



調査 Survey



观点 POV



新闻 News

文 | 安德鲁·格林 (Andrew Green)，马里奥·帕奇 (Mario Paic)



简介

算法是计算或其他解决问题操作（特别是由计算机进行）的过程或所遵循的一组规则。算法在数字世界中无处不在，例如，银行是否愿意放贷，雇主是否会录用面试对象。

在网络世界中，谷歌的 PageRank 算法决定了其搜索引擎结果中的网站如何排名。同时，Facebook 的信息流（News Feed）算法控制着用户在社交网络中看到的内容。

在受众测量领域，算法通常用于进度计划分析软件，该软件由媒体策划人员、购买者和销售者使用，以估算可能看到或听到广告信息的人数。

算法也是更多地用于受众测量的更先进统计技术的核心，包括数据归属、数据融合和受众建模。

仅有调查数据本身，已不足以捕捉快速变化的媒体环境的复杂性，因后者拥有更多的内容选择和更多的访问传播平台。在很多人越来越不愿意参加漫长或乏味的调查时，我们却向受访者要求的越来越多。采用统计技术，我们能够从个体收集较少的数据，并允许他们降低回答的精确度。

这些统计方法的设计和执行，需要高水平的专业知识和技能。对于调查数据的用户而言，在跟上不断变化的受众特征和习惯的竞赛中，他们变得越来越关键。



“统计思维总有一天会像读写一样，成为一个有效率的公民的必备能力。”

H.G. 威尔斯 (H.G. Wells)

情境

想要确定一个人的血型，我们并不需要抽取他的全部血液。同样，我们可以通过抽样调查，并对人群进行询问或跟踪他们的行为，来获得很多信息。良好的市场调查实践要求，有效样本应尽可能地反映被测量人群的变化性。

在受众测量领域，我们知道性别、年龄、文化程度、职业、家庭人数、收入水平以及人们居住的区域都会影响到他们阅读某些报纸或杂志、观看某些节目或收听广播的可能性。因此，这些特征必须在用来测量媒体使用的任何样本中得到忠实反映。

非人口因素也很重要。例如，当询问数字行为时，正确地反映家庭拥有的媒体接收设备（PC、平板电脑、智能手机等）的数量和范围很重要，这些可能与使用水平有关。

然而，虽然对公司和政府而言，这些年来的调查已被证明非常有价值，但它们也有局限性。其中一个局限性在于，除非我们和每个人都进行交谈，并且他们都以完全诚实和完美的回忆来回答，否则他们永远不能完美地代表一个群体。

在市场研究中，经常使用统计加权来纠正这种不平衡。

就本文而言，更重要的局限性在于我们如何从受访者获得尽可能多的信息。人们准备花多少时间回答问题，是有自然限制的。考虑到很少有人拥有完美的回忆，我们也必须注意要清楚地问问题，不要问那些不合理的期望人们会记住的信息。

当今，我们面临的挑战是客户希望我们能收集越来越多的信息。

统计调整有助于以下方面

营销人员希望知道其目标消费者的一切，以最大限度地提高调查投资回报：

- 他们是誰（人口统计学、地理人口统计学、心理学等方面）
- 他们对所问及类别的品牌的看法是什么
- 他们的表现行为如何（购买水平、品牌选择等）
- 他们将来打算购买什么

策划和购买广告宣传活动者，需要找出触及和影响目标受众的最佳方法：

他们在不同的时间，接触到哪些媒体（电视节目、报纸、杂志、广播电台、网站、应用程序、海报板……）？

在一天中的不同时间，他们或多或少地关注或参与哪些媒体？

何时是向人们传达广告信息的最佳时机（对信息的接受度，他们什么时候会在市场上购买……）？



“只有看得见的算法才能信得过。”

但是没有一个受访者会同意回答这么多的问题。而且很多问题都无法准确回答。因此，有两个密切相关的统计技术，可用来帮助解决这个问题，即：数据归属和数据融合。

唐纳德·埃尔文·克努特 (Donald Ervin Knuth)

数据归属

如果调查问卷的回答缺失或不完整，我们可以通过查看类似的调查受访者给出的回答来推断这些回答是什么。这些回应的缺失可能是偶然发生的（人们忘记或忽略了），或是由于设计的原因（我们有太多的问题要问，所以问卷调查在不同的子样本之间分开进行）。

在这种情况下，我们设计两份（或更多）调查问卷，每份问卷都有一些共同的核心问题，但也有一些关于其他主题的单独问题。这些调查问卷可以同时分开用于不同但相似的人群样本，也可以在不同的时间进行（例如：问卷 A 可以进行一个月，然后问卷 B 再进行一个月，等等）。

假设是，我们可以使用每一样本成员的已知人口统计学和其他特征，以及他们对其他一般问题的回答，来匹配回答不同调查问卷的人。

然后我们得到第一组问题的回答，并把它们归为回答第二组问题的相匹配的受访者，反之亦然。这给了我们一个比人群的单个样本所能拥有的更大的回答数据库。

数据归属的示例

在巴西，Ipsos EGM 调查在所有主要地区进行了很长的面对面问卷调查。该调查问卷有两个版本，媒体和人口统计学方面的问题都相同，但涉及有关不同品牌和产品的使用情况以及态度的问题不同。每个版本的调查问卷都在交替的半年内提供，这样我们就可以把所有问题的回答合并到一个为期 12 个月的数据库中。

在澳大利亚，我们的 EMMA (Enhanced Media Metrics Australia) 调查用于测量媒体消费和产品使用情况。由于涵盖了大量的产品类别，所以调查中的所有受访者都会被问及有关最优先产品类别的使用情况（例如，他们驾车吗？他们消费什么样的食物和饮料？）

然后，我们对两个匹配的样本中的每一个，询问关于约一半产品类别的详细品牌问题。然后，将样本每一半的回应归属为样本的另一半，生成一个最终的综合数据集，从而为整个样本提供详细的品牌信息。

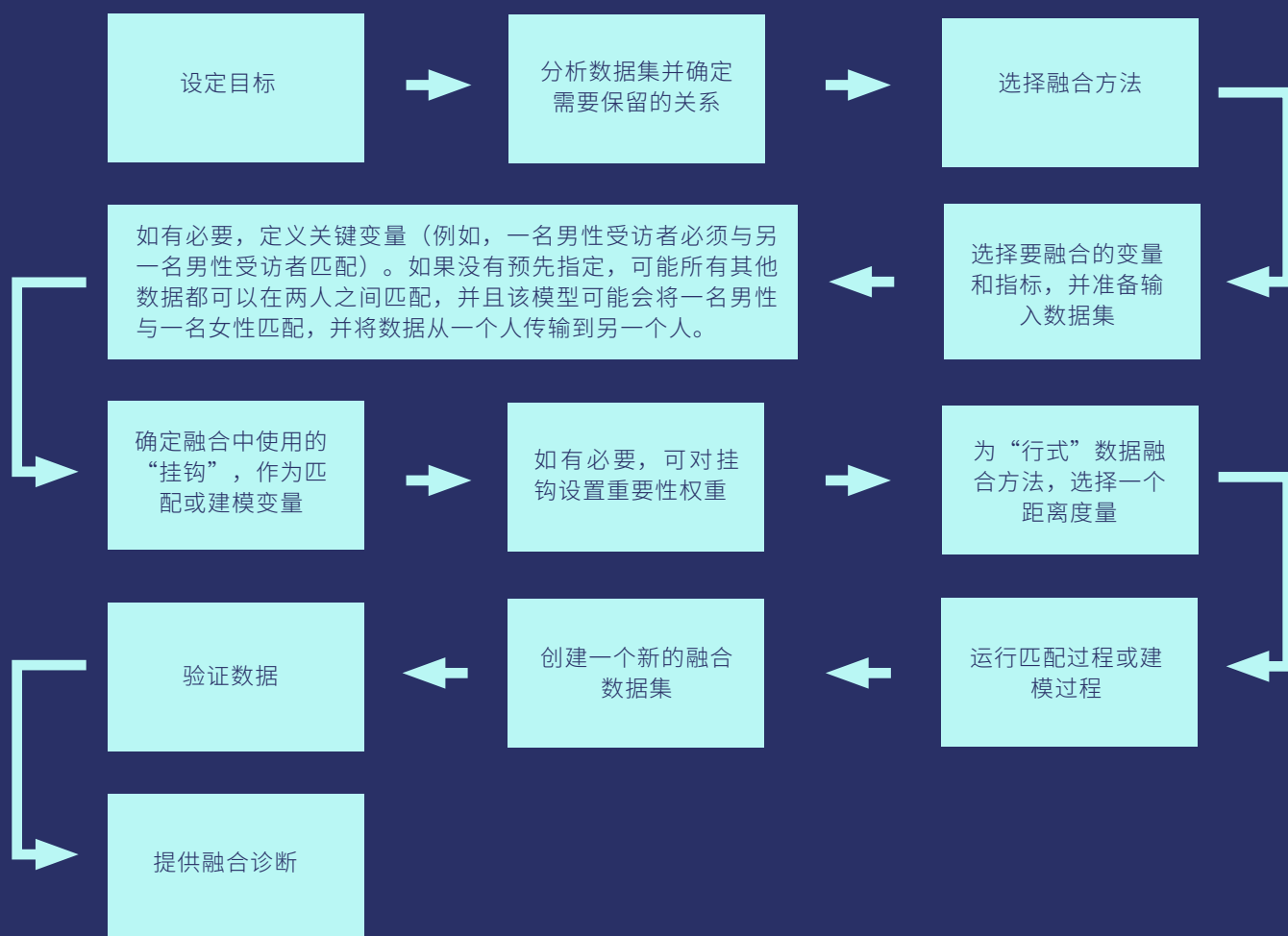
数据融合

与之相关的技术，是两个或两个以上单独调查的数据融合。在这种情况下，只要两项调查的受访者之间有足够的共同信息可以进行匹配，为一个目的而设计的一项调查就能与针对不同目的的一项调查相结合。这样能够生成一个包含所有此前单独信息的看起来像是单一源的数据库。

融合的关键是不同调查共同的“挂钩”的关联，即设计用于连接一项调查中的受访者与另一项调查中的受访者（称为“供体”受访者和“受体”受访者，因为来自一个受访者的数据最终将被“捐赠”给另一项调查的一个相匹配的受访者）。

两项调查中可能都会问及明显问题，包括性别、年龄、受教育水平和居住地区。融合过程寻求的是这些和其他标准的匹配——理想的情况是，举例来说，我们将在一项调查中居住在东南部的年轻、受过良好教育的男性与另一项调查中居住在相同区域的其他受过良好教育的男性进行匹配。

融合不是一种单一的技术——可以根据不同的目标采取不同的方法。但是，任何方法的原则都是相似的，并遵循以下一般步骤：



数据融合的示例

在英国，Ipsos Connect 设计了一种方法，用于组合两组受众数据——AMP（用于收集报纸和杂志的读者情况资料）和 UKOM（用于跟踪网站和应用程序的受众）。

增加融合设计的复杂性是一项要求，以便每一个单独的调查均“保留”其单独的受众估计。此外还需要引入 AMP 调查中的一部分，以捕获个人的跨平台的行为洞察（为了精确测量所有报纸和杂志内容平台的阅读量）。

在澳大利亚，Ipsos Connect 同样建立了一种数据融合方法，将来自 EMMA 调查（调查媒体使用情况）的输出与来自单独样本库的在线受众信息相结合，以便通过单独报纸和杂志的平台来创建总受众估计。

开发的数据集成过程涉及两个数据源之间的约束¹ 融合方法。两个来源的信息包括人口统计学、地理学和互联网活动变量信息，以生成最佳解决方案。

¹不同融合技术的详细说明超出了本文论述范围，选项包括“约束”或“无约束”融合（决定最终结果中必须保留的调查数据），以及我们是采用“行式”还是“列式”方法（处理数据集的不同方法）。

受众建模

受众建模是我们的另一大武器。我们的想法是获取可以帮助预测测量行为的已知信息，并将其组织成一个提供可信结果的过程。

我们希望用这种方法来增强我们的调查数据，至少有三个原因：



扩大我们调查覆盖面的范围。例如，在读者调查中，通过纳入因太小而无法单独在抽样调查中登记的出版物，我们可以增加所涵盖的出版物的数量。



增加我们报告的粒度。报告频率得以增加，因为我们不再需要建立一个足够大的样本来进行可靠报告，而只是受到模型输入数据报告频率（如发行或销售信息）的限制。



提高报告速度。模型不依赖于等待调查数据的收集和和处理，因而可以更快地发布。

受众建模示例

在澳大利亚，我们面临着为 400 多个地区、地方和社区出版物提供可靠读者估计的挑战，其中包括很多销售量不足 5000 份的出版物。要从一个样本获得统计学上有效的读者估计，将需要对不可行的大量受访者进行调查。

由于注意到以前的工作中所展示的报纸销售（发行）与读者层次之间的密切相关性，结合相同分布区域的人口统计资料和出版物本身的信息，我们构建了一个跨报纸分布区域的发行和发行分布模型。

在比利时，Ipsos 开发的新的建模技术正在帮助测量机构 CIM（Centre d' Information sur les Médias）创建世界上首个报纸每日受众数据。该模型由读者调查销售、每日短信样本库销售和来自出版商的每日销售数据组成，然后我们将这些数据捆绑在一起，为报纸提供每日读者信息，为杂志提供特定读者信息。

在英国，我们的“户外路线”研究采用了“多传感器”跟踪设备，以跟踪两周大量人群样本的出行行为和活动。这些数据与从地方主管部门取得的交通和行人流量信息以及该国每块广告板的详细位置信息相结合。借此，我们可以生成经过这些广告板的人群数量和概况的估计，用于规划和购买媒体空间。



“算法就像一个配方。”

Waseem Latif

结论

数字革命，意味着广告主和媒体公司需要比以往更多的关于媒体使用方面的信息。

然而，人们比过去更不愿意参与调查，即使在说服他们参与的时候，他们也不想参与无聊的调查。

随着时间的推移，这一需要更多信息而仅仅依靠调查却更难收集的难题，可能会变得愈加困难。

在受众测量领域……

诸如数据归属、数据融合和受众建模等技术，让我们能够收集和报告更多、更好的数据，使我们能够跟上日益复杂的客户需求。

数据科学的实际应用需要高水平的技能、专业知识以及经验——例如，建立融合和归属的很多决策和选择，需要判断和对情境的深入了解。

我们认为，在受众测量数据的处理方面，它将发挥越来越大的作用。



“统计事关或者应该事关科学调查，
以及如何做得更好……”

乔治 E.P. 博克斯 (George EP Box)