【基础理论与应用研究】　　doi：10.11809/scbgxb2015.01.039

基于 Timed-PageRank 的聚焦爬虫优化研究

李　东[1]，王虎强[2]

（装甲兵工程学院 信息工程系，北京　100072）

摘要：传统的基于 PageRank 算法的网络爬虫在抓取网页时由于只考虑了网页的超链接，势必会使爬虫结果覆盖面广、冗余度高，聚焦爬虫由于其可以有效地过滤与主题无关的链接，只保留有用的链接并将其加入到待抓取的 URL 队列，因此能够有效地降低爬虫冗余；在分析 PageRank 算法的基础上，将网页的时间维数和页面的内容相关度融于其中，提出了基于 Timed-PageRank 的改进算法，并将该算法应用于聚焦爬虫过程中，实践证明该算法能够有效地提高爬虫页面相关度及检索结果的查全率和查准率。

关键词：传统网络爬虫；PageRank 算法；聚焦爬虫；Timed-PageRank 改进算法

本文引用格式：李东，王虎强. 基于 Timed-PageRank 的聚焦爬虫优化研究[J]. 四川兵工学报，2015(1)：141－144.

Citation format：LI Dong，WANG Hu-qiang. Optimization Research on Focused Crawler Based on Improved Timed-PageRank Algorithm[J]. Journal of Sichuan Ordnance，2015(1)：141－144.

中图分类号：TP391.3　　**文献标识码**：A　　**文章编号**：1006－0707(2015)01－0141－04

Optimization Research on Focused Crawler Based on Improved Timed-PageRank Algorithm

LI Dong[1]，WANG Hu-qiang[2]

（Department of Information Engineering，Academy of Armored Forces Engineering，Beijing 100072，China）

Abstract：Traditional web crawler which based on PageRank algorithm only takes the hyperlinks into consideration when scraping the web pages and it is bound to make the results in a wide coverage and high redundancy. Focus crawlers can effectively filter off-topic links，and only save useful links and add them to the URL queue，so it can reduce redundancy effectively. By analyzing the PageRank algorithm，the improved algorithm was proposed based on the Timed-PageRank，and after that we added the time dimension and page content relevance into it. Then we applied the algorithm into focus crawler. Practice proves that the algorithm can effectively improve the relevance，the recall rate and precision of the search results.

Key words：traditional web crawler；PageRank Algorithm；focus crawlers；improved Timed-PageRank Algorithm

随着互联网的普及与发展，信息增长速度迎来了历史高峰，大量信息涌入人们的日常生活，而对于人们所需的信息的检索就有了相当了困难，搜索的结果要么是杂乱无章，要么不能达到令人满意的效果。因此建立特定领域专业化网络检索系统[1]是一个行之有效的措施。而聚焦爬虫就是为了满足这一要求诞生的，它以其爬虫范围紧贴主题，检索效果既全又准的优势赢得了开发人员的青睐。

1　聚焦爬虫

作为搜索引擎的核心组成部分，传统的网络爬虫[2]的主要功能是从网上下载与主题相关的网页，以供后续开发搜索引擎时建立索引并检索的需要。其主要思想是从预先设定好的起始网页开始爬取，将起始网页中的超链接添加到 URL 队列中，反复迭代，根据一定的搜索策略，不断地从当前页面上提取新的 URL 超链接放入队列，直到爬虫结束或是满足一定的爬虫条件而终止，其爬虫流程图如图 1 所示。

因此传统网络爬虫具有一定的局限性：

一是冗余大。传统网络爬虫没有对待爬取的网页进行页面相关度匹配，只是将存在于起始页面的链接加载到 URL 队列里，反复迭代，因此将许多相关度很低甚至是不相关的网页爬取下来，造成了大量的冗余信息。

收稿日期：2014－06－27

作者简介：李东(1962—)，男，副教授，主要从事军事管理信息系统研究。

二是查准率低下。正是因为爬虫冗余大,也就势必造成检索数据库信息量剧增,会检索结果不能很好地达到预期目的。

三是爬虫时间长。传统网络爬虫会爬取许多不相干的网页,势必会增加爬虫时间。

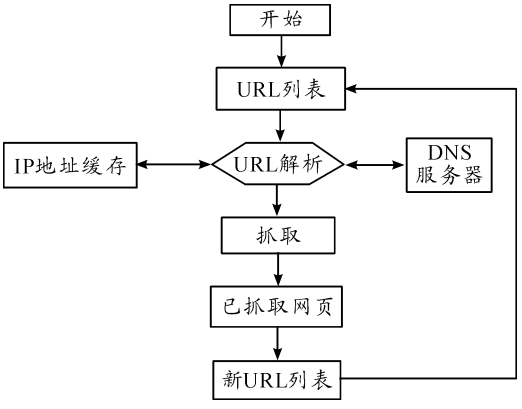

图1 传统网络爬虫流程

聚焦爬虫[3]则不同,它需要根据一定的网页分析算法计算其页面相关度,以此来过滤与主题无关的链接,只保留有用的链接并将其放入等待抓取的 URL 队列。然后,它将根据一定的搜索策略从队列中选择下一步要抓取的网页 URL,并重复上述过程,直到达到系统的某一条件时停止。

此外,聚焦爬虫的另一个特色就是所有被爬虫抓取的网页将会被系统存贮,进行一定的分析、过滤,并建立索引,以便之后的查询和检索;对于聚焦爬虫来说,这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。其爬虫流程图如图2所示。

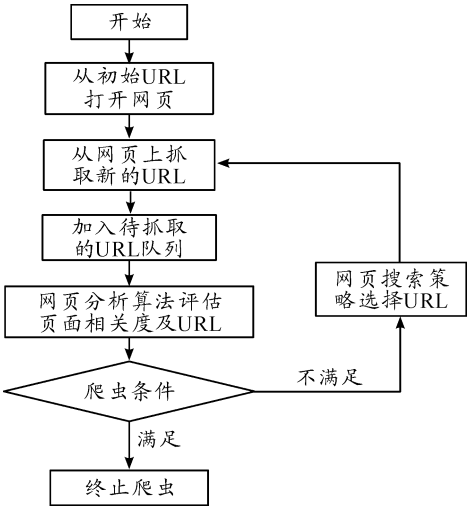

图2 聚焦爬虫流程

因此聚焦爬虫相对于传统网络爬虫具有一定的优越性:

一是冗余小。根据一定的算法有效控制爬虫范围,去糟取精。

二是爬虫结果与主题相关度更加贴近。聚焦爬虫能很好地控制爬取网页与主题的相关度,舍弃相关度地的网页。

三是检索效果佳。因为其爬虫范围不仅小了许多,而且紧贴主题,所以对于检索的效果会更好。

对比上面两幅图可以看出,两种爬虫策略的最大区别在于是否通过算法对页面进行分析,确定其与主题的相关度大小,倘若计算结果大于事先设定的相关度阈值,则将该网页加入到 URL 队列中待爬取,反之则弃之。

2 PageRank 算法

PageRank 算法[4]广泛应用于传统搜索引擎的网络爬虫中,它是基于网页的超链接之间的关系而得出页面排序值的,因此它只考虑与网页有直接或间接关联关系的网页,而并不考虑任何用户的任何查询以及网页的时间维数和页面相关度。

2.1 相关概念

网页 i 的入链:指向网页 i 的来自于其他网页的超链接的个数,即可以通过超链接从其他网页进入到网页 i 中;

网页 i 的出链:从网页 i 出发,指向其他网页的超链接的个数,即可以从网页 i 链接到其他网页中。

2.2 算法原理分析

1)从一个网页指向另一个网页的超链接是一种权威性的隐式传输,即指向它的父链接数越多,表示它的 PageRank 值就越高。

2)网页 i 的 PageRank 值是由所有指向它的父网页的 PageRank 值的总和决定的,而每个父链接的 PageRank 值将分配给所有的子链接,一般取平均分配,即分配系数取 0.5。

为了形象而又直观的理解,文章引入了图论思想:将 Web 抽象为一个有向图 $G=(V,E)$,其中 V 表示有向图的节点,这里的节点指网页,E 表示有向边,这里指超链接。则 Web 上的网页总数 $n=|V|$,网页 i 的 PageRank 值用 $PR(i)$ 表示

$$PR(i)=\sum_{(j,x)\in E}\frac{PR(i)}{O_j} \tag{1}$$

其中 O_j 表示网页 j 的出链个数。

再用矩阵 A 表示有向图的邻接矩阵,按如下规则为每条边赋值

$$A_{ij}=\begin{cases}\dfrac{1}{O_i} & \text{当}(i,j)\in E\\ 0 & \text{其他}\end{cases} \tag{2}$$

假如一个 Web 超链接有向图如图3所示。

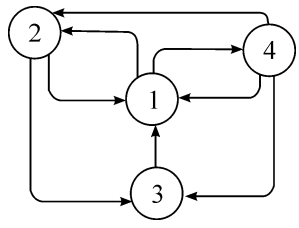

图3 Web 链接有向图

根据式(1)便可求出第1个节点的 PageRank 值

$$PR(1) = \frac{PR(2)}{O_2} + \frac{PR(3)}{O_3} + \frac{PR(4)}{O_4} \quad (3)$$

因为节点 1 有 2 个出链(分别为节点 2 和节点 4),所以根据式(1)可以得出 $A_{12} = A_{13} = 1/2$,同理可求得状态转移矩阵为

$$A = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \end{bmatrix}$$

任何时刻,随机浏览者从任意页面出发,是否选择继续浏览超链接网页,都会给这一链接分配一个由参数 d 控制的微小转换概率,也称为阻尼因子,其取值范围为 0 ~ 1。一般取值 $d = 0.85$。这样一来,对于含有 n 个 Web 网页节点的任意网页的 PageRank 值公式为

$$PR(i) = (1 - d) + d\sum_{j=1}^{n} A_{ji} PR(j) \quad (4)$$

该算法可以赋予任意的 PageRank 值初始值,经过反复迭代,当 PageRank 值不再发生显著变化或是趋于收敛时,算法结束,其迭代[5]求解算法如图 4 所示,当差值向量的 1 阶范数小于网页相关度阈值 ε 的时候就结束迭代。

PageRank-Iterate(G)
P_ε = 任意值
$k \leftarrow 1$
Repeat
$p_k \leftarrow (1 - d) + dA^T P_{k-1}$
$k \leftarrow k + 1$
until $\| P_k - P_{k-1} \| < \varepsilon$
return P_k

图 4　PageRank 迭代求解算法

3　Timed-PageRank 改进算法

3.1　融入时间维度的 PageRank 改进算法

Web 网页的一个显著特征就是网页的动态性,它将不断随着时间的推移而不断的变化,因此它具有良好的时效性,而传统网络爬虫的不足之处就是没有考虑和处理网页检索结果的时效性,因此不能很好地满足用户的检索需求。

Time-PageRank 算法是在 PageRank 算法的基础上,新增了关于网页的一个时间维度[6],其主要思想为:仍延续 PageRank 的随机冲浪和马尔科夫链模型,不同之处就是不再使用阻尼因子 d,而是引入了一个时间函数 $f(t)$〈$f(t) \in [0, 1]$〉来惩罚“陈旧的链接和网页”,这里的自变量 t 指的是当前时间和上次网页更新时间的差值,函数 $f(t)$ 指的是通过超链接访问网页的概率,而 $1 - f(t)$ 就指不通过超链接访问网页的概率。则 *Web* 浏览者就有 2 种选择:以 $f(t)$ 概率通过超链接访问网页;以 $1 - f(t)$ 概率不通过超链接跳到其他网页。

通常,$f(t)$ 的函数值会随着当前时间和上次网页更新时间的差值的增大而减小,对 $f(t)$ 定义如下:

$$f(t) = \begin{cases} 1 & t \leqslant 0.1 \\ 1/\log_2 t & t > 0.1 \end{cases} \quad (5)$$

式(5)中:t 表示当前时间和上次网页更新时间的月份差值。网页越新,t 值越小,其时间权值就越大,这样就能有效地区调整新旧网页的 PageRank 值,调整新的 PageRank 值[7]计算公式如下:

$$PR'(i) = PR(i) \times f(i_t) \quad (6)$$

式(6)中:$PR'(i)$ 为融入时间维度后的 PageRank 值,$PR(i)$ 为传统算法所得的 PageRank 值,$f(i_t)$ 为页面 i 的时间权值。

3.2　融入页面内容相关度的 PageRank 改进算法

聚焦爬虫相比传统网络爬虫的优势就在于爬取的网页尽可能地保证与主题的相关度大,最大限度地降低冗余网页,提高网页检索的查准率。

由于在计算网页相关度[8]的时候,共同词语出现位置的不同,其占的权重也不同[9]。比如标题的权重必然大于内容权值,内容权重必然大于标签权重等。因此计算网页相关度一般从如下几个方面考虑:标题关键字相关;内容共同词语频度。

一般而言,标题关键字的权重要大于内容共同词语频度的权重,这里取 a 为标题权重,b 为内容权重,且满足归一化原理,即 $a + b = 1$。设主题关键字在标题中出现的次数为 m,在内容中出现的次数为 n,则页面 i 与待检索主题相关度 $Sim(i)$ 计算公式为

$$Sim(i) = \frac{a \times m + b \times n}{m + n} \quad (7)$$

综合以上两种 PageRank 改进算法,则新的改进算法将由两部分组成:一是融入时间维度的 PageRank 改进算法,二是融入页面内容相关度的 PageRank 改进算法,假设前者的权重系数为 α,后者的系数为 β,同样满足 $\alpha + \beta = 1$,便可得到改进后的 Timed-PageRank 新算法公式如下:

$$PR(i) = \alpha PR'(i) + \beta Sim(i) = \alpha[(1 - d) + d\sum_{j=1}^{n} A_{ji} PR(j)] \times f(t) + \beta \frac{a \times m + b \times n}{m + n} \quad (8)$$

因此,在进行聚焦爬虫的过程中,要对待爬取的网页进行相关度计算,便可运用式(8)计算其与主题的相关度,看其是否满足爬虫前预先设置好的爬取阈值,若满足,则爬取,否则丢弃。

4　实例分析

为了测试改进后的 Timed-PageRank 算法是否相对于传统的利用 PageRank 算法具有优越性,进行了爬虫测试。实验平台为 PC 机(Pentium(R) 4 CPU +2.93 GHz,2.00 GB 内存),操作系统为 Win7 旗舰版,开发工具为 Myeclipse 10.0.0,使用开源的 Heritrix 程序进行爬虫。选择的爬取起始网页分别为搜狐,新浪和 163 的新闻主页。

爬虫前,可以预先设置相关度阈值,这里取 0.3,在爬虫的过程中,如果待爬取的网页与爬虫主题相关度高于 0.3 时

才能添加到 URL 队列中等待爬取,否则将直接丢弃。

如图 5 所示在 1 000 s 的爬虫时间内,基于 PageRank 算法的传统网络爬虫和基于改进后的 Timed-PageRank 算法的聚焦爬虫所爬取的网页与主题的相关度随爬虫时间的函数关系。

从图 5 中可以看出,爬虫开始时,两种算法网页的相关度都是最高,随着时间的推移,网页的相关度逐渐下降,但都始终高于阈值 0.3。且基于 Timed-PageRank 改进算法的聚焦爬虫的网页相关度总是大于基于 Timed-PageRank 算法的传统网络爬虫。这说明 Timed-PageRank 改进算法明显优于传统的 PageRank 算法,因此基于 Timed-PageRank 改进算法的聚焦爬虫也同样优于传统的网络爬虫,因为其爬取的网页与检索主题的相关度高,这能很好地保障检索结果的查准率,这也将对今后开发专业搜索引擎提供了一个很好的思路。

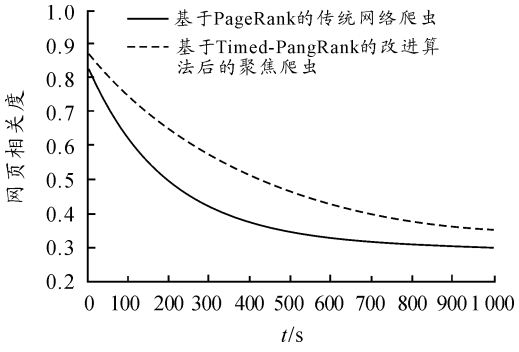

图 5 基于 2 种不同算法的爬虫效果示意图

5 结论

本文在分析了传统网络爬虫没有考虑网页时间维度以及网页相关度的基础上,对传统的 PageRank 算法进行了改进,提出了一种基于 Timed-PageRank 改进算法的聚焦爬虫,不仅增加了时间维度,而且将页面内容相关度融入其中。此算法比较全面地考虑了影响网页相关性的各种因素,能够有效地分析网页相关度,并根据预先设定好的阈值优化爬虫范围,缩小 URL 队列,能够有效地提高检索的查全率和查准率,相对于传统爬虫具有一定的优越性。最后根据实例验证了该算法的可行性和相对于传统算法的优越性,同时也说明了聚焦爬虫优于传统网络爬虫,其在不久的将来定会得到广泛应用。

参考文献:

[1] 李国成. 网络搜索引擎的现状及发展探析[J]. 企业科技与发展,2009(8):25 - 26.

[2] 刘世涛. 简析搜索引擎中网络爬虫的搜索策略[J]. 阜阳师范学院学报:自然科学版,2006,03:59 - 62.

[3] 周立柱,林玲. 聚焦爬虫技术研究综述[J]. 计算机应用,2005(9):1965 - 1969.

[4] 吴信东,库玛尔. 数据挖掘十大算法[M]. 李文波,吴素研,译. 北京:清华大学出版社,2013.

[5] Golub G H, Van Loan C F. Matrix Computations[M]. The Johns Hopkins University Press, 1983.

[6] Li X, Liu B, Yu P S. Time Sensitive Ranking with Application to Publication Search[M]. Conference on Data Mining, 2008.

[7] 张翔,周明全,李智杰,等. 基于 PageRank 与 Bagging 的主题爬虫研究[J]. 计算机工程与设计,2010,14:3309 - 3312.

[8] 邓丹君,周彩兰. 基于内容相关性和时间分析的改进 PageRank 算法[J]. 计算机与数字工程,2011(1):25 - 27.

[9] 陈小飞,王轶彤,冯小军. 一种基于网页质量的 PageRank 算法改进[C]//中国计算机学会数据库专业委员会. 第 26 届中国数据库学术会议论文集(B 辑). 中国计算机学会数据库专业委员会,2009:

(责任编辑 蒲 东)