第 1 章

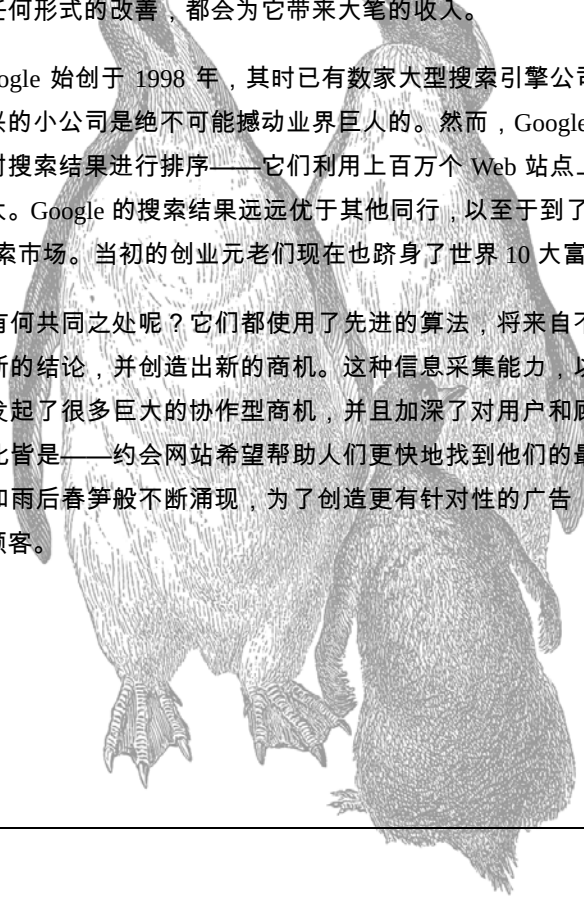
集体智慧导言

Introduction to Collective Intelligence

Netflix 是一家在线的 DVD 租赁公司，这家公司允许人们在线选购影片，并由公司负责送货上门；另外，Netflix 还会根据顾客以往的租片信息为其提供相应的推荐。2006 年底，该公司宣布将 100 万美元的奖金奖励给第一位能够将其推荐系统的精确度提升至 10%的人，并且只要该项竞赛还在进行，公司每年就会将 5 万美元的进步奖授予当时的头名状元。于是，数千个来自世界各地的团队蜂拥而至，截止 2007 年 4 月为止，领先的团队已经成功取得了 7%的成绩。Netflix 的推荐系统可以利用每位老顾客对影片的喜好数据，为以前从未访问过该网站的其他顾客提供推荐，并保证他们能够再次光顾网站。从 Netflix 的角度而言，无论对推荐系统采取任何形式的改善，都会为它带来大笔的收入。

搜索引擎公司 Google 始创于 1998 年，其时已有数家大型搜索引擎公司存在，而且许多人都认为，一家新兴的小公司是绝不可能撼动业界巨人的。然而，Google 的创立者们采用了一种全新的方法对搜索结果进行排序——它们利用上百万个 Web 站点上的链接来决定哪些页面的相关性最大。Google 的搜索结果远远优于其他同行，以至于到了 2004 年，它已占据了 85%的 Web 搜索市场。当初的创业元老们现在也跻身了世界 10 大富豪之列。

上述这两家公司有何共同之处呢？它们都使用了先进的算法，将来自不同人群的数据加以组合，进而得出新的结论，并创造出新的商机。这种信息采集能力，以及对其加以解释的计算能力已经激发起了很多巨大的协作型商机，并且加深了对用户和顾客更好的理解。这样的例子现在比比皆是——约会网站希望帮助人们更快地找到他们的最佳拍档，预测机票价格变化的公司如雨后春笋般不断涌现，为了创造更有针对性的广告，几乎每个人都想更好地了解他们的顾客。



上面提到的，仅仅是集体智慧（collective intelligence）这一令人振奋的新兴领域中少数几个典型的例子，层出不穷的新服务意味着每天都会有新的商机涌现出来。笔者相信，理解机器学习和统计方法在许多不同领域里都会变得愈加重要，而这一点在针对海量信息的解释和组织方面尤为突出（全世界的人们正在不断创造这些信息）。

什么是集体智慧

What Is Collective Intelligence?

人们使用集体智慧这一术语已有十多年之久，随着新型通信技术的出现，这一术语也变得日趋流行和重要。尽管这样的表达也许会让人联想到群体意识或超自然现象，但当技术人员使用这一词汇时，其含义通常是指：为了创造新的想法，而将一群人的行为、偏好或思想组合在一起。

当然，集体智慧的出现可能要早于 Internet。为了从全无关系的一群人中搜集、组合和分析数据，我们不一定要借助于 Web。完成这项工作的一种最为基础的方法，便是使用调查问卷或普查。从一大群人中搜集的答案可以使我们得出关于群组的统计结论：组中的个体成员将会被忽视。从独立的数据提供者那里得出新的结论，是集体智慧所真正关注的。

这里有一个众所周知的例子，是关于金融市场的。在金融市场里，价格并不是由某个个体或某种协作力量所决定的，它是由许多独立个体的交易行为所共同决定的，所有人的行为都建立在这样一种信念基础之上：他们相信当前的交易会为他们带来最大的利益。尽管乍一看这似乎违背直觉，但在未来的市场上，大量的参与者都是根据他们对未来价格的信心而进行契约交易的，这样的市场在价格预测的效果方面，往往被认为要比独立进行预测的专家们表现得更好。这是因为，市场将知识、经验和成百上千人的意志组织在一起，形成了一种不依赖个人观点的预测。

尽管寻求集体智慧的方法在 Internet 之前就已经存在，但自从有了 Internet 之后，从数千甚至数百万网民中搜集信息的能力为人们提供了许多新的可能。一直以来，人们都在利用 Internet 来购买所需、搜索信息、寻求娱乐，以及架设自己的 Web 站点。所有这些行为都可以得到监控，并且不必要求用户放下手头的工作来接受询问，而可以借由监控得到的信息提取出有价值的结论。有大量的方法可以用来对这些信息进行加工和解释。这里有两个重要的例子，分别体现了两种彼此对立的做法。

- *Wikipedia* 是一个在线的百科全书，它完全是由用户维护的。任何人都可以新建或编辑

网站上的任何一个页面，同时会有为数不多的几名管理员对一再出现的不当内容进行监控。Wikipedia 拥有的词条比其他任何百科全书还要多，尽管存在一些恶意用户的操作，但是人们普遍认为，Wikipedia 的大多数主题都是准确的。这便是集体智慧的一

个例子：因为每一篇文章都有大量人员在维护。而其最终的结果，则形成了一个任何单一协作团队都无法企及的大型百科全书。Wikipedia 软件并没有对用户贡献的内容进行特殊的智能处理，它只跟踪内容的变更情况，并显示最新的版本。

- 前文提及的 *Google*，是世界上最为流行的 Internet 搜索引擎，也是第一个根据其他网页对当前网页的引用数多少来评价网页等级的搜索引擎。这种评价等级的方法，搜集了数以千计的人对某一页面的评价信息，然后利用这些信息对搜索结果进行排序。这是集体智慧的一个非同寻常的例子。Wikipedia 明确邀请网站的用户提供内容，而 Google 则是从 Web 内容的创建者对自己网站的操作中提取重要的信息，并利用这些信息为 Google 的使用者设定各个网站的分值。

虽然 Wikipedia 是一个巨大的资源库，而且也是展现集体智慧的一个令人印象深刻的例子，但它的存在很大程度上要归功于提供内容的用户，而非软件中的那些智能算法。本书的焦点并不在于提供内容的用户，而在于算法，这其中就包括了 Google 的 PageRank 算法，该算法会搜集用户的数据，对数据进行计算分析，并从中创造出可以增强用户体验的新信息。在获得的这些数据当中，有一部分是明确搜集而来的，比如向用户询问与评价网页级别相关的问题。另一部分则是偶然搜集得到的，比如观察用户的购买行为。对于这两种情况，重要的不仅是搜集和显示信息，还包括以一种智能化的方式对这些信息加以处理，并产生出新的信息来。

本书将告诉你如何利用开放的 API 来搜集数据，同时还会讨论到各种机器学习算法和统计方法。将二者结合起来，就可以借助集体智慧的相关方法，对由自己编写的应用程序搜集得到的数据进行分析；同时，也可以从其他地方搜集数据，并对数据进行试验。

什么是机器学习

What Is Machine Learning?

机器学习是人工智能（AI，artificial intelligence）领域中与算法相关的一个子域，它允许计算机不断地进行学习。大多数情况下，这相当于将一组数据传递给算法，并由算法推断出与这些数据的属性相关的信息——借助这些信息，算法就能够预测出未来有可能会出现的其他数据。这种预测是完全有可能的，因为几乎所有的非随机数据中，都会包含这样或那样的“模式（patterns）”，这些模式的存在使机器得以据此进行归纳。为了实现归纳，机器会利用它所认定的出现于数据中的重要特征对数据进行“训练”，并借此得到一个模型。

为了理解模型得到的过程，我们来看另外一个复杂领域——电子邮件过滤中的一个简单例子。假定我们收到了大量包含“online pharmacy”单词的垃圾邮件。对人而言，我们可以很轻松地识别出其中的模式，并快速确知任何含有“online pharmacy”单词的信息都是垃圾邮

件，应该将其直接移到垃圾箱中。这就是归纳——事实上，我们已经建立起了一个关于垃圾邮件的智力模型。当我们将多条这样的信息报告为垃圾邮件之后，专门设计用以过滤垃圾邮件的机器学习算法应该有能力做出同样的归纳来。

有许多不同的机器学习算法，所有算法都各有所长，适应于不同类型的问题。有些算法，比如决策树，非常的直观，通过眼睛观察就可以完全理解机器执行的推导过程。另有一些算法，比如神经网络，则像一个黑盒，它们虽然也给出最终的结果，但通常要复现蕴含在这些结果背后的推导过程则是非常困难的。

许多机器学习算法都很倚仗数学和统计学。根据笔者早些时候给出的定义，我们甚至可以认为，简单的相关性分析和回归都是机器学习的基本形式。本书并没有假定读者具备许多统计学方面的知识，所以笔者会尝试尽可能直观地解释所用到的统计学知识。

机器学习的局限

Limits of Machine Learning

机器学习并非没有缺点。机器学习算法受限于其在大量模式之上的归纳能力，而一个模式如果不同于算法先前所曾见到过的任何其他模式，那么它很有可能会被“误解”。人类拥有大量的文化知识及经验可以借鉴；不仅如此，人们还具备一种非凡的能力，即：当对新的信息进行决策时，人们能够从中识别出相似的信息来，而机器学习方法却只能凭借已经见过的数据进行归纳，而且归纳的方式受到很大的限制。

我们将在本书中见到的垃圾邮件过滤方法，是以单词或单词组合的出现为依据的，至于这些单词的含义及句式结构，则根本未予考虑。尽管在理论上，构造一个考虑语法的算法是可行的，但在现实中却很少这样做，这是因为为此付出的努力与算法的改进相比很不成比例。理解单词的含义及单词与个人生活的相关性所要求掌握的信息，远比垃圾邮件过滤算法中所能访问到的现有信息还要多。

另外，尽管在解决问题的倾向性上各有不同，但是所有机器学习算法都有过度归纳的可能性。就如生活中的大多数事情一样，基于少数示例的强归纳很少是完全精确的。我们的确有可能会收到友人寄来的一封重要邮件，里面包含“online pharmacy”的字样。在这种情况下，我们须要告诉算法这不是垃圾邮件，或许算法可以作出判断，将来自某位好友的邮件判定为可以接收。究其本质，许多机器学习算法在新信息到来之时都是能够持续进行学习的。

真实生活中的例子
Real-Life Examples

当前 Internet 上有大量站点正在不断地从广大用户当中搜集数据，并利用机器学习和统计方法从中获益。Google 是其中的佼佼者——它不仅可以利用 Web 链接对网页进行排名，而且当其广告被不同的用户点击时，它会持续搜集信息，这使得 Google 可以更加有效地进行广告定位。在第 4 章中，我们将了解到搜索引擎和 PageRank 算法，这是 Google 排名系统的重要组成部分。

其他的例子还包括带有推荐系统的 Web 站点。如 Amazon 和 Netflix 这样的站点，它们利用人们的购买或租赁信息来确定人或物品的相似程度，然后再根据买卖历史来给出推荐。另有一些站点，比如 Pandora 和 Last.fm，则可以利用我们对不同乐队和歌曲的评价来建立定制的广播电台，其中包含了网站认为我们会喜欢的音乐。第 2 章将会讨论构建推荐系统的方法。

市场预测也是集体智慧的一种形式。这其中最为有名的一个例子莫过于Hollywood Stock Exchange（*http://hsx.com*），在那里人们可以进行涉及影片和影星的模拟股票交易。我们可以按照影片的当前价格买卖股票，其对应的价值相当于电影实际首次票房收入的百万分之一。因为价格是通过交易行为来设定的，所以价值不由任何一个个体所决定，而是由群体的行为来确定的，股票的当前价格可以看作是整个群体对电影票房收入数字的预测。通常而言，由Hollywood Stock Exchange所给出的预测往往要优于某位专家所给出的预测。

某些交友网站，比如 eHarmony，利用从参与者那里搜集而来的信息确定交友的最佳配对。尽管这些公司对他们所采用的匹配算法守口如瓶，但是任何一种成功的匹配算法很可能都会涉及一个持续不断的求值过程——算法会反复判断选定的匹配成功与否。

学习型算法的其他用途
Other Uses for Learning Algorithms

本书所介绍的并不是什么新方法，尽管这些例子都是在讨论基于 Internet 的集体智慧的相关问题，但是掌握机器学习算法的知识对于许多其他领域的软件开发者而言也是很有助益的。尤其是在须要处理大量数据的领域里，我们可以从中发掘出值得关注的如下各种模式。

生物工艺学

人类在测序技术和筛选技术（sequencing and screening technology）上的进步已经创造出了许多不同种类的海量数据，比如 DNA 序列、蛋白质结构、化合物筛选及 RNA 表

达。为了找到能进一步理解生物进程的模式，机器学习技术被广泛应用于所有这些类型的数据之中。

金融欺诈侦测

信用卡公司一直都在寻找侦测交易是否存在欺诈行为的新方法。最终，他们使用了像神经网络和归纳逻辑这样的技术，对交易行为进行检验，并捕获不正当的使用方法。

机器视觉

出于军事或监控的目的，从摄像机中进行图片解析是一个活跃的研究领域。许多机器学习技术被用来自动侦测入侵者、辨别车辆，或者识别人脸。尤其值得注意的是无人监控技术的使用，比如能从大数据集中发现有趣特征的独立组元分析技术。

产品市场化

长期以来，对人口统计资料及其发展趋势的理解被认为是一种艺术而不是科学。最近，人们在消费者数据搜集能力方面的增长，为机器学习技术打开了机会之门，比如聚类方法，就能很好地理解存在于市场中的自然划分，并能更好地预测未来的趋势。

供应链优化

许多企业通过其供应链的有效运行及精确预测不同区域的产品需求，来节省数以百万计的成本投入。构造供应链的方法非常多，影响需求的潜在因素也非常多。优化和学习技术时常被用来分析这些数据集。

股票市场分析

自从有了股票市场，人们就一直在尝试利用数学方法来赚取更多的钱。随着参与股市的股民变得越来越有经验，对大量数据进行分析并采用先进技术来侦测模式已经变得很有必要了。

国家安全

全世界的政府机构都在搜集海量信息，对这些数据的分析过程要求计算机对模式进行检测，并将之与潜在的威胁联系起来。

上述这些仅仅是人们现在大量使用机器学习的典型个案。既然有越来越多的信息被制造出来已是大势所趋，那么有越来越多的领域将依赖于机器学习和统计技术并不是没有可能的，

因为信息扩张的规模已经超出了人们利用旧有方法进行处理的能力。

每天可以获得的新信息有多少，显然就会有多少更多的可能性。一旦你掌握了一点机器学习的算法，你就会发现它们的应用随处可见。