第 41 卷 第 1 期
2004 年 1 月

计 算 机 研 究 与 发 展
JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT

Vol. 41 ,No. 1
Jan. 2004

加速评估算法 :一种提高 Web 结构挖掘质量的新方法

张 岭　　马范援
(上海交通大学计算机科学与工程系　上海　200030)
(zhangling @sjtu. edu. cn)

摘　要　利用 Web 结构挖掘可以找到 Web 上的高质量网页 ,它大大地提高了搜索引擎的检索精度. 目前的 Web 结构挖掘算法是通过统计链接到每个页面的超链接的数量和源结点的质量对页面进行评估 ,基于统计链接数目的算法存在一个严重缺陷 :页面评价两极分化. 一些传统的高质量页面经常出现在 Web 检索结果的前面 ,而 Web 上新加入的高质量页面很难被用户找到. 提出了加速评估算法以克服现有 Web 超链接分析中的不足 ,并通过搜索引擎平台对算法进行了测试和验证.

关键词　Web 结构挖掘; *PageRank* ;信息检索 ;搜索引擎 ;加速评估算法

中图法分类号　TP391; TP393

Accelerated Ranking : A New Method to Improve Web Structure Mining Quality

ZHANG Ling and MA Fan-Yuan
(Department of Computer Science & Engineering, Shanghai Jiaotong University, Shanghai 200030*)*

Abstract　A Web structure mining technique helps to find high quality web pages on the Web, and it improves the Web search precision. The current Web hyperlink analysis algorithm ranks each web page based on the hyperlinks pointed to it and on the quality of the source nodes. This approach may cause the polarization towards rank values, and some newly added high quality web pages are difficult to appear in the top of the Web search results. To fix the drawback a new accelerated ranking algorithm is proposed and also tested on the search engine platform.

Key words　Web structure mining; PageRank; information retrieval; search engine; accelerated ranking algorithm

1　引　　言

Web 结构挖掘(也称 Web 链接分析)是 Web 挖掘研究中的一个重要组成部分 ,它通过分析 Web 上超链接结构以找到有价值的网页 ,与传统基于内容相似性进行页面质量分析的方法相比 ,它提高了 Web 检索的精确度. Google[1]搜索引擎所采用的 *PageRank*[2]算法就是一种典型的 Web 结构挖掘算法 ,它递归地计算每个页面的被链接数和链接其他页面数(即 Web 图结构中页面的入度和出度) ,然后给每个页面计算出一个 *PageRank* 值 ,该值代表了一个页面的重要程度. 虽然这种方法提高了搜索引擎的检索质量 ,但它导致了检索的两极分化现象 ,妨碍了新加入到 Web 上高质量网页内容的传播 ,因为新的页面由于链接数目的限制很难出现在搜索结果的前面 ,从而很难被检索用户发现.

为此 ,我们首次提出了一个加速评估算法 ,它有效地解决了以 *PageRank* 为代表的超链接分析算法存在的不足 ,使得 Web 上有价值的内容可以更快的速度传播 ;相反 ,一些已经陈旧的数据的页面评估值也将加速下滑. 这种算法保证了 Web 信息检索中的

收稿日期:2002 - 07 - 04 ;**修回日期**:2003 - 06 - 25
基金项目:上海市基础研究重点基金项目(02DJ14045)

优胜劣汰,确保向用户提供高质量的 URL 链接. 本文将以 *PageRank* 算法为例,介绍加速评估算法的原理和实现.

2　Web 结构挖掘算法

Web 结构挖掘把整个 Web 看做是一个巨大的有向图,每个页面是图中的一个结点. 如果页面 u 包含一个指向页面 v 的超链接,即存在 $link(u, v)$. 这里页面间的超链接构成了一个有向图 G:对于每个页面构成有向图 G 的一个结点;当且仅当 u 中包含指向页面 v 的超链接时,存在着从 u 到 v 的有向边(u, v). 有向图 G 如图 1 所示:

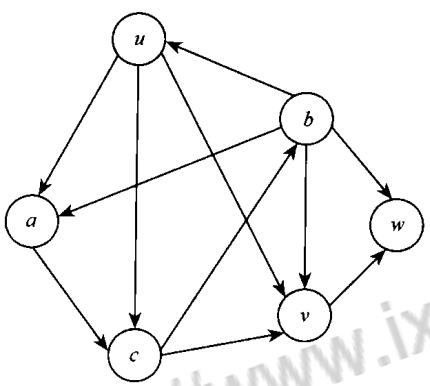


图 1　有向链接图 G

对于结点 v 来说,结点 b, c, u 对于 v 的权值大小有贡献,因为这 3 个结点都存在到 v 的有向边. 指向某一结点的有向边越多,其结点(页面)质量越高. 这种算法的主要缺点是仅仅考虑了链接数量,即所有的链接都是等价的,而没有考虑源结点自身质量的高低. 事实上,Web 上高质量的页面中往往包含有高质量的链接,源结点的质量对于被链接文档质量评价的作用往往高于数量上的影响.

为了解决链接数量和源结点的质量问题,斯坦福大学的 Brin 和 Page 提出了 *PageRank* 算法:某一 Web 文档的 *PageRank* 值等于所有包含指向该文档的源文档的 *PageRank* 值与该文档内超链接总数之比的总和[3],即

$$PR(v) = \varepsilon/n + (1 - \varepsilon) \times \sum_{(u,v)\in G} PR(u)/outlink(u), \qquad (1)$$

其中,ε 是衰减因子,一般取 0.1～0.2; n 为有向图 G 中结点的数量; $outlink(u)$ 为结点 u 包含的超链接数量.

PR 算法的收敛速度较快,对于图 1 的计算,其迭代方程组的解在 15 次迭代之后趋于稳定. 在实际的大规模 Web 页面的计算中,收敛过程也可以在不超过 100 次的递归内完成[4].

PageRank 的本质体现了浏览 Web 的"随机访问模型"(random surfer mode)[5],即用户随机地从某个页面出发,下一个页面被点击的概率即该页面的 *PageRank* 值.

3　加速评估算法

3.1　PageRank 算法的缺陷

通过第 2 节的分析可以看到,Web 结构分析算法如 *PageRank* 等的实质是统计每个页面的有效被链接数,这个页面的评估值与且只与被链接数相关. 比如 Yahoo, MSN 因为有很多网页通过超链接指向了它们,所以这些都具有较高的 *PageRank* 值;而一些小型网站或个人网页因为被链接数目有限,所以它们的 *PageRank* 值较低. 我们注意到这样一个事实:不为大多数人所知的网页完全可能包含极有价值的内容. 加速评估算法正是帮助这些"弱者"提升自己的重要性,用主动的方式加速有价值的网上内容的传播.

目前,一些研究人员也提出了一些改进传统链接分析算法的新思路,比如北京大学计算机系提出利用 HTTP 协议,记录每个页面最近一次的修改时间,在运行分析算法的时候把页面修改时间作为控制参数,给予新修改的页面以较高的权值,而给予老页面以较低权值[6]. 这种方法的主要问题在于,目前大量的网站采用的是动态页面,如 asp, php, jsp 等,它们往往由后台数据库驱动并动态生成. 每个动态页面都是经嵌入脚本运行后输出,它们的时间戳(Last Modifid Timestamp)都是当前时间,所以这种方法有很大的局限性.

3.2　加速评估算法

对于一个搜索引擎来说,一般都是利用 Web Crawler(或称为 Robot、爬行器等)遍历并下载部分或整个 Web. 为了保证索引内容的更新,对于一个特定的 Crawler 来说,它完成一次较完整的页面下载所需的时间一般在半个月(如 Google)到 1 个月不等[7]. 然后索引器对本次下载的数据集进行索引并利用超链接分析算法计算每一个 URL 的 *rank* 值. 然后 Crawler 循环进行下一次的 Web 遍历. 这样,我们可以得到多个数据集,每个数据集中的 URL 由于时间的推移导致它在 Web 上的访问和引用情况也在发生着变化. 加速评估算法的核心思想就是

通过分析基于时间序列的评价值变化情况,预测 URL 在未来一段时期内的期望值并把它作为搜索引擎提供检索服务的有效参数. 用户检索时,搜索引擎将按照预测的 *rank* 值的高低决定一个 URL 在检索结果中的位置.

例如,$D1$ 是 Crawler 开始记录的第 1 次下载的数据集,它包含了 $m1$ 个页面:

$$D1 = \{d_1, d_2, \cdots d_i \cdots d_{m1}\},$$
$$i = 1, 2 \cdots m1,$$

每个页面的 *PageRank* 值组成了另外一个数据集 P,它也包含了 $m1$ 个值:

$$P1 = \{p1_1, p1_2, \cdots p1_i \cdots p1_{m1}\},$$
$$i = 1, 2 \cdots m1,$$

即对 $D1$ 内的页面 d_i,它的 $PageRank(d_i) = p_i$.

对 Crawler 第 j 次下载的数据集 Dj,由于网络连接和 Crawler 爬行策略等原因,它包含的文档数目不一定等于 m,我们设它为 mj 个,有:

$$Dj = \{d_1, d_2, \cdots d_i \cdots d_{mj}\},\ i = 1, 2 \cdots mj,$$
$$Pj = \{pj_1, pj_2, \cdots pj_i \cdots pj_{mj}\},\ i = 1, 2 \cdots mj.$$

考察某一特定页面 d_i 在多个数据集内(设数据集的个数为 n)的 *PageRank* 值,则构成了一个集合 Pi:

$$Pi = \{p1_i, p2_i, \cdots pj_i \cdots pn_i\},$$

Pi 表示了一个文档在某个时间序列内的 *PageRank* 的变化情况,通过分析其变化趋势可以知道该页面在 Web 上重要程度的变化.

加速评估算法(accelerated ranking,AR)对基于时间序列的 URL 的 *AR* 值进行线性拟合,拟合出直线的斜率代表了该 URL 的未来趋势:斜率为正,表示该 URL 重要性在增加;斜率为负则相反. 斜率的绝对值越大表明该 URL 的重要性变化越剧烈. *AR* 算法计算出所有 URL 对应的线性拟合直线斜率,并用下一个统计点的 *PageRank* 值作为当前有效的评估值. *AR* 算法的实质是给予在 *AR* 值呈上升态势的 URL 以额外的奖励;对 *AR* 值呈下降趋势的 URL 以额外的惩罚. 它促进并加速了网络上优质内容的传播和网络内容的优胜劣汰,克服了 *PageRank* 算法对新内容的迟钝性,提高了搜索引擎信息检索的质量.

一个 URL 的 *AR* 值我们定义为

$$AR = PR \times sizeof(D), \qquad (2)$$

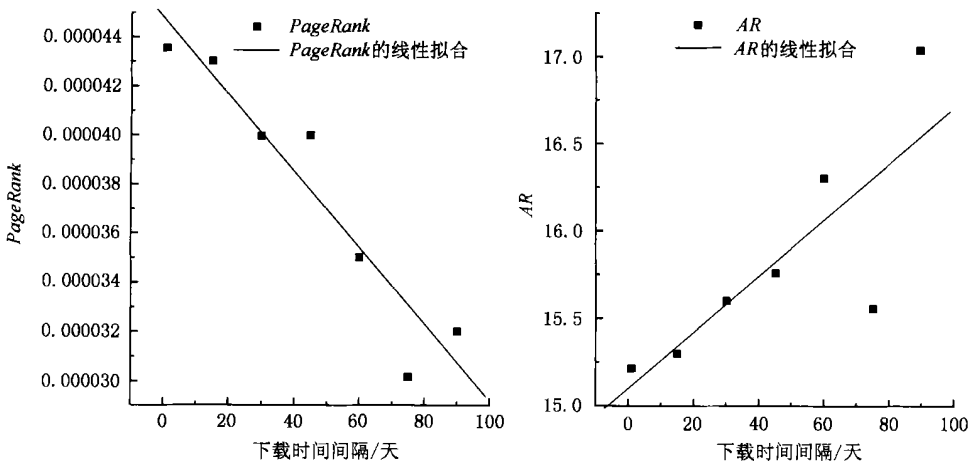
其中 *PR* 是 URL 的 *PageRank* 值,*sizeof*(*D*) 是 Web 文档集页面的总量. 之所以把 *AR* 定义为式(2)中的表达式,是因为 *PageRank* 算法的归一性决定了 *PR* 值的相对性,即在不同集合中 *PR* 的大小不能决定排序位置而必须要考虑到集合的大小. 如果不使用 *AR* 而仅仅用 *PageRank* 值的变化趋势进行拟合,可能产生错误结果. 表 1 是一个 URL 在 7 个文档集中的样本数据:

表 1 某 URL 在多个集合中 PageRank 值比较

下载间隔/天	页面数量	*PR*(*PageRank*)
1	349866	4.35E-5
15	355800	4.30E-5
30	391221	3.99E-5
45	394587	4E-5
60	465822	3.50E-5
75	516958	3.02E-5
90	532458	3.20E-5

图 2(a)是用 *PR* 值对表 1 的数据进行拟合的曲线. 图 2 中横坐标是每次 Crawler 下载的数据集的日期,图 2(a)中每个数据集采样的间隔是 15,图中直线是 7 次采样数据的线性拟合. 从图 2 中可以看出,该 URL 的 *PageRank* 呈下降趋势表示该 URL 的重要程度相对下降,但事实上可能并非如此. 在第 2 节中我们提到 *PageRank* 算法符合随机访问模型,即 $\sum_{i=1}^{n} PR(d_i) = 1$. 当 Web 文档集包含页面数量少时,*PR* 的均值较大;而当页面数量很大时,*PR* 的均值会很小. 因为 *PR* 的下降可能伴随着 URL 总数的增加,所以单纯比较 *PR* 的大小不能正确体现出任意一个 URL 在 Web 中的地位变化情况. 当然,对于采用了限制性爬行策略的一些特定领域的 Crawler,它们各次收集到的 Web 页面数量差别不大,可以仅使用 *PR* 进行预测.

根据式(2),我们用 *AR* 值重新对图 2(a)进行拟合得到图 2(b). 可以看出,虽然该页面的 *PageRank* 值逐渐降低,但实际上这是因为文档集中页面数量的增加所导致,该 URL 的重要程度实际上是呈献了增长的态势. 一个 URL 的 *AR* 值真实地反映了这个页面在整个 Web 文档集合中的重要性.



(a) 使用*PR*值进行线性拟合　　(b) 使用*AR*值进行线性拟合

图 2　某 Web 页面 *PR* 和 *AR* 变化趋势图

3.3　用 AR 加速 PageRank

通过计算表 1 对应页面的 *AR* 值,可以看出,该 URL 在 90 天内其重要程度在不断上升,它的上升态势可能符合不同的函数曲线,我们使用了线性回归的方法对它进行了近似处理,结果如表 2 所示:

表 2　线性拟合出的斜率和误差

参数	值	误差
A	15.10057	0.28675
B	0.01605	0.0053

其中,*B* 为该拟合曲线的斜率:

$$AR = A + B \times Day. \qquad (3)$$

根据式(3)可以计算出在距离第 1 次下载 105 天后,其 $AR = 16.78582$.

我们定义,加速后的 *PageRank* 表达式可以写做:

$$PR_{\text{accelerate}} = \frac{AR_{\text{last}} + B \times D}{M_{\text{last}}}, \qquad (4)$$

其中,AR_{last}是 URL 最近一次的 *AR* 值,*B* 在式(3)中所定义,对于 $B < 0$ 的 URL,其 *PageRank* 值实际上被加速削减了. *D* 为离最近一次页面下载的时间间隔天数(可以根据需要自行定义,一般可取多次下载间隔的平均值),M_{last}是最近一次下载的文档集内文档的数目. 式(4)即为 *AR* 算法的实现表达式,式中参数 *B* 是决定对链接值进行"奖励/惩罚"的关键参数,*B* 值是通过对多个检索集合链接分析后利用线性回归计算得到的.

根据式(4),可以计算出表 1 的 URL 的 *PageRank* 经加速后的值:

$$PR_{\text{accelerate}} = \frac{532458 \times 3.20E^{-5} + 0.01605 \times 15}{532458} = 3.245E^{-5}.$$

对搜索引擎来说,可以把这个预测的 *PageRank* 值当成当前 Web 检索服务里计算页面相关度的参数. 这样,一些有良好发展趋势的网页会更快地出现在搜索结果中.

需要注意的是,用加速评估算法计算出新的 *PageRank* 值可以集成在搜索引擎的检索算法内. 但该修正值在下一次重新进行加速评估的时将不作为评估参数而直接使用加速前的原始数据,这样做的目的是避免多次线性回归产生的误差积累.

4　实　　验

我们利用自己开发的一个分布式 Web Crawler 对 CERNET 内的网页进行了下载. 这个分布式 Crawler 系统由多台机器组成,1 台作为调度器负责 URL 的分配和路由. URL 根据 Hash 函数分配到各 Crawler 机器进行下载. Crawler 与调度器之间利用 Java RMI 协议进行通信. 由 4 台 Crawler 和一个调度器组成的集群在 7 个小时可以下载超过 860000 个网页. 根据 Google 搜索引擎的 *PageRank* 算法的定义,我们重新设计并实现了该算法.

我们于 2002 年 1 月、3 月和 4 月用 Crawler 对 CERNET 进行了 3 次下载. 表 3(a)的左栏是 1 月份用 *PageRank* 算法计算后得到的 *PageRank* 最高的 10 个页面,其余的 3 张图表是这 10 个页面的权重在后几个月的变化情况. 表 3(b)的右栏是使用加速评估算法对第 3 次下载 30 天后的各页面权重的预测,可以发现预测结果和前 3 次实际计算得到的

PageRank 值有所不同,一些 URL 的 *PR* 值被增加或减少了. 对于搜索引擎来说,它在 4 月份实际使用的是经 *AR* 算法预测后的数值进行索引并提供检索服务.

表 3 三次 Web 下载统计结果及用 AR 算法加速后结果

3(a) 使用 AR 算法前

2002.1 Top 10

URL	*PageRank*
http://www.nic.edu.cn	6.05014E-4
http://www.edu.cn	3.26011E-4
http://www.casbic.ac.cn/English.htm	6.48627E-5
http://www.peopledaily.ac.cn	5.01017E-5
http://www.edu.cn/cernetnews	4.49945E-5
http://book.student.tsinghua.edu.cn	2.10553E-5
http://www.pku.edu.cn	1.91704E-5
http://leaf.sdau.edu.cn	1.50519E-5
http://www.sjtu.edu.cn	1.35486E-5
http://www.tsinghua.edu.cn	1.04650E-5

2002.3

URL	*PageRank*
http://www.nic.edu.cn	6.55010E-4
http://www.edu.cn	2.99289E-4
http://www.edu.cn/cernetnews	4.89475E-5
http://www.casbic.ac.cn/English.htm	2.75564E-5
http://www.pku.edu.cn	2.65906E-5
http://www.sjtu.edu.cn	2.26862E-5
http://www.peopledaily.ac.cn	5.40481E-6
http://www.tsinghua.edu.cn	5.11055E-6
http://leaf.sdau.edu.cn	1.09293E-6
http://book.student.tsinghua.edu.cn	N/A

3(b) 使用 AR 算法后

2002.4

URL	*PageRank*
http://www.nic.edu.cn	6.38536E-4
http://www.edu.cn	1.87055E-4
http://www.casbic.ac.cn/English.htm	4.89475E-5
http://www.edu.cn/cernetnews	4.75720E-5
http://www.sjtu.edu.cn	2.31765E-5
http://www.tsinghua.edu.cn	6.48964E-6
http://www.peopledaily.ac.cn	6.44577E-6
http://www.pku.edu.cn	5.63399E-6
http://leaf.sdau.edu.cn	N/A
http://book.student.tsinghua.edu.cn	N/A

2002.5(*AR* 预测)

URL	*PageRank*
http://www.nic.edu.cn	6.17498E-4
http://www.edu.cn	1.26835E-4
http://www.edu.cn/cernetnews	4.60463E-5
http://www.casbic.ac.cn/English.htm	3.74965E-5
http://www.sjtu.edu.cn	2.57772E-5
http://www.tsinghua.edu.cn	4.41979E-6
http://www.pku.edu.cn	1.09126E-6
http://www.peopledaily.ac.cn	-1.17895E-5
http://leaf.sdau.edu.cn	-1.54416E-5
http://book.student.tsinghua.edu.cn	N/A

表 3 中 *PageRank* 栏中"N/A"表示该 URL 未出现在对应的下载数据集中(可能由于 WWW 服务器关闭或其他原因). 从表 3(b)的右栏可以看到,某些 URL 的 *PageRank* 值经 *AR* 算法后其 *PR* 值为负. 虽然负值不符合实际的页面权值,但因为存在着大量页面 *PR* 值的大幅变化,所以负 *PR* 值在相互对比中还是具有参考价值的.

5 结 论

基于超链接分析的方法提高了 Web 搜索的精度,目前已有的算法可以比较好地完成 Web 资源的评定. 同时,单纯使用链接分析方法会带来 Web 文档评估两极分化的现象. 针对 Web 超链接分析中出现的页面两极分化现象,本文提出了基于时间序列分析的加速评估算法. 该算法可以帮助用户逐渐地发掘出有价值的 Web 文档,促使有价值的内容在 Web 上的快速传播,改善了基于 Web 的数据管理.

参 考 文 献

1 J Cho, H Garcia-Molina, L Page. Efficient crawling through URL ordering. The 7th World Wide Web Conference, Brisbane, 1998

2 S Brin, L Page. The anatomy of a large-scale hypertexual web

search engine. The 7th World Wide Web Conference, Brisbane, 1998
3 Taher H Haveliwala. Efficient computing of PageRank. Stanford Database Group, Tech Rep, 1999
4 Monika Henzinger. Link analysis in web information retrieval. IEEE Data Engineering Bulletin, 2000, 23(3): 3～8
5 Dell Zhang, Yisheng Dong. An efficient algorithm to rank web resources. Computer Netwoks, 2000, 33: 449～455
6 Lei Ming, Wang Jianyong *et al*. Improved relevance ranking in web gather. Journal of Computer Science and Technology, 2001, 16(5): 410～417
7 S Lawrence, C L Giles. Accessibility of information on the web. Nature, 1999, 400: 107～109

张　岭　男,1973 年生,博士研究生,主要研究方向为智能信息检索、Web 挖掘、搜索引擎等.

马范援　男,1942 年生,教授,博士生导师,上海交通大学电子商务研究与开发中心副主任,主要研究方向为 Internet 信息获取技术、网络信息挖掘、电子商务.

第九届中国机器学习会议
2004 年 10 月 22～24 日,上海
The 9th China Conference on Machine Learning
October 22～24, 2004, Shanghai

http://www.cs.fudan.edu.cn/ccml2004

第九届中国机器学习会议(CCML2004)由中国人工智能学会机器学习专业委员会和中国计算机学会模式识别与人工智能专业委员会联合主办,复旦大学和上海海运学院联合承办. 该系列会议每两年举行一次,现已成为国内机器学习界最主要的学术活动之一. 此次会议将为机器学习及相关研究领域的学者交流最新研究成果、进行广泛的学术讨论提供便利,并且将邀请国内机器学习领域的著名学者做精彩报告.

征稿范围(不仅限于如下主题)

- 机器学习的新理论、新技术与新应用
- 人类学习的计算模型
- 计算学习理论
- 监督学习
- 非监督学习
- 强化学习
- 多示例学习
- 半监督学习
- 集成学习
- 多策略学习
- 基于案例的推理
- 增量学习与在线学习
- 对复杂结构数据的学习
- 增强学习系统可理解性
- 数据挖掘与知识发现
- 神经网络
- 神经网络集成
- 进化计算
- 人工生命
- 模糊集与粗糙集
- 多 Agent 系统中的学习
- 模式识别
- 信息检索
- 生物信息学
- 语音、图像处理与理解
- 自然语言理解

投稿要求

- 论文必须未公开发表过,一般不超过 6000 字;中、英文稿均可接受;
- 论文应包括题目、作者姓名、作者单位、摘要、关键字、正文和参考文献;另附作者地址、邮编、电话或传真及 E-mail 地址;
- 参选优秀学生论文的稿件请注明(须由在校博/硕士生或本科生)为第一作者;
- 会议鼓励电子投稿,也可邮政投稿:
  若电子投稿,请将 Word、PS 或 PD 格式的文件发到:sgzhou@fudan.edu.cn(超过 1M 的文件请先压缩;请注意接收会议组织机构发出的收稿确认电子邮件);若邮政投稿,请将 3 份打印稿于截稿日期前寄达:上海市邯郸路 220 号复旦大学计算机科学与工程系　周水庚收　邮编:200433.

论文出版

所有录用论文将在《复旦大学学报》(自然科学版)正刊发表. 会后将根据论文及报告质量评选出优秀论文,其中一部分在国际刊物 Asian Journal of Information Technology 的 Special Issue of Selected Papers of CCML '04 发表(中文稿需译为英文,所有选中的稿件都需进行必要的扩展),另一部分在《模式识别与人工智能》正刊发表. 会议还将评出 3 篇优秀学生论文,颁发证书并给予奖励.

重要日期

全文投稿:2004 年 3 月 10 日　　录用通知:2004 年 5 月 10 日　　修改定稿:2004 年 7 月 10 日

阅读此文的还阅读了：

1. 一种基于云计算的Web结构挖掘算法
2. Web结构挖掘算法探讨
3. 一种基于会话聚类算法的Web使用挖掘方法
4. 专业网站零输入导航引擎设计与实现
5. 一种无冗余的Web日志挖掘算法
6. 教育部评估专家冀望我校：深入挖掘研究梅山文化内涵、特色，形成学校人文品牌
7. 深入挖掘教材 提高“纲要”课教学的吸引力
8. 浅谈项目工程质量管理之组织手段利器
9. NPLWAP:一种新的Web序列模式挖掘算法
10. Dijkstra算法在Web结构挖掘的应用
11. 结构挖掘中web有向图模型的改进算法
12. Trawling算法在Web结构挖掘中的应用
13. Web结构挖掘及其算法分析
14. Web结构挖掘算法研究
15. 房地产项目评估新方法探讨
16. 新课标下课程资源处理的艺术

17. Web结构挖掘及HITS算法分析

18. 普适评估算法模型的实现

19. 数据挖掘在医院智能审计与监管平台系统中的应用

20. 以评估为契机,提高高职教学质量

21. 基于Web结构挖掘算法的网站构建

22. Web结构挖掘

23. 打造钢厂特色走质量效益型发展道路

24. 提高农村小学阅读教学质量的探讨

25. 理顺IT:挖掘信息化的最大收益:“信息化应用评估”沙龙在铁山坪举行

26. 计算机行业挖掘结构性机会

27. 法医DNA分析技术新方法与新产品的评估

28. 结构挖掘中web有向图模型的改进算法

29. 基于Web页面链接结构的挖掘算法

30. 一种挖掘Web用户访问模式的新方法MFP

31. 制造纳米硅有新方法(美国)

32. 配电自动化条件下配电系统供电可靠性评估

33. 煤矿安全培训质量的评估与提高的探讨

34. 决策树算法在天气评估中的应用

35. 应用Web结构挖掘的PageRank算法的改进研究

36. 挖掘隐含 找回漏解

37. Web结构挖掘中HITS算法的改进

38. 挖掘现有设备潜力,提高热电经济效益

39. 开发校本课程提高小学生作文能力

40. 一种改进的web挖掘聚类算法

41. 前程锦绣:好莱坞大片与其经济

42. 基于Web结构挖掘算法的网站构建

43. 浅论民族地区基层文化站何以彰显民族特色文化

44. 一种Web流频繁模式挖掘算法

45. 加速评估算法:一种提高Web结构挖掘质量的新方法

46. 改革课堂教学结构 加强知识发生过程的教学——成人教育中试行“数学导学单元教学法”初探

47. 基于PageRank和HITS的Web结构挖掘算法研究

48. 基于Web结构挖掘的HITS算法研究

49. 让学生养成质疑的好习惯

50. 打造钢厂特色 走质量效益型发展道路