

# 基于链接分析的相关排序方法的研究和改进

原福永, 张园园

(燕山大学信息科学与工程学院, 河北秦皇岛 066004)

**摘要:** 搜索引擎的相关结果排序技术是信息检索技术发展中的关键问题, 也是将来研究的热点问题之一。在分析传统的相关排序方法基础上, 介绍了PageRank算法和HITS算法的核心技术, 指出了PageRank算法忽视专业站点、对网页中的超链接评估不当之处, 根据面向主题的思想, 在重新计算链接对网页的影响的基础上, 提出了一种新的关于链接技术的排序方法, 并通过实验对该算法的性能进行分析评价。

**关键词:** 搜索引擎; 相关排序; 链接分析; 算法; 面向主题

**中图分类号:** TP391 **文献标识码:** A **文章编号:** 1000-7024(2007)07-1630-02

## Correlation arrangement method research and improvement based on link analysis

YUAN Fu-yong, ZHANG Yuan-yuan

(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

**Abstract:** The technology of related result sorting in search engine is the key problem during the development of information retrieval technological, and also is one of hot topics which the future will study. In the foundation of analysis tradition correlation arrangement method, the core technology of PageRank algorithm and the HITS algorithm are introduced, and the disadvantage of this algorithm is given, such as neglecting specialized sites, inaccurate judge of hyperlinks, according to the thought of facing the subject. In the foundation of recomputing the influence of the link to the web page, one kind of new about the link technology arrangement method based on the thought of facing subject is proposed. And the performance of this algorithm is carried on the analysis appraisal with the experiment.

**Key words:** search engine; related arrangement; hyperlink analysis; algorithm; facing subject

## 0 引言

随着Internet的出现和迅速发展, 传统的信息检索技术就显出了它的局限性。在信息检索中评价查询和文档的相关度的方法中最经典、最有影响的是Gerald Salton等在30多年前提出的“向量空间模型”(vector space model, VSM)。该模型的基本思想是: 把查询 $q$ 和文档 $d$ 看成由相互独立的词条组构成, 这样查询某个主题的相关度按如下公式计算(余弦表示法)

$$Sim(d, q) = \frac{\sum_{i=1}^m w_i q_i}{|d| \cdot |q|}$$

式中:  $q_i$ —— $q$ 中的第 $i$ 个词条的权重,  $w_i$ —— $q$ 中第 $i$ 个词条在文档 $d$ 中的权重,  $|d|$ 与 $|q|$ ——文档和查询串的长度。但是这种方法在对WWW的页面进行检索的时候返回结果的质量不是很高, 表现在: ①只考虑了普通文本与网页的近似, 忽视HTML标记和超链接等内容。标签是网页作者将网页的不同部分以不同的形式呈现给用户的手段, 因此它能够提示其中文字的重要程度。链接是反映网页之间形成的“参考”、“引用”、“推荐”关系。可以合理的假设, 若一篇网页被较多的其它网页链接, 则说明其内容较重要或较有用; ②不同主题的页面混杂在

一起。一词多义是自然语言里面一种十分普遍的现象, 同一个词在不同领域里面所表达的含义是不一样的, 查询用户希望了解的一般都是某一个领域的信息内容, 而搜索引擎将不同领域的内容混杂在一起提供给用户, 显然这样的页面质量就不是很高; ③低质量的页面大量返回。这里提到的低质量的页面是指: 页面虽然包含查询信息, 但页面本身没有什么内容或者包含的内容价值不是很大。为了提高查找质量, 我们可以把将每个网页认为是一个节点, 每一条超文本链接认为是节点P和Q之间的有向边(从P指向Q), 那么整个Web构成了一个庞大的有向图, 基于对此图的理解出现了两大关于链接结构在搜索算法上的表示和应用, 即PageRank技术和HITS技术。PageRank是由Stanford大学的Google研究小组提出并用于Google开发系统中。HITS(hyperlink-induced topic search)技术是由Cornell大学的J.M.Kleinberg博士等提出并用于IBM公司的Clever系统的开发中。

## 1 两种技术的主要思想

### 1.1 Pagerank 技术

可以用一种“随机冲浪”模型作为它的理论基础, Page-

收稿日期: 2006-03-01 E-mail: kellyuan191@sohu.com

作者简介: 原福永(1958-), 男, 黑龙江鸡西人, 副教授, 硕士生导师, 研究方向为信息检索、智能Agent、网络技术; 张园园(1983-), 女, 河北保定人, 硕士研究生, 研究方向为信息检索、智能Agent。

Rank算法基于以下两个前提:①如果一个页面被多次引用,那么这个页面很可能是重要的;如果一个页面尽管没有被多次的引用,但却被一个重要的页面引用,那么这个页面也可能是重要的;一个页面的重要性被均分并传递到它所引用的页面。②假定用户一开始随机的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。PageRank算法可定义为:假设网页 $u$ 存在 $T_1, T_2, \dots, T_n$ 的连接网页;参数 $d$ 代表“随机冲浪者”沿着链接访问网页的衰减因素,取值范围在0~1之间,根据经验值我们一般取为0.85,  $C(T_i)$ 代表网页 $T_i$ 链向其它网页的连接数量,  $PR(u)$ 定义为网页 $u$ 的链接权值。采用以下公式计算这个权值:  $PR(u) = e(1-d) + d \left( \frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$ , 其中,  $e$ 为  $1/\max$ , 即  $\max$  为所有网页的总和, 每个网页初始分配的权值为总和的倒数。这种算法的特点是它不以站点排序, 而是对单个页面进行排序, 页面的网页级别由链向它的页面的网页级别决定, 但每个链入页面的贡献的值是不同的。也就是说PageRank算法不是面向主题的, 它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”。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该网页的页面等级值也就越高。因而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比较高的排名, 从而影响了搜索结果的相关性与精准性。

## 1.2 HITS 技术

Clever描述两种类型的网页:①权威型网页(authority), 对于一个特定的检索, 该网页提供最好的相关信息;②目录型网页(hub), 该网页提供很多指向其它高质量权威型网页的超链接。一般而言, 好的hub页面指向许多好的权威页面, 好的权威页面是指由许多好的hub页面所指向的页面。这就是hub/authority方法的基本思想。将查询 $q$ 提交给普通的基于相似度的搜索引擎, 搜索引擎返回给多个页面, 取前 $n$ 个作为根集合(root set), 用 $S$ 表示。通过向 $S$ 中加入被 $S$ 引用和引用 $S$ 的页面将 $S$ 扩展成一个更大的集合 $T$ 。以 $T$ 中的hub页面为顶点集合 $V_1$ , 以 $T$ 中的authority页面为顶点集合 $V_2$ ,  $V_1$ 中的页面到 $V_2$ 中的页面链接为边集 $E$ , 这样就形成了一个二分图 $SG=(V_1, V_2, E)$ 。对 $V_1$ 中的任何一个顶点 $v$ , 用 $h(v)$ 表示页面 $v$ 的hub值, 对 $V_2$ 中的任何一个顶点 $u$ , 用 $a(u)$ 表示页面 $u$ 的authority值。首先对 $h(v)$ 和 $a(u)$ 进行初始化, 均置为1。则有

$$a(u) = \sum_{v:(u,v) \in E} h(v) \quad h(v) = \sum_{u:(u,v) \in E} a(u)$$

进行一定次数的迭代递归后, 我们会得到集合中每个网页的hub和authority值。按照这两个不同的权值, 取前 $k$ 个结果返回给用户。这个公式说明如果一个页面有很多hub指向, 那么它的authority权重会相应的增加; 如果一个页面指向许多权威页面, 那么它的hub权重也会相应的增加。HITS技术灵活性强, 并且更加精确, 但是由 $S$ 到 $T$ 的生成代价是昂贵的, 因此检索效率显然不高。

## 2 基于链接算法的改进

通过以上论述, PageRank公式为

$$PR(u) = e(1-d) + d \left( \frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$

对于这个公式我们做两点改进。首先, PageRank对页面的重要性发现起到重要作用, 但是它不是面向某一个主题的。因此, 一个被大量无关于主题的页面群指向的页面的PageRank值就比一个由少量相关于主题的页面群指向的页面的PageRank值高, 这个现象对基于主题的采集来说是不合理的。但是, 对于一个被大量相关于主题的页面群指向的页面的PageRank值高于一个由少量相关于主题的页面群指向的页面的PageRank值这个现象来说, 我们却要加以利用。为此, 我们对PageRank方法进行了改进: 在链接关系的基础上, 加入一定的语义信息权重, 以使得所产生的重要页面是针对某一个主题的, 这就形成了SPageRank算法。SPageRank算法既利用了PageRank发现重要页面的优势, 又利用主题相关性。我们在PageRank算法中引入相关度分析, 得到改进算法

$$SPR(u) = e(1-d) + d \left( \frac{PR(T_1)}{C(T_1)} \cdot \frac{\sum_{i=1}^{k_1} \text{sim}(q, T1)}{\sum_{i=1}^{k_1} \text{sim}(q, Di)} + \frac{PR(T_2)}{C(T_2)} \cdot \frac{\sum_{i=1}^{k_2} \text{sim}(q, T2)}{\sum_{i=1}^{k_2} \text{sim}(q, Di)} + \dots + \frac{PR(T_n)}{C(T_n)} \cdot \frac{\sum_{i=1}^{k_n} \text{sim}(q, Tn)}{\sum_{i=1}^{k_n} \text{sim}(q, Di)} \right)$$

式中:  $u$ ——给定的一个网页, 假设指向它的网页有 $T_1, T_2, \dots, T_n$ ;  $k_1, k_2, \dots, k_n$ ——网页 $T_1, T_2, \dots, T_n$ 中所含的链接数;  $D$ ——衰减因子, 一般取0.85。其次, 我们把SPageRank和HITS算法结合起来形成一种新的迭代来计算网页的hub和authority值。这种改进的基本思想是“出度”(从它连出的链接个数)多的网页具有较高的hub等级; 同理, “入度”(指向它的网页个数)多的网页具有较高的authority等级。为此, 我们对向量 $e$ 修改成网页的出/入度。我们都知道, PageRank算法是一种随机冲浪模型, 我们一般认为, 当随机冲浪者感到厌倦这个网页时总是会选择出度/入度大的网页进行继续冲浪。因此, 这种出度/入度大的网页应该比那些随机的网页有较高的得分。我们这里仅以hub为例说明对 $e$ 的改进。这种用于计算网页的hub值的新算法描述如下:

定义 $N$ 为Web的页面总数; 定义 $SO$ 为所有页面的出度总数; 定义 $O_i$ 为页面 $i$ 的出度数; 则得到向量 $e$ 的计算:  $e_i = O_i \cdot \frac{N}{SO}$ , 把这个向量 $e$ 代入我们改进的面向主题的SPageRank公式得到页面 $u$ 的权重

$$FPR(u) = e(1-d) + d \left( \frac{PR(T_1)}{C(T_1)} \cdot \frac{\sum_{i=1}^{k_1} \text{sim}(q, T1)}{\sum_{i=1}^{k_1} \text{sim}(q, Di)} + \frac{PR(T_2)}{C(T_2)} \cdot \frac{\sum_{i=1}^{k_2} \text{sim}(q, T2)}{\sum_{i=1}^{k_2} \text{sim}(q, Di)} + \dots + \frac{PR(T_n)}{C(T_n)} \cdot \frac{\sum_{i=1}^{k_n} \text{sim}(q, Tn)}{\sum_{i=1}^{k_n} \text{sim}(q, Di)} \right)$$

同样地, 我们可以计算页面的authority值。令 $e_i = I_i \cdot \frac{N}{SI}$ ,  $I_i$ 表示网页 $I$ 的入度总数,  $SI$ 表示所有页面的入度总数。这就是最终改进的基于链接的算法公式, 它可以这样来解释: 假设Web上有一个主题浏览器, FPageRank是它访问页面 $u$ 的概率。它从初始的页面集出发, 按照页面的链接前进, 从不进行back操作。以概率 $d$ 顺着链接的点击访问, 如果他厌倦访问此页面, 则以 $(1-d)$ 的概率从一个新的页面访问, 但是用户总倾向于以有大量出度/入度的页面为新的起点, 因而把这样的页面赋予较高的权重。这样既可以实现面向主题的查找, 又可以提高系统的响应时间。 (下转第1662页)

- [7] 程显毅,于冬梅.基于BDIAgent的Web搜索引擎的研究[J].江苏大学学报(自然科学版),2004,25(6):545-548.
- [8] Witten I H, Paynter G W, Frank E, et al. KEA: Practical automatic keyphrase extraction[C]. Proceedings of the Fourth ACM Conference on Digital Libraries, 1999.254-255.
- [9] 曹树金,杨涛.自动分类在搜索引擎性能优化中的应用[J].情报

科学,2004,22(2):214-219.

- [10] Vlaho Kostov, Eiichi Naito, Jun Ozawa. Cellular phone ring tone recommendation system based on collaborative filtering method [C].Kobe,Japan: Proceeding IEEE International Symposium on Computational Intelligence in Robotics and Automation, 2003. 16-20.

(上接第1631页)

### 3 算法实验分析

#### 3.1 衡量标准

为了综合考察和衡量算法的效率,我们用召回率和准确率作为主要的评价标准。所谓召回率是指一次搜索结果中集中符合用户要求的数目和用户查询相关的总数目之比;所谓准确率是指以一次搜索集中符合用户要求的数目和该次搜索结果的总数之比。即

$$\text{召回率} = \frac{\text{符合用户要求的数目}}{\text{用户查询相关的总数目}}$$

$$\text{准确率} = \frac{\text{符合用户要求的数目}}{\text{搜索结果的总数目}}$$

#### 3.2 实验方法和结果分析

为了验证上述改进算法,用一台CPU为PIII 800,内存为256 MB,操作系统为Windows 2000 Professional的计算机作为测试平台。首先选择气象信息作为主题;收集气象网站25个,并加入了50个无关的网站组成样本集,其中共有大约15 000个页面;其次选定初始的url集合,这里选www.sina.com.cn, www.163.com, www.sohu.com, www.tom.com和www.etang.com作为url的种子,由于改进算法在计算召回率和准确率时,必须知道有多少页面和主题相关,因此本文仍用向量空间模型以及夹角余弦来计算相关度。实验结果如表1所示,3种不同的算法准确率和召回率的比较如图1所示。

表1 实验结果

算法	准确率	召回率
PageRank	47.3 %	51.4 %
SPageRank	65.5 %	73.7 %
FPageRank	62.6 %	89.4 %

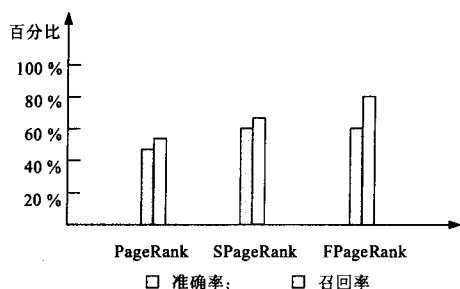


图1 3种不同的算法准确率和召回率的比较

从上述实验结果可以看出,PageRank方法之所以准确率和召回率都最差,是因为它优先采集的都是基于普遍主题重要的页面,并没有向任何一个主题倾斜,因此,它的准确率和召回率都约等于相关主题页面在整个页面集中的比重;SPage-

Rank算法加入了面向主题的思想,相关于某一主题的页面链接到的页面趋向于拥有同一主题;FPageRank算法继承了SPageRank算法面向主题因而采集准确率高的特点,又利用了PageRank较早的发现了相关于主题的重要页面,并且根据较高的出入度提高了查找资源的召回率。从算法代价上分析可以得出,PageRank算法需要计算每个页面的PageRank值,并且对于样本采集来说,为了有效的指导采集,它必须在很短的时间内重新计算PageRank值,以使得它更加准确的反映页面的重要性。而SPageRank和FPageRank都需要引入主题相关度的计算,并且FPageRank算法还要计算每个页面的出入度,这样就增加了复杂性,但是可以通过对具有大量链接的页面建立缓存,从而减少系统的开销。这样,我们认为新算法在的突出显示主题和分析网页链接方面确实优于传统的PageRank算法。

### 4 结束语

算法作为搜索引擎的核心对于提高查询的精度和准确度起着不可忽视的作用。随着Internet的发展,搜索引擎技术的重要性不言而喻。迄今为止,没有任何一种关于页面的排序方法是完美的,因此,搜索引擎的相关排序技术要综合各种排序算法的精华,而提出更好的页面排序方法,更好的为用户服务。

### 参考文献:

- [1] 李盛韬.基于主题的Web信息采集技术研究[D].北京:中国科学院计算机技术研究所,2002.
- [2] 徐宝文,张卫丰.搜索引擎与信息获取技术[M].北京:清华大学出版社,2003.112-116.
- [3] 李晓明,闫宏飞,王继民.搜索引擎-原理、技术和系统[M].北京:科学出版社,2005.183-186.
- [4] 张延红.搜索引擎PageRank算法的改进[J].浙江万里学报,2005,8(4):35-38.
- [5] 曹军. Google的PageRank技术剖析[J].情报杂志,2002,(10):15-18.
- [6] 刘悦,杨志峰,程学旗,等.利用链接分析技术提高搜索引擎查找质量的研究[J].微电子学与计算机,2002,19(5):18-21.
- [7] 宋聚平,王永成,尹中航,等.对网页PageRank算法的改进[J].上海交通大学学报,2003,37(3):397-400.
- [8] Chirita, Daniel, Neidl. Finding husband authorities [Z]. IEEE, 2003.
- [9] 胡亮,许永诚,高文,等.一个高效的层次型搜索引擎模型及应用[J].计算机工程与设计,2005,26(8):2000-2002.
- [10] 耿玉良,陈家琪,王咏梅.中文Web检索中聚类算法的改进[J].计算机工程与设计,2005,26(10):2685-2687.