Good Data Analysis

**概述**

从大量数据中得出真相和洞察力是一项功能强大但容易出错的工作。最好的数据分析员和有数据头脑的工程师在以可靠的数据声明方面享有盛誉。但是他们正在做些什么使他们具有信誉？我经常听到像谨慎和有条理的形容词，但是最谨慎和有条理的分析员实际上会做什么？

这不是一个小问题，特别是考虑到我们经常在Google上收集的数据类型。 我们不仅通常使用非常大的数据集，而且这些数据集非常丰富。 也就是说，每行数据通常具有许多属性。 当您将其与给定用户的事件的时间顺序结合在一起时，可以使用多种方式查看数据。 将其与典型的学术心理学实验进行对比，对于研究人员而言，查看每个数据点是微不足道的。 我们的大型高维数据集带来的问题与在整个科学工作历史的大部分时间内所遇到的问题都大不相同。

本文档总结了认真，有条理的分析人员在大型高维数据集上使用的想法和技术。 尽管本文档重点介绍了日志和实验分析中的数据，但是其中许多技术都得到了更广泛的应用。

本文档的其余部分包括三个部分，涵盖了数据分析的不同方面：

* 技术性：操作和检查数据的想法和技术。
* 流程：有关如何处理数据，要问哪些问题以及要检查哪些内容的建议。
* 心态：如何与他人合作和交流见解。

**技术**

下面介绍一些检查数据的技术。

1. 查看数据分布

大多数从业人员使用统计指标（例如，均值，中位数，标准差等）来传达有关分布的信息。 但是，通常应通过生成直方图，累积分布函数（CDF），分位数-分位数（Q-Q）图等来检查更丰富的分布表示形式。 这些更丰富的表示形式使您可以检测数据的重要特征，例如多峰行为或大量异常值。

1. 考虑离群值

仔细检查异常值，因为它们可能是煤矿中的金丝雀，这表明您的分析存在更多基本问题。 可以将异常值从数据中排除或将它们合并到一个“异常”类别中是很好的，但是您应该确保知道为什么数据最终归入该类别。

例如，查看点击次数最少的查询可能会揭示您漏记了某些元素的点击次数。 查看点击次数最高的查询可能会发现您多记录了点击次数。 另一方面，可能会有一些离群值，您将永远无法解释，因此您需要注意投入此任务的时间。

1. 考虑噪声

随机性的存在会欺骗我们。 有人认为：“ Google有太多数据； 噪音消失了。”这根本不是事实。 您生成的每个数据或数据统计指标都应附带一个有关此估计的置信度的概念（通过置信区间和p值等度量）。

1. 查看样本

每当您生成新的分析代码时，都需要查看基础数据中的示例以及您的代码如何解释这些示例。 如果不执行此步骤，几乎不可能生成任何复杂的工作代码。 您的分析正在从基础数据中提取许多细节，以生成有用的摘要。 通过查看各个示例的全部复杂性，您可以确信摘要是合理的。

如何采样这些样本很重要：

如果要对基础数据进行分类，请查看属于每个类的样本。

如果是更大的类，请查看更多样本。

如果要计算数字（例如页面加载时间），请确保查看极端示例（也许最快和最慢5％；您确实知道分布的样子，对吧？）以及整个空间的点数测量值。

1. 分割数据

切片意味着将您的数据分为多个子组，并分别查看每个子组的指标值。我们通常沿着浏览器，语言环境，域，设备类型等维度进行切片。如果潜在现象在各个子组中的工作方式可能不同，则必须对数据进行切片以确认是否确实如此。即使您不希望切片会产生不同的结果，但查看一些切片以确保内部一致性也可以使您更有信心确定自己在测量正确的事物。在某些情况下，特定切片可能具有不良数据，中断的用户交互，或者在某种程度上根本不同。

每当您对数据进行切片以比较两组时（例如实验vs.对照，甚至“时间A”与“时间B”），您都需要注意混合偏移。混合移位是指每个组的切片中的数据量不同时。可能会导致辛普森悖论和其他混乱。通常，如果两个组中一个切片中的相对数据量相同，则可以安全地进行比较。

1. 考虑实际意义

拥有大量数据时，可能会倾向于只关注统计意义或关注每位数据的细节。 但是您需要问自己：“即使X值确实比Y值高0.1％，这有关系吗？” 如果您无法理解/分类部分数据，则这一点尤其重要。 如果您无法理解日志中的某些用户代理字符串，那么它代表的是数据的0.1％还是10％，对应该调查这些情况的数量会有很大的不同。

另外，有时您的数据量很小。 许多变化在统计上看起来并不重要，但与声称这些变化是“中性的”不同。您必须问自己：“实际上仍然存在显着变化的可能性有多大？”

1. 随着时间维度的数据一致性

您几乎应该总是尝试按时间单位对数据进行切片，因为随着我们的系统随着时间的发展，对基础数据的许多干扰都会发生。 （我们通常使用几天，但其他时间单位也可能有用。）在功能或新数据收集的首次启动期间，从业人员经常仔细检查一切是否按预期进行。但是，随着时间的流逝，可能会出现许多破损或意外行为。

仅仅因为某一天或某几天的时间是异常值，并不意味着您应该丢弃相应的数据。将数据用作挂钩，以确定在丢弃某天之前某几天或某几天不同的原因。

查看每天的数据还可以使您了解数据的变化，最终导致置信区间或具有统计意义的声明。通常，这不应代替严格的置信区间计算，但经常进行较大的更改，您可以发现它们从每天的图表中就具有统计意义。

1. 确认并计算您的过滤

几乎每一个大数据分析都从过滤各个阶段的数据开始。 也许您只想考虑美国用户，网络搜索或带有广告的搜索。 无论如何，您必须：

确认并明确指定您正在执行的过滤。

计算每一步要过滤的数据量。

通常，执行后者的最佳方法是计算所有指标，即使对于要排除的人口也是如此。 然后，您可以查看这些数据来回答诸如“垃圾邮件过滤删除了哪些查询部分？”之类的问题。 （取决于您为什么要进行过滤，这种类型的分析可能并不总是可能的。）

1. 比率应具有清晰的分子和分母

最有趣的指标是基础指标的比率。 通常，有趣的过滤或其他数据选择隐藏在分子和分母的精确定义中。 例如，“查询/用户”实际上是以下哪个？

* 查询/查询用户
* 今天访问过Google的查询/用户
* 具有有效帐户的查询/用户（是的，我必须定义有效帐户）

在这里清楚一点可以避免自己和他人的困惑。

另一个特殊情况是只能对您的某些数据进行计算的指标。 例如，“点击时间”通常是指“在获得点击后的点击时间”。 每当您查看此类指标时，都需要确认过滤条件，并在比较的组之间寻找过滤条件的转变。

**流程**

本节包含有关如何处理数据，对数据要问哪些问题以及要检查什么的建议。

1. 单独的验证，描述和评估

我认为数据分析具有三个相互关联的阶段：

验证：我是否认为数据是自洽的，是否已正确收集并且代表了我的想法？

描述：此数据的客观解释是什么？例如，“用户进行分类为X的查询次数较少”，“在实验组中，X和Y之间的时间间隔增加了1％”和“更少的用户转到结果的下一页”。

评估：根据描述，数据是否告诉我们，对用户，对Google或对整个世界都在发生有益的事情？

通过分离这些阶段，您可以更轻松地与他人达成协议。描述应该是每个人都可以同意的数据。评估可能会引发更多辩论。如果不将“描述”和“评估”分开，则很有可能只会看到希望查看的数据的解释。此外，评估通常会更加困难，因为通常通过与其他功能和指标的严格比较来建立指标的规范值需要大量投资。

这些阶段不是线性进行的。浏览数据时，您可能会在各个阶段之间来回跳动，但是在任何时候都应该清楚自己处于哪个阶段。

1. 确认实验和数据收集设置
2. 检查什么不应该改变
3. 标准第一，定制第二
4. 测量两次或多次
5. 检查可重复性
6. 检查与过去测量的一致性
7. 新指标应首先应用于旧数据/特征
8. 提出假设并寻找证据
9. 探索性分析受益于端到端迭代
10. 注意反馈

**心态**