

一种改进的基于最大流的 PageRank 算法研究

唐 敏

(佳都新太科技股份有限公司, 广东 广州 510665)

摘要: PageRank 是 Google 使用的一种网页排序算法, 它可以对网页进行评估, 并根据这个评估对网页进行排序。评价标准是一个 PR 值, 每个网页都会被赋予这样的一个值, 用来衡量网页的重要性, 并据此给出相应的排名。起初 PageRank 是用于搜索引擎检索结果的排序, 现在也被广泛用于其他很多方面, 比如网页的采集、检索结果聚类分析等。本文将最大流社区发现算法和 PageRank 结合, 提出了一种新的改进算法。

关键词: PageRank 算法, 网页排序, 搜索引擎, 最大流社区算法

中图分类号: TP301.6 文献标识码: A 文章编号: 1673-1131(2013)01-0023-02

0 引言

随着互联网的快速发展, 网上信息资源越来越多, 如何从巨大的信息海洋中寻找最想要的资源呢? 搜索引擎的出现, 给我们带来了可能, 但随之而来摆在搜索引擎面前的一个巨大挑战是如何去提高查询结果的质量。PageRank 根据网站的内外链接的数量及质量来衡量网站的价值, 由于 PageRank 算法是基于内外链接的, 也即几乎每个网页的 PR 值都是被平均地分配到它所指向的所有其他网页, 这就忽略了网页与网页之间的差异, 这种差异主要体现在网页质量上。

基于这样的认识, 本文提出了一种改进的基于最大流的 PageRank 算法。可以利用这个算法实时地评估每个网页的质量, 然后公平地对每个网页的 PR 值作出分配。

1 相关工作与背景知识

1.1 PageRank 算法

PageRank 算法基本思想: 我们可以假设一个漫游者在上网浏览页面时, 他在当前浏览页面中, 每次以相同的概率挑选他接下来想要点击的链接, 那么在这个挑选过程中, 如果遇到

一个没有出链的网页时, 他将随机的跳到任意页面, 这个行为是以一个较小的概率 β 发生。

可以假定 i 为页面 i 的链出页面集 E_i , B_j 为页面 j 的链入页面集, 那么在随机的一个时间节点上, 漫游者出现在页面 j 的概率 $PR(j)$ 为:

$$PR(j) = \beta/n + (1-\beta) * \sum_{i \in B_j} \frac{PR(i)}{|E_i|}$$

漫游者访问的概率分布满足和为 1。概率大的网页将被排在最前面, 也最容易被漫游者访问到。

1.2 最大流问题与社区发现算法

1.2.1 最大流问题

在一有容量的图里面, 找到一条不能超过弧的容量并且可以从发点到收点运送最大数量的单位流量的途径, 这就是最大流问题。给定有弧容量的图如下。

从 B 到 E 的每一个流都须满足下面这三个条件:

(1) 在从 B 至 E 的任一个流中, 从每个顶点 x ($x \neq B, x \neq E$) 出去的单位数量必须等于到达 x 的单位数量, 即

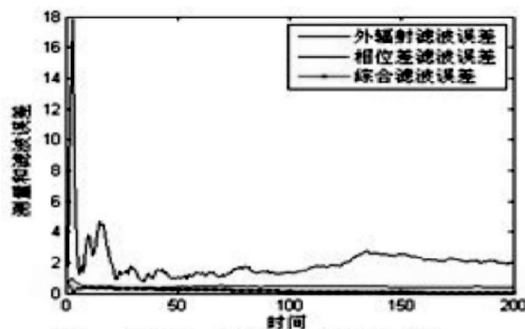


图 5 三种目标跟踪滤波误差

图 5 对比了分别利用外辐射源法、相位差法与多信息融合定位法得到的目标定位滤波误差图。

4 结语

通过对比外辐射源法、相位差法与多信息融合定位法的目标定位滤波误差仿真实验结果, 如上图 8 所示, 得出多信息融合定位法的定位误差显著小于外辐射源法和干涉仪相位差法, 且收敛速度较快, 即多信息融合定位法的定位精度高, 稳定性好, 具有很大的实用价值和应用前景。

参考文献:

[1] 应文, 李冬梅, 胡德秀. 一种基于时差变化率的单站无源定

位方法[J]. 指挥信息系统与技术, 2011, 2(1): 9-12

[2] 杨莘元, 郑思海. 基于运动辐射体 TOA 和 DOA 测量的单站被动定位算法[J]. 电子学报, 1996, 24(12): 66-69

[3] 李宗华, 肖予钦, 周一宇, 等. 利用频域和空域信息的单站无源定位跟踪算法[J]. 系统工程与电子技术, 2004, 26(5): 613-616

[4] 郁春来. 利用空频域信息的单站无源定位与跟踪关键技术研究[D]. 长沙: 国防科学技术大学电子科学与工程学院, 2008

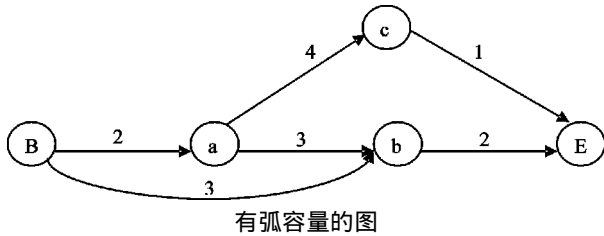
[5] 张芹. 固定单站无源定位系统的可行性研究[J]. 电子科技大学工程硕士学位论文集, 2007

[6] 罗鹏飞. 统计信号处理[M]. 北京: 电子工业出版社, 2009: 127-166

项目基金: 电子信息控制重点实验室项目资助(9140C100603110C1001), 国家自然科学基金面上项目“分数阶非线性动力系统的确定性数学理论及应用研究”(项目编号 11171238)
作者简介: 周正松(1987-), 男, 重庆人, 在读硕士, 主要研究方向为雷达信号处理、不确定性处理的数学; 罗懋康(1956-), 男, 重庆人, 教授, 博士生导师, 主要研究方向为不确定性处理。

$$\sum_{y \in X} f(x, y) - \sum_{y \in X} f(y, x) = 0 \quad (x \neq s, x \neq t) \quad (1)$$

其中 X 是图中所有顶点的集合, 经弧 (x, y) 运输的单位数量用 $f(x, y)$ 表示。



(2) 弧 (x, y) 的容量 $c(x, y)$ 小于经弧 (x, y) 运输的流的单位数量, 即

$$0 \leq f(x, y) \leq c(x, y) \quad (x, y) \in E \quad (2)$$

其中 E 是所有弧的集合。

(3) 流入收点的单位流量的净数与从发点流出的单位流量的净数 V 是相等的, 即

$$\sum_{y \in X} f(B, y) - \sum_{y \in X} f(y, B) = V \quad (3)$$

$$\sum_{y \in X} f(y, E) - \sum_{y \in X} f(E, y) = V \quad (4)$$

对于满足上述这三个条件的数值 $f(x, y)$ 的集合, 其是对应于从 B 至 E 的流。

鉴于此, 我们可以知道, 如能找到这样的一个数值 $f(x, y)$ 的集合, 它是满足上述这四个条件的 $(x, y) \in E$, 则这些数值对应于从 B 至 E 的流。因此当且仅当这个集合满足关系式 (1) 及 (4) 时, 数值 $f(x, y)$ 的集合是一个流 $(x, y) \in E$ 。

上述最大流问题从其本质上考虑, 其实就是求 V 的最大值。这是一个线性规划问题, 这里用一个比较直观的方法求解, 称为最大流算法。

1.2.2 最大流社区发现算法

算法的基本思想: 用户感兴趣社区的页面的一些实例是从一个种子页的集合开始的, 之后系统开始爬行网页, 且过程中利用种子页去查找更多的网页。最大流算法这时被用到从种子页和其它页分离社区 C , 该算法可能会被重复地执行以便查找到目标社区。

最大流社区算法过程, 当一个社区 C 被查找到以后, 它将参照每个网页指向社区内部页面的边数来对社区内部的网页进行评级, 一些评级高的非种子页将被加入种子集合。算法不断迭代, 直到启发式规则决定算法是否停止运行。

2 改进的基于最大流的 PageRank 算法

我们知道 PageRank 算法是基于内外链接的, 由于它只是考虑到了链接, 并没有考虑链接指向的内容。PageRank 算法中, 页面 i 所链出的页面, 对其平均分配 PageRank 值, 这就使得很多与当前页面不相关的链接页面也被分配到一定 PR 值, 这是不公平的。那么如何解决呢?

2.1 考虑的方法

基于分块的考虑:

网页都是由多个块组成的, 如有导航块, 广告块, 版权块等, 根据网页中每一块的重要程度, 给予它所发出的链接一个权值, 比如版权块发出的链接得到的 PageRank 值就会较小。

基于网页内容的考虑:

假设页面 i 当中, 里面的有一个链接指向页面 j , 并假定 $\text{Similarity}(i, j)$ 为页面 i 与 j 的文档相似度, 由此可知页面 i 转向它指向页面集合的概率 Similarity_{ij} , 并进行归一化就可以得

Similarity_{ij} , 这个值是在 Web 有向图中页面 i 指向 j 的转移概率:

$$\text{Similarity}_{ij} = \frac{\text{Similarity}(i, j)}{\sum_{k \in F_i} \text{Similarity}(i, k)}$$

这里的链接转移概率 Similarity_{ij} 可以综合考虑基于分块的分配的权值, 采取的方法可以把原 PageRank 公式中的 $\frac{1}{|F_i|}$ 用 Similarity_{ij} 代替, 这样即可加入对网页内容的考虑因素进行综合考虑。

2.2 改进的基于最大流的 PageRank 算法

由上述的相似度度量可以得到一个关于 Page 集合的有向图, 其中每条链接的权值是当前页面转移到其所指向链接的概率。

相对于下面介绍的最大流社区发现算法, 这个 Page 集合的有向图中的边的权值均小于 1, 可以通过设置一个常量系数 k (如 k 取 Page 集合文档总数) 来得到符合最大流社区发现算法有向图的边容量 $C(u, v)$, 其表示如下:

$$C(u, v) = k * \text{Similarity}_{uv}$$

应用最大流社区发现算法得到相关种子页的社区集合 M (CommunitySize 为社区大小), 对于每一个社区中的每一个网页, 用 Cited 表示社区其它网页引用它的次数, 也称为社区对它的推荐度。我们可以这么认识推荐度, 推荐度指在社区总网页数中, 其它的网页引用该网页的网页数目所占的比例, 也可以简单地理解评价, 而这种评价是其他网页给予的。在这里我们用 Recommend_Degree 表示推荐度, 并通过以下公式求得:

$$\text{Recommend_Degree} = \frac{\text{Cited}}{\text{CommunitySize}}$$

最后得到新的 PageRank 值为:

$$\text{NewPageRank} = \alpha * \text{PageRank} + \beta * \frac{\text{Cited}}{\text{CommunitySize}}$$

这里, $\alpha + \beta = 1$, α, β 用于调节推荐度与 PageRank 值的比例。

3 结语

这里加入社区因素只是通过比例进行调节。当然我们可以先运行社区发现算法, 找出相关种子页的社区集合, 并记录下各个网页的社区关系。然后再对 PageRank 进行反复的迭代, 并且可以将这个社区关系加入到 PageRank 的反复迭代过程中, 这样也许可以对网页的 PR 值作比较好的修正, 这在后续的实验验证中将继续研究。

参考文献:

- [1] 蒋永辉, 吴洪丽. 新的 PageRank 优化算法[J]. 计算机工程与应用, 2012(6)
- [2] 范聪贤, 刘秋菊, 徐汀荣. 应用 Web 结构挖掘的 PageRank 算法的改进研究[J]. 计算机工程与应用, 2010, 46(9): 127
- [3] 陈小飞, 王铁彤, 冯小军. 一种基于网页质量的 PageRank 算法改进[J]. 计算机研究与发展, 2009, 46(6): 381-387
- [4] 秦拯, 张玲, 李娜. 改进的 PageRank 在 Web 信息搜集中的应用[J]. 计算机研究与发展, 2006, 43(6): 1044-1049
- [5] 黄德才, 戚华春. PageRank 算法研究[J]. 计算机工程, 2006, 34(18): 161

作者简介: 唐敏(1980-), 男, 湖南湘潭人, 高级工程师, 研究方向为移动互联、搜索引擎优化、视频通话技术。