

# 搜索引擎 Google 的 PageRank<sup>TM</sup> 技术

孙 莉

(中国人民大学信息资源管理学院 04 级研究生 北京 100872)

**摘 要** Google 搜索引擎以其独到的功能成为现今使用最为广泛的搜索引擎之一,其专利网页级别 (PageRank<sup>TM</sup>) 排名算法作为 Google 搜索引擎的核心技术,在 Google 搜索引擎中起着重要的作用。文章先从 Google 搜索引擎的特点和功能说起,再从距阵特征值的角度阐述 PageRank<sup>TM</sup> 算法排名的原理。

**关键词** 搜索引擎 Google PageRank<sup>TM</sup> (网页级别) 链接

Google 的优势在于它的信息量大以及它所提供的快速搜索速度和高命中率的结果。Google 目前可检索的网络页面数达 30 亿个,每天处理的搜索请求高达 1.5 亿次,几乎占全球所有搜索量的 1/3。这些都是基于 Google 的复杂文本匹配运算法则及其搜索程序所使用的 PageRank<sup>TM</sup> 系统 (网页级别技术)。

## 1 Google 检索的特点及其功能

### 1.1 Google 检索的特点

(1) Google 所支持的语言由 96 种界面语言和 35 种搜索语言组成,其中包括简体中文和繁体中文,还将不断地增加新的支持语言。

(2) Google 网站只提供搜索引擎功能,没有广告,给检索者带来视觉新感受。

(3) Google 并非只使用关键词或代理搜索技术,它的专利网页级别技术 PageRank<sup>TM</sup> 能够提供高命中率的搜索结果。

(4) Google 智能化的“手气不错”功能,提供可能最符合要求的网站,使用户以最方便快捷的方式寻找到最佳网站。

(5) Google 的“网页快照”功能,为用户贮存了大量的应急网页,方便用户查找信息。

### 1.2 Google 检索的功能

(1) Google 提供简单及高级搜索功能。在高级搜索中,用户可根据需求限制搜索范围,限制某一搜索必须包含或排除特定的关键词或短语。它还允许用户定制搜索结果页面所含信息条目数量,可从 10 到 100 条任选,并且提供网站内部查询和横向相关

信息增值服务是通过网上信息服务、专题分析研究服务、专题检索代理及针对特定用户群的需要进行创造性的信息产品深度加工,使得网上信息资源的利用在量与质方面获得提高。服务方式主要有:电子剪报服务、电子文献传递服务、专题库服务、专利检索、科技查新等。

### 3.6 用户教育服务

用户教育也是合作参考咨询一项重要的服务内容。它是利用网络技术,将用户指导教程链接到图书馆主页,为用户自学提供方便的服务。

## 4 结束语

合作参考咨询是参考咨询发展的必然趋势,目前我国在这方面的研究还不是很成熟,本文也只是从概念上进行分析,具体的操作方式及目前图书馆界悬而未决的技术规范、质量控制、补偿机制、信息安全、知识产权等问题,还需要在实践中进一步研究。

## 参考文献

- 1 吴建中. 21 世纪图书馆新论 (第二版). 上海: 上海科学技术文献出版社, 2003
- 2 顾明远. 教育大辞典 (第三卷). 上海: 上海教育出版社, 1991
- 3 詹德优主编. 信息咨询理论与方法. 武汉: 武汉大学出版社, 2004
- 4 黄如花编著. 网络信息的检索与利用. 武汉: 武汉大学出版社, 2002
- 5 焦玉英王娜. 国内合作参考咨询服务发展研究. 中国图书馆学报, 2005 (1)
- 6 董勇. 国内大学城现状、存在问题及发展趋势. 浙江教育学院学报, 2005 (2)
- 7 肖时占. 网络环境下数字参考咨询服务的现状及问题研究. 图书馆, 2004 (3)

(本文作者现为武汉大学信息管理学院 04 级在读硕士研究生)

查询。

(2) Google提供了多种检索入口。Google的检索界面中提供了所有网站、图像、网上论坛、网页目录、新闻等检索入口,用户可以根据自身需要定制检索入口。

(3) Google提供分类目录查询。如果想寻找某些专题网站,可以使用 Google的分类目录。Google采用的是 OpenDirectoryProject的公共网页目录。分类的网站目录信息比较集中,类目比较明确,在某一目录门类中进行搜索往往能有更高的命中率。

## 2 PageRank™ 技术

某些搜索引擎采用关键字匹配的方法来定位用户提问(即检索词),这样可能得到许多与用户检索无关的结果,同时用户需要在众多的匹配中寻找有用的信息,这就增加了检索的难度并且降低了检索的效率。PageRank™帮助 Google搜索引擎根据页面的重要程度来排序网页。如果和用户的提问相匹配的网页有着较高的网页级别,它将被列为搜索结果页面的顶端。Google不可能出售 PageRank™。企业能够购买搜索结果页面的广告空间,但是他们不能够买一个更高的 PageRank™来提高他们网站的命中率。

### 2.1 PageRank™ 概念

PageRank™(网页级别)是 Google用于评测一个网页“重要性”的一种方法。在揉合了诸如 Title标识和 Keywords标识等所有其他因素之后,Google通过 PageRank™来调整结果,使得那些更具“重要性”的网页在搜索结果中令网站排名获得提升,从而提高搜索结果的相关性和质量。

简单来说,Google通过下面几个步骤来实现网页在其搜索结果页中的排名:

- (1)找到所有与搜索关键词匹配的网页;
- (2)根据页面因素如标题、关键词密度等排列等级;
- (3)计算导入链接的锚文本中的关键词;
- (4)通过 PageRank™得分调整网站排名结果。

### 2.2 PageRank™ 的决定因素

PageRank™是基于从许多优质网页中链接网页的原理,根据优质网页的回归关系,来判定所有网页的重要性。其理论为:若 B 网页设置有链接 A 网页的链接(B为 A 的导入链接时),说明 B 认为 A 有链接价值,是一个“重要”的网页。当 B 网页级别(重要性)比较高时,则 A 网页可从 B 网页这个导入链接分得一定的级别(重要性),并平均分配给 A 网页

上的导出链接。PageRank™反映了一个网页的导入链接的级别重要性。所以说 PageRank™是由一个网站的导入链接的数量和这些链接的级别(重要性)所决定的。

### 2.3 PageRank™ 算法原理

WEB中的网页通过导入链接与导出链接建立了网页之间的相互关系。怎样确定它们相互之间的关系呢?通常用行列阵的形式来表达这种链接关系。设页面 i 链接到另一张页面 j 的时,将其系数定义为 1,反之则定义为 0。则行列阵 A 的成分  $a_{ij}$  可以用,

$$a_{ij} = \begin{cases} 1 & \text{从页面 } i \text{ 向页面 } j \text{ 有链接的情况} \\ 0 & \text{从页面 } i \text{ 向页面 } j \text{ 无链接的情况} \end{cases}$$

若网页数用 N 来表示的话,这个行列阵就成为 NN 的方阵。这相当于在图表示理论中的邻接行列。也就是说,Web 的链接关系可以看作是采用了邻接关系有向图表  $S_n$ 。

PageRank™的行列阵是先把这个邻接行列倒置后,再将各列(column)矢量的总和变成 1(全概率),把各个列矢量除以各自的链接数(非零要素数)。这样行成的行列被称为“推移概率行列”,它含有 N 个概率变量,各个行矢量表示状态之间的推移概率。倒置的理由在于 PageRank™并非重视出链的数量而是重视入链的数量。PageRank™的计算,就是求属于这个推移概率行列最大特性值的固有矢量(优固有矢量)。

这是因为,当线性变换系 t 渐近时,我们能够根据变换行列的“绝对价值最大的特性值”和“属于它的固有矢量”将其从根本上记述下来。换句话说,用推移概率行列表示的概率过程,是反复对这个行列进行乘法运算的一个过程,并且能够计算出前方状态的概率。求特性值和固有矢量的值是严密分析的一种基础的数学手段。我们能够自由地给矢量的初始值赋值,但是因为不断地将行列相乘,得到的矢量却会集中在一些特定数值的组合中。我们把那些稳定的数值的组合称为固有矢量,把固有矢量中特征性的标量(scalar)称为特性值,把这样的计算方法总称为分解特性值,把解特性值的问题称为特性值问题。

举一个简单的例子来说明上面的原理,有一个假设前提,即所有的网页形成一个牢固的链接图(即每个网页能从其他网页通过超链接达到)。

假如有 6 个网页,它们之间的关系如图 1 所示:

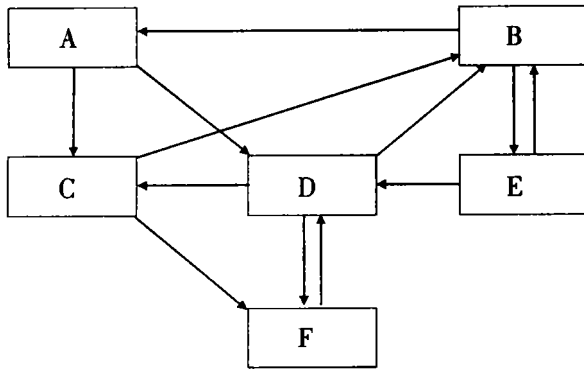


图 1 网页关系示意图

表 1 网页链接关系表

链接源页面	链接目标页面
PageA	PageC、PageD
PageB	PageE、PageA
PageC	PageB、PageF
PageD	PageC、PageB、PageF
PageE	PageB、PageD
PageF	PageD、PageE

根据表 1 建立一个上叙链接关系的链接行列 M,它是一个 66 的距阵,仅有要素 0 和 1 的位图行列。横向查看 i 行表示从网页 i 正向链接的页面。

表 2 网页链接行列

	A	B	C	D	E	F
A	0	0	1	1	0	0
B	1	0	0	0	1	0
C	0	1	0	0	0	1
D	0	1	1	0	0	1
E	0	1	0	1	0	0
F	0	0	0	1	1	0

```
M = [0, 0, 1, 1, 0, 0;
      1, 0, 0, 0, 1, 0;
      0, 1, 0, 0, 0, 1;
      0, 1, 1, 0, 0, 1;
      0, 1, 0, 1, 0, 0;
      0, 0, 0, 1, 1, 0; ]
```

PageRank™ 式的推移概率行列 N,是将 M 倒置将各个数值除以各自的非零要素后得到的。横向查看第 i 行非零要素表示有指向网页 i 链接的网页 (网页 i 的反向链接源),各纵列的值相加的和为 1 (全概率)。

```
N = [0, 1/2, 0, 0, 0, 0;
      0, 0, 1/2, 1/3, 1/2, 0;
      1/2, 0, 0, 1/3, 0, 0;
      1/2, 0, 0, 0, 1/2, 1/2;
      0, 1/2, 0, 0, 0, 1/2;
      0, 0, 1/2, 1/3, 0, 0; ]
```

又因 PageRank™ 的矢量 R (各个的页面的等级的队列),存在着  $R = cMR$  的关系 (c 为定量)。在这种情况下, R 相当于线形代数中的固有矢量, c 相当于对应特性值的倒数。为了求得 R,只要对这个正方形行列 M 作特性值分解就可以了。

运用 matlab6.0 编程语言,根据 PageRank™ 原理,编写如下程序段:

```
%设置计时器
tic
%根据 PageRank™ 的定义,将从文件 i 链接到文件 j 的链接状态的推移概率行列定义为 N(i, j)
N = [0, 1/2, 0, 0, 0, 0;
      0, 0, 1/2, 1/3, 1/2, 0;
      1/2, 0, 0, 1/3, 0, 0;
      1/2, 0, 0, 0, 1/2, 1/2;
      0, 1/2, 0, 0, 0, 1/2;
      0, 0, 1/2, 1/3, 0, 0; ]
%计算全部 N 的特性值和固有矢量列的组合
[V, D] = eig(N)
%保存与绝对值最大的特性值对应的固有矢量到 EigenVector
EigenVector = V(:, find(abs(diag(D)) == max(abs(diag(D)))));
% PageRank 是将 EigenVector 在概率矢量上标准化后得到的值
PageRank = EigenVector/norm(EigenVector, 1)
%输出计算时间
toc
利用 matlab 软件运行以上程序后得到如下结果:
M =
0    0.5000    0    0    0    0
0    0    0.5000    0.3333    0.5000    0
0.5000    0    0    0.3333    0    0
0.5000    0    0    0    0.5000    0.5000
0    0.5000    0    0    0    0.5000
0    0    0.5000    0.3333    0    0
```

V =

0.2683 0.1515 - 0.2153i 0.1515 + 0.2153i - 0.2566 0.2921 + 0.3018i 0.2921 - 0.3018i

0.5365 0.0761 + 0.3301i 0.0761 - 0.3301i 0.1453 - 0.1330 + 0.2121i - 0.1330 - 0.2121i

0.3040 0.0110 + 0.2628i 0.0110 - 0.2628i - 0.4942

0.5007 0.5007

0.5097 - 0.6189 - 0.6189 0.8048 - 0.3310 - 0.0180i - 0.3310 + 0.0180i

0.4292 0.0200 - 0.4736i 0.0200 + 0.4736i - 0.1247 - 0.4218 - 0.0347i - 0.4218 + 0.0347i

0.3219 0.3604 + 0.0960i 0.3604 - 0.0960i - 0.0747 0.0929 - 0.4611i 0.0929 + 0.4611i

D =

1.0000 0 0 0 0 0

0 - 0.4297 + 0.4790i 0 0 0

0 0 - 0.4297 - 0.4790i 0 0 0

0 0 0 - 0.2833 0 0

0 0 0 0 0.0713 + 0.2894i 0

0 0 0 0 0 0.0713 - 0.2894i

EigenVector =

0.2683

0.5365

0.3040

0.5097

0.4292

0.3219

PageRank =

0.1132

0.2264

0.1283

0.2151

0.1811

0.1358

elapsedtime =

0.1700

在 table6.0 的输出中,特性值被表示为对角行列 D 的对角成分,各个特性值相对应的固有矢量被表

示为行列 V 对应列的列矢量,也就是说  $M * V = D * V$  成立。如果包含复数特性值的话这里的特性值有 6 个,其中绝对值最大的特性值是  $\lambda = 1$ ,与之相对应的固有矢量为实矢量:

EigenVector =

0.2683

0.5365

0.3040

0.5097

0.4292

0.3219

即行列 V 的第 1 列。这个求得的固有矢量中概率矢量(要素的和等于 1 的 N 次元非负矢量)没有被标准化,只是矢量的大小等于 1。用算式来表达就是,  $\sum_{i=1}^N p_i = 1, (\sum_{i=1}^N p_i)^2 = 1$ 。对概率矢量进行标准化后

PageRank =

0.1132

0.2264

0.1283

0.2151

0.1811

0.1358

以上就是 PageRank™ 的排位了。全部值之和为 1。完成整个计算仅仅花了 0.17 秒的时间。

对求得的 PageRank™ 排序得到:

基于 PageRank 的网页排序

表 3

名次	PageRank™	网页	出链	入链
1	0.2264	PageB	A, E	C, D, E
2	0.2151	PageD	B, C, F	A, E, F
3	0.1811	PageE	B, D	B, F
4	0.1358	PageF	D, E	C, D
5	0.1283	PageC	B, F	A, D
6	0.1132	PageA	C, D	B

由此可知, PageRank™ 的名次和反向链接的数目是基本一致的。无论链接多少,正向链接都几乎不会影响 PageRank™,相反地有多少反向链接却是从根本上决定 PageRank™ 的大小。但是,仅仅这些并不能说明第 1 位和第 2 位之间的显著差别(同样地,第 3 位、第 4 位和第 5 位之间的差别)。可见 PageRank™ 并不仅仅是通过反向链接数来决定的,

还与链接页面的质量有着关系。

#### 2.4 PageRank™算法

在上述 PageRank™算法原理中有一个重要的假设:即所有的网页形成一个牢固的链接图(即每个网页能从其他网页通过超链接达到)。但是在现实的网络中,并不完全是这样的情况。当一个页面没有出链的时候,它的 PageRank™就不能被分配给其它的页面, Lawrence Page 和 Sergey Brin 称这种页面为悬摆链(dangling page)。同样道理,只有出链接而没有入链接的页面也是存在的。但 PageRank™并不考虑这样的页面,因为没有流入的 PageRank™而只流出的 PageRank™,从对称性角度来考虑是很奇怪的。同时,有时候也有链接只在一个集合内部旋转而不向外界链接的现象。PageRank™称之为 rank-sink。在现实中的页面,无论怎样顺着链接前进,仅仅顺着链接是绝对不能进入的页面群总归是存在的。

PageRank™为了解决这样的问题,提出用户的随机冲浪模型:PageRank™将“时常”固定为 13%来计算。用户在 87%的情况下沿着链接前进,但在 13%的情况下会突然跳跃到无关的页面中。用公式表示为:

$$M' = c * M + (1 - c) * [1/N]$$

其中,  $[1/N]$  是所有要素为  $1/N$  的  $N$  次正方行列,  $c = 0.85$  ( $= 1 - 0.15$ )。  $M$  也同样是推移概率行列。也就是说,根据 PageRank™的变形,原先求行列  $M$  的特性值问题变成了求行列  $M$  的固有有向量特性值问题。

#### 3 Google面临的问题

Google搜索引擎以其良好的用户界面以及快捷的搜索速度成为搜索引擎业界竞相模仿的对象,也成为众多的网络用户在浩瀚的互联网信息汪洋中寻找所需资料的首选引擎。PageRank™排名算法是 Google搜索引擎的核心。当然,PageRank™排名算法并不是完全没有漏洞,也存在着一些不尽如人意的地方。

##### 3.1 信息增长过快与数据更新问题

由于 Google的数据量过于庞大,数据库存有 10 多亿个网站和 42 亿多个网页的数据量,且随着网络信息的爆炸式增长,数据还在不断增长。要对如此庞大的数据进行更新需要耗费大量时间,况且这不

仅仅是软件技术能解决的问题,还要依赖存储设备及其技术的发展。

##### 3.2 查全率与查准率的问题

Google的基本搜索使用方便,只需在搜索栏里输入要搜索的关键词,点击“Google搜索”即可搜索到出现该关键词的网页;但 Google很可能会将 2 个字以上的词语分割开来,例如在查找有关“应用技术”方面的信息时,Google 不仅将含有“应用技术”的网页显示出来,还会将分别含有“应用”和“技术”的网页罗列。如果要使“应用技术”不分开,则应在输入的时候加上引号,这样 Google 就会将“信息系统”作为一个整体进行搜索,因此,Google 的查准率比较低,得到的结果冗杂,很难在短时间内找到精确的信息。

##### 3.3 动态生成网页的搜索

目前,所有搜索引擎中负责搜索网页的蜘蛛软件在技术上仍然还不能做到俘获动态网页(含多媒体内容)的水平,担心会被变化无穷的动态系统的黑洞吸走。Google 在这方面的研究虽然走在前面取得了一些突破,但仍然无法彻底解决这一问题。

#### 参考文献

- 曹军. Google 的 PageRank™技术剖析. 情报杂志, 2002(10)
- 张海涛,董洲. 搜索引擎 Google 的检索功能及 PageRank™技术分析. 情报科学, 2002(8)
- 朱俊卿. 搜索引擎 Google 研究. 现代图书情报技术, 2002(1)
- 张谦. 从 PageRank 的技术优势看 Google 的软件文化理念. 图书馆论坛, 2004(6)
- 汪利利,郑慧. 搜索引擎 Google 与雅虎中国的比较. 河海大学常州分校学报, 2004(6)
- 许涛,吴淑燕. Google 搜索引擎及其技术简介. 现代图书情报技术, 2003(4)
- Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web, 1998
- S Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In Proceedings of the Seventh World Wide Web Conference, 1998
- <http://blog.xdanger.com/archives/2003/12/19/000051.html>