第140封信 | 图论和计算机科学

吴军

第140封信 | 图论和计算机科..
08:42 4.07MB

信件朗读者：宝木

小师弟，你好！

不少读者听了我解释Google的一些做事哲学，特别是信奉算法和机器的作用之后，在留言中问我这样一件事：Google是否下载和存储所有的网页?如果这样做，成本是否太高了?如果不是办不到的话。如果不这样做，似乎也无法在很短的时间内找到成千上万条结果。

对于这个问题简单的回答是肯定的。当然背后的原理和做法就不是一两句话能说清楚的了。因此我们花两天时间介绍这方面的知识。此外，这个问题本身和Google的一道面试题直接相关。

今天我们先介绍一些背景知识和基本原理，明天我们再介绍Google是如何找到和下载所有网页的。即使你不学习计算机，了解这些知识对你了解互联网也是有好处的。

找到所有的网页要用到一个被称为图论的数学工具，它是离散数学的一个重要分支，也是计算机科学的数学基础。我们之前介绍过的有关二进制的布尔代数也是离散数学的一部分。

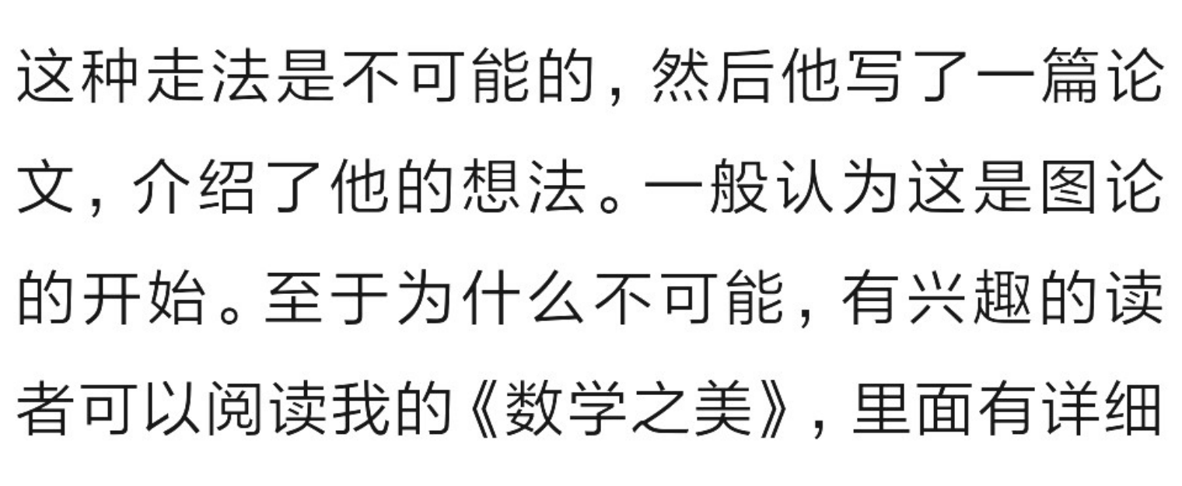
很多人问我，数学不好是不是学不好计算机？其实你中学和大学数学没学好并没有太大关系，因为计算机需要的是离散数学，和几何、代数、微积分都没有什么关系。不过，如果缺乏逻辑性，是搞不了计算机的。

顺便提一句，用Google Trends来搜索一下"离散数学"这个词，可以发现不少有趣的现象。比如，北京、武汉、哈尔滨、合肥和长沙这些城市对这一数学主题最有兴趣，而这和中国大学的分布是一致的。

接下来我们就讲讲图论。我不知道你在中学的时候是否做过这样一道智力题，讲的是大数学家欧拉的故事。

1736年，欧拉来到普鲁士的哥尼斯堡（Konigsberg）。这是大哲学家康德的故乡，不过现在是俄罗斯的加里宁格勒了。欧拉发现当地居民有一项消遣活动，就是试图将下图中的每座桥恰好走过一遍并回到原出发点，但是从来没有人成功过。

欧拉后来发明了一种数学工具证明了这种走法是不可能的，然后他写了一篇论文，介绍了他的想法。一般认为这是图论的开始。至于为什么不可能，有兴趣的读者可以阅读我的《数学之美》，里面有详细的讲解。



哥尼斯堡的七座桥

欧拉的工具简单地讲，就是把地图抽象成一种仅仅包含了节点和连接这些节点的弧组成的图。

举一个例子，如果我们把中国的城市当成节点，连接城市的铁路当成弧，那么全国的铁路网就是图论中所说的图。

图论是生活中的一个抽象的概念或者说是工具，围绕它，计算机科学家们设计了很多算法，然后把很多实际问题抽象出来，用图论的算法解决。

比如我们在图论中有一个"中国邮差问题"，即一个邮差怎么不重复（或者最少重复）地挨家挨户把信送到。这本身是一个实际问题，而它又可以被抽象出来——每一家是图中的节点，道路是连接节点的弧。顺便说一句，解决这个问题的数学家是我的论文委员会主席高德曼（Alan Goldman）教授，他在约翰·霍普金斯大学教我博弈论的课程。当然，这个问题解决之后，很多类似的问题都迎刃而解，这就是理论的意义。

关于图的算法有很多，但最重要的是图的遍历算法，也就是如何从一个节点出发，通过弧访问图的各个节点。以中国铁路网为例，我们怎样从北京出发，访问中国所有的城市？当然，你要尽可能地避免兜圈子，也要避免一个城市被过多地访问。

遍历算法总的来讲有两种。第一种叫做深度优先算法（在计算机领域大家称它为DFS，Depth First Search的首字母缩写）。

这种方法简单地讲就是一条道走到黑。从北京出发，随便找一个相连的城市，比如石家庄，作为到下一个要访问的城市，再从石家庄出发，找一个相连的城市，比如说济南，然后走下去，经过南京、上海、杭州等等，最后走到了深圳，前面再也没有城市了，然后再往回找，比如退回到广州，看看是否有与广州相连的、尚未访问的城市，比如长沙还没有访问，接着就去长沙，就这样走下去。只要图是相连的，一定能走完所有的城市。

当然，为了避免同一个城市被访问多次或者漏掉哪个城市，你需要准备一个小本本，记录已经访问过的城市，或者直接在去过的城市插一面小红旗，看到红旗，下次再遇到就跳过去。此外，为了知道每一次往后退，退回到哪个城市，需要用我们前面介绍过的堆栈来管理路径。

第二种方法是先从北京出发，先把所有直接相连的城市走遍，比如天津、济南、石家庄、沈阳、呼和浩特和北京直接相连，先访问这些城市。然后再从与天津相连的城市开始做第二轮的走访，之后，再走访和济南相连的城市。最后走到离北京最远的深圳。

这种走法因为是尽可能广地访问各个城市，被称为广度优先算法（在计算机领域大家称它为BFS，Breadth First Search的首字母缩写），它可以保证访问到全部的城市。当然，为了防止你走冤枉路，这种算法也需要一个小本本做记录，或者在访问过的城市上插红旗。此外，为了知道下一轮广度优先的遍历从哪个城市开始，需要用我们前面介绍过的队列来管理路径。

当然，铁路是可以双向开通的，而在城市里的有些道路则是单向的。在图论中为了区别这两种情况，并且找到相应的算法，一个联通的图也被分为了无向图，也就是双向联通的，和有向图两种。在北京有很多单行线的地方，对应的就是有向图。

讲到这里可能有人会问，城市间的旅游和下载网页有什么关系？关系很大，因为如果你要是把一个网页看成是图中的节点，把它们之间的超链接（Hyperlinks）看成是弧，整个互联网不就是一个有向图吗?

事实上，当你在一个网页中，将鼠标挪到那些蓝色、带有下划线的文字上，会显示出藏着的对应的网址，当你点击的时候，浏览器通过这些隐含的网址跳转到相应的网页，这个经验大家都有。自动下载网页，用的就是这个原理。

我在Google经常问候选人的一个面试问题就是，如何建立一个互联网的网络爬虫，将互联网上所有的网页发现并且下载下来？所谓网络爬虫，也称为"机器人"（Robot)，它是一个程序，可以从任何一个网页出发，用图的遍历算法，自动地访问到每一个网页并把它们存起来。

这个问题是一个开放的问题，它从理论上讲并不很复杂，但是在工程上有非常多的技巧。通过这个问题，很容易考察一个人的工程素养和对互联网的了解。明天我就来详细讲解这个问题，我先把问题给你，这样你就有一天的思考时间。

最后总结一下今天的内容：

1.很多实际问题都有共性，将这些共性提炼出来，就形成了理论，而一个理论问题解决后，很多实际问题就迎刃而解了。

2. 好的理论其实都很简单，你已经看到了图论这个数学工具的原理一点也不复杂。

思考题：1. 思考一下为什么哥尼斯堡七桥问题无解；2. 思考一下实际问题抽象化的意义。

祝近安

吴军

2018年7月24日