一种评价搜索引擎信息覆盖率的模型及其验证

孟 涛,闫宏飞,李晓明
(北京大学计算机科学技术系,北京 100871)

摘 要: 搜索引擎的网页搜集子系统通常以 WWW 的网页构成的有向图结构为依据,循着网页间的链接进行搜集从而扩大信息覆盖面. 本文针对这种信息覆盖能力,建立量化模型从多个角度考察搜集系统对 WWW 信息资源的覆盖程度. 文章首先分析了网页搜集不完全性的若干因素,在指出信息覆盖率的研究意义后提出了三类重要的信息覆盖率概念,然后围绕其中的数量和质量覆盖率展开研究工作. 在建立"采样 - 权值计算 - 验证"的覆盖率评测模型之后,以北大"燕穹"网页信息博物馆为考察对象并获得其网页数据,用不同的方式对中国 Web 进行采样;然后分别采用 PageRank 和 HITS 两种网页权值算法算出其中的重要网页作为样本,从量和质的角度考察"燕穹"系统的信息覆盖率,得到合理的数量和质量覆盖率值,从而验证了"燕穹"系统信息覆盖率结论的合理性和该信息覆盖率评测模型的可靠性.

关键词: 搜索引擎; 信息覆盖率; 采样; 权值计算; 验证; 数量覆盖率; 质量覆盖率
中图分类号: TP393 **文献标识码**: A **文章编号**: 0372-2112 (2003) 08-1168-05

An Evaluation Model on Information Coverage of Search Engines

MENG Tao, YAN Hong-fei, LI Xiao-ming
(Department of Computer Science & Technology, Peking University, Beijing 100871, China)

Abstract: Search engines usually get web pages by using links between them. With already massive and ever increasing of web pages, they can only crawl and index a portion of the whole web pages. A model to evaluate their information coverage percentages is presented. We analyze main factors why crawlers can't cover all web information, and put up three kinds of benchmarks to measure the coverage of a search engine. The paper gives out an evaluation model for two of three benchmarks as follows: First, sampling WWW to get many web pages, which are used to check the coverage percentage of quantity through generating random IPs or breadth first search. Second, selecting high-qualified pages as samples of important pages, by HITS or PageRank algorithms. Finally, we submit the samples to page database of search engines, and get the coverage percentage. In our research work, we get experimental data from WebInfoMall system of Peking University and compute the coverage percentages of quantity and quality. Using different sampling approaches and algorithms, we get the same results, which can prove our model is right and all the results are exact.

Key words: search engine; information coverage percentage; sampling; weight computing; check; quantity coverage percentage; quantity coverage percentage

1 引言

WWW(World Wide Web)自诞生以来,经过十多年的发展,已成为人类社会信息资源的重要组成部分,越来越多的信息实体选择 Web 作为载体. 当前,WWW 上约有 875 万个网站,25 亿静态网页,5500 亿动态网页,而且网页数正以每天 750 万的速度净增长[1,2,12]. 在中国国内 Web 上,截至 2002 年 1 月,约有 5343 万静态网页[4],277,100 个 Web 站点,CN 下注册域名 127,319 个[3].

搜索引擎一直是人们从浩瀚的网络资源中快速查找信息的重要工具之一. 优秀的搜索引擎总会尽量多的搜集 WWW 网页,但没有一个能搜尽所有网页. 当前搜集网页数量最多[8]的 Google 系统和 WiseNut 系统,搜到的网页数分别是 2,073,418,204[6]个和 1,571,413,207[7]个,不到总数的 80 %.

因此,研究搜索引擎的网页搜集系统对 WWW 信息资源的覆盖程度,是不断改进这种搜集性能的依据,也是评价搜索引擎整体性能的关键因素之一. 另一方面,基于搜索引擎技术开发的网络信息博物馆,旨在从 WWW 的角度记录人类社会的历史发展进程. 它借助网页搜集系统的工作,记录 WWW 在时间和空间两维上的每一点. 因此,研究信息覆盖能力对验证该博物馆网页资源的有效性有着决定性的意义.

本文的研究工作基于上述目的展开,针对北京大学计算

收稿日期:2002-08-12;修回日期:2003-02-20
基金项目:国家重点基础研究发展规划(973)项目(No. G1999032706);北京大学 985 项目

机网络与分布式系统实验室开发的天网 WWW 搜索引擎[5]及以此为基础的网上信息博物馆"燕穹"系统[9],创建量化模型采取多种方法从多个角度计算其对中国 Web 的信息覆盖率.

2　模型概述

2.1　网页搜集的不完全性

如果把网页看作一个节点,它的内容中存在其他网页的 URL,再把这种链接关系看作连接节点的边,则整个 WWW 构成了一张弱连通的有向图,如图 1 示:

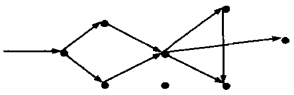


图 1　WWW 有向图结构

搜集系统就是以这种有向图结构为基础,从一个点出发循着有向边到达其他点,从而遍历 Web. 这种搜集机制存在下列问题导致无法遍历到所有节点:

® 结构缺陷:有些节点的入度为 0. 这类网页数量约占总体的 10 %[10].

® 有向路径缺失:从起始节点不存在到该点的链接路径. 只有约 21.3 %的点能作为起始点到达共约 90 %的节点[10].

® 资源限制:搜集系统的资源限制(磁盘容量和时间限量等)可能导致部分网页直到搜集过程终止都没有被搜集[11].

因此,搜集系统只能尽量搜集以接近实际网页总数. 信息覆盖率正是对这种接近程度的衡量.

2.2　几类重要的覆盖率

本文将静态网页定义为 URL 静态存在于其他网页中的网页,将动态网页定义为需要用户递交数据才能动态生成的网页,仅在前者范畴内考察搜集系统的信息覆盖率,它含有多层内涵:

数量覆盖率搜集系统搜到的网页在绝对数量上占 WWW 的比例,这是衡量搜集系统覆盖 WWW 信息能力的一个全局标准.

质量覆盖率 搜集系统搜到的重要网页占 WWW 中该类网页总量的比例. 当质量覆盖率很高时,它覆盖了当前社会信息资源在每个重要主题上在 WWW 上的投影.

可视信息覆盖率 由于搜集技术不成熟,搜集系统当前搜集存储的网页内容中相当一部分在日后将不可见,例如图片. 考察对这类信息的覆盖率,可得知多大比例能够在若干年后重新被浏览.

在本文的研究中,将对前面的两种进行详细的讨论和量化分析.

2.3　信息覆盖率评测模型

定义 WWW 为有向图 $G = (V, E)$,系统搜集到的网页集 $G' = (V', E')$ 是 G 的弱连通子图,数量覆盖率如式(1)所示:

$$C = |V'| / |V| \qquad (1)$$

其中,分子是搜到的网页数,已知;分母是 WWW 中的网页总数,未知. 获得 G 的子图 $G_0 = (V_0, E_0)$ 作为其样本,考察 V 与 V_0 交集中的点占 V_0 的比例 C_0,可作为 C 的近似值.

类似的,考察 WWW 中重要网页构成的子图 G_i 以及它的样本 S,检查属于 S 与 V 交集 S_0 的点占 S 的比例,即质量覆盖率 C_i 的近似值,如式(2)所示:

$$C_i = |S_0| / |S| \qquad (2)$$

因此,对 WWW 采样获得普通网页和重要网页两类样本,验证系统对其覆盖比例,就可得到数量和质量覆盖率,总的流程如图 2 所示,验证时要去除已不存在的网页.

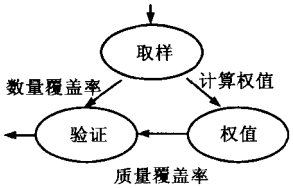


图 2　系统工作模型

3　数量覆盖率

对 WWW 采样的方法可有多种:基于全局角度,可用随机产生 IP 地址的方法;基于局部角度,本文提出了采用广度优先搜集采样的方法.

3.1　随机 IP 法

在 Edward T O 的工作[13]中,提出了通过随机产生 IP 地址对 WWW 采样的方法. 由此,本文在研究中将国内的所有 IP 分段表示成 $Ai.Bi.Ci.Di \Rightarrow Ai\ !\ Bi\ !\ Ci\ !\ Di$ 的形式,i 是分段的编号,假设第 i 个分段中的 IP 地址数为 Ti,找到从 IP 地址集到某个整数集合的可逆映射 F,使得位于第 t 个分段的 IP 地址 $a.b.c.d$ 的象如式(3)所示.

$$F(a.b.c.d) = \sum_{i=0}^{t-1} Ti + (a - At) * 256^3 + (b - Bt) * 256^2 + (c - Ct) * 256 + (d - Dt) \qquad (3)$$

表 1　选取不同数量中国随机 IP 地址并验证 HTTP 服务

编号	1	2	3	4	5	6	7	8	9	10
随机 IP 数	100K	200K	300K	400K	500K	600K	700K	800K	900K	1M
存在 http 数	95	172	252	336	418	478	570	652	717	817
覆盖 IP 数	4	7	9	13	16	21	28	36	43	47
扩展数量	1084	*	2584	*	3859	*	5484	*	*	8094
扩展后覆盖数	174	*	596	*	841	*	1204	*	*	1899

将 IP 地址映射到整数后,随机选取若干整数作逆映射便得到一个 IP 地址集. 由于约 93.6 %[14]的网站通过 80 端口提供 HTTP 服务,顺次扫描这些 IP 地址便得到一个存活的 IP 地址集合. 再经反向域名解析,就可得到一个 URL 集作为 WWW 网页样本.

用该方法对国内的所有 IP 地址作处理,然后对"燕穹"系统作验证,得到的统计结果在表 1 前三行中. 对数据用最小二乘法处理,得到数量覆盖率约为 5.7 %.

这种方法无法区分网站的大小和实际存在情况,误差很大. 改进途径是提取样本中网页包含的 URL 加入该样本(去除重复). 从表 1 中抽取 5 组数据,改进后的数据如表 1 后两行所

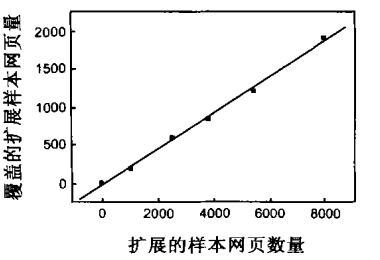


图 3　对随机 IP 地址数与存在数作线性拟合

示;再经线性拟合,结果如图 3 示,信息覆盖率约为 23.5 %.

3.2 广度优先法

为了使网页样本在逻辑结构上与 WWW 更接近,本文提出了一种用广度优先遍历进行采样的办法,它基于局部角度.

具体的步骤是:创建一个队列,反复针对队列头的 URL 搜集其对应网页,然后将头指针后移,新提取出的 URL 加入队列尾部(去除重复).当进入过队列的 URL 足够多时,验证得到它的一个存活子集即可作为 WWW 网页样本.

为减小误差,取多个起始 URL 得到多组样本,验证后的实验数据如表 2 所示,表中的五组覆盖率的均值和方差分别为 41.6 %和 0.0230;前者即为所测的数量覆盖率.

搜集系统通常是在 WWW 的最大弱连通子图内遍历,广度优先采样只能反映对这一子图的覆盖比例.由于这个子图包含 WWW 中约 90 %[10]的点,因此结论应乘以 90 %变成约 37.4 %才更准确.样本容量越大,覆盖率就越逼近此值,这从第 4 组大容量样本的结果可以看出.

表 2 广度优先取样所得到的数量覆盖率

样本编组	1	2	3	4	5
种子 URL	www.21cn.com	www.etang.com	net.cs.pku.edu.cn	www.pku.edu.cn	www.sina.com.cn
扩展 URL 数	66891	211629	154314	723866	174078
覆盖数量	26732	87562	67178	273066	74370
数量覆盖率	40.0 %	41.4 %	43.5 %	37.7 %	42.7 %

该结果之所以与随机 IP 法所测相差较大,是因为后者更容易得到一些循着链接无法到达的网页,导致结果比广度优先法低;但通过链接扩展,WWW 的逻辑结构在样本中得到反映,结果就会上升,这在文中做第一级扩展之后已经初步得到证实.

可以预测到,如果结合以上两种采样方法的优点,通过随机 IP 法产生 URL 集合作为广度优先法的种子 URL 集合,当样本容量足够大时,数量覆盖率值应该和上面的估计相符合,在 37.4 %附近.

4 质量覆盖率

考察搜集系统的质量覆盖率,需要获得一组重要网页样本.在本文的研究中,为保证样本的随机性和客观性,采取了对网页集做链接分析而获得重要网页的办法.

4.1 网页重要性评测方法

通常而言,可以从三个角度来分析网页的重要性(权值),讨论如下:

® 网页的 URL:网页的权值可从 URL 中得到体现.例如,URL 的域名越短,相对根目录的层次越浅,权值越大.

® 网页作为 WWW 结构的节点:网页 A 链向网页 B,通常表示 A 的作者认可 B.这种认可的增多意味着 B 权值上升.

® 网页的内容:可以对用户递交给搜索引擎的查询词与网页内容进行匹配,从满足用户关心度的角度分析网页的重要性.这种方法在引文[18]中有详细讨论.

Google 系统成功使用的 PageRank 算法[15]和 IBM 研究院提出的 HITS 算法[16]正是以上三种方法的综合利用,它们的有效性已经得到公认.其中,HITS 将网页权值分为目录型权值和权威型权值两种,前者衡量从它到达重要网页的可能性,后者衡量它包含重要信息的多少.基于这两种权值算法,本文提出了下面两类质量覆盖率确定方法.

4.2 广度优先法

PageRank 算法要求所处理网页集的链接结构与 WWW 类似,这在广度优先遍历的过程中伴随初始网页集的增大可得到满足.本文从数十万网页中用 PageRank 算法选取权值靠前的网页集作为重要网页样本.

假定网页 A 被 $T1\cdots Tn$ 链接到,$C(A)$ 是 A 的出度,根据 PageRank 算法,A 的权值如式(4)所示:

$$PR(A) = (1-d) + d\sum_{i=1}^{n} PR(Ti)/C(Ti) \quad [15] \qquad (4)$$

其中,d 是用户在 Ti 继续浏览的平均概率.构造矩阵 R,若网页 i 链到 j,则 $R[i,j] = 1/C(i)$,否则 $R[i,j] = 0$.设样本网页的权值为 W,那么式(4)可以写成:

$$W = d * W \times R + E \qquad (5)$$

因此,可用迭代法求 W 值:反复计算 $W(i+1) = d * W(i) \times R + E$,直到 W 收敛,E 为常数向量,d 实际为收敛因子.

根据上述原理,以在数量覆盖率计算中得到的五组样本作为初始样本,其链接关系以二元组〈$id1$, $id2$〉形式保存,用 PageRank 算法计算选取权值在前面约 5 %的网页作为重要网页样本对"燕穹"系统作验证,得到覆盖率数据如表 3 所示.五个质量覆盖率值的均值和方差分别是 48.0 %和 0.0579.在表中,重要网页所占比例之所以不一样,是因为对样本网页都作了存在性验证.

表 3 广度优先法初始取样经 PageRank 算法所得到的质量覆盖率

样本编组	1	2	3	4	5
种子 URL	www.21cn.com	www.etang.com	net.cs.pku.edu.cn	www.pku.edu.cn	www.sina.com.cn
初始扩展数量	66891	211629	154314	723866	174078
PageRank 取数	3523	8497	8702	33436	8408
所占比例	5.3 %	4.0 %	5.6 %	4.6 %	4.8 %
覆盖 URL 数	1542	4032	4754	17717	3456
质量覆盖率	43.8 %	47.5 %	54.6 %	53.0 %	41.1 %

鉴于样本网页的重要性缺乏严格评判标准,我们选取上述样本中的第 2、4 组,改变重要网页所占比例,得到质量覆盖率随之变化的情形如表 4 和 5 所示. 再对这两组数据作简单指数拟和,结果如图 4 所示. 可见,重要性标准越苛刻,重要网页样本容量越小,所测得质量覆盖率越高;反之越低,直到等于数量覆盖率.

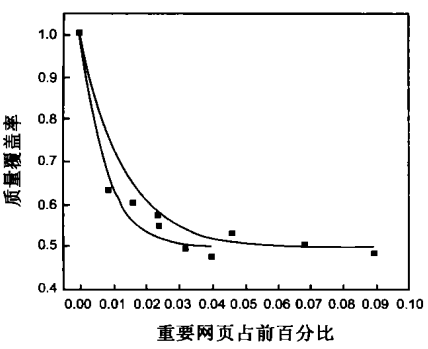


图 4　对第 2、4 组样本中重要标准与质量覆盖率数据作简单指数拟合

具体而言,当重要标准降至约占前约 5 %时候,两条曲线都开始逐渐变得平缓,因此该点的覆盖率值 48.0 %无疑最适合作为所测的合理搜集系统质量覆盖率值.

表 4　对第二组样本,改变重要性标准所得不同覆盖率数据

重要网页个数	1811	3434	5117	6808	8497
所占比例	0.86 %	1.62 %	2.42 %	3.22 %	4.02 %
覆盖网页数	1145	2068	2788	3363	4032
质量覆盖率	63.2 %	60.2 %	54.5 %	49.4 %	47.5 %

表 5　对第四组样本,改变重要性标准所得不同覆盖率数据

重要网页个数	17321	33436	49198	64450	79155
所占比例	2.39 %	4.62 %	6.80 %	8.90 %	10.9 %
覆盖网页数	9903	17717	24718	31126	36780
质量覆盖率	57.2 %	53.0 %	50.2 %	48.3 %	46.5 %

4.3　主题查询法

网页根据其内容而属于人类社会信息资源的某一主题类别,基于这一点,我们可以通过主题查询进行 WWW 网页采样,然后用 HITS 算法计算其中重要网页作为重要网页样本. 这类似于文献[18]中 HITS 算法选择网页集合所采用的方法.

具体步骤是:递交关于主题 T 的查询词 Q 给搜索引擎,得到其返回的网页集 R,再对 R 作扩展,加入其链入和链出的网页,得到网页集合 S. 而 S 中网页之间具有较强的链接关系[18],适合用 HITS 算法计算权值.

考虑 S 中某元素 P 的目录型权值 $H(P)$ 和权威型权值 $A(P)$,若 $F1, \cdots, Fm$ 链向 P, P 又链向 $T1, \cdots, Tn$,则根据 HITS 算法有:

$$A(P) = \sum_{i=1}^{m} H(Fi)\ ;\ H(P) = \sum_{i=1}^{n} A(Ti) \tag{6}$$

设 S 中网页的目录型权值为 H,权威型权值为 A,构成的有向图邻接矩阵为 M,那么上式(6)可以写成:

$$A = M^T \times H;\ H = M \times A \tag{7}$$

即有 $A = M^T \times M \times A$,可用 QR 算法[17]求 $M^T \times M$ 的特征向量 A. 本文选择了八个主题的查询词递交给天网搜索引擎,对应的 S 如表 6 所示. 在确定重要网页样本占 S 的比例界限时,我们选取 15 %,与搜索引擎响应查询返回的网页数量比例相当.

表 6　选取八个查询主题,通过主题查询法进行取样后所得数据

样本编组	1	2	3	4	5	6	7	8
查询词	北京大学	考研	股票	江泽民	Linux 教程	联想集团	三个代表	世界杯
返回数量	1802	1802	1802	1802	1355	1802	1802	1802
扩展数量	11667	19379	13403	11006	26548	15498	11608	20823

对具有这两类较高权值的网页,实验数据分别如表 7 和 8 所示,八组样本的平均质量覆盖率分别为 43.0 %和 52.9 %,方差分别为 0.106 和 0.1269.

表 7　对目录型权值较高的重要网页进行取样并验证 WebInfoMall 质量覆盖率

样本编组	1	2	3	4	5	6	7	8
Hub 取数	1671	1701	1532	1040	1508	1428	1859	1961
覆盖数量	958	956	578	312	772	514	793	651
覆盖率	57.3 %	56.2 %	37.7 %	30 %	51.2 %	36.0 %	42.7 %	33.2 %

表 8　对权威型权值较高的重要网页进行取样并验证 WebInfoMall 质量覆盖率

样本编组	1	2	3	4	5	6	7	8
Auth 取数	1400	1572	1124	930	1197	1047	1581	1985
覆盖数量	943	930	385	301	692	631	927	1052
覆盖率	67.4 %	59.2 %	34.3 %	32.4 %	57.8 %	60.3 %	58.6 %	53.0 %

在 HITS 算法中,所有链接被视为平等,而实际并非如此,这从它的 AnchorText 与查询词的匹配度可以体现,该原理在 Soumen 和 Byron 的工作[19]中有论述.

若 A 链向 B, $F(A, B)$ 是 A 中 B 的 Anchor Text 与 Q 的匹配度,则 HITS 算法可修正为:

$$A(P) = \sum_{i=1}^{m} H(Fi)\ F(Fi, P) \tag{8}$$

$$H(P) = \sum_{i=1}^{n} A(Ti)\ F(P, Ti) \tag{9}$$

用这种改进后的 HITS 算法对 S 进行处理,求得两种质量覆盖率均值分别为 46.2 %和 50.3 %. 从该结果看到,广度优先法和主题查询法所求得的质量覆盖率几乎完全符合,"燕穹"系统的质量覆盖率约为 50 %. 如果提高重要网页的标准,该值会更高.

5　结论

本文针对搜索引擎搜集子系统对 WWW 的信息覆盖能力,建立了一个信息覆盖率的量化研究模型. 在这个模型中,提出两套 WWW 网页的采样办法,采取了两种典型的网页权值算法,分别从量和质的角度计算搜集系统的信息覆盖率.

运用这个模型,我们针对中国的 Web 进行采样,对北大

“燕穹”网页信息博物馆所存储的国内网页数据作评估.结果显示,它在数量上约覆盖了国内网页总数的 37 %,而在质量上约覆盖了总数的 50 %.

对于相同的覆盖率,采用不同的取样和权值计算方法,得到的结果能够很好的符合,证明了该覆盖率模型的正确性和所得“燕穹”博物馆信息覆盖率的准确性.

本文不足之处在于网页重要的标准不够严密.如文中所述,在实验中通过改变样本网页的重要性标准得到可靠的质量覆盖率数据,这也是对此不足的一些修正.

参考文献:

[1] OCLC Online Computer Library Center, Inc. Web Characterization [DB/OL]. http://wcp.oclc.org/stats/size.html. 2002 - 08 - 10.

[2] Cyveillance, Inc. White Papers. Sizing the Internet [DB/OL]. http://www.cyveillance.com/web/corporate/white_papers.htm. 2000 - 07 - 10.

[3] 中国互联网络信息中心.第九次中国互联网络发展状况统计报告 [R]. 北京,中国,2002 - 01.

[4] 闫宏飞.关于中国 Web 的大小、形状和结构 [J].计算机研究与发展,2002,38(8):1 - 10.

[5] http://e.pku.edu.cn/ [Z/OL].

[6] http://www.google.com/ [Z/OL]. 2002 - 05.

[7] http://www.wisenut.com/ [Z/OL]. 2002 - 05.

[8] Greg R Notess. Search Engine Statistics: Database Total Size Estimates [DB/OL]. http:// www.searchengineshowdown.com/stats/sizeest.shtml,2002 - 03.

[9] http://net.cs.pku.edu.cn/~webg/infomall/ [Z/OL].

[10] Andrei Broder, Ravi Kumar, Farzin Maghoul. Graph Structure in the Web [DB/OL]. http:// www9.org/w9cdrom/160/160.html [Z/OL].

[11] Junghoo Cho, Hector Garcia-Molina, Lawrence Page. Efficient crawling through URL orering [A]. Proc of the 7 Int'l Conf on the World Wide Web [C]. Australia, 1998.

[12] Bright Planet. The Deep Web: Surfacing Hidden Value [DB/OL]. http://www. brightplan etcom/deepcontent/tutorials/DeepWeb/index.asp,2002 - 08.

[13] Edward T O'Neil, Patrick D McClain. A Method for Sampling the World Wide Web [DB/OL]. http://net.cs.pku.edu.cn/~webg/refpaper/html/Sampling%20Methodology.htm.

[14] Woodruff A, P Aoki. An investigation of documents from the World Wide Web [A]. Proc of Fifth International World Wide Web conference [C]. France, 1996.

[15] Sergey Brin, Lawrence Page. The anatomy of a large-scale hypertextual Web search Engine [A]. Proc of the 7 Int'l Conf on the World Wide Web [C]. Australia, 1998.

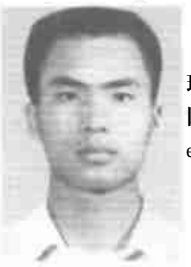
[16] Members of the Clever Project. Hypersearching the Web [J]. Scientific American, http://www.sciam.com/issue.cfm? 1999 - 06.

[17] 徐萃薇.计算方法引论 [M].北京:高等教育出版社,1985.

[18] Jon M Kleinberg. Authoritative sources in a Hyperlinked Environment [J]. Journal of the ACM 46, 1999.

[19] S Chakrabaryi, B Dom. Automatic resource compilation by analyzing hyperlink structure and associated text [A]. Proc of the 7 Int'l Conf on the World Wide Web [C]. Australia, 1998.

作者简介:



孟 涛 男,1980 年出生于湖北省公安县,现为北京大学计算机系博士研究生,研究方向为网络与分布式处理. Email: mengtao@net.cs.pku.edu.cn.

闫宏飞 男,1973 年出生于黑龙江省哈尔滨市,2002 年在北京大学计算机系获得博士学位,现为北京大学讲师,研究方向为网络与分布式处理. Email: yhf@net.cs.pku.edu.cn.

李晓明 男,1957 年出生于湖北省荆州地区,现为北京大学教授,博士生导师,研究方向为并行与分布式处理、Internet 与 Web 技术. Email: lxm@pku.edu.cn.