PageRank 算法与 HITS 算法比较研究

赵悦阳

（中国医科大学附属盛京医院图书馆 沈阳 110004）

〔摘要〕 对 PageRank 算法和 HITS 算法的基本思想和具体算法进行介绍，从算法思想、权重的传播模型、处理的数据量及用户等待时间几方面对两种算法进行比较，并分析其各自的优缺点。

〔关键词〕 PageRank 算法；HITS 算法；比较

Comparison Research on PageRank Algorithm and HITS Algorithm *ZHAO Yue - yang, Library of Shengjing Hospital of China Medical University, Shenyang 110004, China*

〔**Abstract**〕 The paper introduces two popular algorithms PageRank and HITS from their basic idea and concrete algorithms, with the comparison on the algorithmic ideas, distribution modules of the weights, data process amount and user waiting times, it analyzes individual advantages and shortcomings.

〔**Keywords**〕 PageRank algorithm; HITS algorithm; Comparison

随着互联网的不断发展，人们越来越多地在互联网上发布和获取信息，网络已经成为信息制造、发布、加工和处理的主要平台。传统的互联网应用技术大多是基于文档内容的，与经典的信息检索技术和数据库技术有着密切的联系。但是互联网中特有的许多问题，诸如超大规模的非结构化文档数量、良莠不齐的网页质量、包含在文档中的大量多媒体信息，甚至含糊或不规范的用户查询表示等，都使得经典的信息检索技术和数据库技术在互联网环境中很难有效地应用。同时，互联网又包含了传统数据环境所没有的另一种丰富信息，即互联网的超链接拓扑结构。网页间的超链接一方面引导网页浏览的过程，另一方面也反映了网页创建者的一种判断。即有理由认为如果网页 A 存在一条超链接指向网页 B，那么网页 A 的作者是认为网页 B 包含了有价值的信息。充分利用互联网的链接结构信息对互联网应用技术的研究将具有极为重要的意义。

〔收稿日期〕 2010 - 06 - 29

〔作者简介〕 赵悦阳，硕士，实习研究员，发表论文 5 篇。

1 PageRank 算法简介

1.1 基本思想

Sergey Brin 和 Lawrence Page 在 1998 年提出了 PageRank 算法，PageRank（TM）是美国 Google 公司的登记注册商标。该算法是基于“从许多优质的网页链接过来的网页，必定还是优质网页”的回归关系，来判定所有网页的重要性。PageRank 算法有效地利用了互联网所拥有的庞大链接结构的特性。从网页 A 导向网页 B 的链接被看作是页面 A 对页面 B 的支持投票，Google 根据这个投票数来判断页面的重要性。但是 Google 不仅仅只看投票数（即链接数），对投票的页面也进行分析，“重要性”高的页面所投票的评价会更高。根据这样的分析，得到了高评价的重要页面会被给予较高的 PageRank（网页等级），在检索结果内的名次也会提高。PageRank 是 Google 中表示网页重要性的综合性指标，而且不

会受到各种检索的影响。或者说 PageRank 就是基于对“使用复杂的算法而得到的链接构造”的分析，从而得出的各网页本身的特性。当然重要性高的页面如果和检索词句没有关联同样也没有任何意义。为此 Google 使用了精炼后的文本匹配技术，使得能够检索出重要而且正确的页面。

1.2 具体算法

具体的算法就是用行列阵的形式来表达链接关系。从页面 i 链接到另一张页面 j 时，将其成分定义为 1，反之则定义为 0。即行列阵 A 的成分 aij 可以采用如下表示方法：

aij〕1 if（从页面 i 向页面 j「有」链接的情况）

aij〕0 if（从页面 i 向页面 j「没有」链接的情况）

文件数用 N 来表示的话，这个行列阵就成为 N × N 的方阵。相当于在图表理论中的“邻接行列”，只是 PageRank 的行列阵是把这个邻接行列倒置后（行和列互换），为了将各列（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）。这样做成的行列被称为“推移概率行列”，含有 N 个概率变量，各个行矢量表示状态之间的推移概率。倒置的理由是 PageRank 并非重视“链接到多少地方”而是重视“被多少地方链接”。PageRank 的计算就是求解属于这个推移概率行列最大特性值的固有矢量。这是因为当线性变换系 $t\to\infty$ 渐近时，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。

1.3 PageRank 计算举例

举例逐次计算 PageRank。图 1 表示含有链接关系的 7 个 HTML 文件，并且这些文件间的链接关系只是闭合于这 1 – 7 的文件中。即除了这些文档以外没有其他任何链接的出入。所有的页面都有正向和反向链接（即没有终点），实际应用中不可能获取一个页面的所有反向链接，但是该页面的所有正向链接都可获得。

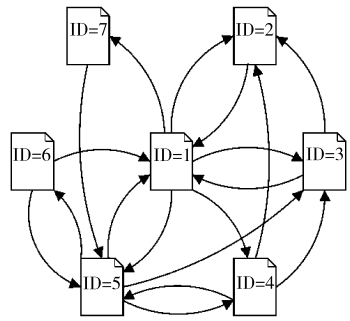

图 1 表示页面间互相链接关系的推移图

根据这张推移图构造的邻接列表，见表 1。即根据各个链接源 ID 列举链接目标的 ID。

表 1 表示页面间链接关系的邻接列表

链接源 ID	链接目标 ID
1	2，3，4，5，7
2	1
3	1，2
4	2，3，5
5	1，3，4，6
6	1，5
7	5

以这个邻接列表中所表示的链接关系的邻接行列 A 是如下所示的 7 × 7 的正方行列，一个仅有要素 0 和 1 的位图行列（Bitmap Matrix）。横向查看第 i 行表示从文件 i 正向链接的文件 ID。

A = [1, 2, 3, 4, 5, 6, 7
1 0, 1, 1, 1, 1, 0, 1;
2 1, 0, 0, 0, 0, 0, 0;
3 1, 1, 0, 0, 0, 0, 0;
4 0, 1, 1, 0, 1, 0, 0;
5 1, 0, 1, 1, 0, 1, 0;
6 1, 0, 0, 0, 1, 0, 0;
7 0, 0, 0, 0, 1, 0, 0;]

PageRank 式的推移概率行列 M，是将 A 倒置后将各个数值除以各自的非零要素后得到的，即如下所示的 7 × 7 正方行列。横向查看第 i 行非零要素表示有指向文件 i 链接的文件 ID（文件 i 的反向链接源），各纵列值相加的和为 1（全概率）。

M = [0, 1, 1/2, 0, 1/4, 1/2, 0;
1/5, 0, 1/2, 1/3, 0, 0, 0;

1/5, 0, 0, 1/3, 1/4, 0, 0;
1/5, 0, 0, 0, 1/4, 0, 0;
1/5, 0, 0, 1/3, 0, 1/2, 1;
0, 0, 0, 0, 1/4, 0, 0;
1/5, 0, 0, 0, 0, 0, 0;]

表示 PageRank 的矢量 R（各个页面等级数的队列）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 只要对这个正方行列 M 做特性值分解即可。

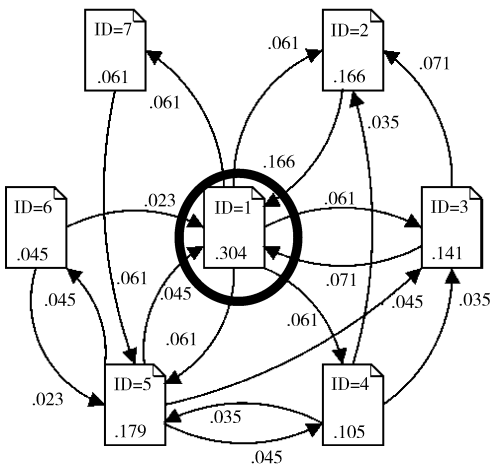

图 2 表示页面间互相链接关系的推移图（加入 PageRank 值）

流入量 =（ID = 2 发出的 Rank）+（ID = 3 发出的 Rank）+（ID = 5 发出的 Rank）+（ID = 6 发出的 Rank）= 0.166 + 0.141/2 + 0.179/4 + 0.045/2 = 0.30375。

2 HITS 算法简介

2.1 基本思想

HITS 算法是康奈尔大学（Cornell University）的 JonKleinberg 博士于 1998 年首先提出的，英文全称为 Hypertext - induced Topic Search，主要思想是网页的重要程度与所查询的主题相关。IBM 研究院在 Clever 系统中使用 HITS 算法计算一个网页的重要性。它首先引入一种网页，称为中心型（Hub）网页（good source of links），它本身可能并不重要，即没有几个网页指向它，但提供了指向某个主题最为重要的站点链接的集合。权威型（Authority）网页对于一个特定的检索提供最好的相关信息；中心型网页提供很多指向其它高质量权威型网页的超链。由此可以在每个网页上定义“目录型权值”和“权威型权值”两个参数。HITS 算法的基本思想是好的 Hub 型网页指向好的 Authority 网页，好的 Authority 网页是由好的 Hub 型网页所指向的网页。

2.2 执行算法

将查询 q 提交给基于关键字查询的检索系统，从返回结果页面的集合中取前 n 个网页（如 n = 200），作为根集合（Root Set），记为 S，则 S 满足以下 3 个条件：S 中的网页数量较少；S 中的网页是与查询 q 相关的网页；S 中的网页包含较多的 Authority 网页。将 S 扩展为基本集合（Base Set）T，T 包含由 S 指出或指向 S 的网页。可以设定一个上限如 1 000 ~ 5 000 个网页。

开始权重传播，这是一个递归的过程，用于决定 Hub 与权威权重的值，具体操作如下。

为基本集中的每个页面赋予一个非负的 Authority 权重 ap 和非负的 Hub 权重 hp，并将所有的 a 和 h 值初始化为同一个常数，如 ap = 1，hp = 1。Hub 与 Authority 的权重可按如下公式进行迭代计算：

$$a_p = \sum_{q|q\to p} h_q \quad (1)$$

$$h_p = \sum_{q|q\to p} a_q \quad (2)$$

式（1）反映了若一个页面由许多好的 Hub 所指，则其 Authority 权重会相应增加（即增加为所有指向它的页面的现有 Hub 权重之和）。式（2）反映了若一个页面指向许多好的 Authority 页，则 Hub 权重也会相应增加（即权重增加为该页面链接的所有页面的 Authority 权重之和）。每次迭代后使用下面公式进行规范化处理，保证一致性：

$$\sum_p (a_p)^2 = 1 \quad (3)$$

$$\sum_p (h_p)^2 = 1 \quad (4)$$

当 a 和 h 值没有收敛时，转向（2）。实验证明经过大约 10 ~ 15 次迭代计算，a 和 h 值将趋于稳定，迭代结束。此时可设置阀值 T，将所有 a 和 h 大于 T 的网页挑选出来，排序输出查询结果。

实践证明该算法对于许多查询具有良好的查准率和查全率。

3 PageRank 算法与 HITS 算法比较

3.1 算法思想

显而易见两者均是基于链接分析的搜索引擎排序算法，并且在算法中两者均利用了特征向量作为理论基础和收敛性依据。但两种算法的不同点也非常明显。从算法思想上看虽然均同为链接分析算法，但两者之间还是有一定的区别。HITS 的原理如前所述，其 Authority 值只是相对于某个检索主题的权重，因此 HITS 算法也常被称为 Query - dependent 算法。而 PageRank 算法独立于检索主题，因此也常被称为 Query - independent 算法。PageRank 算法的提出者把引文分析思想借鉴到网络文档重要性的计算中来，利用网络自身的超链接结构给所有的网页确定一个重要性的等级数。当然 PageRank 并不是引文分析的完全翻版，根据因特网自身的性质等，它不仅考虑了网页引用数量，还特别考虑了网页本身的重要性。简单地说重要网页所指向的链接将大大增加被指向网页的重要性。

3.2 权重的传播模型

HITS 首先通过基于文本的搜索引擎获得最初的数据，网页重要性的传播是通过 Hub 页向 Authority 页传递，而且 Kleinberg 认为 Hub 与 Authority 间是相互增强的关系；而 PageRank 基于随机冲浪（Randomsurfer）模型，可以认为它将网页的重要性从一个 Authority 页传递给另一个 Authority 页。

3.3 处理的数据量及用户端等待时间

表面上看 HITS 算法对排序的网页数量需求较少，所计算的网页数量一般为 1 000 ~ 5 000 个，但由于需要从基于内容分析的搜索引擎中提取根集并扩充基本集，这个过程需要耗费相当长的时间。而 PageRank 算法处理的数据数量上远远超过了 HITS 算法。据 Google 介绍目前已收录的中文网页已达 33 亿以上，但由于其计算在用户查询时已由服务器端独立完成，不需要用户端等待。从用户端等待时间来看 PageRank 算法应该比 HITS 要短。

4 PageRank 算法与 HITS 算法的优缺点

4.1 PageRank 算法

PageRank 算法的优点在于对互联网上的网页给予一个全局的排序，而且计算过程可以离线完成，这样有利于迅速响应用户的请求。其缺点在于主题无关性，没有区分页面内的导航链接、广告链接和功能链接，容易导致对广告页面的过高评价；即使包含有用的信息，一个新网页的评分通常都比较低，其原因在于时间短暂，很少有链接指向它；对于实时的搜索引擎算法还需要更快的计算方法。

4.2 HITS 算法

HITS 算法的优点在于更好地描述了互联网的组织特点，而且所需迭代次数更少，收敛速度很快，这是因为它是对于互联网的一个很小的子集进行分析，减少了时间复杂度。但其也存在下述缺点：中心网页之间的相互引用以增加其网页评价，当一个网站上的多篇网页指向一个相同的链接，或者一个网页指向另一个网站上的多个文件时会引起评分的不正常增加，这会导致易受“垃圾链接”的影响；网页中存在自动生成的链接；主题漂移，在邻接图中经常包括一些和搜索主题无关的链接，如果这些链接自身也是中心网页或权威网页就会引起主题漂移；对于每个不同的查询算法都需要重新运行一次来获取结果。这使得它不可能用于实时系统，因为对于上千万次的并发查询这样的开销实在太大。

参考文献

1 Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: bringing order to the web〔EB/OL〕. 〔2010 - 04 - 30〕. http://ilpubs.stanford.edu:8090/422/.

2 Kleinberg JM. Authoritative Sources in a Hyperlinked Environment〔C〕. 9th Annual ACM - SIAM Symposium on Discrete Algorithms, 1998: 668 - 677.