市场调研结合起来可以给出一个更全面的图景。

10.3 人工评估

人工评估指的是一个公司雇佣人工判断者（或者叫作*评估者*）来完成某些任务。这些结果会用于之后的分析。这在搜索和推荐系统中是一种常用的评估方法。简单的评估问题可以是“你偏好选项 A 还是 B？”或者“这个图片含有色情吗？”问题也可以逐渐变得更复杂，比如，“请标注这个图片”，或者“这个结果和搜索词有多相关”。更复杂的评估任务可能需要详细的说明，以保证评估结果能够被校准。一般来说，多个评估者会被分配做同一个任务，因为评估者间可能有不同意见。你可以用各种投票或者解决分歧的机制来得到高质量的汇总标签。例如，从类似于 Mechanical Turk（Mechanical Turk 2019）的收费系统获得的数据质量由于奖励机制和报酬金额不同而参差不齐，这令质量控制和分歧解决更加重要。

人工评估的一个局限性是，评估者一般来说不是你的最终用户。评估者执行分配给他们的任务（通常批量进行），然而你的产品是你的最终用户在生活中自然而然接触到的。另外，评估者可能不了解最终用户的当地情境。例如，搜索词“5/3”对于很多评估者来说是一个算术运算的搜索，会期待得到 1.667 的答案，但居住在“五三银行”（商标为“5/3”）附近的用户寻找的是有关这个银行的信息。这个例子说明了评估个性化推荐系统有多难。然而，这个局限性也可以成为一个优势，因为评估者可以被训练，以检测出用户无法感知或识别的垃圾信息或者有害体验。我们最好把人工评估提供的校准过的标签数据看作是对从真实用户收集的数据的补充。

基于人工评估的指标可以当作评估 A/B 实验的额外指标（Huffman 2008）。
130 再以搜索排序的改动为例，对于给定的搜索词，你可以要求评估者对从对照组或实验组得到的结果评分，并将这些评分汇总来比较哪一种变体更好；或者用一个并排实验，将对照组和实验组的搜索结果并排显示，要求评估者判断哪一边更好。例如，必应和谷歌的大规模人工评估项目足够快到可以和线上对照实验的结果一起使用来决定是否推出这个改动。

人工评估的结果还可以用于调试：你可以通过详细查验结果来了解这些改