

# 原创优先的搜索引擎排序算法

郝金隆, 王成良

(重庆大学软件学院, 重庆 400044)

**摘要:** 现有的搜索引擎排序算法大多根据网页之间的链接关系进行排序, 没有考虑原创和转载文章之间的优先次序。该文提出一种适用于专业搜索引擎的新型排序算法, 在排序时优先考虑原创文章, 使原创文章的搜索排名得以提高, 有助于新兴原创技术网站的发展, 提高网络竞争的公平性。

**关键词:** 搜索引擎; 排序算法; 原创优先; 互联网

## Ranking Algorithms of Originals Promotion for Search Engine

HAO Jin-long, WANG Cheng-liang

(College of Software, Chongqing University, Chongqing 400044)

**【Abstract】** The existing algorithm of search engines merely works on interlinks among Web pages while putting aside ordering of original and copied articles, and rising original technical websites are put into great disadvantages in competition with those earlier websites. In order to develop these rising technical websites and enhance fair competition on the Internet, this paper presents a new ranking algorithm which gives priority to those original articles and promotes their search rankings and therefore can be employed by professional search engines.

**【Key words】** search engine; ranking algorithm; originals promotion; Internet

### 1 概述

随着网络规模的爆炸性增长, 网上的信息正以几何级的速度增加, 方便快捷的搜索引擎成为人们查找信息的重要工具。据中国网络信息中心(CNNIC)调查, 在互联网用户经常使用的网络服务中, 搜索引擎占了 51.5%<sup>[1]</sup>。由此可见, 搜索引擎已成为互联网应用的重要组成部分。而搜索引擎的核心技术——排序算法也变得极为重要, 一个合理的排序算法可为互联网营造一个公平的竞争环境。

现有的搜索引擎网页排序技术主要有PageRank 算法<sup>[2-3]</sup>和HITS(Hypertext-Induced Topic Search)算法<sup>[4-5]</sup>。两者都是利用网页和超链接组成的有向图, 根据网页之间的相互链接关系进行递归运算来确定网页的排名<sup>[6]</sup>。利用现有的算法搜索同一篇文章, 会有转载站点排在首发站点之前的可能, 这无疑是不公平的。如果能够优先考虑原创文章, 将使搜索引擎排序算法变得更为合理。

### 2 现有的搜索引擎排序算法

#### 2.1 PageRank 算法

PageRank 算法的基本思想来源于“从许多优质网页链接过来的网页, 必定还是优质网页”这一回归关系。网页 A 链接到网页 B, 就认为网页 A 为网页 B 投了一票。

PageRank 最初的算法描述如下: 网页 A 的 PageRank 值为  $PR(A) = (1-d) + d(PR(T_1)/PC(T_1) + \dots + PR(T_n)/PC(T_n))$ , 其中,  $d$  为阻尼系数, 且  $0 < d < 1$ ;  $T_1, T_2, \dots, T_n$  表示链接到 A 的所有  $n$  个网页;  $PR(T_i)$  表示  $T_i$  的 PageRank 值;  $PC(T_i)$  表示  $T_i$  页面上的总链接数。而用户点击页面上链接的概率, 则由页面上的链接数确定, 即式(1)中的  $PR(T_i)/PC(T_i)$ , 阻尼系数  $d$  的引入是为了降低这一概率, 因为用户不可能无限地点击链接, 常常会随机转到其他页面。

由此可见, 一个页面的 PageRank 值取决于链接到此页面

的 PageRank 值以及那些页面各自的链接数。对各个页面赋初值后, 通过递归计算可得到页面的 PageRank 值。

#### 2.2 HITS 算法

HITS 算法是在 20 世纪 90 年代末提出的一种链接分析算法, 它将网页的质量评估结果反映在对每个网页给出的 2 个评价数值——内容权威度(authority)和链接权威度(hub)上。内容权威度与网页自身提供的内容质量相关, 被越多网页所引用的网页, 其内容权威度越高; 相对应地, 链接权威度与网页提供的超链接的质量相关, 引用越多内容质量高的网页, 其链接权威度越高。

HITS 算法的具体实现是一个“迭代-收敛”过程, 由于算法流程比较复杂, 因此本文不作详细描述。需要说明的是, 网页 A 的链接权威度由其链向的网页的内容权威度决定, 而内容权威度由链向的网页的链接权威度决定。

#### 2.3 现有算法的不足之处

以上 2 种排序算法都是通过分析网页链接来进行排序的, 但它们都对新兴原创网站极为不利, 即没有考虑内容相同的原创文章和转载文章之间的排序问题。例如, 有人在工作和学习过程中积累了很多实际经验, 为了与大家分享, 专门建设了一个网站将经验公布于众。但当网站的文章大部分被其他网站转载时, 根据现有的算法, 此网站的排名可能很低, 这很不公平。为此, 本文提出了原创优先的搜索引擎排序算法。

### 3 原创优先的排序算法

#### 3.1 网上专业文章的分类

网上发布的专业技术文章可分为 4 类: (1) 标明是原创的;

**作者简介:** 郝金隆(1983 - ), 男, 硕士研究生, 主研方向: 搜索引擎, 软件工程; 王成良, 教授、博士

**收稿日期:** 2007-10-30 **E-mail:** hjl\_100@sohu.com

(2)标明是转载的,但没有指定首发地址;(3)标明是转载,且说明了首发地址;(4)既没有标明是原创的,也没有标明是转载的。对于第(1)类,一般可以确定是原创的,但也不排除故意将转载文章标为原创的情况;对于第(2)类、第(3)类,可以确定不是原创,其中,第(3)类中指出的转载地址一般是首发地址,但也不可能出错;第(4)类则最难确定。

### 3.2 技术网站的3个属性

对每一个技术网站都设定以下3个参数: $P_1(0 < P_1 < 1)$ ,表示第(1)类文章的原创可信度,即该网站上发表的文章如果标明是原创,则确实是原创文章的可信度; $P_2(0 < P_2 < 1)$ ,表示第(3)类文章指明的转载地址的可信度,即该网站上发表的文章标明是转载且指明转载地址时,转载地址即为首发地址的可信度; $P_3(0 < P_3 < 1)$ ,表示第(4)类文章的原创可信度,即此网站中没有指明是原创还是转载的文章的原创可信度。

这3个属性并非指绝对的可能性,而是一个相对的概念。例如,网站A、网站B的 $P_1$ 属性分别为0.3和0.5,并不代表A、B中标明原创文章确是原创的可能性分别为30%和50%,只能说明网站A中标明是原创文章确是原创的可能性小于网站B,如果没有其他因素影响,对于同一篇文章,在A中标明为原创时的原创可能性是B中的3/5。

### 3.3 文章原创可能性的判定方法

假定存在 $n$ 个站点 $W_1, W_2, \dots, W_n$ ,某文章被这 $n$ 个网站引用,各网站的3个属性依次为 $P_{1h}, P_{2h}, P_{3h}(h=1, 2, \dots, n)$ 。

对于声明是原创文章,如站点为 $W_i$ ,则取:

$$K_i = P_{3i} + (1 - P_{3i}) \cdot P_{1i}$$

对于既没有声明是原创,又没有声明是转载文章的站点 $W_j$ ,则取:

$$K_j = P_{3j}$$

对于声明是转载文章的站点 $W_k$ ,则取:

$$K_k = 0$$

如果此文章在 $W_k$ 指定了转载站点 $W_l$ ,但在 $W_l$ 的相同文章没有声明为转载,则令

$$K_l = K_l + (1 - K_l) \cdot P_{2k}$$

最后作正规化处理:

$$R_i = K_i / \sum_{j=1}^n K_j$$

即可得到此文章在各个站点首发的可能性。

### 3.4 后续处理

得到此文章在各个站点的原创可能性后,还需要据此对这 $n$ 个站点的3个属性进行修正。

首先构造函数:

$$f(n) = \begin{cases} 2^{n-1} & 0 < n < 1 \\ 2^{(1-n)} & n = 1 \end{cases}$$

显然, $f(n)$ 单调递增,且在 $[0,1]$ 上的取值范围是 $[0.5,1]$ ,在 $[1,+\infty]$ 上的取值范围是 $[1,2)$ 。

对于声明是原创的站点 $W_i$ ,根据计算所得原创可能性大小,增加或减小此站点标明是原创的可信度 $P_{1i}$ :取 $T_i = f(R_i \cdot \sqrt[n]{n})$ ,则

$$P_{1i} = \begin{cases} P_{1i} \cdot (1 - c_2 + T_i \cdot c_2) & T_i < 1 \\ P_{1i} + (1 - P_{1i}) \cdot c_2 \cdot (T_i - 1) & T_i = 1 \end{cases}$$

当 $R_i$ 大于 $\frac{1}{\sqrt[n]{n}}$ 时, $T_i > 1$ ,则 $P_{1i}$ 增加,反之, $T_i < 1$ , $P_{1i}$ 减小。

对于既没有声明是原创,又没有声明是转载的站点 $W_j$ ,

根据计算所得的此文章的原创可能性大小,增加或减小此站点无说明的原创可信度 $P_{3j}$ :取 $T_j = f(R_j \cdot \sqrt[n]{n})$ ,则

$$P_{3j} = \begin{cases} P_{3j} \cdot (1 - c_2 + T_j \cdot c_2) & T_j < 1 \\ P_{3j} + (1 - P_{3j}) \cdot c_2 \cdot (T_j - 1) & T_j = 1 \end{cases}$$

当 $R_j > \frac{1}{\sqrt[n]{n}}$ 时, $T_j > 1$ ,则 $P_{3j}$ 增加,反之, $T_j < 1$ , $P_{3j}$ 减小。

对于声明是转载的站点 $W_k$ ,增加此站点标明是原创的可信度 $P_{1k}$ 和无说明的原创可信度 $P_{3k}$ :

$$P_{3k} = P_{3k} + (1 - P_{3k}) \cdot c_3$$

如果声明了转载地址 $W_l$ ,则根据转载地址的原创可能性,增加或减小此站点的转载地址可信度 $P_{2k}$ :取 $T_l = f(R_l \cdot \sqrt[n]{n})$ ,则

$$P_{2k} = \begin{cases} P_{2k} \cdot (1 - c_2 + T_l \cdot c_2) & T_l < 1 \\ P_{2k} + (1 - P_{2k}) \cdot c_2 \cdot (T_l - 1) & T_l = 1 \end{cases}$$

当 $R_l$ 大于 $\frac{1}{\sqrt[n]{n}}$ 时, $T_l > 1$ ,则 $P_{2k}$ 增加,反之, $T_l < 1$ , $P_{2k}$ 减小。

在后续处理的公式中, $c_1, c_2, c_3$ 都是优化因子,在实际使用时需要对其调整以优化计算结果。其中, $c_1$ 是 $n$ 的开方次数,取值范围为 $(1, +\infty)$ ,用于确定何时增加属性,何时减少属性。如 $n=100$ ,即文章被100个站点引用, $c_1$ 取值为2,表示只有当计算所得的原创可能性大于0.1时,才增加此站点标明原创的可信度,反之减少。而 $c_2, c_3$ 是3个属性调整的比例,取值范围为 $(0, 1)$ ,用于确定每次的调整范围。

## 4 实验分析

### 4.1 实验方法

首先随机选取100篇计算机软件类的文章,然后通过Google搜索每一篇文章,各取其前50个站点,再去重和非专业计算机类的站点,记录最后所得的站点。之后以开源搜索引擎nutch<sup>[7]</sup>为基础建立搜索引擎,配置nutch,使之只抓取这些站点的网页,然后建立索引。网页索引建立完成后,利用原创优先的排序算法对这些站点中所有内容相同的文章进行迭代运算,再搜索这100篇文章,对使用算法前后的排序结果进行比较。在迭代时,每个站点 $P_1, P_2$ 的初始值设为0.5, $P_3$ 的初始值设为0.1, $c_1$ 为2,并不断调整 $c_2, c_3$ 的值,以得到较好的实验结果。

### 4.2 实验结果

实验共得到671个站点。当 $c_2=0.09, c_3=0.05$ 时,得到的统计结果如表1所示。

表1 实验统计结果

排名	使用原创优先算法前	使用原创优先算法后
首发站点排名第1	32	58
首发站点排名前3	63	73
首发站点排名前5	84	93

### 4.3 结果分析

上述实验结果表明,使用原创优先的搜索引擎排序算法后,原创文章的搜索排名有了很大的提高,但仍有27篇文章在首发站点的排名没有进入前3名。其中,11篇是由于根据原有算法使其排名太低造成的;7篇是因为其首发站点根本不在这671个站点之中;另外9篇可能是由于在调整优化因子时只对 $c_2, c_3$ 进行了调整,而没有调整 $c_1$ ,得出的结果不是本算法的最优结果,因此,还需要进一步优化算法,才能得到更好的实验结果。

(下转第92页)