

Google 的 PageRank 技术分析

王玉珍*

摘 要：讨论了 PageRank 的计算方法, 并对 PageRank 值的计算公式进行了修正, 最后给出了改进的建议。

关键词：PageRank 搜索技术 计算方法

中图分类号 TP393 文献标识码 B 文章编号: 1002-2422(2007) 05-0013-03

Analysis of Technology about Google's PageRank

Wang Yuzhen

Abstract: The paper deeply analyzes the calculation method of PageRank, then mends the calculation formula of PageRank, at last the improving advice is put forward.

Keyword: PageRank Technology of Seeking Calculation Method

1 PageRank 的定义

定义 1(PageRank) 令 u 为一个网页, $N(v)$ 表示从网页 v 向外的链接数目, $B(u)$ 表示链接到网页 u 的网页集合, $R(u)$ 表示网页 u 的

PageRank 值, C 为规范化因子, 作用是保证所有网页的 PageRank 总和为常量 (例如为保证总的 PageRank 值为 1, 可以通过网页 PageRank 总和的倒数求得)。

$$R_{(u)} = C * \sum_{B(u)} R_{(v)} / N_{(v)} \quad (\text{定义 1})$$

必须注意的是定义 1 有一个假设前提, 即所有的网页形成一个牢固的链接图 (即每个网页能从其他网页

```
<body>
<APPLET CODE = Applet1.class WIDTH = 150 HEIGHT = 50>
</APPLET>
</body>
</html>
```

转换后的 HTML 如下:

```
<html>
<head>
<title>my first java Applet</title>
</head>
<body>
<!-- CONVERTED_APPLET -->
<!-- HTML CONVERTER -->
<OBJECT
classid= clsid:CAFEEFAC-0014-0000-0001-ABCDEFEDCBA
WIDTH = 150 HEIGHT = 50
codebase= http://java.sun.com/products/plugin/autodl/install-1_4_0_01-
win.cab#Version=1,4,0,10 >
<PARAM NAME = CODE VALUE = Applet1.class >
<PARAM NAME= type VALUE= application/x-java-applet;jpi-version=1.4.0_01 >
<PARAM NAME= scriptable VALUE= false >
<COMMENT>
<EMBED
```

```
type= application/x-java-applet;jpi-version=1.4.0_01
CODE = Applet1.class
WIDTH = 150
HEIGHT = 50
scriptable=false
pluginspage= http://java.sun.com/products/plugin/index.html#download >
<NOEMBED>
<NOEMBED>
<EMBED>
<COMMENT>
</OBJECT>
<!--
<APPLET CODE = Applet1.class WIDTH = 150 HEIGHT = 50>
</APPLET>
-->
<!-- END_CONVERTED_APPLET -->
</body>
</html>
```

参 考 文 献

[1](美)Hendricks M 著. Java Web 服务编程指南. 北京: 电子工业出版社, 2002-02.

[2]傅雯彬, 蔡予书著. Java Script 动态网页设计实务. 北京: 中国铁道出版社, 2001.

[3](美)Darby C 著. Java 网络编程指南. 北京: 电子工业出版社, 2002-08.

* 王玉珍 兰州商学院信息工程学院副教授(730020), 从事管理信息系统与计算机应用研究 收稿日期 2007-03-15

通过超链接达到)。从定义 1 可以看出,网页的 PageRank 是一个由网络的超链接结构所产生的一个网页重要性等级值,所有的网页的 PageRank 值都可以根据其他网页的 PageRank 值和链接的数量来计算得到,即所有链接到它的网页的 PageRank 值除以各自向外的链接数的商进行求和。

2 PageRank 的计算

2.1 计算步骤

前面给出的定义 1 本身是一个 PageRank 的计算公式,利用这个公式,可以计算网页集合中所有网页的 PageRank 值。计算步骤如下:(1)假设 S 为整个网页的总和;(2)给每个网页的 PageRank 都赋以 $1/S$,因为所有的网页的 PageRank 值开始是未知的,所以进行平均的分配;(3)根据定义 1 进行计算。然后对得到的值再次利用定义 1 计算,反复地计算,直到计算得到的 PageRank 值收敛于一个相对固定的数()。也就是说,根据超链接结构计算出的所有网页的重要性等级趋于稳定,这时停止计算。

算法 1 描述如下:

```

forall u s: R(u)0=1/|s|
while( |R(u)j-R(u)j-1| > )
for each u s:

```

$$R'_{(u)j} = R'_{(u)j-1} + \sum_{v \in B(u)} R_{(v)j-1} / N_{(v)}$$

$$C = 1 / \sum_s R'_{(u)j}$$

for each u s:

$$R_{(u)j} = C * R'_{(u)j}$$

2.2 修正的 PageRank 的定义和计算

前面所定义的 PageRank 有一个假设前提,就是所有的网页形成一个牢固的链接图。但是实际的网络超链接环境没有这么理想化,存在着两个主要问题:等级沉没(rank sink)和等级泄漏(rank leak)。整个网页图中的一组紧密链接的网页如果没有外出的链接就产生等级沉没。一个独立的网页如果没有外出的链接就产生等级泄漏。rank sink 中,不在 sink 中的网页的等级值变成了 0,即意味着不能判断出这些网页的重要性,如图 1 的网

页 5 到网页 1 的链接被删除,那么使得网页 4、5 产生了沉没(sink)。一个随机的访问者访问的时候将在 4、5 间陷入其中,没有出去的链接可寻,1、2 和 3 的 PageRank 值都成了 0,网页 4 和 5 的 PageRank 值都变成 0.5,而 rank leak 将丢失其它所有的等级。

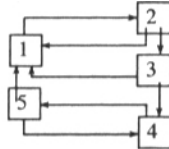


图 1 网页通过超链接的相互关系

针对这些问题,对于 rank leak 可以删除所有的 leak 节点,这样 leak 节点将没有了 PageRank;另外一种方法是假设 leak 节点对所有的指向它的链接有相应的返回链接。为了解决 sink,引进了一个 rank source(等级源)来不断地补充每个网页的 PageRank 值,以使得 PageRank 的分配不完全依赖于链接。修正的 PageRank 定义为:

$$R_{(u)} = d * \sum_{v \in B(u)} R_{(v)} / N_{(v)} + C * E_{(u)} \quad (\text{定义 2})$$

计算中可以 $\forall u s: E_{(u)} = 1/|s| \dots$ ($0 < d < 1$)来实现计算。

S.Brin 和 L.Page 更加直观的定义为:

$$R_{(u)} = d * \sum_{v \in B(u)} R_{(v)} / N_{(v)} + (1-d) \quad (\text{定义 3})$$

其中定义一个衰减系数 d ($0 < d < 1$)。在这个修正的定义中,网页的 PageRank 值仅 d 部分在它所链接到网页中分配,剩下的部分用在整个网络的所有网页中分配, d 的值通常为 0.85 左右。这样一个网页的 PageRank 值的 85% 分配到所链接的网页,另外的 15% 被分配到全部的网页中去。可以看出简单定义其实是 $d=1$ 的时候的特例。Arvind Arasu 等将上述定义更加准确地表达为:

$$R_{(u)} = d * \sum_{v \in B(u)} R_{(v)} / N_{(v)} + (1-d) / m$$

(定义 4)

定义 4 中 m 表示网页节点的总数,作用在于 $1-d$ 的 PageRank 在所有网页中分配。针对实际应用,可以将定义 4 规范化修正如下:

$$R_{(u)} = C * d * \sum_{v \in B(u)} R_{(v)} / N_{(v)} + (1-d) / m \quad (\text{定义 5})$$

算法 2 描述如下:

```

forall u s: R(u)0=1/|s|
while( |R(u)j-R(u)j-1| > )
for each u s:

```

$$R'_{(u)j} = d * R'_{(u)j-1} + \sum_{v \in B(u)} R_{(v)j-1} / N_{(v)} + (1-d) / |s|$$

$$C = 1 / \sum_s R'_{(u)j}$$

for each u s:

$$R_{(u)j} = C * R'_{(u)j}$$

利用算法 2 可以解决图 1 中的等级沉没(rank sink)问题,上面的算法用程序实现后在计算机上运行 60 次,得到的收敛值为(其中 $d=0.85$, $=0.0001$) $P(1)=0.13138$, $P(2)=0.14330$, $P(3)=0.09183$, $P(4)=0.32736$, $P(5)=0.30613$ 。

3 PageRank 的改进

在用户使用搜索引擎时,其每次点击 URL 的动作都被引擎服务器记录下来,用户 IP(及其它用户信息)与相应的页面 ID 一起保存,这样,在下次计算 PageRank 时,就得到了一个新的向量(点击向量),其每一个分量是对应页面的点击次数与所有页面点击次数之和的比,记为 b ,称为全局个性化向量,因为它代表了所有用户对页面选择的合力。

在迭代过程中 b 作为页面重要性的一部分,迭代公式修正为:

$$R_{i+1} = cM * R_i + (1-c) * (1/N)_{N \times 1} + b$$

其中 M 表示矩阵,即如果存在网页 j 到网页 i 的链接,则置矩阵中元素 m_{ij} 的值为 $1/N_j$,否则置 0。

把上式看作对方程 $x_{i+1} = Hx_i + b$ 的求解,其收敛性不变;但是如果 b 的某些分量值较大时,所需的迭代次数将增加。因为可能存在如下情况:

$$b_1 = b_2 = b_k = 0, m_{1j} = 0, m_{1k} = 0, m_{1i} = 0$$

此时即出现了 Rank Sink 问题。

集体个性化向量的来源有两部分:一是服务器保存的用户点击信息;二是对网络上的新页面,其点击次数为 0,此时我们给这样的页面一个随机

全自动包装码垛生产线 PLC 控制系统设计*

蒋继红** 卢志珍 岳满林*** 高安邦****

摘 要：介绍了全自动包装码垛生产线工艺流程，设计了以日本欧姆龙公司 C200H 型 PLC 为基础的全自动包装码垛生产线控制系统，重点介绍了控制系统的硬件配置、输入输出分配和软件设计。

关键词：全自动包装码垛生产线 PLC 控制系统 硬件 软件

中图分类号 TP273*.5 文献标识码 B 文章编号：1002-2422(2007)05-0015-03

Control System Design of Automatic Bagging and Palletizing Line Based on PLC

Jiang Jihong Lu Zhizhen Yue Manlin Gao Anbang

Abstract: The technological process of automatic bagging and palletizing line is introduced, and a set of automatic bagging and palletizing line control system based on OMRON C200H PLC is designed. Then the emphasis is given to the hardware disposition of control system, I/O allocation and software system.

Keyword: Automatic Bagging and Palletizing Line PLC Control System Hardware Software

石化、粮食、医药等行业的散装物料的包装码垛离不开自动化包装机械。国内新一代包装码垛生产线的设计制造将打破国外产品在我国自动包装行业的垄断地位，并实现用具有自主知识产权的高技术产品武装我国支柱产业的目标。全自动包装码垛生产线是集机、电、仪于一体的高技术产品，它主要应用于化工、粮食、食品及医药等行业中的粉、粒、块状物料（如塑料、化肥、合成橡胶、粮食等）的全自动包装，即对包装过程中的称重、供货、装袋、折边、封袋、倒袋整形、批号打印、检测、转位编组、码垛、托盘和垛盘的输送等作业全部实现自动化。以 PLC 为基础的全自动包装码垛生产线，控制系统简单、便于维护、适应性强，自动化程度高，节约人力，可极大提高生产效率。

1 全自动包装码垛生产线工艺流程

全自动包装码垛生产线的机械系统主要包括全自动称重单元、包装单元、输送检测单元、码垛单元。其主要工艺流程如下：物料自储料斗进入包装秤的给料装置，通过粗、细

给料，实现粗、细两级加料。当秤斗中的物料重量达到最终设定值时，称重终端发出停止加料信号，待空中的飞料全部落入秤斗后此次称重循环结束，此时电子包装秤等待装袋机的投料信号。当自动装袋机完成上袋后，发出讯号，使称重箱打开卸料翻门，向包装袋内投料，卸料后称重箱关闭翻门，装袋机张开夹袋器，包装袋通过夹口整形机和立袋输送机进入自动折边机，包装袋经折边后，进入缝口机，当设在缝口机旁边的光电开关检测到包装袋后，缝切机开始工作，缝合包装袋，当包装袋离开缝切机后，缝切机停止，并自动切断缝合线。包装袋经过倒袋整形机进入金属检测机及重量复检机，若检测不合格，在包装袋通过自动捡选机时将被剔除，而合格的包装袋则顺利通过自动捡选机，再经喷墨打印机、过渡输送机、缓停机等设备，将包装袋输送到码垛单元，由转位机根据码垛工艺要求将料袋依次按 2 袋直-3 袋横和 3 袋横-2 袋直 循环做转位处理。包装袋便以 2 袋直或 3 袋横的形式进入编组机，最后由码垛机将包装袋堆码

PageRank 的结果。最好的结果当然是在所有用户都正常使用 的情况下。

更多新的技术的应用也将会给搜索引擎技术带来新的活力，搜索引擎一定会有一个更美好的未来。

4 结 束 语

搜索引擎的发展与完善还有很长的路，研究 PageRank 是为了在此基础上提出更有创意的排序算法，同时，

参 考 文 献

[1] 曹军. Google 的 PageRank 技术剖析[J]. 西安: 情报杂志, 2002, (10.)

* 中国高等教育学会“十一五”规划教研重大课题(批准号:06AIP0090046);江苏省教育科学“十一五”规划 2006 年度课题(立项号:11513037);山东省教育科学“十一五”规划 2006 年度课题(立项号:11GG41);黑龙江省教育厅 2006 年科学技术研究计划项目(项目编号:11513037)。

** 蒋继红 淮安信息职业技术学院机电系讲师(东南大学工程硕士研究生)(223200),研究方向为机电一体化技术开发应用。

*** 岳满林 哈尔滨理工大学硕士研究生(150080),研究方向为机电一体化技术开发应用。

**** 高安邦 淮安信息职业技术学院特聘教授(原哈尔滨理工大学教授、硕士学位研究生导师)(223003,150080),研究方向:机电一体化技术开发应用 收稿日期 2007-01-06