基于索引属性改进的 OPIC 算法①

田文波, 徐洪珍, 卢群乐

(东华理工大学 信息工程学院, 南昌 330013)

摘 要: 搜索结果的可靠性是影响广大网民使用搜索引擎的一项重要指标. 在开源 OPIC 算法的基础上, 提出了 TS 算法, 该算法通过基于索引属性的改进, 加入网页的创建时间和网页深度两个属性, 使得网页的评分不仅仅局限于网页的链接, 实现了网页评分因素的多元化. 而网页最后的排序分值在原有 OPIC 算法的基础上与网页创建时间成正比, 与网页深度成反比, 有效改善了 OPIC 算法偏重历史网页的缺点, 使搜索结果更加合理. 最后, 对 TS 算法进行效果演化, 经过与传统搜索结果的分析对比, 显示本算法返回的结果具有更高的可靠性.

关键词: 搜索引擎; 索引属性; OPIC 算法; PageRank; 可靠性

Improvement of OPIC Algorithm Based on Indexed Properties

TIAN Wen-Bo, XU Hong-Zhen, LU Qun-Le

(Information Engineering Institute, East China Institute of Technology, Nanchang 330013, China)

Abstract: The reliability of search results is an important index affecting the majority of Internet users use search engines. TS algorithm is proposed based on the open source OPIC algorithm. The algorithm is based on improved indexing attribute and to create two properties includes Webpage creation time and Webpage depth added to TS algorithm to realize the diversification Webpage score. This can improve the disadvantages of the OPIC algorithm emphasis on the history of the page, to make the search results more reasonable. Finally, it is proved by experiments that the result of TS algorithm is more reliable.

Key words: search engine; property index; OPIC algorithm; PageRank; reliability

随着互联网的不断发展, 搜索引擎已经成为人们进入互联网的第一途径, 而网页的排序算法是一个搜索引擎的核心内容, 但是由于商业搜索引擎的封闭性, 他们不可能把所有的算法公之于众, 然而开源搜索引擎由于它的开放性和非营利的特点在各行各业中得到了广泛的运用, 所以本文选择 Apache 软件基金会下的一个完全开源的软件项目 Nutch 来实现本文的算法, Nutch 采用了基于链接的 OPIC 算法, 该算法对存在时间较长的网页, 由于其有较多的链接, 所以评分比较高, 而对于较新的网页评分往往很低, 这就是 OPIC 算法偏重历史网页的缺点, 本文通过加入网页的其他属性来影响网页的排序分值, 避免了传统 OPIC 算法仅仅依靠网页链接来决定排序分值所带来的弊端, 使影响排序结果因素多样化, 从而使排序结果更加可靠.

1 OPIC算法[3]

OPIC 的字面含义是“在线页面重要性计算”, 可以将其看做是一种改进的 PageRank 算法, 在算法开始之前, 每个互联网页面都给予相同的现金(cash), 每当下载某个页面 P 后, P 就将自己拥有的现金平均分配给页面中包含的链接页面, 自己的现金清空[4]. 而对于待爬取 URL 队列中的网页, 则根据其手头拥有的现金金额多少排序, 优先下载现金最充裕的网页, OPIC 从大的框架上与 PageRank 思路基本一致, 区别在于: PageRank 每次需要迭代计算, 而 OPIC 策略不需要迭代过程. 所以计算速度远快与 PageRank, 适合实时计

① 基金项目:国家自然科学基金(61262001);江西省青年科学家培养对象计划(20142BCB23017);江西省发明专利产业化技术示范项目(20143BBM26115);江西省自然科学基金(20114BAB201043);江西省科技支撑计划项目(20112BBE50048);江西省教育厅科技计划项目(GJJ12382)

收稿时间:2014-11-14;收到修改稿时间:2015-01-26

算使用. OPIC 算法伪代码如下所示

```
OPIC:
for each I let C[i]:= 1/n ;
for each I let H[i]:= 0 ;
let G:=0 ;
do forever
begin
    choose some node I ;
    H[i] += C[i];
    For each child j of i,
    Do C += C[i]/out[i];
    G += C[i];
    C[i] := 0 ;
end
```

对于每个页面 OPIC 算法有两个值 *cash* 和 *history*, 分别用 *C[i]*表示页面 *i* 当前的 *cash*, *H[i]*表示在 OPIC 算法从开始到最后被抓取页面 *i* 获得的 *cash* 的总和, *C[i]*的初始值为 1/*n*(*n* 为网页总数), *H[i]*初始值为 0. 网页 *j* 的链接为包含于网页 *i* 中的一个链接, 网页 *j* 的 *cash* 值 *C* 为网页 *i* 的 *cash* 值 *C[i]*除以网页 *i* 的外向链接总数, 为了优化计算效率, 定义变量 *G*, 在每一步中 *G*=|*H*|, 经过不断的在线运算按照 *cash* 值的大小加载到 Nutch 的未抓取队列, 最后按照贪心法的抓取策略, 优先抓取 *cash* 值高的网页.

这项技术的主要缺点和 PageRank 一样是旧的页面等级会比新页面高. 因为即使是非常好的新页面也不会有很多外链.

2 OPIC算法的改进

OPIC 算法只是从页面链接的角度为页面评分, 这种评分由于考虑的因素比较单一,Nutch 索引属性如图 1 所示.

status	markers	parseStatus	moc	baseUrl
	2 (BLOB)	(BLOB)		http://www.fmprc.go
	2 (BLOB)	(BLOB)		http://www.fmprc.go
	2 (BLOB)	(BLOB)		http://www.fmprc.go
	2 (BLOB)	(BLOB)		http://www.fmprc.go
	2 (BLOB)	(BLOB)		http://www.fmprc.go
	2 (BLOB)	(BLOB)		http://www.fmprc.go
	2 (BLOB)	(BLOB)		http://www.fmprc.go
	2 (BLOB)	(BLOB)		http://www.fmprc.go
	2 (BLOB)	(BLOB)		http://www.fmprc.go

图 1 数据库中索引图

图 1 是 Mysql 中的索引, 属性 score 表示基于页面属性出站 outlinks 和入站 inlinks 的数量来通过 OPIC 算法得出的分值, 一般来讲对于搜索引擎来讲他的创建时间越近, 说明它的越接近实际情况, 这一点的重要性对于新闻来讲更加突出, 所以新抓取的页面理应获得更高的 score 分值, 这可以有效地克服 PageRank 算法和 OPIC 算法的缺点, 有效提高新建网页的分值, 另外网页的深度(Webpage depth)也是决定网页可靠性的重要因素, 由于一般比较重要的信息, 站长都希望放在首页或者前几页, 如果放的位置很“偏僻”也可以在一定程度上判断这个网页不重要, 而这种偏僻直接体现在网页深度上, 随着网页层次变多, 网页的质量也在下降, 这种情况理应反映在网页的分值上, 由于新的网页一般放在首页, 所以这在一定程度上也能提高新建网页的排序结果[5].

2.1 时间属性

由于时间在数据库中网页创建时间 *T*, 设网页 *i* 的创建时间设为 T_i, 通过分析在网站爬行时网页的创建时间都是在一个基准时间 T_c 后, 我们将时间 T_i 减去基准时间 T_c 为标准时间间隔 t_i, 并且将最终时间天转化为分钟, 表示公式(1)如下:

$$t_i=(T_i-T_c)*24*60 \quad (1)$$

所以, t_i 越大说明网页的创建时间越短, 反之, 则创建时间越长, 但是由于时间数据分散, 所以采用 Log 函数做标准化处理.

数据标准化也就是统计数据的指数化便于不同单位或量级的指标能够进行比较和加权.

采用 Log 函数转换的方法使时间值转化为(0, 1)的区间内.

$$t_i'=\log(t_i) \ / \ \log(t_{max}) \quad (2)$$

其中, t_i'为标准化后的时间属性值大小, t_i 为网页创建时间的分钟数, t_{max} 是网页 t_i 的最大值

2.2 网页深度属性

在爬虫抓去网页时, 网页的的质量很大程度上取决于网页的深度, 一般来说一级页面上的链接的质量应该好于二级页面上的链接,以此类推, 页面层次越大, 页面上的链接质量就会相应的降低, 链接到的页面的的分值应该相应的降低, 对于网页的深度值我们可以根据网页的链接层次取得, OPIC 算法是用 Java 实现的, 所以计算网页深度属性的 Java 代码如下:

```
private static long webwd(String str,String find) {   long
count = 0;
    int len = find.length();
```

```
    int index = 0;
for(int i=0; i<str.length();i=len+index) {
    if((index = str.indexOf(find, i)) > -1)
        count ++;
          else
          break; }
          return count; }
```

函数参数 str 为网页 i 的网址 baseurl, 参数 find 为网址层次符号'\', 通过执行上述函数可以得到网页深度的初始值 d_i, d_i 越大说明说明网页层次越大, 网页的质量就会降低, 由此可以看出网页的质量和网页的深度成反比.

2.3 TS 算法

针对上文中提到的 OPIC 算法的缺陷, 我们在新算法中加入时间属性和网页深度属性, 改进后的算法应该在基于链接排序的基础上加入上文中提到的两个属性对结果的影响,在下文中我们称改进后的算法为 TS 算法.

为两个属性值根据他们的重要程度, 分别为赋予两个属性值不同的权重值 w_t w_d. 由于网页深度 d_i 与 TS_i 负相关, 而网页创建时间 t_i'与可靠性正相关 我们用 TS_i(trusted score)表示:

$$TS_i=OS_i- w_d*d_i+ w_t* t_i' \quad (3)$$

其中, TS_i 为网页 i 的 TS 算法的网页得分, OS_i 是网页 i 的 OPIC 的网页得分, w_d 是对网页深度因素的权重, d_i 是网页深度值, w_t 是时间属性的权重, t_i'是时间属性值, 在上式中 d_i 和 t_i'不是具体的创建时间和网页深度, 而是经过数据标准化, 使它可以匹配 OS_i 的大小.

权重 w_t, w_d 需要通过实验去决定他们的大小, 必须在合理分配各个属性在可靠性中的比重的前提下去确定各个系数的大小, 以至于能让他们比较合理地影响 TS_i 的大小, 进而影响搜索排序的结果, 通过实验来验证改进后的搜索算法[6].

3 实验结果

我们以抓去外交部的新闻网站(http://www. fmprc. gov.cn/mfa_chn/), 以搜索"和平发展"为例, 搜索结果如图 2 所示.

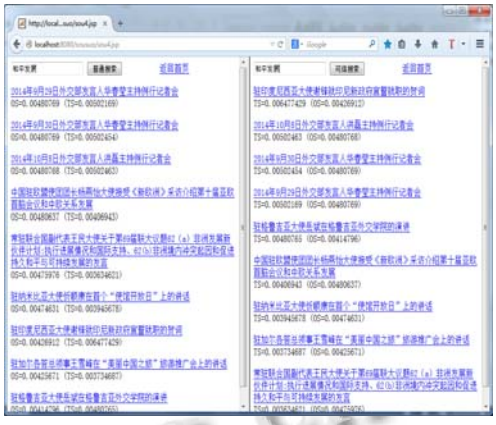

图 2 搜索结果页面

4 实验分析

根据搜索结果分析两种排序策略的优越性, 研究网页创建时间和网页深度对排序结果的影响.

4.1 OS 和 TS 数据分布

我们先列出两种搜索算法搜索结果的变化趋势图, 我们通过分析初始得分 OS 值和可靠性 TS 值数据, 结果如图 3 所示.

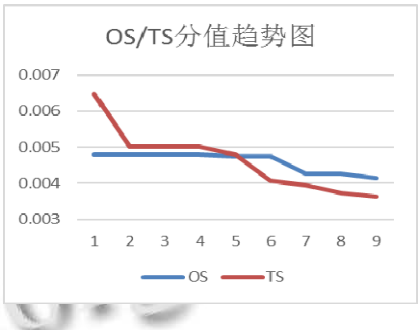

图 3 OS 值和 TS 值分布图

图中横坐标为搜索结果的排序值, 纵坐标为相应分值, 蓝色折线为 OPIC 算法的 OS 值, 红色折线为 TS 算法 TS 值, 由它们的分布来看 TS 算法搜索结果更具有区分性, 而 OPIC 算法搜索结果的差异性较小, 在此看来 TS 算法搜索的结果各网页之间有更强的区分度, 而 OS 的分值相似的网页分值相同的情况较多.

4.2 两种算法分值随着时间 t 的分布

我们再分别从两种算法搜索结果网页创建时间 t 的数据分布情况, 我们以他们的搜索结果排序不变, 分别查看它们的创建时间, 经过数据的标准化后, 分布结果如图 4 所示.

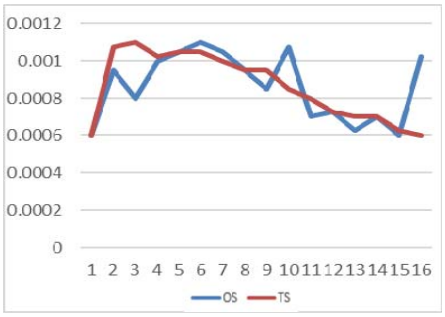

图 4 两种搜索返回结果网页的 *t* 分布图

图 4 中横坐标为搜索结果的排序值, 纵坐标网页时间 *t*, 由图可知 TS 算法的时间分布基本按照由大到小的顺序分布, 排序值越小, *t* 值越大, 说明网页的创建时间距离当前时间越短, 而仅仅依靠链接的 OPIC 算法的搜索结果在时间 *t* 上是无序的, 所以 TS 算法的排序结果从时间上看更加具有时效性[7].

4.3 两种算法分值随着网页深度的分布

我们再分别从两种算法搜索结果网页深度 d 的数据分布情况, 我们以他们的搜索结果排序不变, 分别查看它们的网页深度, 经过数据的标准化后, 分布结果如图 5 所示.

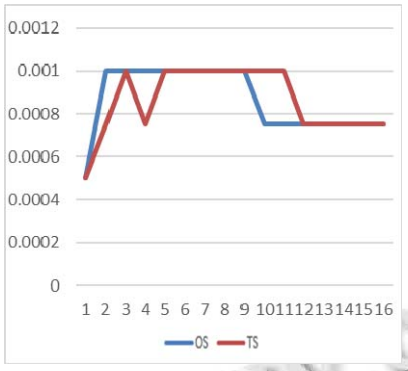

图 5 两种搜索返回结果网页 *d* 分布图

图 5 中横坐标为搜索结果的排序值, 纵坐标网页深度 *d* 值由图可知, TS 算法在排序值较小时, 相应的网页深度小于 OPIC 算法的网页深度, 而在排序值较大时, TS 算法的网页深度大于 OPIC 算法的网页深度. 这就说明相比较 OPIC 算法, TS 算法的结果网页深度较小的排序靠前, 网页深度较大的结果比较靠后, 所以从这一方面来看 TS 算法比原始 OPIC 算法的搜索结果更加可靠[8].

综上分析, TS 算法的搜索结果比原有的 OPIC 算法的搜索结果的质量在时效性和网页深度上有明显改善, 比较合理地克服原来 OPIC 算法比较偏重旧网页而忽视新网页的缺点[9].

5 结语

针对 OPIC 算法在网页评分上存在的缺陷, 本文利用索引属性改进算法, 进而使得搜索结果更多考虑到时间的有效性和网页深度对搜索结果的影响, 而且最后的搜索结果在区分性上也得到明显改善, 这尤其对新闻搜索来说至关重要, 通过结果的验证, 基本完成了本文初期的设想, 并且经过实验分析得到了可靠性比较高的搜索结果.

参考文献

1 张贤,周娅.基于 Lucene 网页排序算法的改进.计算机系统应用,2009,18(2):155–158.
2 杨劲松.搜索引擎 PageRank 算法的改进.计算机工程,2009,35(22):35–37.
3 Abiteboul S, Preda M. Adaptive On-Line Page Importance Computation. http://www2003.org/cdrom/papers/refereed/p007/p7-abiteboul.html#foot170.
4 李耀芳,张涛.基于 Nutch 的垂直搜索引擎系统.计算机系统应用,2011,20(9):193–196.
5 李亚楠,王斌,李锦涛.搜索引擎查询推荐技术综述.中文信息学报,2010,24(6):75–84.
6 于天恩.搜索引擎开发权威经典.北京:中国铁道出版社,2008.
7 李璐旸.面向网络文本的信息可信度研究[硕士学位论文].哈尔滨:哈尔滨工业大学,2011.
8 袁津生,李群,蔡岳.搜索引擎原理与实践.北京:北京邮电大学出版社,2008.
9 赵洁,肖南峰,钟军锐.Web 使用挖掘在信息管理中的应用.计算机工程,2009,35(24):33–35.