谷歌背后的数学

卢昌海

在如今这个互联网时代，有一家家喻户晓的公司，它自1998年问世以来，在极短的时间内就声誉鹊起，不仅超越了所有竞争对手，而且彻底改观了整个互联网的生态。这家公司就是当今互联网上的第一搜索引擎：谷歌(Google)。

在这样一家显赫的公司背后，自然有许许多多商战故事，也有许许多多成功因素。但与普通商战故事不同的是，在谷歌的成功背后起着最关键作用的却是一个数学因素。

本文要谈的就是这个数学因素。

谷歌作为一个搜索引擎，它的核心功能顾名思义，就是网页搜索。说到搜索，我们都不陌生，因为那是凡地球人都会的技能。我们在字典里查个生字，在图书馆里找本图书，甚至在商店里寻一种商品等，都是搜索。如果我们稍稍推究一下的话，就会发现那些搜索之所以可能，并且人人都会，在很大程度上得益于以下三条：

1. 搜索对象的数量较小——比如一本字典收录的字通常只有一两万个，一家图书馆收录的不重复图书通常不超过几十万种，一家商店的商品通常不超过几万种等。

2. 搜索对象具有良好的分类或排序——比如字典里的字按拼音排序，图书馆里的图书按主题分类，商店里的商品按品种或用途分类等。

3. 搜索结果的重复度较低——比如字典里的同音字通常不超过几十个，图书馆里的同名图书和商店里的同种商品通常也不超过几十种。

但互联网的鲜明特点却是以上三条无一满足。事实上，即便在谷歌问世之前，互联网上的网页总数就已超过了诸如图书馆藏书数量之类传统搜索对象的数目。而且这还只是冰山一角，因为与搜索图书时单纯的书名搜索不同，互联网上的搜索往往是对网页内容的直接搜索，这相当于将图书内的每一个字都变成了搜索对象，由此导致的数量才是真正惊人的，它不仅直接破坏了上述第一条，而且连带破坏了二、三两条。在互联网发展的早期，象Yahoo那样的门户网站曾试图为网页建立分类系统，但随着网页数量的激增，这种做法很快就“挂一漏万”了。而搜索结果的重复度更是以快得不能再快的速度走向失控。这其实是可以预料的，因为几乎所有网页都离不开几千个常用词，因此除非搜索生僻词，否则出现几十万、几百万、甚至几千万条搜索结果都是不足为奇的。

互联网的这些“不良特点”给搜索引擎的设计带来了极大的挑战。而在这些挑战之中，相对来说，对一、二两条的破坏是比较容易解决的，因为那主要是对搜索引擎的存储空间和计算能力提出了较高要求，只要有足够多的钱来买“装备”，这些还算是容易解决的。套用电视连续剧《蜗居》中

某贪官的台词来说，“能用钱解决的问题就不是大问题”。但对第三条的破坏却要了命了，因为无论搜索引擎的硬件如何强大，速度如何快捷，要是搜索结果有几百万条，那么任何用户想从其中“海选”出自己真正想要的东西都是几乎不可能的。这一点对早期搜索引擎来说可谓是致命伤，而且它不是用钱就能解决的问题。

这致命伤该如何治疗呢？药方其实很简单，那就是对搜索结果进行排序，把用户最有可能需要的网页排在最前面，以确保用户能很方便地找到它们。但问题是：网页的水平千差万别，用户的喜好更是万别千差，互联网上有一句流行语叫做：“在互联网上，没人知道你是一条狗”(On the Internet, nobody knows you're a dog)。连用户是人是狗都“没人知道”，搜索引擎又怎能知道哪些搜索结果是用户最有可能需要的，并对它们进行排序呢？

在谷歌主导互联网搜索之前，多数搜索引擎采用的排序方法，是以被搜索词语在网页中的出现次数来决定排序，出现次数越多的网页排在越前面。这个判据不能说毫无道理，因为用户搜索一个词语，通常表明对该词语感兴趣。既然如此，那该词语在网页中的出现次数越多，就越有可能表示该网页是用户所需要的。可惜的是，这个貌似合理的方法实际上却行不大通。因为按照这种方法，任何一个像祥林嫂一样翻来复去倒腾某些关键词的网页，无论水平多烂，一旦被搜索到，都立刻会“金榜题名”，这简直就是广告及垃圾网页制造者的天堂。事实上，当时几乎没有一个搜索引擎不被“祥林嫂”们所困扰，其中最具讽刺意味的是：堪称互联网巨子的当年四大搜索引擎在搜索自己公司的名字时，居然只有一个能使之出现在搜索结果的前十名内，其余全被“祥林嫂”们挤跑了。

就是在这种情况下，1996 年初，谷歌公司的创始人，当时还是美国斯坦福大学研究生的佩奇 (Larry Page) 和布林 (Sergey Brin) 开始了对网页排序问题的研究。这两位小伙子之所以研究网页排序问题，一来是导师的建议（佩奇后来称该建议为“我有生以来得到过的最好建议”），二来则是因为他们对这一问题背后的数学产生了兴趣。

网页排序问题的背后有什么样的数学呢？这得从佩奇和布林看待这一问题的思路说起。在佩奇和布林看来，网页的排序是不能靠每个网页自己来标榜的，无论把关键词重复多少次，垃圾网页依然是垃圾网页。那么，究竟什么才是网页排序的可靠依据呢？出生于书香门第的佩奇和布林（两人的父亲都是大学教授）想到了学术界评判学术论文重要性的通用方法，那就是看论文的引用次数。在互联网上，与论文引用相类似的显然是网页链接。因此，佩奇和布林萌生了一个网页排序的思路，那就是通过研究网页间的相互链接来确定排序。具体地说，一个网页被其它网页链接得越多，它的排序就越靠前。不仅如此，佩奇和布林还进一步提出，一个网页越是被排序靠前的网页所链接，它的排序就也应该越靠前。这一条的意义也是不言而喻的，就好比一篇论文被诺贝尔奖得主所引用，显然要比被普通研究者所引用更说明其价值。依照这个思路，网页排序问题就跟整个互联网的链接结构产生了关系，正是这一关系使它成为了一个不折不扣的数学问题。

思路虽然有了，具体计算却并非易事，因为按照这种思路，想要知道一个网页 W_i 的排序，不仅要知道有多少网页链接了它，而且还得知道那些网页各自的排序——因为来自排序靠前网页的链接更“值钱”。但作为互联网大家庭的一员，W_i 本身对其它网页的排序也是有贡献的，而且基于来自排序靠前网页的链接更“值钱”的原则，这种贡献与 W_i 的排序有关。这样一来，我们就陷入了一个“先有鸡还是先有蛋”的循环之中：想要知道 W_i 的排序，就得知道与它链接的其它网页的排序，而想要知道那些网页的排序，却又首先得知道 W_i 的排序。

为了打破这个循环，佩奇和布林采用了一个很巧妙的思路，即分析一个虚拟用户在互联网上的漫游过程。他们假定：虚拟用户一旦访问了一个网页后，下一步将有相同的几率访问被该网页所链接的任何一个其它网页。换句话说，如果网页 W_i 有 N_i 个对外链接，则虚拟用户在访问了 W_i 之后，下一步点击这些链接中任何一个的几率均为 $1/N_i$。初看起来，这一假设并不合理，因为任何用户都有偏好，怎么可能以相同的几率访问一个网页的所有链接呢？但如果我们考虑到佩奇和布林的虚拟用户实际上是对互联网上全体用户的一种平均意义上的代表，这条假设就不象初看起来那么不合理了。那么网页的排序由什么来决定呢？是由该用户在漫游了很长时间（理论上为无穷长时间）后访问各网页的几率分布来决定，访问几率越大的网页排序就越靠前。

为了将这一分析数学化，我们用 $p_i(n)$ 表示虚拟用户在进行第 n 次浏览时访问网页 W_i 的几率。 显然，上述假设可以表述为（请读者自行证明）：

$$p_i(n+1) = \Sigma_j\, p_j(n) p_{j\to i}/N_j$$

这里 $p_{j\to i}$ 是一个描述互联网链接结构的指标函数 (indicator function)，其定义是：如果网页 W_j 有链接指向网页 W_i，则 $p_{j\to i}$ 取值为 1，反之则为 0。显然，这条假设所体现的正是前面提到的佩奇和布林的排序原则，因为右端求和式的存在表明与 W_i 有链接的所有网页 W_j 都对 W_i 的排名有贡献，而求和式中的每一项都正比于 p_j，则表明来自那些网页的贡献与它们的自身排序有关，自身排序越靠前（即 p_j 越大），贡献就越大。

随机矩阵在谷歌算法里占据了重要的地位

为符号简洁起见，我们将虚拟用户第 n 次浏览时访问各网页的几率合并为一个列向量 p_n，它的第 i 个分量为 $p_i(n)$，并引进一个只与互联网结构有关的矩阵 H，它的第 i 行第 j 列的矩阵元为 $H_{ij} = p_{j\to i}/N_j$，则上述公式可以改写为：

$$p_{n+1} = Hp_n$$

这就是计算网页排序的公式。

熟悉随机过程理论的读者想必看出来了，上述公式描述的是一种马尔可夫过程 (Markov process)，而且是其中最简单的一类，即所谓的平稳马尔可夫过程 (stationary Markov process)[1]，而 H 则是描述转移概率的所谓转移矩阵 (transition matrix)。不过普通马尔可夫过程中的转移矩阵通常是随机矩阵 (stochastic matrix)，即每一列的矩阵元之和都为 1 的矩阵（请读者想一想，这一特点的“物理意义”是什么？）[2]。而我们的矩阵 H 却可能有一些列是零向量，从而矩阵元之和为 0，它们对应于那些没有对外链接的网页，即所谓的“悬挂网页”(dangling page)[3]。

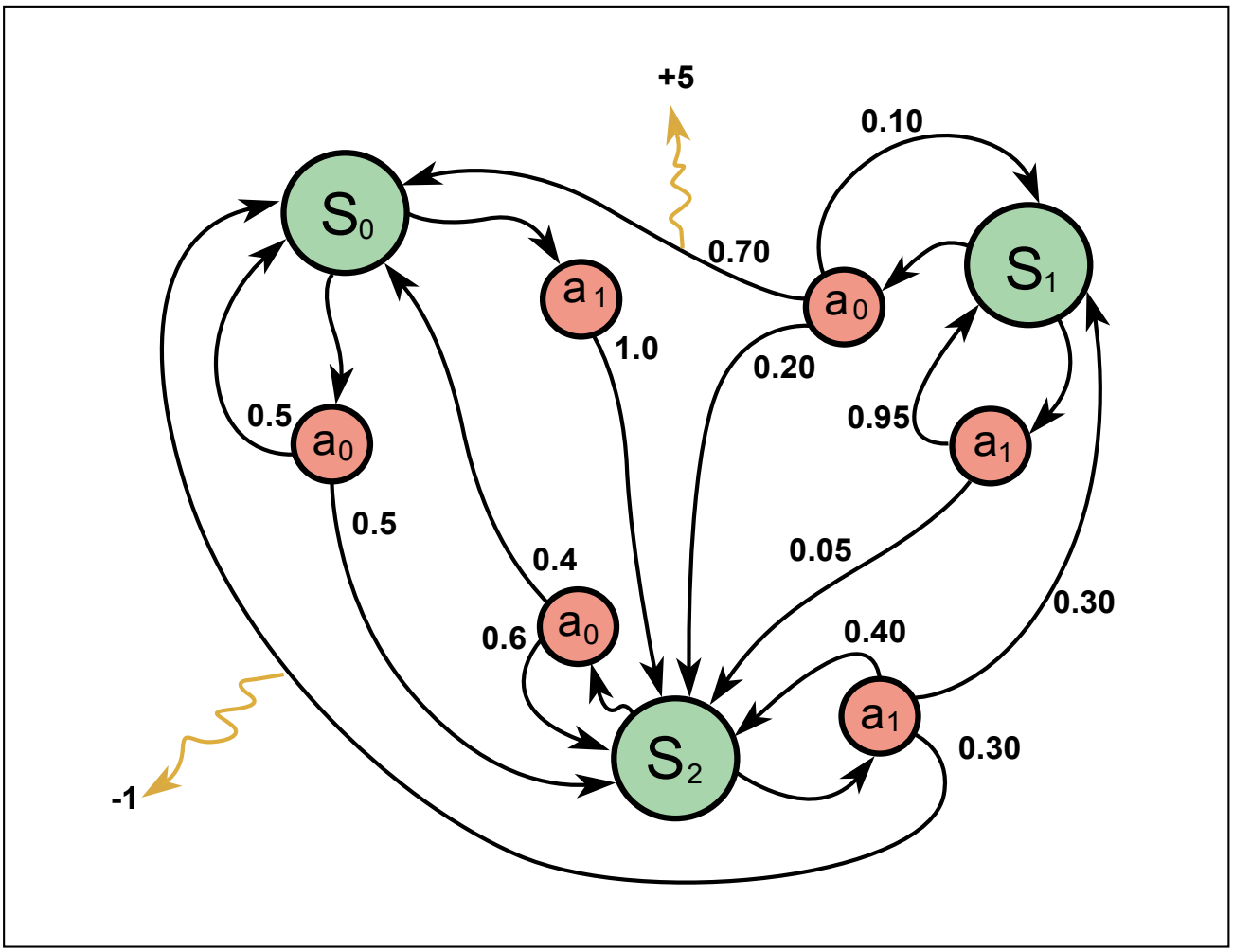
上述公式的求解是简单得不能再简单的事情，即：

$$p_n = H_n p_0$$

其中 p_0 为虚拟读者初次浏览时访问各网页的几率分布（在佩奇和布林的原始论文中，这一几率分布被假定为是均匀分布）。

如前所述，佩奇和布林是用虚拟用户在经过很长（理论上为无穷长）时间的漫游后访问各网页的几率分布，即 $\lim_{n\to\infty} p_n$，来确定网页排序的。这个定义要想管用，显然要解决三个问题：

1. 极限 $\lim_{n\to\infty} p_n$ 是否存在？
2. 如果极限存在，它是否与 p_0 的选取无关？
3. 如果极限存在，并且与 p_0 的选取无关，它作为网页排序的依据是否真的合理？



马尔可夫过程 (Markov process) 在谷歌算法里占据了重要的地位。我们称离散的马尔可夫过程为马尔可夫链；其在各个时刻的状态转变由一个概率矩阵所控制。这是一个马尔可夫链的例子。

如果这三个问题的答案都是肯定的，那么网页排序问题就算解决了。反之，哪怕只有一个问题的答案是否定的，网页排序问题也就不能算是得到满意的解决。那么实际答案如何呢？很遗憾，是后一种，而且是其中最糟糕的情形，即三个问题的答案全都是否定的。这可以由一些简单的例子看出。比方说，在只包含两个相互链接网页的迷你型互联网上，如果 $p_0 = (1, 0)^T$，极限就不存在 (因为几率分布将在 $(1, 0)^T$ 和 $(0, 1)^T$ 之间无穷振荡)。而存在几个互不连通（即互不链接）区域的互联网则会使极限——即便存在——与 p_0 的选取有关（因为把 p_0 选在不同区域内显然会导致不同极限）。至于极限存在，并且与 p_0 的选取无关时它作为网页排序的依据是否真的合理的问题，虽然不是数学问题，答案却也是否定的，因为任何一个“悬挂网页”都能象黑洞一样，把其它网页的几率“吸收”到自己身上（因为虚拟用户一旦进入那样的网页，就会由于没有对外链接而永远停留在那里），这显然是不合理的。这种不合理效应是如此显著，以至于在一个连通性良好的互联网上，哪怕只有一个“悬挂网页”，也足以使整个互联网的网页排序失效，可谓是“一粒老鼠屎坏了一锅粥”。

为了解决这些问题，佩奇和布林对虚拟用户的行为进行了修正。首先，他们意识到无论真实用户还是虚拟用户，当他们访问到“悬挂网页”时，都不可能也不应该“在一棵树上吊死”，而是会自行访问其它网页。对于真实用户来说，自行访问的网页显然与各人的兴趣有关，但对于在平均意义上代表真实用户的虚拟用户来说，佩奇和布林假定它将会在整个互联网上随机选取一个网页进行访问。用数学语言来说，这相当于是把 H 的列向量中所有的零向量都换成 e/N（其中 e 是所有分量都为 1 的列向量，N 为互联网上的网页总数）。