第 35 卷 第 6 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.6 Computer Engineering March 2009
·软件技术与数据库· 文章编号：1000—3428(2009)06—0050—02 文献标识码：A 中图分类号：TP312

一种基于本体的 PageRank 算法的改进策略

姚文琳，刘 文
(中国海洋大学信息科学与工程学院，青岛 266100)

摘 要：介绍 Google 等搜索引擎应用的 PageRank 算法的定义、特点及缺陷。针对 PageRank 算法在基于 Ontology 的海洋文档检索系统应用中的问题对其加以改进，增加了文本文档的判断和主题相关性的判断，提出 IPageRank 算法。介绍海洋文档检索系统，并将改进的 IPageRank 算法应用于该系统中进行验证。

关键词：PageRank 算法；IPageRank 算法；主题相关；本体

Improved Strategy of PageRank Algorithm Based on Ontology

YAO Wen-lin, LIU Wen
(School of Information Science and Engineering, Ocean University of China, Qingdao 266100)

【Abstract】This paper introduces the definition, the characteristic and the limitation of PageRank algorithm used by Google and other search engine. It develops a system, ocean search system, which is based on Ontology and PageRank algorithm. However, because of PageRank algorithm's insufficiencies in this system, a new algorithm is proposed and applied in the system's consummation. It introduces the functions of the ocean search system and takes an illustration and test to the new algorithm.

【Key words】PageRank algorithm; IPageRank algorithm; theme-relativity; Ontology

随着 Internet 网上信息资源的飞速增长，人们不仅要求搜索引擎查找的信息全面、准确，还要求它能对专业领域的信息进行集中查询。现代的搜索引擎都在搜索结果排列中引入关键字串匹配程度的概念，旨在向用户提供最重要、最有用的网页。应用于搜索引擎中的算法大多依据主题的相关性或页面的权威性来评价 URL 的重要性。比较常用的页面权威值计算方法有 HITS 算法等。排名算法中最有影响力的当数 PageRank 算法。

1 PageRank 算法介绍

搜索引擎 Google 最初是斯坦福大学的博士研究生 Sergey Brin 和 Lawrence Page 实现的一个原型系统，现在已经发展为 WWW 上最好的搜索引擎之一。Google 的体系结构与传统搜索引擎的最大不同在于对网页进行了基于权威值的排序处理，使最重要的网页出现在结果的最前面。Google 通过 PageRank 元算法计算出网页的 PageRank 值，从而决定网页在结果集中的出现位置，网页的 PageRank 值越高，在结果中出现的位置越前。

(1)PageRank 算法[1]基于的 2 个前提

前提 1：一个网页被多次引用，则它可能很重要；一个网页虽然没有被多次引用，但是它被重要的网页引用，则它也可能很重要，一个网页的重要性被平均地传递到它所引用的网页。这种重要的网页称为权威(authoritive)网页。

前提 2：假定用户一开始随机地访问网页集合中的一个网页，以后跟随网页的向外链接向前浏览网页，不回退浏览，浏览下一个网页的概率就是被浏览网页的 PageRank 值。

(2)PageRank 算法定义

算法基于“从许多优质的网页链接过来的网页必定还是优质网页”的回归关系判定所有网页的重要性。认为当某个网页链接到另一个网页时，它就对该网页“投了一票”。一个网页的得票越多，则它的重要性就越高。进一步说，投票网页的重要性也决定了票本身的重要程度。Google 通过计算网页得票得到页面重要性。计算 PageRank 值时，每票的重要性都要考虑在内。

PageRank 算法描述如下：将网络看作一个有向图：$G=(V, E)$，其中，V 是节点(网页)集；E 是边(当且仅当存在从页面 i 到页面 j 的链接时存在从节点 i 到节点 j 的边集)。PageRank 算法的思想在于一个页面重要，或者有链接指向它的页面多，或者有链接指向它的页面重要或者两者兼而有之。

其初始定义公式表示如下：

$$PR(K)=\sum_{i=1}^{n} PR(T_i) \Big/ N(T_i)$$

其中，$N(T_i)$表示 T_i 的出度；$PR(T_i)$表示 T_i 的 PageRank 值。

由图 1 可知：$PR(B)=PR(A)/N(A)+PR(C)/N(C)$。

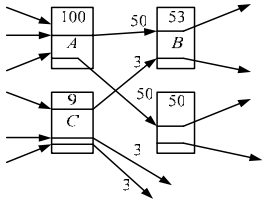

图 1 PageRank 概念图

为了完善算法，保证叠代过程的收敛，引入了阻尼因子 c[2](c 通常被设置成 0.85)，定义公式变为

基金项目：国家自然科学基金资助项目(60403012)；山东省自然科学基金资助项目(Y2005G06)

作者简介：姚文琳(1967－)，女，副教授，主研方向：信息获取，语义 Web，信息分类、过滤和存储；刘 文，硕士研究生

收稿日期：2008-08-16 **E-mail：**apple_liuwen@hotmail.com

$$PR(K)=(1-c)+c\bullet\sum_{i=1}^{n}PR(T_i)\Big/N(T_i) \quad (1)$$

PageRank 算法不以站点排序，而是对单个页面进行级别排序，页面网页级别由一个个页面各自独立决定；页面的网页级别由链向它的页面的网页级别决定，每个页面链入的贡献值是不同的。由式(1)可知，T_i 页面中链出数越多，即 $N(T_i)$ 越大，$PR(T_i)/N(T_i)$越小，它对当前页面 K 的贡献就越小。页面 K 链入页面越多，$PR(K)$的网页级别也越高；阻尼系数 c 的使用减少了其他页面对当前页面 K 的排序贡献。但是在应用的过程中发现 PageRank 算法也有一定的缺陷和不足。

2 PageRank 算法的缺陷

PageRank 算法的优点是网页的 PageRank 值由网络链接的结构决定，与检索的内容无关，因此，检索期间消耗很小。它的缺陷在于主题相关性问题。互联网上的资源涵盖了上百万甚至更多的主题，很多情况下搜索都是基于语义的文档资料查询，查询用户寻找的是一些具有特定主题的文档资料。然而，PageRank 算法是单纯根据一个网页上被链接的站点数量和质量来给该网页分配一个绝对的“重要性值”。同时将链接页面的页面等级考虑在内，如果指向一个网页的外部链接页的页面等级越高，则该链接页面传递给此网页的页面等级值就越高。“页面等级值”并非针对查询词语，由式(1)可知，一个网页即使只是在内容中偶然提到了一个和查询主题偏离的关键词语，也会因其居高的页面等级值而获得一个较高的排名，从而影响了搜索结果的相关性与精准性。因此，需要对 PageRank 算法进行改进，以改善搜索的相关性和精准性。

3 对 PageRank 算法的改进

在查询过程中，PageRank 算法根据 *PR* 值决定页面出现的排名，会出现许多与查询内容不相符的页面，所以，首先需要对查询出的内容是否是文档进行可能性的判断：

(1)对文档进行可能性的判断

```
Procedure Judge(int i)
If(deg(i))
C(Ui) = 1 ;
Else C(Ui) = d ; //(d 为较小的常数)
```

其中，$deg(i)$代表页面中是否具有文档内容形式，代表查询的第 i 个页面；$C(U_i)$代表对页面内容的判断因子。当页面为非文档内容时，$C(U_i)$为较小常数，否则，$C(U_i)$为 1。

由式(1)可知，页面的 PageRank 值为 $PR(U_i)$，引入对文档可能性的判断可以对 PageRank 算法进行如下改进：

$$PR'(U_i)=PR(U_i)\times C(U_i) \quad (2)$$

(2)主题相关性判断

为了进一步提高搜索结果页面的可能性，引入主题相关性的判断对上面筛选出的页面进行主题相关的计算。计算公式如下：

$$Simular(Q,B_i)=\sum_{i=1}^{n}M_i\Big/\sqrt{\sum_{i=1}^{n}M_i^2} \quad (3)$$

其中，$Simular(Q, B_i)$表示主题 Q 与文档 B_i 的相似度；$M_i = Freq_i \times WF_i$；$Freq_i$ 表示主题 Q 在文档 B_i 中出现的次数；WF_i 表示在 WWW 上包含主题 Q 的文档的估计值。

(3)对 PageRank 算法的改进

根据式(1)~式(3)，在链接关系的基础上，加入页面与查询主题的相关性，使所产生的 PageRank 值高的页面是针对用户查询主题的，这就形成了 IPageRank 算法。改进算法公式如下：

$$IPageRank(K)=(1-c)+c\bullet\sum IPageRank(T_i)\times Simular(T_i)\Big/\sum Simular(T_i)$$

对 IPageRank 算法的解释如下：假设 Web 上有一个主题的浏览者，首先 PPR 算法对查询的页面进行筛选，选择论文或者文档的页面进行排序，*IPageRank*(*K*)是访问页面 *K* 的概率。从初始的页面集开始，按照页面的链接前进，在每一个页面中，浏览者感兴趣的概率与主题呈正比关系，与主题相关的页面 *IPageRank* 值相对较高，这就使与主题联系不大而 PageRank 值高的页面的 *IPageRank* 值相对较低。

4 在基于 Ontology 的文档检索系统中的应用

(1)海洋信息文档检索系统

基于 Ontology 的海洋信息文档检索系统主要包括：查询模块，Ontology 管理模块，语义标注抽取模块和信息的获取模块 4 个部分，图 2 是系统框架。

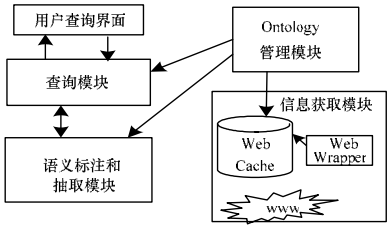

图 2　文档检索系统结构

将改进的算法应用于该系统的信息获取模块中。工作流程如下：

1)利用网络蜘蛛[3]获取网页信息。

2)通过式(2)进行文档的可能性判断。分析上面所获取的网页信息，进行文档的可能性判断，去除不感兴趣的网页地址信息。

3)利用式(3)进行主题相关性的评价，与主题无关的网页会因获得相对较小的 *PageRank* 值而得到一个比较靠后的排名。

4)计算出相应的改进了的 *PageRank* 值，根据新的 *PageRank* 值对结果进行排序。

(2)试验数据

利用网络蜘蛛对爬到的 10 000 多个网页进行分析，对某一主题进行查询，例如：哲学，用 *PageRank* 算法可以搜索到 371 个有效的网页信息。在进行文档可能性判断和主题相关性判断之后，用改进的算法计算 *IPageRank* 值，得到 57 个有效的信息提供给语义标注和抽取模块。

搜索得到了网页新的排名，与主题相关且相似度高的网页的 *PageRank* 值相对较高，而那些原本与主题无关但又排名较高的网页会得到一个较低的排名，从而满足了基于 Ontology 的海洋信息检索系统的检索要求。

5 结束语

从实验数据中可以确定，在基于 Ontology 海洋信息文档检索系统中应用 *IPageRank* 算法增加了系统查询的准确性。但是，由于在文档可能性判断和主题相关性判断时不可避免地使用更多的时间和资源，因此下一步需要探索更加有效的排序算法和主题相关的算法。 (下转第 54 页)