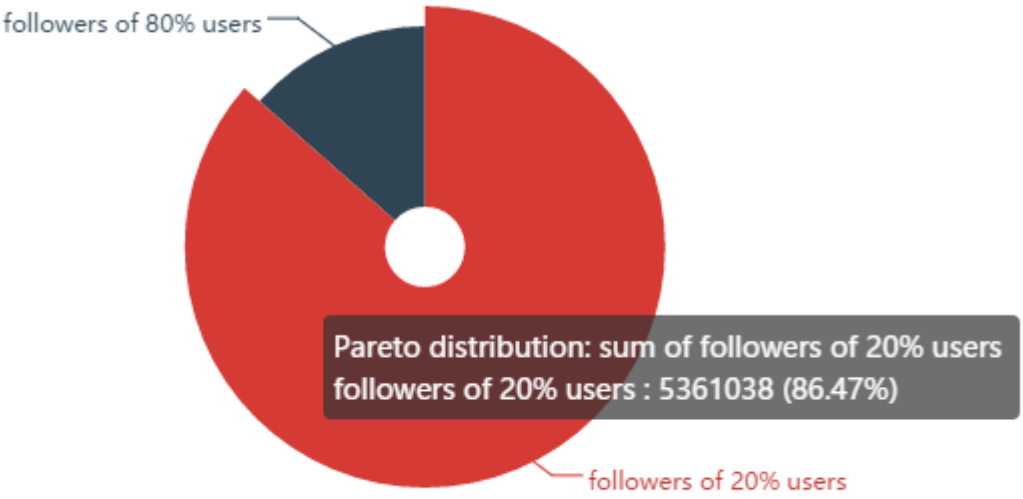
Github 的演变壮大，是否与这条曲线有关？

三、 Github 数据分析

Github 允许用户关注其他用户或被其他用户关注。我们统计用户的关注人数，以此代表这位用户在 Github 上受欢迎的程度。按照关注人数排序，我们绘制了以下图表：



可以看到，少数的用户获取了多数的关注者。这样的分布是否符合帕累托分布呢？考虑到卡方拟合检验的复杂性，我们这里采取了较为简单的验证方式，即直接验证二八定律。

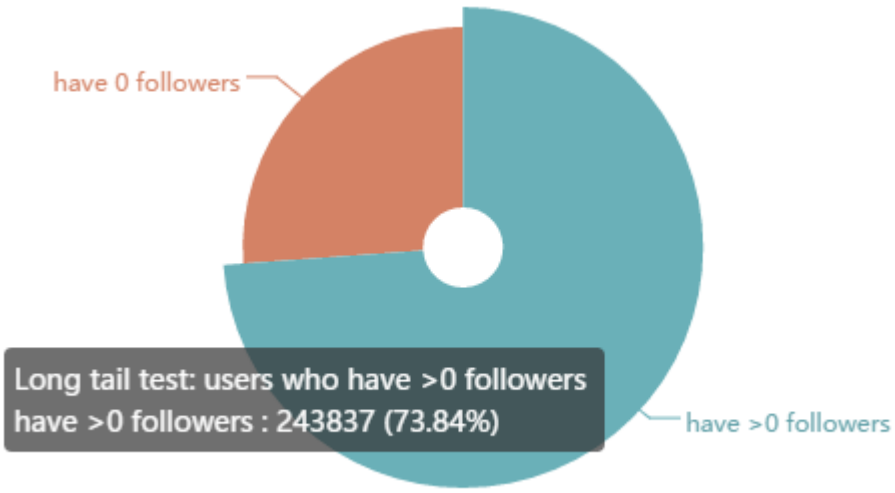


而我们的结论是，前 20%的用户拥有 86.47%的关注者数量，甚至超过了 80/20 的估计值。具体的数据如下：

数据视图

```
Pareto distribution: sum of followers of 20% users
followers of 20% users  5361038
followers of 80% users  839169
```

这是否意味着热门人物完全占领社区而不需要关注留下的长尾呢？为了展现长尾的长度，我统计了拥有至少一个关注者，即至少参与了这场社交运动的用户数：

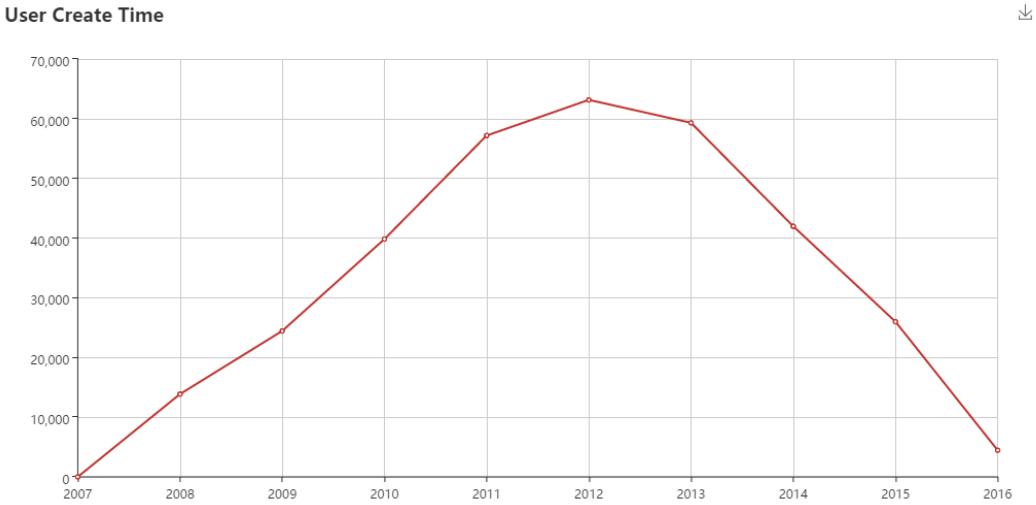


具体数据如下：

```
Long tail test: users who have >0 followers
have >0 followers       243837
have 0 followers        86389
```

也就是说，有 73.84%的用户至少被关注过一次，对于一个代码托管仓库，这样的成绩令人惊讶。Github 并不是极为优秀的少数用户的狂欢，而是属于大众的交流社区。尽管少数大神的光芒耀眼夺目，但有绝大多数的用户都获得过他人的赏识，这条长尾令人深思。

之前我们说过，长尾的实现有两个条件：第一个是更多的内容，在我们这次的数据分析中，主要体现是用户的数量，因此，我们统计 Github 用户每年的新增情况：



可以看到，尽管增速放缓，但 Github 用户人数一直保持正增长。

Github 的开放注册政策和免费代码托管政策无疑为自己带来了大量的用户，尽管无门槛的注册可能让 Github 拥有不少僵尸账号，但正是这样的开放促成了 Github 社区的活跃。少数人的优秀让人钦佩，但 Github 的普及和协作社交机制的形成，更是因为 Github 拥有了这样多元、各有特色、满足不同人认同点的用户群体，在成熟的技术之外，这样的开发也在一定程度上促成了 Github 的发展。

第二个要求则是更多获得利基产品的途径。Github 灵活的用户搜索功能，让查看任何一位用户的公开信息变得没有任何困难。这样的搜索优化使用户更容易找到自己欣赏的用户，并能方便地关注他们的动态。

四、长尾——未来的关注点？

开放和便捷的搜索在一定程度上促成了 Github 的成长，尽管 80/20 定律在这样的社区表现还非常明显，但长尾已经逐渐出现，并展现出令人惊讶的长度。随着互联网的发展和软件技术的大众化，这样的尾巴必定会越来越长。未来的 Github，是否更应该将文化中心和宣传重点放在这一部分的长尾上呢？

（备注：第三部分统计图表为原创生成，未经授权请勿转载）