基于页面链接结构 Page Rank 算法的改进——有向访问模型

李立耀

(福建师范大学福清分校数学与计算机科学系,福建福清 350300)

摘 要:互联网上的信息每天都以指数量级的速度爆炸性增长,面对如此浩瀚的资源,从 web 中的大量信息中准确并且有效的提取用户所需要的信息成为了 Internet 的用户的迫切需要。web 信息检索系统可以利用 web 页面的这种特殊的链接结构关系来改进检索的算法,以提高检索的精度。链接结构分析显著地提高了检索结果的相关性。在充分分析基于链接结构的算法的基础上,本文提出了一个更接近真实情形的模型——有向访问模型,它假定访问者将根据与查询相关的概率模型来指导下一步的访问,它能够真实地描述用户在浏览网页时的行为。

关键词:链接结构;信息检索;数据挖掘;随机访问模型;有向访问模型;Page Rank;

中图分类号:TP393 文献标识码:A 文章编号:1008-3421(2006)02-0004-07

引 言

Internet 自 60 年代以来得到了迅猛的发展,近几年更是以惊人的速度增长,联网主机量每年翻一番,Internet 站点每半年翻一番。互联网上的信息每天都以指数量级的速度爆炸性增长,面对如此浩瀚的资源,从 web 中的大量信息中准确并且有效的提取用户所需要的信息成为了 Internet 的用户的迫切需要。随着计算机硬件的发展,检索的效率已经不在是一个主要问题。检索的瓶颈是如何提高检索的质量,包括查全率和查准率。事实上,现在的搜索工具大多能在几秒钟内响应用户的查询。问题是它们检索出来了大量的文档,其中只有一小部分是用户所需要的内容。而且,与用户查询最相关的文档往往并不是出现在检索结果的前面。信息检索是计算机科学的一个重要子类,它的目标是从大量的文档中找到与用户查询相关的文档。web 页面的链接结构信息反映了页面作者对其他页面内容的评价。web 页面中的链接总是指向作者认为对用户可能有用的页面。web 信息检索系统可以利用 web 页面的这种特殊的链接结构关系来改进检索的算法,以提高检索的精度。链接结构分析显著的提高了检索结果的相关性,因此,大多数的搜索工具声称采用了某种链接分析算法。然而,目前实际所使用的链接分析算法大多使用了简化的模型,因此,所得到的检索结果也具有一定的局限性。本文在分析链接结构算法的基础上提出了一个更接近实际情形的模型——有向访问模型,并且给出了基于此模型的链接分析算法,它具有更好的检索效果。这对于提高 web 信息检索的质量具有重要的意义。

1 Internet 的信息分布

Internet 上的信息资源随着 Internet 的发展而呈现出的特点是:信息量大而且分散、自治性强、信息资源多种多样、不一致和不完整性。这些特点对网络软件的性能提出了很高的要求。网络的快速发展给信息挖掘带来了挑战。WWW 上信息呈现爆炸性的指数增长,同时伴随着上网经验不足、不太晓得如何查找信息的新用户的加入。用户很可能最大程度的运用超链接来在网上冲浪,他们通常从以下两类网站开始:

第一类是目录系统,其典型代表是 Yahoo!(http://www.yahoo.com),它通过有专业知识的网页编辑人员对网上的网页进行精选,建立一个索引目录,来给用户提供服务。这类通过手工维护得很好的系统的优点是提供的网页准确率高,可以有效的覆盖所有热门的主题,但它们的缺点是过于主观,而且需要高昂的代价来建立和维护,更新改进

收稿日期:2005-12-21

作者简介:李立耀(1970-)男,福建平潭人,讲师

的慢,同时不能很好的覆盖所有深奥的主题。

第二类是搜索引擎系统,比如 Google(http://www.google.com),它通过程序自动地从网上搜集和分析网页,建立索引,为用户服务。这类通过关键词匹配实现查找的自动更新的搜索引擎优点是涵盖的网页数量巨大,但通常返回太多的低质量相关性不大的结果。

2　Web 数据挖掘

Web 数据挖掘 (Web Data Mining) 是数据挖掘技术在 Web 环境下的应用, 是从数据挖掘发展过来的集 Web 技术、数据挖掘、计算机技术、信息科学等多个领域的一项综合技术。Web 数据挖掘是指从大量的 Web 文档集合中发现蕴涵的、未知的、有潜在应用价值的、非平凡的模式(Pattern)。它所处理的对象包括: 静态网页(文字、多媒体信息等)、Web 数据库、Web 页面的内部结构、Web 结构、用户使用记录等信息,通过对这些信息的挖掘,可以得到仅通过文字检索所不能得到信息。

Web 信息检索是从信息检索技术发展过来的, 它最本质的特征是系统对 Web 文档集合和用户的需求集合的匹配与选择。

Web 挖掘大致分为 3 类: Web 内容挖掘(Web Content Mining)、Web 结构挖掘(Web Structure Mining)、Web 用户使用记录挖掘(Web Structrue Mining)、Web 用户使用记录挖掘(Web Usage Mining)。Web 内容挖掘是指从 Web 上的文件内容及其描述信息中获取潜在的、有价值的知识或模式的过程; Web 结构挖掘是从 WWW 的组织结构和链接关系中推导知识,主要是通过对 Web 站点的结构进行分析、变形和结构,将 Web 页面进行分类,以利于信息的搜索; Web 用户使用记录挖掘就是对用户访问 Web 时在服务器留下的访问记录进行挖掘,挖掘的对象是在服务器上的日志信息,也称为 Web 日志挖掘。

3　Web 结构分析算法

查询器从用户那里得到查询条目,然后搜索出相关的页面。但是传统的 IR(信息检索)方法在对检索结果的评价方面效率低下。第一,由于 web 页面十分庞大,其中包含有数量巨大、质量不等、类型不同的信息。许多含有查询条目信息的页面质量不高或者与查询根本无关。第二,许多 web 页面本身并没有足够的描述自身的信息,这样传统的仅检索文本内容的 IR 技术就不能有效的进行检索。例如,如果需要查询‘搜索引擎’这个条目,实际上大多数主要的搜索引擎的页面并不包含这样的关键词。这样传统的信息检索技术便不能找到所需的结果。此外,某些页面可能会被人为地添加某些条目,从而误导纯粹基于页面内容的搜索引擎。

Web 的链接结构包含许多重要的信息,它能够用来帮助过滤或者评价 web 页面。例如,从页面 A 指向页面 B 的链接可能是页面 A 的作者对页面 B 的推荐。一些利用链接结构的算法被提出。它不仅被用来辅助基于关键词的检索,还被用于像 Yahoo 的自动层次分类或区别网上的社区等等。由于利用了比页面内容更多的信息,从总体上来说,这些算法的效果较传统的 IR 算法好。由于 web 链接的影响范围较广,因此,要利用链接结构来误导访问者远比更改页面的内容困难。故此类算法也较传统的 IR 算法更为健壮。

两种常见的基于链接结构的技术——Page Rank 和 HITS,我们主要来看 Page Rank 技术

3.1　Page Rank

Page Rank 算法是一种评价页面重要性的算法,它定义了一个全局的评价模式。它通过页面之间的链接关系来定义页面的重要性。一个页面的重要性依赖于其他页面对它的链接,同时它指向其它页面的链接也影响了其它页面的重要性。这种算法是递归的。

3.1.1　web 的链接结构

web 页面中包含有两种链接,一种是指向其它页面的链接,一种是由其它页面指向自身的链接。可以称前者为前向链接,后者为后向链接。不可能找到一个页面的所有后向链接,但是可以从页面中发现它所有的前向链接。Web 页面可能包含数目不同的后向链接。例如, sina 可能包含数以万计的后项链接,但某些个人主页可能仅有几个后向链接。一般来说,具有更多后向链接的页面可能比仅有少数链接的页面更为重要。例如,简单的引用计数曾被用来预测 Nobel 奖的获得者。Page Rank 利用一种更复杂的方法来进行引用计算。

简单的引用计数并不能说明所有的问题。例如,一个 web 页面仅有一条后向链接,但这条链接是从 sina 的主页上链接过来的。它可能只拥有一条后向链接,但这条链接具有很高的重要性。因此,这个页面应该比那些具有很多从不重要的页面上链接过来的后项链接的页面更重要。Page Rank 试图解决这种问题。

3.1.2 Rank 的传播

基于上面的考虑,Page Rank 可以描述为:如果一个页面的后向链接的重要性之和越高,页面的重要性就越高。这个描述既考虑了页面具有很多后向链接的情况,同时也考虑了页面具有一些重要性很高的后向链接的情况。

3.1.3 Page Rank 的定义

令 u 是一个 web 页面,F_u 是页面 u 所指向的页面的集合,B_u 是指向页面 u 的页面的集合,$N_u=|F_u|$ 是页面 u 的前向链接的数目。令 c 是用来规范化的因子,因此所有页面的总的 rank 值是一个常数

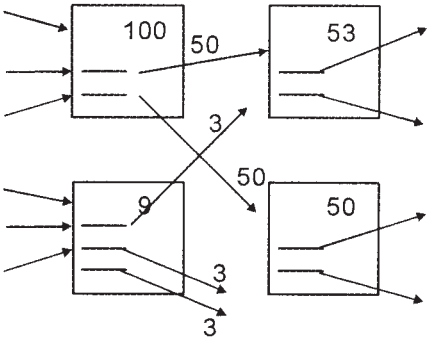

图 1 Page Rank 的传播

R 是简化了的 Page Rank,$R(u)=c\sum_{v\in B_u}\frac{R(v)}{Nv}$ 它将上一节所描述的内容进行了形式化的定义。一个页面的重要性被分解到它所有的前向链接中,同时也构成了它所指向的页面的重要性。由于一些页面没有前向链接,因此它们的重要性值将会从系统中消失,故 $c<1$。这个方程式递归的,但从任意的 rank 集合开始,经过反复的迭代直到它收敛,可以计算其值。图 1 说明了rank 值从一对页面传播到另一页面。图 2 显示了一个页面集合的相容的稳定解。

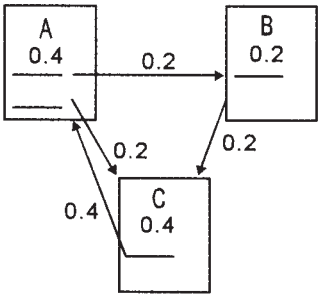

图 2 简化的 Page Rank 的计算

从另一种角度描述,令 A 是一个行与列都和 web 页面相关的对称矩阵。如果有从 u 到 v 的链接,则令 $A_{u,v}=1/N_u$,否则 $R=cAR$。令页面的 rank 矢量 $A_{u,v}=0$,故 R 是 A 具有特征值 c 的特征矢量。事实上,这里需要的是 A 的最大特征矢量。只需将 A 反复应用于任何初始矢量即可。

简化的 Rank 函数存在一些问题。考虑两个互相指向的页面,并没有指向其它任何页面。假设有也指向其中任何一个页面的页面。由于它们没有其他前向链接,在迭代过程中,它们将不断积累 rank 但却没有将 rank 传递出去。这个循环构成了一个陷阱,称之为 rank sink。为了解决这个问题,引入:

定义 1:令 $E(u)$ 是相应于 rank 源的 web 页面的矢量,则 web 页集合的 Page Rank 值 R' 满足:$R'(u)=c\sum_{v\in B_u}\frac{R'(v)}{N_v}+cE(u)$ 这里 c 被最大化,且 $\|R'\|_1=1$($\|R'\|_1$ 表示)。这里 $E(u)$ 是与 rank 源相对应的 web 页面的矢量。如果 E 的每个元素都是正的,c 必须能够是方程平衡。因此,这种方法对应于一个衰减因子。在矩阵描述

中有 $R'=c(AR'+E)$。由于 $\|R'\|_1=1$，故可重写为 $R'=c(A+E\times l)R'$

其中 l 是所有的都存在的矢量，故 R' 是 $A+E\times l$ 的一个特征根。

3.1.4　随机访问模型

上面定义的 Page Rank 在图中随机访问时存在直觉地偏向。简化的模型相应于 web 图中随机浏览的稳定的概率分布。这可以被视为一个随机访问者模型。这个随机访问者将随机地点击连续的链接。但是，如果一个真正的 web 访问者陷入了一个 web 页面的循环，它不可能永远在其中循环点击。相反，他会跃出来访问集合外的其他页面。附加的因子 E 可以被视为来说明这种行为。访问者可能会跳出，以分布 E 来选择其它随机页面。

这样 E 便成为用户定义的参数。在大多数情况下，令 E 对所有的 web 页面具有相同的值 α。Page Rank 矢量 R 与随机访问的稳定概率分布成正比。因此，页面的 Page Rank 与随机访问者访问它的频率成正比。

3.1.5　计算 Page Rank

如前面所述，Page Rank 的计算等价于计算矩阵 A 的最大特征矢量。一种计算最大特征矢量的方法称为“幂迭代”。（只有图市非周期的才能保证这种方法是收敛的。事实上，web 大多是非周期的）在这种方法中，任意的初始矢量与给定矩阵反复相乘，直到它收敛于主特征矢量。对 Page　Rank 的计算过程如下：

（1）$S\leftarrow$ 任意随机矢量

（2）$R\leftarrow A^T\times S$

（3）如果 $\|R-S\|<\varepsilon$，则终止。R 是 Page Rank 矢量。

（4）$S\leftarrow R$，跳转到（2）。

图 3 说明了一个简单图的 Page Rank 计算。容易验证这个 rank 指派满足 Page　Rank 的定义。例如，结点 2 的 rank 值为 0.286 且有 2 条前向链接，它的一半 rank（0.143）传递给结点 1，另一半传递给了结点 3。由于结点 3 没有其他后向链接，其 rank 值仅来自于结点 2，为 0.143。结点 1 从结点 2 处得到了 0.143，加上从结点 3 处得到的 0.143/2，再加上从结点 5 处得到的 0.143/2，总共为 0.286。因为结点 1 有 3 个后向链接，故它具有更高的 rank 值。由于每个访问结点 1 的访问者必须访问结点 2，故结点 2 具有同样的 rank 值。所有的 rank 值之和等于 1。

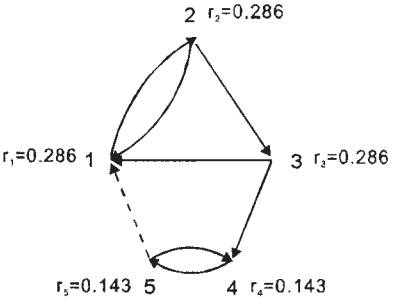

图 3　简化的 Page Rank 的计算

3.1.6　实际的 Page Rank

只有在强连通的时候，简单 Page　Rank 算法才是有效的。但是 web 远不是强连通的。因此，在实际的 web 中存在两种问题：rank 汇聚和 rank 泄漏。

一个内部连通的页面集合，如果没有任何指向外部的链接，这样就形成了 rank 汇聚。如果一个页面没有任何指向其它页面的链接，这样就形成了 rank 泄漏。虽然 rank 泄漏只是 rank 汇聚的一种特殊情况，但是它会引起完全不同的问题。在 rank 汇聚的情形中，如果一个结点不在汇聚结点集合中，那么这个结点就只能得到 rank 值 0。这意味着不能说明这样的结点的重要性。；如，在图 3 中，将结点 5 到结点 1 的链接删去，会使结点 4 和结点 5 形成汇聚。一个随机访问者在访问图时，最终会陷入到结点 4 和结点 5 的循环中。例如，结点 1，2，3 的 rank 值为 0，结点 4 和 5 的 rank 值为.05。

另一方面，任何传递到 rank 泄漏结点的 rank 值将永远从系统中消失。例如，在图 3 中，如果移去结点 5 及其相关的所有链接，结点 4 将变成一个 rank 泄漏结点。这个泄漏点将导致所有的 rank 收敛于 0。即随机访问者最终会到达结点 4，并且将永远停在那里。

解决这个问题可以采取两种方法。第一，将所有出度为0的泄漏结点都删去。因此，泄漏结点的不到Page Rank。另一种方法是假设泄漏结点都含有指向它们的后项链接所指的结点的链接，这样通过具有高rank指访问的泄漏结点同样具有高rank值，具有低rank指访问的泄漏结点也同样具有低rank值。第二，在Page Rank的定义中引入衰减因子d（$0<d<1$）。在修改的定义中，页面rank值的一部分分布于它所指的结点中。剩余的rank值被平均分布到web中的所有页面中。因此，修改的定义为 $R(u)=d\sum_{u\in B_v}R(v)/N(v)+(1-d)/m$，其中$m$是图中的结点总数。简单的Page Rank实际上是这里的一种特殊情形（$d=1$）。

在随机访问模型中，修改的定义说明了访问者偶尔会不遵循链接而跳转到web上的一个随机页面上（而不是从现在的页面所链接的页面中选取一个）。衰减因子d说明了这种情形发生的频率。

图4说明了将图3中把结点5到结点1的链接删除后所得到的修改的Page Rank（其中$d=0.8$）。结点4和结点5比其他结点具有更高的rank值，这说明访问这可能会向结点4和结点5移动。但是，其它结点也具有不为0的rank值。

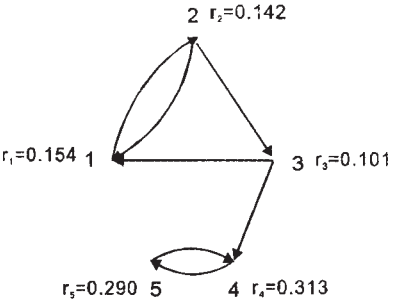

图4 修改的Page Rank的计算

3.1.7 计算问题

为了使“幂迭代”的计算切合实际，不仅需要收敛于Page Rank，而且要求这一过程在有限步骤中完成。理论上说，矩阵的幂迭代的收敛依赖于其特征值的间隙，即所给矩阵的特征值的绝对值之差。文献说明了幂迭代相对较快（大约在100步迭代左右）。

事实上，搜索算法利用的是由Page Rank得到的页面的rank值大小的相对顺序，而不是Page Rank值本身。因此，一旦页面的顺序变得稳定了便可以停止幂迭代运算。实验表明，由Page Rank决定的顺序比Page Rank值本身的收敛要快得多。

4 改进的算法

4.1 Page Rank：随机访问

Page Rank算法假定访问者从一个页面跳转到另一个页面市，在每一跳中队每一个链接的选择具有相同的概率。为了减少叶结点或死循环的影响，访问者偶尔会跳转到一个具有较小概率β的随机页面。若将web视为有向图，其中结点代表web页面，而边代表web页面之间的链接。令W是结点的集合，$N=|W|$，F_i是页面i所指向的页面的集合，B_i是指向页面i的页面的集合。对于没有指向其它页面的页面，为其添加一跳指向图中所有结点的链接。利用这种方法，可以将由于没有指向其它页面链接而导致的rank统一分布到所有页面中。经过足够多的步数后，访问者在页面j的概率是

$$P(j)=\frac{(1-\beta)}{N}+\beta\sum_{i\in\beta_j}\frac{P(i)}{|F_i|} \quad (4.1)$$

Page Rank的评价指数便定义为这个概率：$PR(j)=P(j)$。由于方程4.1是收敛的，它必须经过反复迭代直到$P(j)$。一般地，$P(j)$的最初分布是一致的。Page Rank和转移矩阵Z的主特征根是相等的。转移矩阵Z：

$$Z=(1-\beta)\left[\frac{1}{N}\right]_{N\times N}+\beta M \quad (4.2)$$

其中如果存在从i到j的边，则$M_{ji}=\frac{1}{|F_i|}$；否则$M_{ji}=0$。方程(4.1)的一个迭代等价于迭代t时计算$x^{t+1}Z=Zx^t$，

其中 $x'_j = P(j)$。在收敛时，有 $x^{T+1} = x^T$ 或 $x^T = Zx^T$。这说明 x^T 是 Z 的一个特征根。并且，由于 Z 的列是规范化的，所以 x 有一个特征根 1。

4.2　有向访问模型

我们考虑一个具有更高智能的访问者，他根据页面的内容和查询的条目从一个页面跳转到另一个页面。因此，页面的概率分布为：

$$P_q(j) = (1-\beta)P'_q(j) + \beta \sum_{i \in \beta_j} P_q(i \to j) \quad (4.3)$$

其中 $P_q(i \to j)$ 是访问者在页面 i 并且进行查询 q 时跳转到页面 j 的可能性。 $P'_q(j)$ 是当不顺着链接时访问者选择跳转的地方。$P_q(j)$ 是页面的概率分布和相应的与查询相关的 Page Rank 指数 $(QD-PageRank_q(j) \equiv P_q(j))$。和 Page Rank 一样，QD-Page Rank 是由方程(4.3)从一些初始分布反复迭代评估得到的。和转移矩阵 Z_q 的主特征根一致，其中：

$$Z_{q_{ji}} = (1-\beta)P'_q(j) + \beta \sum_{i \in \beta_j} P_q(i \to j)$$

虽然 $P_q(i \to j)$ 和 $P'_q(j)$ 是任意分布的，但是这里主要考虑当两种概率分布都由 $R_q(j)$ 导出时的情形。$R_q(j)$ 是页面 q 和查询 j 的相关度的一种度量：

$$P'_q(j) = \frac{R_q(j)}{\sum_{k \subset W} R_q(k)} \qquad P_q(i \to j) = \frac{R_q(j)}{\sum_{k \in F_i} R_q(k)} \quad (4.4)$$

这就是说，当页面的多个指向其它页面的链接中选择时，有向的访问者可能会选择那些指向内容可能和查询相关（通过 R_q）的页面的链接。与 Page Rank 相似，当页面的所有指向其它页面的链接的相关度为 0 时，或者根本就没有指向其它页面的链接时，给页面添加指向网络中所有页面的链接。在这样的页面中，访问者可能会通过分布 $P'_q(j)$ 来选择一个新的页面来访问。

当查询有多个条目时，$Q = \{q_1, q_2, \ldots\}$，访问者通过某种概率分布 $P(q)$ 选择一个 q，并且利用这个条目来指导访问（通过许多步）。然后通过概率分布再选择一个 q 来指导下面的访问。最终的访问页面分布，即 $QD-PageRank_Q$ 为：

$$QD-PageRank_Q(j) \equiv P_Q(j) = \sum_{q \in Q} P(q)P_q(j) \quad (4.5)$$

对标准的 Page Rank 算法，Page Rank 矢量即为矩阵 Z 的主特征根。对单一条目的 $QD-PageRank_q$ 矢量同样是矩阵 Z_q 的主特征根。但是，$QD-PageRank_Q$ 矢量并不是矩阵 $Z_Q = \sum_{q \in Q} P(q)Z_q$（相应于方程 4.5 得到的）的主特征根。事实上，Z_Q 的主特征根对应于一个随机访问者，他在每一步都通过 $P(q)$ 来选择一个新的查询所得到的 QD-Page Rank。但是，$QD-PageRank_Q$ 和这种单步访问所得到的 Page Rank 大致相等。令 x_q 是矩阵 Z_q 的规范化的主特征矢量（注意：x_q 的元素 j 即是 $QD-PageRank_q(j)$），因此满足 $x^i = T^i x^i$。由于 x_q 是 Z_q 的主特征向量，故有 $\forall q,r \in Q: \|Z_q x_q\| \geq \|Z_q x_r\|$ 因此，有 $Z_q \sum_{r \in Q} x_r \approx kZ_q x_q$。假设 $P(q) = 1/|Q|$，考虑 $x_Q = \sum_{q \in Q} P(q)x_q$（见方程 4.5）有：

$$Z_Q x_Q = \left(\sum_{q \in Q} \frac{1}{|Q|} Z_q\right)\left(\sum_{q \in Q} x_q\right) = \frac{1}{|Q|}\sum_{q \in Q}\left(Z_q \sum_{r \in Q} x_r\right) \approx \frac{1}{|Q|}\sum_{q \in Q}(kZ_q x_q) = \frac{k}{|Q|}\sum_{q \in Q} x_q = \frac{k}{n} x_Q$$

因此，x_Q 约等于 Z_Q 的一个特征矢量。由于 x_Q 等价于 $QD-PageRank_Q$ 且 Z_Q 描述了一个单步访问者的行为，因此 $QD-PageRank_Q$ 大致与单步访问者所得到的 Page Rank 相同。当由 Z_q 定义的随机访问者之间具有很大或很小相似性时，约简最接近于精确值。

对相关性函数 $R_q(j)$ 的选取是任意的。在最简单的例子中，$R_q(j) = R$，与查询的条目和文档无关，QD-Page Rank 简

化为 Page Rank。最简单的与内容相关的函数是当条目 q 出现在页面 j 中时 $R_q(j)=1$，否则为 0。有许多复杂的函数可以使用，例如 TFIDF 信息检索矩阵，等等。现在所使用的大多数文本评价函数可以很容易的用于上面的模型。

5 结论

有向访问模型是将页面内容和页面链接结构一起考虑的概率模型。由于该模型比 Page Rank 算法的随机访问模型更接近真实的情形，因此，改进的算法比 Page Rank 算法具有更好的效果，而且算法的复杂度大致与原算法相当。

参考文献:

[1] 韩家炜，孟小峰等.Web 挖掘研究[J].计算机研究与发展，2001，38(4)：405-414.
[2] 雷鸣等.第三代搜索引擎与天网二期[J].北京大学学报(自然科学版)，2001，35(9)：734-740.
[3] 王继成，潘金贵等.Web 文本挖掘技术研究[J].计算机研究与发展，2000，37(5)：513-520.
[4] 王继成，萧 嵘等.Web 信息检索研究进展[J].计算机研究与发展，2001，38(2)：187-193.
[5] 王 奇等.信息检索中基于链接的网页排序算法[J].华东理工大学学报，2000，26(10)：455-458，465.
[6] 阳小华.Web 站点的超链接结构挖掘[J].计算机工程与应用，2001，37(8)：64-65.
[7] 邹 涛，王继成等.文本信息检索技术[J].计算机科学，1999，26(9)：72-75.

The Improvement for the Page Rank Algorithm Based on Page Link Structure——Directive Visit Model

LI Liyao
(Fuqing Branch of Fujian Normal University, 350300)

Abstract:With the indexically explosive increase of the information from Internet,it is essential for the Internet users to extract neccessary information effectively and accurately from the vast amount of information from web. To improve the accuracy of retrieval system for a better algorithm of retrieval. Link structure analysis can highlight the relevance of retrieval result. On the basis of a full analysis on link-based algorithm, this paper suggests a Directive Visit Model, a model much closer to what is real, which supposes that visitors' further visit be instructed and guided by a probabilistic model that is relevant to browsing. It is expected to be able to describe more accurately the web browsing behavior of users.

Key Words: Link structure; Information retrieval; Data excavation; Random visit model; Directive visit model; Page Rank

(责任编辑：薛世平)