一种基于分层结构的网页排序算法

王大震[1],庄 重[1],张 帅[2]
(1.湖北工业大学 计算机科学与技术学院,湖北 武汉 430068;
2.中山大学 信息科学与技术学院,广东 广州 510006)

摘 要:链状解析算法已经被广泛应用于网络信息检索。然而,当前的链状解析算法通常用于平面链状图,忽略了网络的分层结构。这会导致两个问题的出现:链状结构越来越少和比较偏向于上层网页。文中提出了一个能够解决这两个问题的新排序算法,叫做分层排序,这种算法可以用于网络中的分层结构和链状结构。在实验结果中显示分层排序算法持续超过了其它知名排序算法,其中包括网页排序算法、块排序算法和层次排序算法。

关键词:链状解析;分层网络图;分层自由遍历模式
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2009)09-0035-03

A Page Ranking Algorithm Based on Hierarchical Structure

WANG Da-zhen[1], ZHUANG Zhong[1], ZHANG Shuai[2]
(1.College of Computer Science and Technology, Hubei University of Technology, Wuhan 430068, China;
2. Information Science and Technology Institute, National Sun Yat-sen University, Guangzhou 510006, China)

Abstract: Link analysis algorithms have been extensively used in Web information retrieval. However, current link analysis algorithms generally work on a flat link graph, ignoring the hierarchical structure of the Web graph. They often suffer from two problems: the sparsity of link graph and biased ranking of newly-emerging pages. In this paper, propose a novel ranking algorithm called hierarchical rank as a solution to these two problems, which considers both the hierarchical structure and the link structure of the Web. Experimental results show that hierarchical ranking algorithm consistently outperforms other well-known ranking algorithms, including the PageRank, BlockRank and LayerRank.

Key words: link analysis; hierarchical web graph; hierarchical random walk model

0 引 言

链状解析算法在网络搜索中扮演着重要的角色,网络链状结构揭示了网页人度的关系。现存的链状解析算法通常面临两大问题,第一个是网络图的链状分布要符合权值规律并且稀疏链状矩阵使得大多数网页不能获得正确的人度排序。第二个问题是新出现的网页很少得到别的链接,导致它不能得到正确的人度。因此提出了一个分层网络结构来解决这些问题。

1 研究现状

1.1 链状解析研究现状

链状解析技术被广泛地应用到计算网页的人度中,比如 HITS[1]和 PageRank[2]。在文献[3]中,作者将网络图分成几块,通过区分不同块之间的链接或相同块内部之间的链接来分配不同的权值,然后再用 PageRank 算法计算。还有文献[4]中提出一个双层马尔可夫模型用来计算网页排名。在稍后的实验中这种算法被称为"LayerRank"。有些算法将网页内容也作为权衡排名依据[5]。

1.2 网状结构的研究现状

现在有许多发掘网络分层结构的研究,文献[6]中提出利用网络图的分层结构来加快 PageRank 的计算速度。还有一些现有的研究通过网络分层结构模拟网络。在文献[7]中提出了一个网络分层模式,在这种模式中每个网页所在的"区域"都有一个定值;并且网页只能与同一个区域内的其它网页链接。

2 分层网络图

在分层结构中,网络包含网页、目录、主机和域。因此,整个网络图可以大概分为几个层次结构,例如网页层、目录层、主机层和域层。通过几个不同的层次,

收稿日期:2008-12-31;修回日期:2009-03-30
基金项目:湖北省中青年人才计划项目(Q20074006)
作者简介:王大震(1975-),男,博士,副教授,硕士生导师,研究方向为分布式计算、信息检索。

每层之间的超级链接可分为两种:内部链接和外部链接。同样,两个不同超级结点中的网页之间的链接叫做外部链接。而且,链接也分为两种,一种是链入一个网页,另一种是从这个网页链出。通过详细分析分层网络结构,得出以下数据(见表 1):

表 1 链接分布表

层	内部链接	外部链接
域	7 342 029 (97%)	227 324 (3%)
主机	6 506 478 (86%)	1 062 875 (14%)
目录	2 956 546 (39%)	4 612 807 (61%)
网页	0 (0%)	7 569 353 (100%)

如表 1 所示,在两个不同层次中内部链接和外部链接的比例是不同的。以主机层为例,有 86% 的链接是内部链接,意味着本地连接率非常高。因此建立了一个两层网络模式(见图 1)。

如图 1 所示:

上层图:上层图包括 m 个称为超级结点的顶点,超级结点之间的边称为超级边。超级边遵循以下原则:如果在超级结点 S_i 和 S_j 中至少有一到多个网页,那么 $E_{i,j}$ 就表示 S_i 到 S_j 的有向超级边。从 S_i 指向 S_j 的这条边的权值就是从 S_i 中的网页链接到 S_j 中网页的链接条数。

下层图:用 $P = \{P_0, P_1, \cdots, P_n\}$ 来表示超级结点 S 中的所有网页,通过 URL 的关系将这些网页放到一个超级结点中。图 2 就是在主机层上用这种方式建立的图。

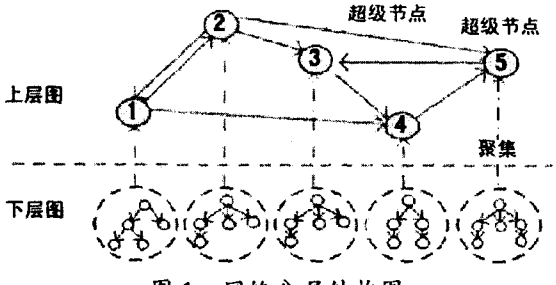

图 1 网络分层结构图

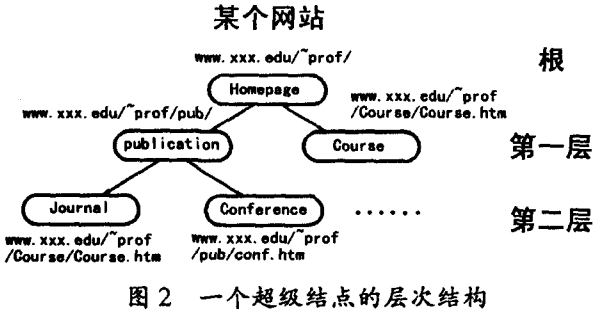

图 2 一个超级结点的层次结构

3 分层排序

3.1 分层自由遍历模式

1)在每次浏览初期,用户随机选择一个超级结点。

2)用户在阅读完超级结点中的一个网页时,他会选择以下三种可能的行为之一:

(a) 继续阅读当前超级结点中的一个网页时,并且是顺着这个超级结点的分层链状结构读下去。

(b) 跳到与这个超级结点链接的另一个超级结点中。

(c) 结束浏览。

通过以上分层自由遍历模式,分两步计算分层排序:第一步通过超级结点之间的链接关系计算超级结点的入度。第二步通过超级结点内部的链接关系计算网页的入度。

3.2 计算超级结点的入度

把上层图看作一个矩阵。假设整个网络含有 m 个超级结点,一个 $m * m$ 的邻接矩阵被定义为 A 并且 $A[i,j]$ 代表从超级结点 i 到 j 的链接的权值。邻接矩阵 A 用来计算每个超级结点 S_i 的入度,当浏览一个超级结点时,用户随机选择当前这个超级结点的一个矩阵并且跳到这个链接所指向的超级结点的概率为 $1-\varepsilon$。当用户统一从超级结点集合跳到一个超级结点的概率为 ε,所以,超级结点的入度方程为:

$$SI_i = \frac{\varepsilon}{n} + (1-\varepsilon)\sum_{j:1_{ji}\in E} SI_j \cdot A[j,i] \qquad (1)$$

其中 $\varepsilon(0<\varepsilon<1)$ 是一个参数。在实验中设 $\varepsilon = 0.15$。

3.3 计算网页入度

在获得一个超级结点的入度时,文献[8]中提出了一个计算网页入度的方法,一个超级结点中所有网页的入度之和就等于这个超级结点的入度。

3.3.1 构建权值树状结构

如图 3 所示,将一个超级结点内部的分层树状结构模拟为一个有向权值树结构。在这个权值树结构中,每条边是指从双亲结点到孩子结点并且通过网页的特征来衡量边的权值。超级结点 S_i 中的一个网页 P_j,从它到它双亲结点的边的权值 ω_j 可以计算为:

$$\omega_j = \theta \cdot \mathrm{link}(p_j) + (1-\theta)\mathrm{index}(p_j) \qquad (2)$$

是否为索引页也是分配不同权值的方法。用分层规则来判断是不是索引页。

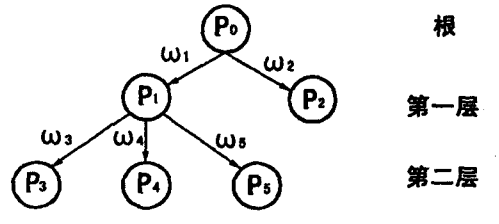

图 3 分层权值树结构

$$\mathrm{index}(p_j) = \begin{cases} 1 & \text{if } p_j \text{ is index page} \\ \alpha & \text{if } p_j \text{ is other page} \end{cases} \qquad (3)$$

参数 α 是 0～1 之间的系数因子,它的入度在后面介绍。

link 用来计算链入到网页 P_j 的链接数量,公式:

$$\text{link}(p_j) = \beta \frac{\text{OIL}(p_j)}{\sum_{p_k \in s_i} \text{OIL}(p_k)} + (1-\beta) \frac{\text{IIL}(p_j)}{\sum_{p_k \in s_i} \text{IIL}(p_k)} \quad (4)$$

其中 β 是给予链出和链入不同权值的系数因子,$\text{IIL}(P_j)$ 是网页 P_j 的内部超链接数量,$\text{OIL}(P_j)$ 是网页 P_j 的外部链接数量。

3.3.2　用 DHC 计算网页入度

基于以上分层权值结构,一个网页在超级结点中的入度可以用 DHC[9]算法从根网页到底层网页进行计算。每个网页 P_j 有一个权值 w_{ij},它表示在超级结点中网页 P_j 的入度。

$$w_{ij} = \prod_{p_k \in \{\text{nodes from } n_j \text{ to root}\}} \gamma \times \omega_k \quad (5)$$

其中参数 γ 是分散热量系数因子。

最后,一个网页 P_j 在整个网络图中的入度定义为 PI_j,计算方程如下:

$$PI_j = SI_i \times w_{ij} \quad (6)$$

其中网页 P_j 属于超级结点 S_i,SI_i 是超级结点 S_i 的入度。根网页的入度等于超级结点的入度。

4　实　验

选择爬虫收集的 50 000 张网页集作为搜索的数据库。在实验中选择了 5 个不同领域的关键词进行搜索,对每个关键词检索出的结果选取前 50 条记录,每隔 10 条记录进行一次查准率分析,然后将搜索结果和链状解析算法和网状解析算法进行查准率的对比,如图 4 所示。

通过对比,不难发现分层排序算法,比已知的链状解析算法和网状解析算法查准率更高。

5　结束语

在 PageRank,BlockRank 等知名算法的基础上,给出了基于网状结构的分层排序算法。考虑到网络既有分层结构又有链状结构,提出了分层自由遍历模型,这种模型能够模拟用户进行网上冲浪的行为。基于该模型,提出了一个分层排名算法来计算网页的入度。该排名算法可以显著提高网络搜索的效率,并且对于新出现的网页能够分配合理的排名。在今后的工作中,将进行试验的大型网站收集来评价我们的算法。

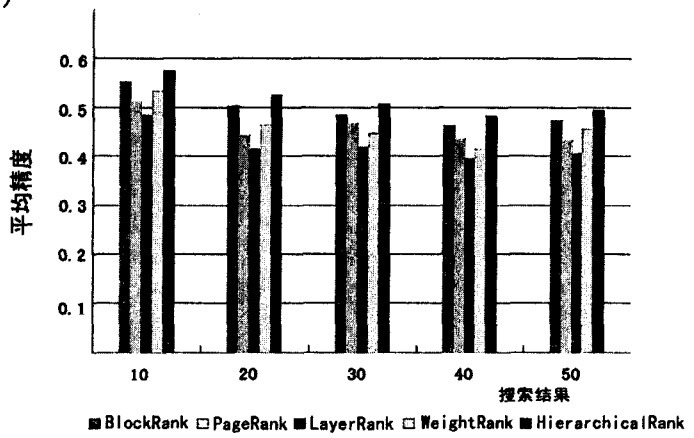

图 4　排序结果对比

参考文献:

[1] 何晓阳,吴　强,吴治蓉. HITS 算法与 PageRank 算法比较分析[J]. 情报杂志,2004,17(3):77－80.

[2] 吴淑燕,许　涛. PageRank 算法的原理简介[J]. 图书情报工作, 2003,26(2):364－367.

[3] 白似雪,刘华斌. 基于页面分块模型的 PageRank 算法研究[J]. 南昌大学学报:工科版, 2008,12(1):34－36.

[4] Wu J, Aberer K. Using a Layered Markov Model for Decentralized Web Ranking[R]. [s.l.]: ACM Press,2004.

[5] 李绍华,高文宇. 基于层次分类的页面排序算法[J]. 计算机工程, 2007,19(2):56－60.

[6] 钱功伟,倪　林,曹　荣. 基于网页链接和内容分析的改进 PageRank 算法[J]. 计算机工程与应用, 2007,43(3):71－74.

[7] Laura L, Leonardi S, Caldarelli G, et al. A Multi－Layer Model for the Web Graph[C]// In 2nd International Workshop on Web Dynamics. Honolulu:[s.n.], 2002.

[8] Eiron N, McCurley K S, Tomlin J A. Ranking the Web Frontier[C]// In Proceedings of the 13th International World Wide Web Conference. [s.l.]:ACM Press,2004:309－318.

[9] Zhou D, Weston J, Gretton A, et al. Ranking on Data Manifolds[M]. Cambridge, MA: MIT Press, 2004.

(上接第 34 页)

[5] Stutzle T, Hoos H H. Max－Min ant system[J]. Future Generation Computer System,2000,16(19):889－914.

[6] 黄国锐,曹先彬,王熙法. 基于信息素扩散的蚁群算法[J]. 电子学报,2004,32(5):865－868

[7] 张　然,贾瑞玉,钱光超,等. 带佳点交叉算子的非均匀窗口蚁群算法[J]. 计算机技术与发展,2007,17(12):68－70.

[8] 张　亮,孙力娟. 蚁群算法和免疫算法的融合及其应用[J]. 计算机技术与发展,2006,16(3):31－33.

[9] 段海滨. 蚁群算法原理及其应用[J]. 北京:科学出版社, 2005.

[10] 郑　松,侯迪波,周泽魁. 动态调整选择策略的改进蚁群算法[J]. 控制与决策,2008,23(2):225－228.