Web搜索算法研究综述

蒋卫星[1]，金 瓯[1,2]，张 彬[3]

（1. 中南大学 信息科学与工程学院，湖南 长沙 410083；
2. 湖南金融货币识别与自助服务平台技术工程中心，湖南 长沙 410004；
3. 衡阳师范学院 计算机科学系，湖南 衡阳 421008）

摘 要：介绍了PageRank和HITS两种最常见的算法，对基于链接结构分析的Web搜索算法的研究进展进行了综述，主要包括：介绍了独立于查询的各种改进算法以及基于查询主题的有关算法，并分析上述算法的优缺点及其改进策略或方法，以及Web搜索算法的关键技术和应用，最后是关于Web搜索算法存在的问题和研究展望。

关键词：Web搜索；链接分析；PageRank；HITS

中图分类号：TP393 **文献标识码**：A **文章编号**：1673－629X(2007)04－0178－04

A Survey of Algorithm Research in Web Searching

JIANG Wei-xing[1], JIN Ou[1,2], ZHANG Bin[3]

(1. School of Information Science and Engineering, Central South University, Changsha 410083, China;
2. Hunan Engineering Center for Currency Recognition & Self－service, Changsha 410004, China;
3. Department of Computer Science , Hengyang Normal University, Hengyang 421008, China)

Abstract: Two algorithms of PageRank and HITS are firstly introduced. Then the research advances of Web searching algorithms based on linkage analysis technology are deeply described, including all sorts of improved algorithms query－independent and the corresponding algorithms based on the topic of query, and at the same time the merit and disadvantage of above－mentioned algorithms are analyzed. Finally, the existent problems and research prospective are discussed.

Key words: Web searching; linkage analysis; PageRank; HITS

0 引 言

传统的Web搜索页面排序技术可以分为二类：

(1) 基于人工分类的页面排序技术：它通过手工的方式把Web上的文档子集分成相应的类别和关键词集合。这种方法的缺点在于：它只能应用于少量的数据并且速度比较慢，另外由人工方法形成的类别和关键词常常是不充分的或不完整的。Yahoo搜索引擎就是这种技术的典型应用。

(2) 基于使用信息的页面排序技术：它通过部署在Web上的一些服务程序来收集用户的查询、用户查询后访问的网页和每个网页的浏览时间等信息。这种方法成功的关键在于需要为每个查询收集大量的信息，缺点在于它只能在少量数据集上使用，容易受到广告、垃圾邮件等无关信息的影响。

随着Web页面不断增加以及网络规模不断扩大，上述页面排序技术被一种基于网页链接分析的页面排序技术所代替。这种技术假设：a. 关于某一主题的页面是相互链接的；b. 权威的页面倾向于指向其它的权威性页面。PageRank算法就是基于假设b来对页面排序的。文中论述的Web搜索算法也是基于页面链接分析技术来进行的。

1 PageRank及其衍生算法

1.1 PageRank算法

PageRank算法[1]是一种与查询不相关的、针对全球Web页面排序的、最早应用链接分析技术到搜索引擎中的算法。它的基本思想是试图为全球的所有可以搜索的网页赋予一个量化的价值分数，该价值分数由指向每个网页的所有网页的价值分数决定。每个网页的价值分数 PR 可用公式表示如下：$PR(i) =$

收稿日期：2006－06－22

基金项目：国家863计划资助项目(2003AA1Z2190)；国家“十五”科技攻关资助项目(2003BA104C)

作者简介：蒋卫星(1973－)，男，湖南长沙人，硕士，研究方向为计算机应用技术、网络信息系统；金 瓯，教授，博士生导师，研究方向为嵌入式系统、信息处理。

$C\sum_{j\to i}[PR(j)/N(j)]$,其中 $PR(j)$ 表示所有链接向网页 i 的网页 j 的价值分数,$N(j)$ 表示所有链接向网页 i 的网页 j 的总数,C 为小于 1 的规范化因子。如果要计算网页集合中所有网页的 PR 值,必须利用这个公式进行反复迭代。假设 N 为所有网页的总数,则初始化每个网页的 PR 值可以都赋以 $1/N$,当迭代达到一定的次数 PR 值将会收敛于一个相对固定的值。

前述公式存在等级沉没(Rank Sink)[1] 问题。等级沉没是由于一组没有外出链接的紧密链接网页不断积聚 PR 值而不分配 PR 值给其它网页,使得其它网页的 PR 值不断下降的现象。所以通常采用的公式是下面的形式:$PR(i) = d\sum_{j\to i}[PR(j)/N(j)] + (1-d)$,其中 $d(0 < d < 1)$ 是一个衰减系数,一般取 0.85。上述公式 PR 值的计算体现了如下的思想:即一个具有很高价值的网页应该被很多高价值网页所链接。

1.2 PageRank 的改进算法

PageRank 算法是一种独立于用户查询的、离线的被实践证明具有快速响应能力和很高成功率的算法,然而它仍存在一些明显的缺陷,例如:PageRank 计算独立于用户查询,没有考虑用户查询的具体要求,从而不能够应用于特定主题获取信息,算法过分强调网页的链入链接而贬低链出链接、忽视专业站点以及偏重旧网页等[2]。自从 1998 年至今,针对原始 PageRank 算法的弱点,人们已经提出了许多改进算法。它们大致可以分为下面几种类型:(1) 针对传统 PageRank 算法为网络上每一个超链接赋予相同权重的弱点而提出的基于不同权重的 PageRank 算法[3];(2) 把一个网页分成不同的语义块的块级别 PageRank 算法[4];(3) 基于 Web 图分区的 PageRank 算法[5];(4) 考虑用户点击的 URL 位置和点击次数的个性化 PageRank 算法[6];(5) 针对网页可以分类为目录、主机或域名级别网页而提出的层次结构的 PageRank 算法[7];(6) 针对传统的随机网上冲浪(surfer)模型的缺陷,提出的把网页分成主机内的链接和主机间链接的两层 PageRank 算法[8,9];(7) 加速算法效率的 PageRank 算法:包括并行、分布式以及加速迭代等算法[10~13]。

2 查询相关算法

由于网络的结构具有分散性、多元性、动态变化性等特点,为了提高效率通常的 Web 查询算法是查询无关的。如果在 Web 搜索算法中利用了查询信息,算法的执行就更能体现用户预期的目标和限制条件,能够得到更精确的结果,从而提高 Web 搜索算法返回结果的准确性。常见的查询相关算法包括 HITS 及其改进算法、HillTop[14]、TSPR (Topic - Sensitive PageRank)[15] 以及 SALSA[16] (The Stochastic Approach for Link - Structure Analysis)和 indegree 等。

2.1 HITS 及其改进算法

与 Brin 和 Page 提出的独立于查询的 PageRank 算法不同的是,由 Kleinberg[17]提出的 HITS 算法是一种查询相关的算法,其算法模型由权威性网页(authority)和中心网页(hub)组成。其中:authority 为表达某一主题的高质量页面,它被很多链接所指向,而 hub 页面是指一个或多个 Web 页面,它提供了指向权威页面的链接集合。authority 和 hub 具有相互强化的关系,即一个好的中心性网页应该指向很多好的权威性网页,而一个好的权威性网页则应该被很多好的中心性网页所指向。

HITS 算法首先利用一个传统的文本搜索引擎获取一个与主题相关的网页根集合(root set),然后把根集扩充成一个包括所有引用根集中页面和被根集中页面引用的页面基本集合(base set),并为每个页面 i 引入两个权值:权威度 a_i 和中心度 h_i,最后执行一个传播页面权值的迭代过程。设基本集合 T 中包含 N 个页面,则可将 T 表示成一个 $N\times N$ 的矩阵 $\boldsymbol{A}$,若页面 i 和页面 j 存在链接则 $\boldsymbol{A}$ 中元素 $A_{ij}=1$,否则 $A_{ij}=0$。设向量 $\boldsymbol{a}=(a_1,a_2,\cdots,a_N)$ 代表所有基础集合的权威度,而向量 $\boldsymbol{h}=(h_1,h_2,\cdots,h_N)$ 代表所有的中心度,初始时两向量所有元素都置为 1。设 α 和 β 是缩放因子,则每个页面的权威度 a_i 和中心度 h_i 可以通过下面的迭代过程得到:$\boldsymbol{a}=\alpha\beta\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{a}$ 和 $\boldsymbol{h}=\alpha\beta\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{h}$,为了不使计算溢出每次迭代后须对向量 $\boldsymbol{a}$ 和 $\boldsymbol{h}$ 进行规范化,Kleinberg 已经证明经过若干次迭代后,向量 $\boldsymbol{a}$ 和 $\boldsymbol{h}$ 将收敛于矩阵 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 和 $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 的主特征向量。

文献[18] 指出 HITS 采用矩阵 $\boldsymbol{K}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 的主特征向量来确定权威度,算法的稳定性由 $\boldsymbol{K}$ 的特征间隙 δ 决定。δ 为第一大特征向量(主特征向量)和第二大特征向量之间的数值差。当 δ 很小时,一个很小的扰动可能引起结果发生剧烈的变化;当 δ 足够小时,还可能产生主特征向量和第二特征值向量交换位置的"翻转"现象。为此,文中提出了一种根据多个特征向量构成的子空间来确定页面的权威权重的 HITS 算法,该算法加强了 HITS 的稳定性,因为权威权重依赖于跨越 k 个特征向量的子空间,而并不仅仅是某个特征向量。

HITS 算法由于 hub 页面的多主题性而使得主题存在"漂移"现象,由于权威度和中心度的相互强化关系可能使得两台主机之间存在不必要的强化关系,导航链接、广告链接等无关链接和无关页面也将会影响算法的精度[19]。为此,文中分别提出了对应的改进措

施:将 hub 权重的计算由它所指向的 authority 权重之和改为 authority 权重的平均值;通过公式 $a_p = \sum_{q \to p} h_q / k$ 和 $h_p = (\sum_{q \to p} a_q / m) / n$ 来改进两台主机的相互强化关系,其中 k 表示与 q 同主机并指向 p 的页面数,m 表示与 q 同主机并被 p 所指向的页面数,n 表示 q 的总数;通过链接文本匹配、主题词次数统计和导航链接地址分析等方面来消除无关链接的影响。

2.2 HillTop 算法

康柏系统研究中心的 Krishna Bharat 和多伦多大学的 George A. Mihaila 于 2001 年 1 月提出了 HillTop[14]算法并申请了专利,该算法是 2003 年 12 月 Google 搜索引擎升级的主要技术。HillTop 算法利用专家网页来匹配查询词和链接关系以及区分网页超链接的质量,专家网页采用倒排文件的方式建立索引,利用专家网页的原因在于网页之间的链接关系的重要性相差悬殊,尤其对于网上社区的这种链接关系传统的 PageRank 算法将出现较大的偏差。HillTop 算法的实现需运行一个周期较长的针对频繁关键词、关键短语和查询术语的批处理过程。

HillTop 算法基本过程可以分为二个步骤:第一个步骤根据查询寻找“专家网页”,专家网页是关于一定主题并且指向许多非隶属网页且至少有一个短语(phrase)包含了查询中的所有术语(term)的网页。专家网页得分 $\text{Expert\_Score} = 2^{32} * S_0 + 2^{16} * S_1 + S_2$,其中 $S_i = \text{Sum}_{|\text{key phrasesp with k-iquery terms}|} \text{Level Score}(p) * \text{Fullness Factor}(p, q)$,$k$ 表示查询 q 术语的个数,Level Score(p) 为短语类型得分,可根据短语处在标题、页面头部和锚(archor) 文本等层次分别给不同权重值:如 16、6 和 1,而完整性因子 Fullness Factor(p,q) 表示短语 p 的术语覆盖查询 q 中术语的个数。第二个步骤给顶部专家网页链向的目标网页打分,这个过程综合了它与所有相关专家网页的链接关系。设每个指向目标网页 T 的专家网页 E 形成边(E,T),且对每个查询关键词 w 令 occ(w,T) 表示包含 w 和(E,T) 的 E 中可区分的关键短语(key phrases) 的个数,则 $\text{Edge\_Score}(E, T) = \text{Expert\_Score}(E) * \text{Sum}_{|\text{query keywords w}|} \text{occ}(w, T)$,而 $\text{Target\_Score} = \text{Sum}_{|\text{all edges incident on this query}|} (\text{Edge\_Score})$。

2.3 TSPR 算法

为了考虑查询主题对查询结果的影响,在 PageRank 算法的基础上 Haveliwala 提出了 TSPR[15]算法,其后也成为了 Google 的核心算法。TSPR 算法主要由两个过程组成:第一个过程根据基本主题集离线生成有偏的 PageRank 向量,查询主题可以利用 ODP(open directory project)的主题。设 L_j 为 ODP 中类别为 c_j 的链接集合,基于主题 c_j 的 PageRank 向量为 $PR_j(i)$,则当页面 $i \in L_j$ 时,$PR_j(i) = 1/|L_j|$,当页面 $i \notin L_j$ 时,$PR_j(i) = 0$。第二个过程在查询时执行,它将计算查询属于各个主题类别的概率。对于一个给定的查询 q 和类别 c_j,其概率 $P(c_j | q) = P(c_j) \cdot P(q | c_j) / P(q)$。对于 Web 文档 d,其查询敏感重要性得分 $s_{qd} = \sum_j P(c_j | q) \cdot PR_j(d)$。TSPR 算法考虑了页面链接和主题相关的复合影响。

3 Web 搜索算法关键技术

随着 Web 页面以惊人的速度增长,如何高效、准确地找到用户所需要的 Web 页面信息,是 Web 搜索算法要重点考虑的问题。当前的 Web 搜索算法正在变得越来越复杂,Web 搜索算法将集成多个领域的关键技术(例如:数据库技术、Web 数据挖掘技术、分布式并行技术以及 XML 技术等),同时它也将呈现多样化的发展方向,除了常规的 Web 页面搜索外,还将考虑用户的多种需求,例如针对图片、视频、音乐或语义 Web 页面等的搜索。现在已经出现针对语义 Web 和元数据搜索的搜索引擎 Swoogle。

从 Web 搜索算法的准确率看,为了满足建立在 Web 动态变化之上的用户需求,Web 搜索算法需综合利用现有的 Web 文档归类、Web 信息抽取、链接分析等技术。采用聚类算法对网页进行初步的文档归类在不影响实时性能的前提下可以有效地提高页面排序的准确率;Web 信息抽取技术利用抽取的查询主题信息也能在一定程度上提高搜索页面的准确性;深入地分析 Web 页面的链接关系、类型及结构对于页面排序及搜索算法的准确性也是至关重要的。从 Web 搜索算法的性能看,如何研究高效基于主题的 Web 爬虫、如何利用 P2P 技术来提高搜索算法的深度以及搜索的效率、如何建立快速的 Web 索引、如何提高集群的计算能力以及如何减少页面排序过程中矩阵的迭代次数等技术都是非常关键的,尤其是分布式和并行计算技术。据最新结果统计,Google 矩阵的当前尺寸已达到 4.2×10^9,因此特征向量的计算量将会显得非常巨大,分布式计算已经成为 Web 搜索算法一项必要的技术。分布式运算的理论基础是图论,可以把 Web 看作一个超大的连通图,并把连通域分成弱连通域和强连通域两种类型。

随着 XML 技术的成熟和语义 Web 技术的发展,Web 页面将从人可理解的页面变成机器可理解的页面,因此 XML 和语义 Web 技术也必将成为 Web 搜索算法关键技术之一。

4　结束语

当前的 Web 搜索算法已经从 Yahoo 的第一代文本搜索算法发展到第二代以 PageRank 和 HITS 为代表基于链接分析的搜索算法。Web 搜索算法仍然是一个值得深入研究的课题,这个领域仍有许多需要解决的问题:(1) 在综合利用各种算法和技术的基础上,如何在查询的准确性、查全率和实时响应性能方面达到平衡将是一个需综合考虑的问题;(2) 如何防止欺骗和垃圾链接对页面排序的影响也是非常重要的问题;(3) 面对用户的不同需求,提供个性化的 Web 搜索算法也正在变得越来越重要,同时 Web 搜索算法也将面对多样性的问题;(4) 如何利用网上社区发现技术从而在一定程度上改善 Web 搜索算法返回页面的准确度和查全率也将是非常关键的问题;(5) 语义 Web 的发展将有可能推动 Web 搜索算法朝着新一代的方向发展。

参考文献:

[1] Page L, Brin S ,Motwani R, et al. The PageRank citation ranking: Bringing order to the Web[R]. Database Group, Stanford University,1998.

[2] 宋聚平, 王永成, 尹中航, 等. 对网页 PageRank 算法的改进[J]. 上海交通大学学报,2003,37(3):397 - 400.

[3] Xing W, Ghorbani A. Weighted PageRank Algorithm[C]// 2nd Annual Conference on Communication Networks and Services Research. Fredericton, N.B., Canada: IEEE Computer Society,2004:305 - 314.

[4] Cai Deng, He Xiaofei, Wen Ji - Rong, et al. Block - level Link Analysis [C] // In Proc. of the SIGIR' 04 Conf. Sheffield, United Kingdom: ACM Press, 2004:440 - 447.

[5] Bradley J T, de Jager D V, Knottenbelt W J, et al. Hypergraph Partitioning for Faster Parallel PageRank Computation [C]//EPEW'05. Versailles, France:[s.n.], 2005:155 - 171.

[6] 李　凯, 赫枫龄, 左万利. PageRank - Pro:一种改进的网页排序算法[J]. 吉林大学学报:理学版, 2003,41(2):175 - 179.

[7] Xue Gui - Rong, Yang Qiang, Zeng Hua - Jun, et al. Exploiting the Hierarchical Structure for Link Analysis[C]// Proceedings of the 2005 ACM SIGIR Conference. Salvador, Brazil: [s.n.], 2005:186 - 193.

[8] Kamvar S D, Haveliwala T H, Manning C D, et al. Exploiting the Block Structure of the Web for Computing PageRank [R]. US:Stanford University, 2003.

[9] Wu Jie, Aberer K. Using a Layered Markov Model for Decentralized Web Ranking[R]. Lausanne: Swiss Federal Institute of Technology, 2004.

[10] Wang Yuan, DeWitt D J. Computing PageRank in a Distributed Internet Search System[C]// Proceedings of the 30th VLDB Conference. Toronto, Canada:[s.n.], 2004:420 - 431.

[11] de Jager D. PageRank: Three Distributed Algorithms[D]. London: the Imperial College of Science, Technology and Medicine, University of London, 2004.

[12] Gleich D, Zhukov L, Berkhin P. Fast Parallel PageRank: A Linear System Approach[C]// WWW2005. Chiba, Japan: [s.n.],2005:1 - 8.

[13] 陈再良, 凌　力, 周　强. dPageRank——一种改进的分布式 PageRank 算法[J]. 计算机应用,2006,26(1):21 - 24.

[14] Bharat K. Hilltop: A Search Engine based on Expert Documents [EB/OL]. 2001. http://www.cs.toronto.edu/~georgem/hilltop/, University of Toronto.

[15] Haveliwala T H. Topic - Sensitive PageRank[C]// In Proceedings of the eleventh international conference on World Wide Web. [s.l.]:ACM Press,2002:517 - 526.

[16] Lempel R, Moran S. SALSA: The Stochastic Approach for Link - Structure Analysis[J]. ACM Transactions on Information Systems,2001,19(2):131 - 160.

[17] Kleinberg J. Authoritative sources in a hyperlinked environment[C]// Proceedings of the 9th ACM - SIAM Symposium on Discrete Algorithms. New Orleans: ACM Press, 1997: 668 - 677.

[18] 石　晶, 龚震宇, 裘杭萍, 等. 一种更稳定的链接分析算法——子空间 HITS 算法[J]. 吉林大学学报:理学版,2003, 41(1):49 - 53.

[19] 宋建康, 张礼平. Web 结构挖掘算法探讨[J]. 华东理工大学学报, 2003,29(5):537 - 540.

(上接第 177 页)

加速比。最后还分别采用两种方法进行了并行计算实验,实验结果表明这种以计算机的实际计算能力的大小来动态分配任务的策略实现比较简单,并且具有很好的并行效果。

参考文献:

[1] 孙家昶,张林波,迟学斌,等. 网络并行计算与分布式编程环境[M]. 北京:科学出版社,1996.

[2] Wilkinson B,Allen M. 并行程序设计[M]. 陆鑫达 等译. 北京:机械工业出版社,2001.

[3] 张信一,李代平,章　文. 基于 Win32 平台上的 PVM 并行程序设计[J]. 计算机应用研究 2004(5):102 - 104.

[4] 张建军,蒋廷耀,郭志鑫. PVM 中动态负载平衡的设计和实现[J]. 计算机工程,2005(7):63 - 64.

[5] 张信一,李代平,罗伟刚. 并行程序开发平台的可视化实现[J]. 计算机应用研究,2004(11):266 - 269.