> ……但很自然的是，过度的希望总伴随着极度的沮丧。我们确信那本贵重的秘籍就放在了某间六边形小室中的某个书架上，但也同样确信我们无法获得这秘籍。这样的确信简直令人无法忍受。

正如在虚构的巴别尔图书馆中一样，万维网图书馆中的许多信息也仍然是无法获得的。实际上，早期的搜索引擎在减轻用户的挫折感方面几乎没有做什么事情：我们可以按雅虎上的主题层次结构进行整理来执行搜索，或者通过筛选搜索引擎所返回的众多（通常是数以千计）的页面并单击这些页面链接来人工确定出与查询最为相关的页面。某些使用者还求助于古代查询者们所使用的最早的搜索技巧——口头传述以及专家建议。他们从朋友那儿获知有价值的网站，并且链接上那些业已花费了许多个钟头来努力搜索的同事们所推荐的站点。

当 1998 年**链接分析**（link analysis）技术出现于信息检索的图景之中的时候，所有的一切都改变了[40,106]。最为成功的搜索引擎的开发者们开始通过利用万维网超链接结构的内在附加信息的链接分析技术来改进搜索结果的质量。网络搜索得到了极大的改善。网络搜索者们笃信不移地使用他们所偏爱的谷歌和 AltaVista 这样的引擎，并使它们的地位不断提升。实际上，在 2004 年，许多网上冲浪者都直率地承认了他们对当今搜索引擎的痴迷、依赖和上瘾。我们将在下面加入数名谷歌迷的评论[117]，以表达我们对于万维网图书馆可访问性提高的喜悦之情，而正是链接分析引擎使得这一提高成为可能。顺便说一句，在 2004 年 5 月，谷歌保有了搜索市场的最大份额，有 37% 的搜索者都在使用谷歌，其后有 27% 的搜索者使用的是雅虎的联合引擎，包括 AltaVista、AlltheWeb 和 Overture[㊀]。

- “它并不是我的个人主页，不过也差不多了。我用它来自我搜寻（ego-surf）[㊁]；我用它来阅读新闻；无论何时我希望了解任何事情时，我就会用它。”——马特·格罗宁（Matt Groening），《辛普森一家》（The Simpsons）的创作者与执行制片人。
- “我简直无法想象没有谷歌新闻的日子。遍布全球的数以千计的新闻来源，能保证让任何一个连上因特网的人都可以了解现在所发生的事情。在这里所获得的各种观点数量之繁多，真令人难以置信。”——迈克尔·鲍威尔（Michael Powell），美国联邦通信委员会主席。
- “谷歌是我的反应迅速的搜索助手。在抓紧时间赶工期的时候，我会用它来检查外国人名的拼写，获取某个特殊的军用装置的图像，找到某个公众人物的确切的发言，核对统计数据，翻译短语，或者是研究某个公司的背景。它就是信息检索中的瑞士军刀。”——加里·特鲁多（Garry Trudeau），漫画家，《杜恩斯比利》（Doonesbury）的创作者。

现在，几乎所有主要的搜索引擎都将类似于谷歌所用的链接分析评分和更为传统的信息检索评分结合使用。在本书中，我们将记录**网络**信息检索的诸多方面之一——链接分析

㊀ 这些市场份额数据由 comScore 公司所汇编，这家公司对美国网民在 2004 年 5 月利用主要搜索引擎进行的搜索次数进行了统计。请见 http://searchenginewatch.com/reports/article.php/2156431 上的文章。

㊁ 指在搜索引擎中输入自己的名字进行搜索。——译者注

评分。例如，在一个文档后面给出的 97% 意味着该文档被认为与用户的查询之间存在着 97% 的相关度。（请参见美国联邦通信委员会的搜索引擎 http://www.fcc.gov/search-tools.html，其搜索功能由 Inktomi 提供，外界曾一度知道后者使用了向量空间模型。请试着输入一个诸如 taxes 这样的查询，并注意在右侧所给出的相关性评分。）向量空间模型的另一个优点即相关性反馈，它是一个信息检索调整方法，也是向量空间模型的一个自然拓展。相关性反馈使用户能够选择检索到的文档中的一个有用子集，然后查询会与这个附加的相关性反馈信息一同被重新提交，从而检索出一个经过修正的、通常更为有用的文档集合。

向量空间模型的缺点之一是它的计算开销。在查询时，必须计算每个文档和查询之间的距离度量（也称为相似性度量）。而诸如 LSI 这样的高级模型还需要进行开销高昂的大矩阵奇异值分解[82,127]。该矩阵以数值的方式表示了整个文档集。随着文档集的增大，矩阵分解的开销将使模型不再具有可行性。这一计算开销还暴露了另一个缺点——向量空间模型无法很好地扩展。它们的成功仅局限于小文档集。

《理解搜索引擎》

迈克尔·贝里（Michael Berry）和莫雷·布朗恩（Murray Browne）所著的内容丰富的小书《理解搜索引擎：数学建模与文本检索》（Understanding Search Engines：Mathematical Modeling and Text Retrieval）[23]对向量空间模型特别是 LSI 进行了很好的阐释，并且包含了数个示例和例程。数学专业的读者应该能够颇为愉悦地阅读该书并了解线性代数算法在传统信息检索中的应用。

1.2.3　概率模型搜索引擎

概率模型（probabilistic model）试图对用户找到某个特定相关文档的概率进行估计。检索得到的文档根据它们的相关几率（该文档与查询相关的可能性除以不相关的可能性所得的比值）进行排名。概率模型以递归的方式运行，并要求算法能够猜测得到初始参数，然后逐次尝试改善这一初始猜测，以得到最终的相关几率的排位。

不幸的是，概率模型的构建和编程有可能十分困难。它们的复杂度上升得很快，令许多研究者望而却步，也限制了其扩展性。概率模型还要求做出若干不现实的简化假设，如检索项之间以及文档之间的独立性等。例如，在本书中，最可能跟在 information 之后出现的词是 retrieval，但是独立性假设却认为任何词都会以等概率出现在 information 后面。不过在另一方面，概率框架可以自然地纳入先验偏好，因此这些模型确实有希望做到针对单个用户的偏好来调整搜索结果。例如，可以将用户的查询历史结合到概率模型的初始猜测

万维网是自组织的。传统文档集通常是由经过训练（而且酬金也常常很高）的专家来收集和分类的。但是在万维网上，任何人都可以发布一个网页并任意链接至别处。没有什么标准，也没有什么看守对内容、结构和格式加以管制。数据是易变的：快速的更新、破损的链接以及消失的文件，等等。美国在 2002 年的一项有关“链接无效”的研究中指出，在两本信息技术类期刊所引用的 URL 中，有多至 50% 的 URL 将在 4 年内变成无法访问[1]。数据是异质的，以多种格式、语言和字符集的形式存在。而且这些易变、异质的数据通常会被发布多次。此外，对网页也不存在编辑审稿流程，这意味着其中可能充满了错误、谎言和非法言论。进一步地，自组织为那些鬼鬼祟祟的**垃圾信息制造者**（spammer）们打开了一扇门，他们利用万维网在商业方面的潜力牟利。**垃圾信息制造者**这个名字最初指的是那些发送大量广告电子邮件的人。垃圾邮件发送者只需单击一下发送按钮，就可以在数秒钟内将他们的广告信息发给数千名潜在客户。随着网络搜索和在线零售业的发展，这个名字的含义已经得到了扩展，包含了那些利用欺诈网页生成技术、针对特定查询来获得网络搜索结果中高排名的人。垃圾信息制造者利用极小的文本字体、隐藏文本（白色背景上的白色文本）和误导性的**元标签**（metatag）描述来愚弄早期的网络搜索引擎（如那些使用传统信息检索中布尔模型的引擎）。万维网的自组织还意味着产生网页的目的大相径庭。某些页面针对的是上网购物的人，而其他的则针对网上的研究者。实际上，搜索引擎必须能够回答许多类型的查询，例如，业务性查询、导航性查询和信息咨询性查询。万维网的所有这些特性组合在一起，使得网络搜索引擎的工作成为了赫拉克勒斯的任务[⊖]。

但万维网也是超链接在一起的。这个作为范内瓦·布什的记忆扩展机基础的链接性，是网络搜索引擎能够利用的仅有的优点。超链接使得上网冲浪这个新的全民消遣得以成为现实，但更加重要得多的一点，则是超链接使得有针对性的、有效的搜索得以实现。本书所讲述的，便是网络搜索引擎利用万维网蔓延无极的链接结构中有用的附加信息来改善搜索结果质量的种种方式。因此，我们在书中只会关注网络信息检索过程诸多方面中的仅仅一个方面，但我们相信这是最激动人心也是最重要的一个方面。但是，万维网链接结构所带来的好处也伴随着副作用。最有趣的副作用与那些卑鄙的垃圾信息制造者有关。垃圾信息制造者们很快就嗅到了主要搜索引擎使用链接分析的蛛丝马迹，并马上着手制作垃圾链接。垃圾链接制造者们细心地手工设计出了超链接策略，以期增加其页面的访问量。由此产生了搜索引擎和垃圾链接制造者之间的有趣的猫和老鼠的游戏，而对于这个游戏，包括作者在内的许多人都是乐于观赏一番的。请见第 43 页和第 52 页上的杂谈。

对任何文档集，尤其是对与万维网相关的文档集，信息检索的另一挑战则是查准率。尽管可以访问的信息量一直在增长，但用户查阅文档的能力却并未相应增加。用户很少会去查看检索所得的前 10 个或 20 个文档之后的内容[94]。用户缺乏耐心就意味着搜索引擎的查准率必须增长得和文档数量一样快。网络搜索引擎所特有的另一个悖论与其性能度量和对比有关。尽管传统搜索引擎可以通过在研究者熟悉的、得到充分研究的、受控的文档集上进行测试来加以对比，但对于网络引擎而言，这一做法却不现实。即使是一个小型网页集，对于研究者们而言也已经大得无法加以归类、计数并给出数十个查询的查准率和查全率的分子和分母部分的估计值。两个搜索引擎的对比除了根据速度和存储需求等最基本的比较指标外，通常是通过用户满意度研究和市场份额指标来完成的。

⊖ 指极为困难甚至被认为是不可能完成的任务。——译者注

MATLAB 爬虫的 m 文件

在得到克里夫·莫勒尔（Cleve Moler）的允许之后，我们将在此展示他的 MATLAB 蜘蛛的五脏六腑。如果你是一名程序员或充满好奇心的读者，而且不会见到蜘蛛或 MATLAB 代码就大吐不止的话，那就请安心地详加剖析吧。至于那些对此反胃、讨厌代码的读者，则应该向下跳到 2.2 节。

MATLAB 6.5 或更高版本中包括了两条命令：urlread 和 urlwrite，这两条命令使我们能写出简单的 m 文件以进行万维网上的爬行。以下的 m 文件 surfer.m 从一个 root 页面开始网络爬行，并一直继续直到爬过了 n 个页面。该程序将生成两个输出，即 n 个爬过的页面的 URL 列表 U，和一个包括了这 n 个页面的链接结构的稀疏二值邻接矩阵（adjacency matrix）L。（这个 L 矩阵与第 4 章中的 PageRank 矩阵 $\boldsymbol{H}$ 有关。）接着就可以用 urlwrite 命令将每个所得的 URL 的内容存入一个文件，之后该文件将被送至搜索引擎的索引建立模块进行压缩。（该 m 文件可从克里夫的著作《MATLAB 数值计算》（Numerical Computing with MATLAB）[132] 的网站 http://www.mathworks.com/moler/ncmfilelist.html 上下载。）

```
function [U, L] = surface (root, n);

%  SURFER  函数用来生成一部分万维网的邻接矩阵。
%  [U, L] = surfer(root, n)由 URL root (根地址) 开始，并跟随网络链接行进，
   直至得到一个 n 乘 n 阶的链接的邻接矩阵。输出结果 U 是一个访问过的 URL 的单元
   数组，而 L 是一个稀疏矩阵，如果 url(i) 链接到了 url(j)，则 L(i, j) = 1。
%  例: [U, L] = surfer ('http: //www.ncsu.edu', 500);
%  该函数目前有两个缺陷。(1) 寻找链接的算法十分简单。我们仅仅是查找字符串
   'http: '; (2) 由于试图由一个可访问但速度非常慢的 URL 读取数据，因此函数可
   能需要不可接受的长时间才能结束。在某些情况下，可能必须依靠操作系统来终止
   MATLAB。这些 URL 中的关键词可以被添加到 surfer.m 中的 skip 列表中。
%  初始化
U = cell (n, 1);
hash = zeros (n, 1);
L = logical (sparse (n, n));
m = 1;
U (m) = root;
hash (m) = hashfun (root);

for j = 1: n
    % 尝试打开一个网页
```

因特网档案馆计划

在1996年，一个称为因特网档案馆（Internet Archive）的非营利性组织担负起了一项艰巨的任务，即将万维网的内容——网页、图像、视频文件、音频文件等一一加以存档。这一计划将把那些旧版本的网页、现在已经消失的网页以及现有的网页都进行存档。例如，如果想看看作家卡尔·梅耶（Carl Meyer）个人主页的先前版本，就可以连上因特网档案的“网页时光旅行机”（Wayback Machine）（http://web.archive.org/）。输入卡尔当前主页的网址 http://meyer.math.ncsu.edu/，网页时光旅行机将返回存档的各个版本以及页面的更新日期。对档案馆站点的一个临时性补充是安娜·帕特森（Anna Patterson）的“追忆”（Recall）搜索引擎测试版。由于该引擎被设计用于对存档进行搜索，因此它具有一些新特性，如检索项相关性随时间变化的时序图等。（也许这些特性将会成为主流引擎的普通功能，因为帕特森现在正在为谷歌工作。）因特网档案馆的一个目的是保证外围网页上的信息不会被永远丢失掉，因为在这些页面中存在着有价值的趋势和文化遗留。档案馆还使我们能系统地跟踪万维网的演化。当然，随着因特网档案馆计划的不断发展并获得资助，它将不可避免地获得无可争议的索引之王的称号，并保有世界上最大的文档集。

2.3　查询处理

不同于爬虫和索引建立模块，搜索引擎查询模块的操作依赖于用户。查询模块必须实时处理用户的查询，并在若干毫秒的时间内返回结果。在2003年1月，谷歌称它每天都为2.5亿搜索者提供服务，而 Overture 和 Inktomi 各自服务了1.67亿和8千万用户[156]。谷歌喜欢将其处理时间保持在半秒钟以下。为了如此迅速地处理一个查询，查询模块需要访问预先计算好的索引，如内容索引和结构索引。

考虑使用如下由2.2节复制过来的倒排文件的例子。

- 项1（aardvark）-3，117，3961
  ⋮
- 项10（aztec）-3，15，19，101，673，1199
- 项11（baby）-3，31，56，94，673，909，11114，253791
  ⋮
- 项 m（zymurgy）-1159223

假设用户输入了一个不大常见的查询 aztec baby，搜索引擎将假设这两项是通过布尔“与”结合的。之后，查询模块将查阅项10aztec 和项11baby 的倒排列表。“切题”或相关

内容得分仅由内容索引及其倒排文件便能计算得到，且与查询相关。另一方面，欢迎度得分则仅由结构索引便能计算得到，而且一般是查询无关的。本书的剩余部分均集中于欢迎度得分，因此我们将对欢迎度得分的描述与计算都加以推后。目前我们仅仅指出，万维网上的每个页面都有一个欢迎度得分，它独立于用户的查询，且给出了页面在搜索引擎的整个页面索引中的一个全局性欢迎度。然后这个欢迎度评分与内容评分——比方说通过相乘一一而结合，来给出对给定查询的每个相关页面的一个总评分。

坎贝尔大法官关于索引的动议

约翰·坎贝尔（John Campbell，1799—1861）是一名苏格兰律师和政治家，他在 1859 年成为了大不列颠大法官。在其著作《首席大法官生平》（Lives of the Chief Justices）[45] 第 3 卷的序言中，坎贝尔大法官写道：

“我认为给每一本书都赋以一个索引极为重要，因此这促使我提议向国会提交一份法案，以剥夺出版没有附加索引的著作的作者所享有的版权，而且更进一步地，要因为他的违法行为而对其课以罚金。”

不幸的是，他的法案从未颁布过，这可能是由于国会议员及其选民希望逃避创建一个好的索引所需要承担的责任和工作。

网络团体和对网页建立索引的组织或个人也已经提出了与坎贝尔大法官类似的诉求。万维网联盟（World Wide Web Consortium，W3C）已经在推进 HTML 文档的更为严格的结构（例如 XML 文档和 RSS 代码等），以允许搜索引擎的索引者能更为精确、迅速地从文档中抽取关键性元素。另一方面，没有统一构架这一点，却已被普遍认同为万维网力量的一个来源，而且这也是万维网众多创造性用法的主要贡献者。在一次确定统一结构与自由之间均衡原则的尝试中，美国前总统比尔·克林顿（Bill Clinton）签署了《全球电子商务框架》（Framework for Global Electronic Commerce），提倡对网络立法和规范采取一种自由放任的态度。

网络图的图谱

巨大的网络图和图 3.1 中那清晰明了的微型图之间几乎没有多少相似之处。万维网的结点和弧构成了一个纷乱杂陈的乱摊子，想要从中理清头绪并以能够吸引注意力的同时又以有意义的方式来加以呈现的话，这个图简直令人头痛欲裂。幸运的是，许多研究者已经在这些方面取得了若干成功。《网络空间图谱》（The Atlas of Cyberspace）[62] 给出了 300 余幅五彩缤纷、蕴含丰富信息的网络活动的图谱。我们获允在此展示一幅超链接图它是斯坦福大学塔玛拉·明茨纳（Tamara Munzner）的毕业作品。她用三维双曲空间产生出了图 3.2 左边的图。玄英（Young Hyun）[⊖] 在其 Java 程序 Walrus 中实现了明茨纳的思路（尽管这个软件画出来的图看起来更像是水母）。图 3.2 中的右侧是玄英所作的一幅具有 535 102 个结点和 601 678 条链接的图谱。

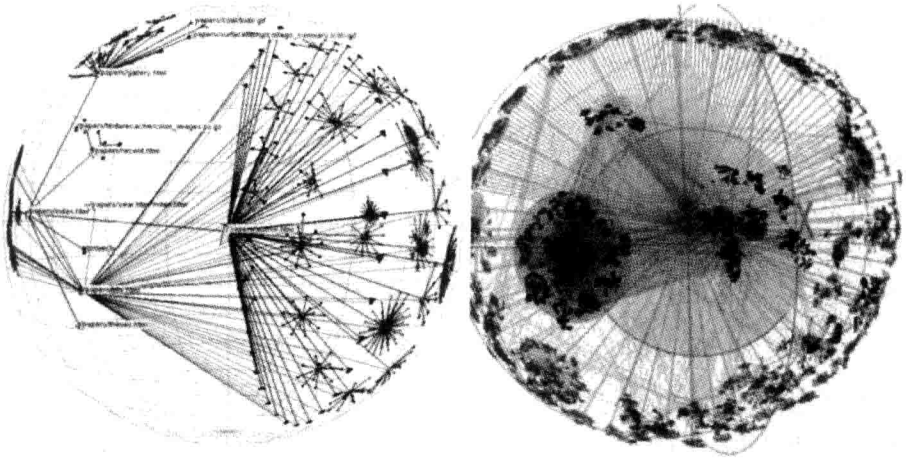

图 3.2　明茨纳和玄英制作的一个万维网子集的图谱

3.2.1　PageRank

在 1998 年之前，网络图在很大程度上仍是一个未得到开发的信息资源。虽然如克莱因伯格以及布林和佩奇这样的研究者们已经认识到了这个图的潜能，但大多数人还搞不清网络图到底与搜索引擎的结果有什么关系。我们可以将超链接视为一种推荐来理解这两者之间的联系。由我的主页指向你的主页的超链接，就是我对你的主页的一种认可。因此，具有更多推荐（由入链所体现）的页面肯定比只具有少数入链的页面要更为重要。但是，类似于诸如文献引用或举荐信等其他推荐系统，推荐者本身的地位也同样重要。例如，唐纳德·特朗普（Donald Trump）的个人推荐可能在应征职位时比 20 个毫不出名的教师或同事的 20 封推荐信更加管用。但另一方面，如果面试官知道唐纳德·特朗普在夸赞雇员

⊖ 按英文音译。——译者注

方面相当随意而慷慨，而且他（或他的秘书）这辈子已经写了超过 40 000 封推荐信的话，那么他的推荐的权重就会猛然下降了。因此，那些几乎不加区分地进行推荐的推荐者，他所具有的代表其地位的权重必须被调低。实际上，每个推荐的权重都应当由该推荐者所作出的推荐的总数加以调节。

事实上，这正是谷歌 PageRank 欢迎度评分的运作方式。这一 PageRank 评分十分有名，甚至可以说是恶名远扬（请见第 52 页和第 112 页上的杂谈）。关于这个问题已经货真价实地发表了数百篇文章，而在关注于 PageRank 的方法论、机制和计算的众多出版物中，本书无疑是最早者之一。在后续的章节中，我们将揭示使得 PageRank 如此流行的众多原因，但研究 PageRank 评分的最具说服力的理由之一，则是谷歌自己已经承认了 PageRank 对谷歌那相当成功的技术所具有的影响。据谷歌网站（http://www.google.com/technology/index.html）所言，“（谷歌）软件的心脏就是 PageRank……（它）仍然在为（我们）所有的网络搜索工具提供着基础。”

简短来说，**PageRank 的论点就是，如果一个网页被其他重要的页面所指向，那它就是重要的**。听起来像是循环定义，不是吗？在第 4 章中我们将看到，这句话可以被形式化为一个美丽而简单的数学公式。

谷歌工具条

两个页面的 PageRank 得分之间的对比，可以反映出这两个页面的相对重要性。但是，谷歌非常谨慎地将页面的确切 PageRank 得分保护在其索引之中，而且它也确实有足够的理由如此行事。（请见第 52 页上的杂谈。）不过谷歌确实十分厚道地提供了一个公开的访问方式，能够获取其 PageRank 评分的一个非常粗略的近似。这些近似值可通过谷歌工具条获得，工具条位于浏览器上，用一个条形图显示了当前页面 PageRank 评分的近似值。所显示的 PageRank 是一个 0 到 10 之间的整数，最重要页面的 PageRank 为 10。工具条将对你所访问的每个页面自动更新显示，因此，它肯定将你正在查看的页面信息发送到了谷歌服务器。谷歌的隐私条款声称，它不会采集可以直接说明你个人身份的信息（例如你的姓名或电子邮件地址），也不会出售任何信息。对那些仍然关心隐私权的人，谷歌允许用户禁用 PageRank 特性，同时仍然保持工具条的其他功能。有一种途径可以不用通过工具条就可以获得近似的 PageRank 评分——访问 http://www.seochat.com/seotools/PageRank-search/，输入一个查询，并观察结果后面的 PageRank 条形图。通过定位到高 PageRank 评分的页面（例如 www.espn.com，其得分为 9/10）和低 PageRank 评分的页面（例如 http://www.csc.ncsu.edu:8080/nsmc2003/，得分为 0/10），读者便能够对 PageRank 有一个直观感受。谷歌主页（www.google.com）的 PageRank 评分为 10 分，这可能是自动设置的。谷歌将那些被识别为垃圾信息制作者所制作的页面的 PageRank 设置为 0[160]，这个值在垃圾信息制作者之间以令人闻风丧胆的 **PR0** 所著称。

形成了鲜明的对比。修改和保存谷歌矩阵的元素并不可行，因为即使 $\boldsymbol{H}$ 非常稀疏，但其巨大的规模以及缺乏有规律性的结构也阻止了我们使用直接法。与之相反，像迭代方法这一类的不需要矩阵的方法则更受欢迎。

幂法就存储方面的需求而言也是友好的。除了稀疏矩阵 $\boldsymbol{H}$ 和悬挂结点向量 $\boldsymbol{a}$ 之外，就只需要保存 $\boldsymbol{\pi}^{(k)\mathrm{T}}$ 了。这个向量是完全稠密的，即我们必须保存 n 个实数。就谷歌而言，$n=81$ 亿，因此我们应当能够理解谷歌对于存储所持的节俭态度了。如 GMRES 或 BICGSTAB 等其他的迭代方法尽管更快，但却需要存储多个向量。例如，一个重启式 GMRES（10）方法在每个迭代中都需要存放 10 个长度为 n 的向量，这等价于整个 $\boldsymbol{H}$ 矩阵所需的存储空间，因为 $nnz(\boldsymbol{H}) \approx 10n$。

使用幂法来计算 PageRank 向量的最后一个理由与所需的迭代次数有关。布林和佩奇在 1998 年的论文中指出，仅需 50～100 次迭代幂法就能够收敛，并给出精确 PageRank 向量的一个令人满意的近似，这一点也被其他研究者所进一步确认。回忆一下，由于 $\boldsymbol{H}$ 矩阵是如此稀疏，以至于幂法的每次迭代都只需 $O(n)$ 的计算量。而我们很难找到一个方法的计算量能比 $50O(n)$ 的幂法更好。运行时间和计算量与问题规模成线性（或准线性）关系的算法非常迅速，但也非常稀少。

自然，下一个问题就是：为什么应用于 $\boldsymbol{G}$ 的幂法仅需约 50 次迭代就可以收敛？在 $\boldsymbol{G}$ 的结构中是否存在某种东西能够暗含这一迅速的收敛？马尔可夫链理论给出了答案。一般而言，应用于某个矩阵的幂法的*渐进收敛速率*（asymptotic rate of convergence）依赖于绝对值最大的两个特征值 λ_1 和 λ_2 的比值。确切来说，渐进收敛速率就是 $|\lambda_2/\lambda_1|^k \to 0$ 的速率。对于如 $\boldsymbol{G}$ 这样的随机矩阵，$\lambda_1=1$，因此 $|\lambda_2|$ 就决定了收敛与否及快慢的情况。由于 $\boldsymbol{G}$ 也是素的，因此 $|\lambda_2|<1$。一般情况下，如果仅仅是为了估计渐进收敛速率的话，那要数值计算出一个矩阵的 λ_2 所需要的计算量，而这却一般是人们所不愿意付出的。幸运的是，对于 PageRank 问题，容易证明[127:第502页,90,108]，如果矩阵的谱分别为 $\sigma(\boldsymbol{S})=\{1,\mu_2,\cdots,\mu_n\}$ 和 $\sigma(\boldsymbol{G})=\{1,\lambda_2,\cdots,\lambda_n\}$，则

$$\lambda_k = \alpha\mu_k \quad (\text{其中},k=2,3,\cdots,n)。$$

（该命题的一个简短的证明将在本章结尾处给出。）此外，万维网的链接结构使得 $|\mu_2|=1$（或至少是 $|\mu_2|\approx 1$）的情况很可能会发生，这意味着 $|\lambda_2(\boldsymbol{G})|=\alpha$（或 $|\lambda_2(\boldsymbol{G})|\approx\alpha$）。因此，凸组合参数 α 就能够解释上述的仅 50 次迭代的收敛速度。在论文中，谷歌的创始人布林和佩奇使用了 $\alpha=0.85$，而根据最近一次所公开的信息，谷歌仍在使用该值。$\alpha^{50}=0.85^{50}\approx .000296$，即在第 50 次迭代时，我们可以期望近似 PageRank 向量的精度大约为**小数点后 2～3 位**。对于需要进行排名的谷歌来说，这样的精度显然已经足够。仅就数学方面的因素而言，要区分 PageRank 中的元素，可能需要小数点后 10 位的精度（请见 8.3 节），但是当 PageRank 评分和内容评分结合起来之后，高精度就变得不那么重要了。

谷歌矩阵的次主特征值（subdominant eigenvalue）

对于谷歌矩阵 $\boldsymbol{G}=\alpha \boldsymbol{S}+(1-\alpha)\, 1/n\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$，

$$|\lambda_2(\boldsymbol{G})| \leqslant \alpha。$$

- 对于 $|\lambda_2(\boldsymbol{S})|=1$（由于网络图的可约性常会导致这种情况的产生），有 $|\lambda_2(\boldsymbol{G})|=\alpha$。因此，方程（4.6.1）的 PageRank 幂法的渐进收敛速率即为 $\alpha^k \to 0$ 的速率。

现在，我们可以对 4.3 节中那 6 个问题给出肯定性的回答了。通过随机性和素性调整，应用于 $\boldsymbol{G}$ 的幂法能保证收敛到一个唯一的、称为 PageRank 向量的正向量，而与初始向量无关。由于所得的 PageRank 向量为正的，因此没有我们不希望看到的 0 比 0 的平局。为了产生精度大约为小数点后 τ 位的 PageRank 评分，需要大约 $-\tau/\log_{10}\alpha$ 次迭代。

PageRank 幂法的 MATLAB m 文件

这个 m 文件是方程（4.6.1）中所给 PageRank 幂法的一个 MATLAB 实现。

```
function [pi, time, numiter] = \hbox {PageRank} (pi0, H, n, alpha,
epsilon);

% \hbox {PageRank} 函数用于计算一个 n 乘 n 阶的马尔可夫矩阵 H 的 \hbox
  {PageRank} 初量，起始
% 向量为 pi0（行向量），比例参数为 alpha（标量）。使用幂法计算。
% 例：[pi, time, numiter] = \hbox {PageRank}(pi0, H, 1000,.9,1e-8);
% 输入：pi0 =第 0 次迭代时的初始向量（行向量）
%       H =按行归一化的超链接矩阵（n 乘 n 阶的稀疏矩阵）
%       n =H 矩阵的大小（标量）
%       alpha = \hbox {PageRank} 模型中的比例参数（标量）
%       epsilon =收敛精度（标量，例如 1e-8）
% 输出：pi = \hbox {PageRank} 向量
%       time =计算 \hbox {PageRank} 向量所需的时间
%       numiter =达到收敛所需的迭代次数
% 初始向量通常设为一个均匀分布的向量：pi0 =1/n* ones (1, n)。
% 注意：MATLAB 按列存放稀疏矩阵，因此在 H 的转置 H'上执行某些操作可能更为快
  捷。
```

SEO 之间的永恒的战争之中。由看得见的网页内容到隐藏着的元标签，由链接到锚文本，由服务器到链接农场（同样请见第 52 页上的杂谈），激战蔓延于整个万维网之上。

利用垃圾主题词和隐藏技术[84]，SEO 在对抗早期的搜索引擎上取得了胜利。在垃圾主题词技术中，垃圾主题词常常多次出现于页面主体中，也包括在标题、元标签、锚文本和 URL 文本中。隐藏技术利用配色方案和两面派手法来欺骗搜索引擎，例如人眼是不能看到在白色背景上用白色文本显示的垃圾信息的，因此不会向搜索引擎发出对藏有垃圾信息的页面的投诉；而两面派则是指对普通用户返回包含了垃圾信息的网页，而对搜索引擎爬虫却返回无垃圾信息页面的技术。只要制作者可以清楚地辨认出网络爬行者，他们就可以用一个干净的、无垃圾信息的页面将其打发走。由于这些方法对于网页制作者而言十分易于使用，因此搜索引擎不得不增加其蜘蛛和索引建立者的 IQ，来采取报复措施。许多蜘蛛和索引建立者经训练后会忽略元标签，因为到了 20 世纪 90 年代晚期，已经很少有元标签还保存着准确的页面信息。它们还会忽略重复的关键词。但是，两面派则更难以反制。搜索引擎请求用户的帮助来制止两面派。例如，谷歌会请上网者来作出仲裁，一旦他们发现那些重定向到新页面的可疑网页，就立刻吹响裁判哨。

到 1998 年，搜索引擎在锦囊中增加了链接分析。从此以后，垃圾信息制造者们仅靠垃圾主题词和两面派手法已无法再继续愚弄使用链接分析的引擎，也无法为自己骗得不公正的高排名了。而垃圾信息制造者和 SEO 们也开始通过学习链接分析的工作原理来适应这个新情况。SEO 群体一直都相当活跃——它的成员们不时地召开会议，撰写论文和书籍，开设博客，并出售他们的秘密。最为著名也最使人受益的 SEO 论文是由克里斯·莱汀斯（Chris Ridings）所写的《解读 PageRank：所有那些你一直想知道的 PageRank 的事情（PageRank explained：Everything you've always wanted to know about PageRank）》[143]和《PageRank 揭秘（PageRank uncovered）》[144]。这些论文提供了获得 PageRank 并防止出现 PageRank 溢出之类情况的实用策略。搜索引擎不断调整他们的算法，以保持对 SEO 玩家领先一步的优势。尽管搜索引擎将不道德的 SEO 视为对头，某些网络分析者却称之为网络食物链中的不可或缺的一环，因为他们推动了创新、研究和发展。

杂谈：搜索引擎如何赚钱？

我们经常都被问到这个问题。这是个很好的问题。搜索引擎提供了免费而不受限制的服务，因此那几十亿美元的搜索收入到底是从何而来？搜索引擎有多个收入来源。首先，某些搜索引擎会向网站制作者收取一笔入选费。某些缺少耐性的制作者希望能够得到保证，以使得他们的新站点很快（在一两天内）就被索引，而不是等上一两个月才被蜘蛛找到并加入到待爬行的 URL 列表中。搜索引擎在收取一小笔费用后就能提供这一保证，而如果再稍微多花点钱，还能保证制作者的站点能被更频繁地——可能是每月二次——被重新索引。

大多数搜索引擎还通过向感兴趣的有关方出售分析数据来进行创收。搜索引擎每一天都收集了海量的用户数据。这些数据可以用于改善搜索质量并预测用户需求，此外还以归总数据的形式被出售给各种公司，例如，对当前流行的查询词或商务搜索的比例感兴趣的搜索引擎优化公司，可以直接通过搜索引擎购买到这些信息。

尽管搜索引擎并不会将获取搜索的能力出售给个人用户，但它们确实会向公司出售这些搜索服务。例如，网景（Netscape）公司向谷歌付费来将谷歌作为其浏览器的默认搜索工具。GoTo（已被 Overture 收购，而后者现在是雅虎的一部分）一度曾把它对每个查询项所给出的前 7 个结果出售给雅虎和 AltaVista，而后者则把这 7 个结果放在了它们的结果中的最前面。

尽管有这些收入来源，但迄今为止搜索引擎最为有利可图且增长最为迅速的收入源依然是广告。据估计，在 2004 年，搜索收入中有 30 亿美元来自于广告。谷歌于 2004 年 6 月 21 日发布的 IPO[⊖]公报非常清晰地体现了该公司对于广告的依赖：广告收入占其 2003 年收入的 97% 强。许多搜索引擎出售横幅条广告服务，这些广告将出现在它们的主页以及结果返回页中。其他搜索引擎则出售有偿排名广告。这些有争议的广告允许公司通过付费而获得最靠前的排名。许多网络分析者指责有偿排名广告污染了搜索结果。但是，使用这一技术的搜索引擎（GoTo 就是个最好的例子）则反驳称这一排名方法对于商用搜索而言十分优秀。由于近期的调查估计所有搜索中有 15% ~30% 在本质上而言都是商业搜索，因此 Overture 这样的引擎为此类查询提供了一种有价值的服务。另一方面，还有许多的搜索行为是为了帮助研究工作的，因此有偿排名引擎的结果令这些用户感到沮丧。

谷歌对广告和排名采取了一种不同的方法。他们将无偿的搜索结果显示于一个主列表中，而有偿排名站点则作为“赞助链接”单独出现在一旁。谷歌通过单击付费的广告方案来显示赞助链接。公司选择与其产品或服务有关的一个关键词，然后给出一个投标价格，每当一名搜索者单击了他们的链接，他们便会按此价格向谷歌支付费用。例如在罗利的一个自行车店可能会为每次“自行车罗利”的查询投标 5 美分，而只有当搜索者实际单击了他们的广告时，才真正需要付费。但是另一个公司可能会对相同的查询投标 17 美分，则该公司的广告就可能出现在更靠前的位置，因为尽管会进行某些微调和优化，但赞助广告一般都是按投标费用从高到低排列的。

单击付费广告是市场营销中的一个创新。传统上很少进行广告投入的小商户们，现在也正在网络广告上投入多得多的金钱，因为单击付费广告的效费比很高。如果一个搜索者单击了链接，那这就表明了他或她的一种购买意向，而其他形式的广告如广告牌或邮件广告则无法传递这一信息。有趣的是，如同万维网上的许多其他东西一样，单击付费广告终将演变为竞争者之间的另一个战场，也只不过是个时间问题。如果没有保护（可以通过软件程序的形式购得这样的保护），那些天真的购买了单击付费广告的公司将会很轻易地被其竞争者捣乱。竞争对手只要不断地单击这类公司的广告，就可以让后者买更多的单，并耗尽其广告预算。

4.7　谷歌矩阵的谱定理及其证明

在本章中，我们将谷歌矩阵定义为 $\boldsymbol{G}=\alpha\boldsymbol{S}+(1-\alpha)\,1/n\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$。但是，在下一章的 5.3 节中，我们将扩展这一定义，以包括更为一般的谷歌矩阵，其中人为制造的矩阵 $\boldsymbol{E}$ 将由均匀分布矩阵 $1/n\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$ 变为 $\boldsymbol{e}\boldsymbol{v}^{\mathrm{T}}$，其中，$\boldsymbol{v}^{\mathrm{T}}>0$ 为一个概率向量。在本节中，我们给出这一更为一般性的谷歌矩阵的次主特征值的有关定理和证明。

⊖ Initial Public Offerings 的缩写，即首次公开募股。——编辑注

$$\begin{aligned}\boldsymbol{\pi}^{(k+1)\mathrm{T}} &= \boldsymbol{\pi}^{(k)\mathrm{T}}\boldsymbol{G} \\ &= \alpha\,\boldsymbol{\pi}^{(k)\mathrm{T}}\boldsymbol{S} + (1-\alpha)\,\boldsymbol{\pi}^{(k)\mathrm{T}}\boldsymbol{e}\boldsymbol{v}^{\mathrm{T}} \\ &= \alpha\,\boldsymbol{\pi}^{(k)\mathrm{T}}\boldsymbol{H} + (\alpha\,\boldsymbol{\pi}^{(k)\mathrm{T}}\alpha + 1 - \alpha)\boldsymbol{v}^{\mathrm{T}}。\end{aligned} \tag{5.3.1}$$

与式（5.3.1）相比，第40页的式（4.6.1）是使用原来的无区分对待的跳转矩阵 $\boldsymbol{E} = 1/n\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$ 时的情况。因此现在不过是每次迭代中所加的常数向量由 $\boldsymbol{e}^{\mathrm{T}}/n$ 变为了 $\boldsymbol{v}^{\mathrm{T}}$，因此我们在第4章中关于幂法的结论差不多全都仍然适用。具体来说，渐进收敛速率、稀疏向量-矩阵乘法、最小存储与易于编程等性质都得到了保留。但是，PageRank 向量本身却的确发生了改变。不同的个性化向量将给出不同的 PageRank 排名[158]，即 $\boldsymbol{\pi}^{\mathrm{T}}$（$\boldsymbol{v}^{\mathrm{T}}$）是 $\boldsymbol{v}^{\mathrm{T}}$ 的函数。

一旦认识到 $\boldsymbol{v}^{\mathrm{T}}$ 的用处，我们就可以放开手脚了。想想看吧，为什么我们每个人都必须屈从于同一个网页排名呢？那个单一的、全局的、查询无关的排名 $\boldsymbol{\pi}^{\mathrm{T}}$（它使用了 $\boldsymbol{v}^{\mathrm{T}} = 1/n\boldsymbol{e}^{\mathrm{T}}$）根本没有体现我的个性和偏好。作为美国人，难道我们就不能有权去拥有自己个人的排名向量——一个知道我们自己对万维网网页和话题的偏好的排名向量？如果你喜欢浏览新闻时事的网页，那只要简单地偏置一下你的 $\boldsymbol{v}^{\mathrm{T}}$ 向量，使得新闻和时事网页 P_i 的 v_i 值较大，而其他所有网页的 v_j 值接近于0，然后再计算出这个量身定做的 PageRank 向量。政客们现在可以在他们的竞选承诺中再多加一条了："每个车库里都有一辆车，每个家里都有一台计算机，每个上网者都有一个个性化向量 $\boldsymbol{v}^{\mathrm{T}}$。"

这看来就是谷歌引入个性化向量的最初动机[38]。但是，它使得与查询无关、与用户也无关的 PageRank 变得依赖于用户，而且计算负担也更重了。为每位用户定制排名在理论上听起来很棒，但实际却无法计算。请记住，谷歌需要花费数日才能计算出对应于一个 $\boldsymbol{v}^{\mathrm{T}}$ 向量——那个一视同仁的个性化向量 $\boldsymbol{v}^{\mathrm{T}} = 1/n\boldsymbol{e}^{\mathrm{T}}$——的仅仅一个 $\boldsymbol{\pi}^{\mathrm{T}}$。

许多人都将个性化引擎视为搜索的未来。部分地受到这一事实的推动，若干研究者对不可能进行计算的这一断言采取了无视的态度，而创造出了准个性化的 PageRank 排名系统[58,88,91,99,142]。我们称之为准个性化，因为这些系统并非给出针对每个用户所定制的排名，而是针对不同的用户群来给出这一排名。

此类系统之一便是塔赫·哈维利瓦拉（Taher Haveliwala）所创造的产品。当时他还是斯坦福的一名研究生。他将标准的、查询无关的 PageRank 加以改变，生成了一个对主题敏感的 PageRank[88,89]。他产生了有限数量的 PageRank 向量 $\boldsymbol{\pi}^{\mathrm{T}}$（$\boldsymbol{v}_i^{\mathrm{T}}$），每个向量都偏向于某个特定主题 i。哈维利瓦拉在进行实验时从开放目录专案（Open Directory Project，ODP）的网页分类中选择了16个顶级主题。例如，假设 $\boldsymbol{\pi}^{\mathrm{T}}$（$\boldsymbol{v}_1^{\mathrm{T}}$）是第一个 ODP 主题为 Arts 的 PageRank 向量，而 $\boldsymbol{\pi}^{\mathrm{T}}$（$\boldsymbol{v}_2^{\mathrm{T}}$）是第二个 ODP 主题为 Business 的 PageRank 向量。$\boldsymbol{\pi}^{\mathrm{T}}$（$\boldsymbol{v}_1^{\mathrm{T}}$）将会偏向于 Arts，因为 $\boldsymbol{v}_1^{\mathrm{T}}$ 仅对那些包含了艺术内容的页面给出了较高的概率，而其余概率则接近于0。这16个偏置的 PageRank 向量被预先计算出来，然后在查询期间，通过模仿用户的兴趣及查询的含义，这些向量被迅速组合起来，这就是诀窍所在。哈维利瓦拉用这16个偏置的 PageRank 向量的凸组合来形成他自己的对主题敏感、与查询相关的 PageRank 向量，即

$$\boldsymbol{\pi}^{\mathrm{T}} = \beta_1 \boldsymbol{\pi}^{\mathrm{T}}(\boldsymbol{v}_1^{\mathrm{T}}) + \beta_2 \boldsymbol{\pi}^{\mathrm{T}}(\boldsymbol{v}_2^{\mathrm{T}}) + \cdots + \beta_{16} \boldsymbol{\pi}^{\mathrm{T}}(\boldsymbol{v}_{16}^{\mathrm{T}})。$$

式中，$\sum_i \beta_i = 1$。比如查询 science project ideas 落在了 Kids and Teens（ODP 的第 7 类主题）、Reference（ODP 的第 10 类主题）和 Science（ODP 的第 12 类主题）上。于是符合逻辑的做法，便是对与这些主题相关联的 PageRank 向量赋予更多的甚至是全部的权重，所以与其他系数相比，β_7，β_{10}和β_{12}应取大的值。哈维利瓦拉使用一个贝叶斯分类器来计算实验中的β_i值，但也有其他选择。当所有这些都完成后，这个主题敏感的欢迎度评分便与第 1 章中的传统内容评分结合为一体。当然，如果希望获得更精细的个性化层级，可以使用多于 16 个的主题来更好地使排名偏重于用户的查询和兴趣。

这个小小的个性化向量 $\boldsymbol{v}^{\mathrm{T}}$ 却似乎具有更为严重的潜在作用。有人推测，谷歌可以利用该个性化向量来控制那些所谓的链接农场的垃圾信息制造行为。请见第 52 页上的杂谈《SearchKing 对谷歌》。

Kaltix 个性化网络搜索

谷歌没用多长时间就认识到了个性化搜索的价值。实际上，一家名为 Kaltix 的新兴个性化搜索公司在开张后仅仅三个月，就被谷歌抢了过去。Kaltix 科技由格伦·杰（Glen Jeh）、斯潘达·卡姆瓦尔（Sepandar Kamvar）和塔赫·哈维利瓦拉于 2003 年夏天所创办，当时这三人正从斯坦福计算机科学系休学。在那个夏天，这些 Kaltix 的小伙子们每天工作 20 个钟头，名副其实地干得连手指骨头都露了出来，甚至有时需要在过劳的手腕上敷上冰袋才能入睡。勤奋的工作终有回报。谷歌于 2003 年 9 月收购了 Kaltix，而这三个人也搬到了谷歌总部去继续他们的项目。在 2004 年 3 月，谷歌实验室发布了个性化搜索测试版（http://labs.google.com/personalized）。通过在一个感兴趣领域的多层列表中勾选多选框，用户便能建立起自己的档案。针对该档案，谷歌将生成一个个性化向量。当查询被输入个性化搜索框后，结果就将以标准的排名列表的方式被给出，不过除此之外，用户还可以通过一个滚动条来调整定制的层级，以增强个性化向量的效果。

个性化 PageRank 幂法的 MATLAB m 文件

第 42 页上 PageRank 幂法的 MATLAB 实现使用了一个均匀分布的个性化向量 $\boldsymbol{v}^{\mathrm{T}} = \boldsymbol{e}^{\mathrm{T}}/n$。这个 m 文件只是简单地修改了一行代码，便实现了一个更为一般的 PageRank 幂法，能够允许个性化向量作为输入参数而被改变。因此，以下的 m 文件实现了应用于 $\boldsymbol{G} = \alpha \boldsymbol{S} + (1-\alpha)\boldsymbol{e}\boldsymbol{v}^{\mathrm{T}}$ 的 PageRank 幂法。

者。不过，搜索引擎优化公司 AutomatedLinks 发来的一封电子邮件则说服了她，并支付了一笔22美元的年费来使用该公司的排名提升服务。（请见第43页上关于搜索引擎优化的杂谈。）AutomatedLinks 的全部工作都专注于提高其客户页面的 PageRank 排名（以及在其他搜索引擎中的排名），而它则是通过链接农场来做到这一点的。既然优化商已经知道当指向客户页面的重要入链数增加时，其 PageRank 也会增加，那么他们就会增加指向客户网页的这类重要链接。链接农场拥有多个与重要话题有关的互联结点，它们具有高 PageRank。然后这些互联的结点便会连向客户页面，从而在实质上将部分 PageRank 与客户页面进行了分享。霍尔顿投给 AutomatedLinks 的22美元的投资，为她带来了每月超过26 000名的访客，以及数千美元的收入。

大多数链接农场用链接交换方案（link exchange program）或互惠链接策略（reciprocal linking policy）来提高客户页面的排名，不过也有其他方法能达到相同的目的[28, 29]。当然，对于注重排名正确性的搜索引擎而言，链接农场非常棘手。搜索引擎使用了若干种方法来发现链接农场。首先，它们请求上网者担当告发者，由他们来报告任何可疑的页面；其次，它们使用算法去识别出万维网中具有高密度互惠链接的联系紧密的子图；第三，它们通过人工来检查算法的结果，以确定嫌疑网站是否遵守规则。谷歌威胁将嫌疑站点以及与之互联的邻居站点们排除在索引之外，或者是降低他们的排名，以此劝阻制造垃圾链接的行为。

谷歌这一贬低链接农场 PageRank 的做法，在2002年至2003年间还在司法领域激起了波澜。搜索引擎优化公司 SearchKing 从2001年2月一直到2002年8月期间都经营得顺风顺水，其中部分原因便是由于它拥有高 PageRank，并且与客户分享了这个高 PageRank。拥有高 PageRank 的客户能拥有更多的流量，因此他们很乐意向 SearchKing 付费购买其排名提升服务。但是，在2002年8月之后的数月间，SearchKing 的总裁鲍勃·马萨（Bob Massa）眼看着他的谷歌工具条（请见第28页上文本框中的内容）所报告的 PageRank 估计值从 PR8 掉到了 PR4，然后又从 PR2 掉到了 PR0。他的客户当然也受到了牵连。客户们抱怨不断，而许多人就干脆跳船逃生了。盛怒的鲍勃·马萨在2002年10月17日开始行动，在俄克拉荷马西区联邦地区法院对谷歌提起了诉讼。SearchKing 的法律团队对谷歌提出了指控，要求赔付75 000美元的收入损失及审理费用，恢复公司及客户页面之前的 PageRank，并公开2002年8月至10月间谷歌所使用的 PageRank 算法源代码。对阵双方都深知这个案子的重要性。案件的结果将成为优化公司与搜索引擎之间关系的一个案例。到2002年12月30日，谷歌准备好了一个有力、令人信服且经过仔细研究的答复，来应对 SearchKing 所提请的临时禁令动议，此外还进一步加上了一个拒绝受理的动议。谷歌的答复中主要有两个论点。首先，谷歌称 PageRank 实际是一种意见和看法，即该公司对于网页价值的一个评判。这些意见受到第一修正案的保护。实际上，谷歌的辩护团队引述了一个与之类似的与排名有关的先例，即征信所进行的排名。在1999年杰斐逊县 R-I 学区对穆迪投资服务公司（Jefferson County School District #R-I vs. Moody's Investors Service, Inc.）一案中，正是同一个法院裁定认为，虽然穆迪对学校学区的低信用排名可能会伤害到该学区之前在公众眼中所具有的居住和就读的价值，但排名不过是一种意见的表达，因而受到第一修正案的保护。类似地，谷歌辩护团队辩称：

“谷歌所赋的 PageRank 值无法由客观证据来证明为真或假。SearchKing 怎么可能“证明”其“真正”的排名就应该是4或6或8？可以确定的是，SearchKing 应该不是在做出

和

$$(\boldsymbol{I}-\boldsymbol{S})^{\#}=\boldsymbol{X}\begin{pmatrix}0 & 0\\ 0 & (\boldsymbol{I}-\boldsymbol{C})^{-1}\end{pmatrix}\boldsymbol{X}^{-1}。$$

矩阵 C 由与特征值 $\lambda_k\neq 1$ 相对应的**若当块**（Jordan block）$\boldsymbol{J}_*$ 组成，而在 $(\boldsymbol{I}-\boldsymbol{C})^{-1}$ 中对应的块为 $(\boldsymbol{I}-\boldsymbol{J}_*)^{-1}$。结合这一结果与定理 6.1.3 就能清楚看出，当 $\alpha\to 1$ 时，$\boldsymbol{\pi}^{\mathrm{T}}(\alpha)$ 的敏感性由 $(\boldsymbol{I}-\boldsymbol{S})^{\#}$ 中元素的大小所决定。由于 $\|(\boldsymbol{I}-\boldsymbol{S})^{\#}\|\leqslant\kappa(\boldsymbol{X})\|(\boldsymbol{I}-\boldsymbol{C})^{-1}\|$，其中，$\kappa(\boldsymbol{X})$ 为 X 的条件数（condition number），因此当 $\alpha\to 1$ 时 $\boldsymbol{\pi}^{\mathrm{T}}(\alpha)$ 的敏感性主要由 $\|(\boldsymbol{I}-\boldsymbol{C})^{-1}\|$ 的大小所决定，而它又由 $|1-\lambda_2|^{-1}$（以及 λ_2 的下标）所决定，其中 $\lambda_2\neq 1$ 为 $\boldsymbol{S}$ 的最接近 $\lambda_1=1$ 的特征值。换而言之，若 λ_2 越接近于 $\lambda_1=1$，则当 α 接近于 1 时，$\boldsymbol{\pi}^{\mathrm{T}}(\alpha)$ 就越敏感。

一般来说，具有接近于 1 的次主特征值的随机矩阵，它们表示了近非耦合链（nearly uncoupled chain）[85]（又称为近完全可分解链（nearly completely decomposable chain））。这些链中的状态形成了若干团簇，团簇内部的状态间存在着紧密的链接，而在团簇之间则仅存在稀疏的链接，即我们可以把状态进行如此排序，以使得转移概率矩阵形如 $\boldsymbol{S}=\boldsymbol{D}+\varepsilon\boldsymbol{E}$，其中，$\boldsymbol{D}$ 是一个分块对角阵，$\|\boldsymbol{E}\|\leqslant 1$，而 $0\leqslant\varepsilon<1$ 相对于 1 而言为小值。

由万维网链接结构所定义的链几乎是近非耦合的（基于专业兴趣、宗教兴趣以及地域方面等因素的考虑，将造成大量由紧密耦合的结点所构成的相互间弱链接的团簇），所以可以预计矩阵 $\boldsymbol{S}$ 具有非常接近于 $\lambda_1=1$ 的次主特征值。因此，当 α 增大时，PageRank 向量对于 α 的变动将显得越来越敏感，而当 $\alpha\approx 1$ 时，PageRank 极为敏感。总结以上事实，便可得如下结论。

PageRank 敏感性的总结

作为参数 α 的函数，PageRank 向量 $\boldsymbol{\pi}^{\mathrm{T}}(\alpha)$ 对于 α 值的微小变化的敏感性如下。

- 对于小的 α 值，PageRank 对 α 的微小变动不敏感。
- 当 α 值变大时，PageRank 对 α 的微小扰动变得越来越敏感。
- 对于接近于 1 的 α 值，PageRank 对 α 值的微小改变非常敏感，敏感度由 $\boldsymbol{S}$ 的近非耦合的程度所决定。

平衡措施

较大的 α 值赋予万维网的真实链接结构以更大的权重，而较小的 α 值则增加了人为生成的概率向量 $\boldsymbol{v}^{\mathrm{T}}$ 的影响。由于 PageRank 本身的想法便是利用万维网的链接结构，因此选择接近于 1 的 α 值更为可取。但这也是 PageRank 变得最为敏感的情形，因此对 α 必须适度取值——据报导，谷歌使用了 $\alpha\approx 0.85$ 的这一取值[39,40]。

丁（Borodin）等学者。这三个小组都对旧的 PageRank 向量 $\boldsymbol{\pi}^{\mathrm{T}}$ 和更新后的 PageRank 向量 $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}$ 之差计算出了某个上界[29,113,133]。

$$\| \boldsymbol{\pi}^{\mathrm{T}} - \tilde{\boldsymbol{\pi}}^{\mathrm{T}} \|_1 \leqslant \frac{2\alpha}{1-\alpha} \sum_{i \in U} \pi_i 。$$

式中，U 是所有更新过的页面的集合。（证明将在 6.5 节中给出，请见第 69 页。）这一上界给出了对敏感性的另一种解释：只要 α 不接近于 1，且被更新的页面的 PageRank 值不高，那么更新后 PageRank 值的变化将不会很大。让我们考虑上界中的两个因子 $2\alpha/(1-\alpha)$ 和 $\sum_{i \in U} \pi_i$。

作为示例，我们假设 $\alpha = 0.8$，而所有更新过的页面的旧 PageRank 值的总和 $\sum_{i \in U} \pi_i$ 为 10^{-6}。则常数因子 $2\alpha/(1-\alpha) = 8$，这意味着新旧 PageRank 向量之差的 1 范数 $\| \boldsymbol{\pi}^{\mathrm{T}} - \tilde{\boldsymbol{\pi}}^{\mathrm{T}} \|_1$ 最多为 8×10^{-6}。此时，PageRank 值对于万维网的更新相当不敏感。随着 $\alpha \to 1$，这个上界逐渐变得越来越没有用。这一界限是否有意义，取决于 $\sum_{i \in U} \pi_i$ 能在多大程度上抵消掉 $2\alpha/(1-\alpha)$ 的增加。有两个因素影响了 $\sum_{i \in U} \pi_i$ 的大小：被更新的页面数量，以及这些被更新页面的 PageRank 值。由此也暴露出了上述界限的另一个局限性，即它并没有提供多少帮助，来回答如下这个更有意思也更为自然的问题："当具有高 PageRank 值的页面被更新时，PageRank 将会发生何种改变？"例如，对于如亚马逊网站这样受欢迎的、具有高排名的页面，发生于其上的改变将会如何影响整个排名？6.2 节已经对这一问题给出了更为完整的回答。

PageRank 和垃圾链接

旧的 **PageRank** 向量 $\boldsymbol{\pi}^{\mathrm{T}}$ 和更新后的 PageRank 向量 $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}$ 之差由如下界限所约束：

$$\| \boldsymbol{\pi}^{\mathrm{T}} - \tilde{\boldsymbol{\pi}}^{\mathrm{T}} \|_1 \leqslant \frac{2\alpha}{1-\alpha} \sum_{i \in U} \pi_i 。 \qquad (6.4.1)$$

式中，U 是所有被更新的页面的集合。

- 当 α 取小值并且被更新页面的集合具有较小的总 PageRank 值时，这个界限是有用的。只要 α 不接近于 1 而且被更新的页面不具有高的 PageRank，则更新后的 PageRank 值相比于原来的 PageRank 值就不会有大的改变。
- 另一方面，这个界限并未告诉我们 PageRank 对于受欢迎的高 PageRank 页面的改变的敏感性如何。
- 利用式（6.4.1）中的界限，研究者们[29]对于垃圾链接的有效性做出了如下的陈述：

  "……PageRank 的一个很好的性质是某个团体仅能对整个万维网的总 PageRank 造成非常有限的改变。因此，不论这种改变是通过何种方式发生，非权威性的团体无法显著影响全局的 PageRank。"

这一界限巩固了如下的基本想法，即优化的把戏要么是让若干高 PageRank 的页面指向你的页面，要么是让许多较低 PageRank 的页面指向你的页面。其他的从数学上来说与 PageRank 有关的最优链接策略请见参考文献［12］。