

文章编号:1007-130X(2008)03-0004-03

主题搜索引擎中网络爬虫的搜索策略研究*

Research on the Search Strategy of the Web Spider in Topic-Oriented Search Engines

李 勇, 韩 亮

LI Yong, HAN Liang

(大连海事大学计算机科学与技术学院, 辽宁 大连 116026)

(School of Computer Science and Technology, Dalian Maritime University, Dalian 116026, China)

摘 要:本文对主题搜索引擎中的网络蜘蛛搜索策略进行了详细的分析,在深入分析主题页面在 Web 上的分布特征与主题相关性判别算法的基础上提出了一个面向主题搜索的网络蜘蛛模型,对模型的组织结构进行了详细阐述。作为主题网络蜘蛛搜索策略的核心部分,主题相关性判断算法是网络蜘蛛能够围绕设定主题进行聚焦检索的关键。在 URL 的主题相关性判别过程中引入了链接文本及相关链接属性分析,提出了一种新颖的 URL 主题相关性算法—EPR 算法。

Abstract: Based on our in-depth research on the search strategy in topic-driven engines and topic dependency judgement algorithms, this article presents a design model of the topic-oriented web spider and analyzes its structure in detail. As the key component of the search strategy for the topic-oriented web spider, the topic dependency judgement algorithms ensure a focused web crawling process of the spider. In the process of the dependency judgement between URLs and topics, a novel URL pruning algorithm called EPR is presented based on an analysis of the anchor text and the related properties.

关键词: 搜索引擎; 网络蜘蛛; 搜索策略; 主题提取

Key words: search engine; Web spider; search strategy; topic distillation

中图分类号: TP393

文献标识码: A

1 引言

随着 Web 上多元化信息的增长,传统的搜索引擎即通用搜索引擎已经不能满足人们对个性化信息检索服务日益增长的需要。近年来,面向主题搜索引擎应运而生,以提供分类更细致精确、数据更全面深入、更新更及时的因特网搜索服务^[1]。

所谓主题型搜索引擎,就是以构筑某一专题领域或学科领域的因特网信息资源库为目标,智能地在互联网上搜集符合设定专题或满足学科需要的信息资源^[2]。

在主题搜索引擎中,网络蜘蛛以何种搜索策略访问 Web 以提高效率,是近年来主题搜索引擎研究中的热点问题之一^[3]。而 Web 的动态性、异构性和复杂性要求网络蜘蛛能够高效率地实现 Web 信息提取,以保证信息的实时性和有效性^[4]。

2 网络蜘蛛

作为搜索引擎的基础组成部分,网络蜘蛛起着举足轻重的作用。随着应用的深化和技术的发展,网络蜘蛛越来越多地应用于站点结构分析、内容安全检测、页面有效性分析、用户兴趣挖掘以及个性化信息获取等多种服务中^[5]。

网络蜘蛛在采集 Web 信息时通常从一个“种子集”(如用户查询、种子链接或种子页面)出发,通过 HTTP 协议请求并下载 Web 页面,分析页面并提取链接,然后再以循环迭代的方式访问 Web。网络蜘蛛的搜索策略与搜索引擎的性质和任务密切相关。为了获得较高的 Web 覆盖率,通用搜索引擎网络蜘蛛通常采用图的遍历算法(如广度或深度优先策略)搜索 Web,如图 1a 所示^[6]。

与通用搜索引擎不同的是,面向主题搜索引擎服务于特定人群,其索引的页面内容仅限于特定主题或专门领

* 收稿日期:2007-08-10;修订日期:2007-10-10

基金项目:国家自然科学基金资助项目(60672031)

作者简介:李勇(1961-),男,辽宁大连人,硕士,副教授,研究方向为信息工程与数据库、软件工程、管理信息系统分析与设计、企业资源管理(ERP)、语义网;韩亮,硕士生,研究方向为信息工程与数据库、语义网。

通讯地址:116026 辽宁省大连市大连海事大学计算机科学与技术学院;Tel:13804241026;E-mail:hanliang001@163.com
Address: School of Computer Science and Technology, Dalian Maritime University, Dalian, Liaoning 116026, P. R. China

域,它在搜索过程中无须对整个 Web 进行遍历,只需选择与主题页面相关的页面进行访问,如图 1b 所示^[7]。主题搜索引擎网络蜘蛛在搜索 Web 时,需要对站点的主题相关性作出预测和删选,并对网页的主题相关性作出判断。本文的研究重心就是对主题搜索引擎中的搜索策略进行详细分析,并提出一个优化的主题搜索网络蜘蛛模型。

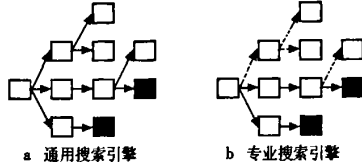


图 1 两类搜索引擎网络蜘蛛搜索顺序比较

3 主题网络蜘蛛搜索策略算法研究

3.1 系统模型

面向主题的网络蜘蛛的设计是以普通爬虫为基础的,实际上它是对一个普通网络蜘蛛进行功能上的扩充和调整,进行面向主题的网页信息提取。在应用需求的推动下,面向主题的 Web 信息提取技术已经成为一个热门的研究课题^[9]。我们设计了一个面向主题的网络蜘蛛的系统模型,如图 2 所示。为实现对面向主题的信息自动采集,我们将主要处理过程分成四大部分:主题选择、Web 信息提取、页面分析与过滤(页面过滤)、URL 与主题相关性计算(链接过滤)。

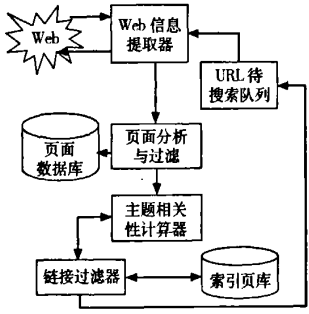


图 2 系统模型

3.2 主题选择

从语义上看,一个主题可以是一个含义,或者称之为一个概念,它可以是一个词语,也可以是一个短语,甚至是一个段落、一篇文章。主题概念的范围可大可小,主题范围可以非常抽象,但此时它的含义非常模糊;它的范围也可以非常具体,而此时它的意义却非常明确。主题选择是面向主题的 Web 信息提取的基础^[10]。

3.3 Web 信息提取

网络蜘蛛系统中,Web 信息提取模块是封装实现了具体 Web 协议的部分,它主要通过各种 Web 协议来自动提取 Web 站点中有效的信息,包括文本、图像、声音、影像等各种文档。涉及的相关协议包括 HTTP、FTP 以及 BBS 等,还可以根据用户的实际需要采集 Web Chat 等特殊的 Web 信息^[11]。为了达到理想的信息获取速度,网络蜘蛛系

统中大多采用了多线程并行信息提取的策略^[12]。

3.4 页面分析与过滤

在本系统中,主题网络蜘蛛主要处理的是 HTML 页面。因此,在页面分析与过滤过程中,我们所做的工作主要包括对 HTML 页面进行语法分析,以提取出网页正文、链接、链接相关标签属性数据及其它相关内容,然后对网页进行主题相关性判别,从而过滤掉主题无关页面,提高主题网络蜘蛛的主题信息提取的准确性。

3.4.1 HTML 语法分析

针对 Web 页面的语法分析基于 HTML 协议,整个语法分析过程可以分解为两个层次的操作:SGML 标记文法层和 HTML 标记层。SGML 文法层将页面分解成正文、标记、转义字符、注释等不同语法成分,而 HTML 标记层维护着当前解析正文的各种状态。这些状态根据特定标记创建或发生改变。

标记文法分析器的基本工作原理是:从标记文法创建状态转换表,根据输入流中的字符切换状态,在特定状态到达时执行相应语义操作^[13]。

通过执行对页面的 HTML 语法分析,可以提取出标题、正文、链接、链接标签属性数据及其它相关内容,以便系统进行页面主题相关性判别和 URL 主题相关性裁剪。

3.4.2 页面主题相关性判断算法

在进行 Web 主题信息提取的实施过程中,所提取的 URL 已经通过了主题相关性判别。尽管如此,所提取的页面内容还是可能与设定的主题相差甚远。这种现象将影响主题页面信息的提取准确率。因此,在页面提取之后,需要对页面进行主题相关性判别,以滤掉主题无关页面。

在本系统中,向量空间模型因为其处理能力强和灵活简便的优点,被用作进行页面的主题相关性判别。

页面的主题相关性判别算法流程如下:

- (1)预处理阶段:在信息提取之前,先将描述主题的多个页面进行关键词的提取和加权,从而得到该主题的特征向量及向量的权重。
- (2)对页面的正文进行分词,去掉停用词,保留关键词,然后按照关键词在文章中出现的频率对关键词加权处理。
- (3)对页面标题进行分词,将得到的关键词与网页正文中的关键词进行合并,并加重权于得到的标题关键词上。
- (4)根据设定主题中的特征向量对得到的页面关键词进行调整和扩充。
- (5) D_1 为主题, D_2 为待判别的页面,则:

$$Sim(D_1, D_2) = \frac{\sum_{k=1}^N W_{1k} * W_{2k}}{\sqrt{(\sum_{k=1}^N W_{1k}^2) * (\sum_{k=1}^N W_{2k}^2)}} \quad (1)$$

- (6)根据 $Sim(D_1, D_2)$ 值的大小和闭值 d 进行比较:如果 $Sim(D_1, D_2)$ 大于等于 d ,则表示页面与主题相关,保留到数据库中;否则判为不相关,丢弃该页面。

3.5 URL 的主题相关性 EPR 算法

面向主题的网络蜘蛛在采集 Web 信息时是面向选定主题的。为了有效地提高主题 Web 信息提取的准确率和效率,系统需要对待采集 URL 进行 URL 与主题的相关性判定,也可以叫做链接过滤或链接预测。按照高预测值优先采集、低预测值被抛弃的原则对发现的 URL 进行剪枝处理,可以大幅度减少采集页面的数量,有效地提高了主题信息搜索的速度和效率。针对主题搜索网络蜘蛛而言,如

何评价链接对于主题的价值,即链接价值的计算方法,是搜索策略中的关键所在。

我们在权衡了性能和效率后,设计了利用链接标签属性信息的 EPR(Enhanced PageRank,简称 EPR)算法来进行 URL 的主题相关性判别。

3.5.1 EPR 算法的目标

通过观察可以发现,尽管 PageRank 方法可以发现 Web 上的重要页面,但其确定的重要页面是针对广泛主题的,而不是面向具体主题的^[14]。

而作为另一个被广泛接受的超链分析算法 HITS,其基于权威页面和中心页面相互加强的设计模型提供了发现权威页面的有效办法。但是,HITS 算法存在一个最大的弱点,就是处理不好主题偏离问题,也就是紧密链接 TKC(Tightly-Knit Community Effect,简称 TKC)现象^[15]。

为此,我们对 PageRank 方法进行了如下改进:在链接关系的基础上加入针对链接的相关主题权重,同时引入链接网页之间主题度相互反馈加强的考虑,以使得所产生的重要页面是针对某一个主题的,这就形成了 EPR 算法。

3.5.2 EPR 算法的产生过程

EPR 方法的基本思想是:基于 Web 主题关联拓扑模型,在链接路径上,一个页面节点传递主题相关度到所链页面节点,并且从其所链页面节点得到主题相关度的反馈加强。

设 u, v, w 为 Web 页面, F_u 是 u 指向的页面集合, B_u 是指向 u 的页面集合,则页面 u 的主题相关度为:

$$S(u) = k_1 \sum_{v \in F_u} f(v, u) * S(u) + k_2 \sum_{w \in B_u} g(w, u) * S(u) \quad (2)$$

式(2)中, k_1, k_2, k_3 为常数, $0 \leq k_1, k_2, k_3 \leq 1$, 而且 $k_1 + k_2 + k_3 = 1$ 。 $f(u, v)$ 是页面 u, v 的函数,称之为传递因子,用来表示页面 u 到页面 v 的主题相关度的传递关系; $g(w, u)$ 是页面 w, u 的函数,称之为反馈因子,用来表示页面 w 到页面 u 的主题相关度的反馈关系:

$$g(u, v) = f(u, v) = \begin{cases} rel(k_4 ur_{lu} + k_5 ur_{lv}), & u \rightarrow v \\ 0, & \text{else} \end{cases} \quad (3)$$

其中, u, v 为给定的页面, k_4, k_5 为常数, $0 \leq k_4, k_5 \leq 1$, 而且 $k_4 + k_5 = 1$, rel 表示页面 u 到页面 v 的超级链接的主题相关度估计值。 $K_4 ur_{lu}$ 表示页面 u 中超级链接的数目。从式中可以发现,反馈因子由页面之间的虚拟文档的相似度与链接的主题相关度联合定义。

设 S 为 n 维向量,对应于 n 个网页的主题相关度,而且 $\|S\| = 1$ 。 M 为页面集对应的有向图对应的邻接矩阵,是 $n * n$ 矩阵。 M 中每个分量 $M_{uv} = f(u, v)$ 。设 P 为 n 维向量,对应于 n 个网页的主题相关度。下面是式(3)的矩阵形式:

$$S = k_1 * M^T * S + k_2 * M * S \quad (4)$$

设 I 为所有元素均为 1 的 n 维向量,由于 $\|S\| = 1$, 所以有 $I * S = 1$ 。式(4)可变换为: $S = (k_1 * M^T + k_2 * M + k_3 * P * I) * S$, 则得到 $S = A * S$ 。因此,将向量 S 的求解问题转换为矩阵 A 的按模最大的特征值与相应特征向量的求解问题。一般可以采用幂法来进行迭代求解。

给定初始向量 $S^{(0)} \neq 0$, 由迭代公式: $S^{(k+1)} = A * S^{(k)}$ 产生向量序列 $\{S^{(k)}\}$ 。可以证明,当 k 充分大时, $S^{(k+1)}$ 收敛, 并且有 $\lambda_1 = S^{(k+1)} / S^{(k)}$ (λ_1 为 A 的按模最大的特征值), 相应的特征向量为 $S^{(k+1)}$ 。给出 EPR 的算法如下:

算法 1 EPR()

```

给定初始向量  $S^{(0)} = 1$ ;
do {
    /* 计算网页的主题相关度  $S^*$  /
     $Normalize(S^{(i+1)})$ ; /* 对向量  $S$  进行归一化 */
     $\delta = \|S^{(i+1)} - S^{(i)}\|$ ; /* 计算差量 */
     $i++$ ; /* 递增循环变量 */
}while ( $\delta > \epsilon$ ); /* 循环计算直至向量收敛 */
输出迭代计算结果;
```

4 结束语

随着人们对个性化信息检索服务需求的日益增长,面向主题的搜索引擎应运而生。在主题搜索引擎中,网络蜘蛛以何种策略访问 Web 能提高搜索效率,是近年来主题搜索引擎研究中的主要问题之一。为此,我们展开了面向主题的网络蜘蛛搜索策略的研究,并设计了一个原型系统。特别地,在 URL 与主题的相关性判定过程中引入了链接文本及相关链接属性分析,提出了一种新颖的 URL 主题相关性算法——EPR 算法。

基于主题的采集的核心问题就是采集时向主题页面群的引导和对无关页面的过滤问题,需要进一步研究出新算法。例如,对于主题页面在 Web 上的分布规律,以及基于 Web 页面的语义分析方法等,还有待进一步发现。而对这些规律和新型算法的有效利用,也还有待进一步的提高。

参考文献:

- [1] 林彤,江志军. Internet 的搜索引擎[J]. 计算机工程与应用, 2000, 36(15):160-163.
- [2] 李蕾. 中文搜索引擎概念检索初探[J]. 计算机工程与应用, 2000, 36(6):1-11.
- [3] Paterson L. HTML4 编程指南[M]. 徐征,冯文镛,陈晓良,等译. 杭州:浙江科学技术出版社,2002:10-45.
- [4] Eichmann D. The RBSE Crawler-Balancing Effective Search Against Web Load[C]// Proc of the 1st Int'l World Wide Web Conf, 1994:113-120.
- [5] McBryan O A. GENVL and WWW: Tools for Taming the Web[C]//Proc of the 1st Int'l World Wide Web Conf, 1994: 70-90.
- [6] Pinkerton B. Finding What People Want: Experiences with the WebCrawler[C]//Proc of the 2nd Int'l World Wide Web Conf, 1994.
- [7] Cowie J, Lehnert W. Information Extraction[J]. Communications of the ACM, 1999, 1(1):80-91.
- [8] 余一妍. Google Linux Cluster 的系统结构分析[EB/OL]. [2005-04-29]. <http://www.woodpacker.org.cn:9081/doc/Googlefs/TR-2005-04.pdf>.
- [9] 李名智. 中文搜索引擎发展的现状、问题及对策[J]. 中国信息导报, 1999(2):30-32.
- [10] 土峰松. 新一代智能搜索引擎—网典[J]. 网络世界, 1999, 13(2):12-21.
- [11] 左远清. 自然语言处理在搜索引擎信息检索中的应用[J].

(下转第 56 页)

显示效果更好,以便医生作出正确的判断。

4.1 图像几何变换

几何变化通常包括图像平移、图像镜像、图像转置、图像缩放和图像旋转等。

图像的旋转计算方法为:假设在 X-Y 坐标体系中,将点 (x, y) 顺时针旋转角度 θ ,得到点 (u, v) ,计算公式为:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad (1)$$

图像的放大是图像处理中常用的技术,医学图像的放大有助于医生发现病因,做出诊断。设图像 x 轴方向的放大比率为 r_x , y 轴方向的放大比率为 r_y ,则坐标变换公式为:

$$\begin{aligned} u &= r_x \times x \\ v &= r_y \times y \end{aligned} \quad (2)$$

4.2 滤波处理

对于未经处理的医学图像,都存在一定程度的噪声干扰。噪声恶化了图像质量,致使图像模糊,给分析、识别过程带来困难,影响医生的诊断。目前,去除图像噪声最流行也是最实用的方法有邻域平均法、中值滤波法和形态学膨胀腐蚀法。

数学形态学是一种新兴的图像处理工具。基本的形态学运算是腐蚀和膨胀。

在 IMAQ 中,可以利用 Morphology 函数对图像进行各种形态学滤波处理。

4.3 边缘检测

物体的边缘是由灰度的不连续性所致。所谓边缘是指信号强度发生急剧变化的位置,它包含了图像的许多特征信息。由边缘点构成的边缘,是区分图像不同区域的关键。因此,边缘检测在医学图像处理中有重要的意义。通过边缘提取可以获得图像的边缘轮廓,使图像更加简明清晰,有利于辅助医生进行诊断。常用的边缘检测算子有 Roberts 算子、Sobel 算子、Prewitt 算子和 Laplacian 算子。

拉普拉斯(Laplacian)算子是一种二阶导数算子,对一个连续函数 $f(x, y)$,它在位置 (x, y) 的拉普拉斯值定义如下:

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \quad (3)$$

数字化以后用二阶差分实现如下:

$$\begin{aligned} L = \nabla^2 f(x, y) &= \nabla^2 x^2 f(i, j) + \nabla^2 y^2 f(i, j) = \\ &= f(x+1, y) + f(x-1, y) + f(x, y+1) + \\ &= f(x, y-1) - 4f(x, y) \end{aligned} \quad (4)$$

某病者‘气管’CT图在 IMAQ Vision 中采用 Laplacian 边缘检测后的结果如图 4 所示。

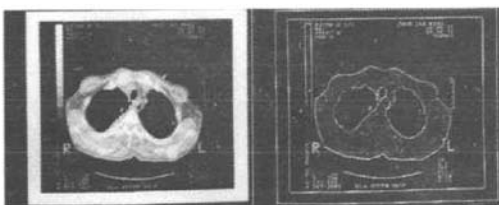


图 4 Laplacian 边缘检测

4.4 伪彩色增强

伪彩色增强是将一个波段或单一的黑白图像变换为彩色图像,从而把人眼不能区分的微小灰度差别显示为明显的色彩差异,更便于解译和提取有用信息。

受人眼对灰度的敏感度影响,医生在阅片时很容易漏诊或很难确诊病灶,特别对于最有希望治愈的早期病变,这样的情况更可能发生。因为人眼对灰度微弱递变的敏感程度远远小于对色彩变化的敏感程度。所以,将一幅灰度图像按照特定的彩色编码表进行伪彩色变换,就可以看到图像更加精细的结构,将更有利于医生进行诊断。伪彩色增强的方法主要有密度分割法、空间域灰度级-彩色变换和频率域伪彩色增强。其基本原理如图 5 所示。图中, $f(x, y)$ 为灰度图像, $R(x, y)$, $G(x, y)$ 和 $B(x, y)$ 为 $f(x, y)$ 映射到 RGB 空间的三色分量。

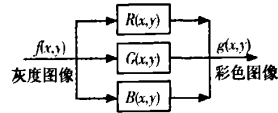


图 5 伪彩色处理原理图

5 结束语

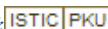
本文以 LabVIEW 为平台,利用医学图像工具包、数据库连接工具包以及 IMAQ 软件包实现了 DICOM 图像在 Windows 平台下的显示、存取和处理,解决了 DICOM 图像在 Windows 平台下无法直接显示的问题,提高了图像的质量,利于医生诊断,对促进虚拟仪器在国内医学事业的发展上有一定的参考价值。

参考文献:

- [1] 谢长生,熊华明,陈 颖. DICOM 图像显示的研究与实现[J]. 计算机工程与科学,2002,24(6):38-41.
- [2] Klinger T. Image Processing with LabVIEW and IMAQ Vision[M]. Upper Saddle River, New Jersey: Prentice Hall PTR,2003.
- [3] 雷振山,赵晨光. 虚拟仪器系统的数据存储技术[J]. 微计算机信息,2006(22):117-119.
- [4] 王立功,刘伟强,于雨华,等. DICOM 医学图像文件格式解析与应用研究[J]. 计算机工程与应用,2006,42(29):210-212.
- [5] NI Corp. NI-IMAQ User Manual[M]. 2001.

(上接第 6 页)

- 现代计算机:下半月版,2002(7):28-29.
- [12] 吴友政,赵军,段湘煌,等. 问答式检索技术及评测研究综述[J]. 中文信息学报,2005,19(3):1-13.
- [13] Shapiro D. Value-Driven Agents: [Ph D Thesis]. [D]. Stanford:Stanford University, 2001:23-178.
- [14] 陈红英,杨宜民. 基于多智能体的网络信息系统的原理与实现[J]. 微电子学与计算机,2005,22(3):57-64.
- [15] Barroso A,Dean J,Hizle U. Web Search for a Planet: The Google Cluster Architecture[J]. IEEE Micro,2003,23(2):22-28.

作者: 李勇, 韩亮, LI Yong, HAN Liang
作者单位: 大连海事大学计算机科学与技术学院, 辽宁, 大连, 116026
刊名: 计算机工程与科学 
英文刊名: COMPUTER ENGINEERING & SCIENCE
年, 卷(期): 2008, 30 (3)
被引用次数: 10次

参考文献(15条)

1. 林彤;江志军 [Internet的搜索引擎](#)[期刊论文]-[计算机工程与应用](#) 2000 (15)
2. 李蕾 [中文搜索引擎概念检索初探](#)[期刊论文]-[计算机工程与应用](#) 2000 (06)
3. Paterson L;徐征;冯文镛;陈晓良 [HTML4编程指南](#) 2002
4. Eichmann D [The RBSE Crawler-Balancing Effective Search Against Web Load](#) 1994
5. McBryan O A [GENVL and WWW.Tools for Taming the Web](#) 1994
6. Pinkerton B [Finding What People Want:Experiences with the WebCrawler](#) 1994
7. Cowie J;Lehnert W [Information Extraction](#) 1999 (01)
8. 余一娇 [Google Linux Cluster的系统结构分析](#) 2005
9. 李名智 [中文搜索引擎发展的现状、问题及对策](#)[期刊论文]-[中国信息导报](#) 1999 (02)
10. 土峰松 [新一代智能搜索引擎一网典](#) 1999 (02)
11. 左远清 [自然语言处理在搜索引擎信息检索中的应用](#)[期刊论文]-[现代计算机](#) 2002 (07)
12. 吴友政;赵军;段湘煌 [问答式检索技术及评测研究综述](#)[期刊论文]-[中文信息学报](#) 2005 (03)
13. Shapiro D [Value-Driven Agents](#) 2001
14. 陈红英;杨宜民 [基于多智能体的网络信息系统的原理与实现](#)[期刊论文]-[微电子学与计算机](#) 2005 (03)
15. Barroso A;Dean J;Hilze U [Web Search for a Planet:The Google Cluster Architecture](#)[外文期刊] 2003 (02)

本文读者也读过(6条)

1. 黄旭, 朱艳琴, 罗喜召, HUANG Xu, ZHU Yan-qin, LUO Xi-zhao [基于内容评价的爬虫搜索策略研究](#)[期刊论文]-[微电子学与计算机](#)2008, 25 (11)
2. 刘淑梅, 夏亮, 许南山, LIU Shu-Mei, XIA Liang, XU Nan-Shan [主题搜索引擎网络爬虫搜索策略的研究与实现](#)[期刊论文]-[计算机系统应用](#)2010, 19 (3)
3. 陈丛丛 [主题爬虫搜索策略研究](#)[学位论文]2009
4. 吴安清 [主题搜索引擎爬行策略的研究](#)[学位论文]2006
5. 黄莉, 王成良, 杨铮, HUANG Li, WANG Cheng-liang, YANG Zheng [面向主题网络爬行的智能隧道穿越算法研究](#)[期刊论文]-[计算机应用研究](#)2009, 26 (8)
6. 张红云, 刘炜, 熊前兴, Zhang Hongyun, Liu Wei, Xiong Qianxing [一种基于语义本体的网络爬虫模型](#)[期刊论文]-[计算机应用与软件](#)2009, 26 (11)

引证文献(13条)

1. 许金玲, 陈旭翔, 赵少娟, 丁必蛟 [基于信令分析的客户网络标签体系搭建](#)[期刊论文]-[电信快报: 网络与通信](#) 2012 (5)
2. 刘淑梅, 夏亮, 许南山 [Postgresql数据库集群在主题网络爬虫的应用](#)[期刊论文]-[计算机系统应用](#) 2010 (12)

3. [汲业](#), [陈燕](#), [杨健](#), [慕容](#) 生活服务领域垂直搜索引擎的设计与实现[期刊论文]-[计算机工程](#) 2010(24)
4. [何毅](#) 建筑院校主题搜索引擎设计与实现[期刊论文]-[吉林建筑工程学院学报](#) 2010(5)
5. [周远超](#), [叶枫](#), [高依旻](#), [张雪洁](#) 水利垂直搜索引擎的研究[期刊论文]-[计算机与数字工程](#) 2012(10)
6. [张安妮](#), [姜华](#), [郝相连](#) 面向主题爬虫改进算法的个性化搜索引擎应用研究[期刊论文]-[海南大学学报\(自然科学版\)](#) 2011(3)
7. [刘淑梅](#), [夏亮](#), [许南山](#) 主题搜索引擎网络爬虫搜索策略的研究与实现[期刊论文]-[计算机系统应用](#) 2010(3)
8. [何毅](#) 基于Web的建筑业主题搜索引擎技术[期刊论文]-[吉林广播电视大学学报](#) 2009(6)
9. [张安妮](#), [姜华](#), [郝相连](#) 面向主题的快速搜索引擎的设计与研究[期刊论文]-[淮阴工学院学报](#) 2011(3)
10. [方东权](#), [吴天吉](#), [李翠霞](#) “三农”信息资源整合与服务平台的设计与实现[期刊论文]-[中国农学通报](#) 2009(4)
11. [李园伟](#) 面向高校主题搜索引擎的的爬行器设计[期刊论文]-[电脑知识与技术](#) 2011(16)
12. [韩国辉](#), [陈黎](#), [梁时木](#), [唐小棚](#), [王亚强](#), [于中华](#) Na(i)ve Bayes分类器制导的专业网页爬取算法[期刊论文]-[中文信息学报](#) 2010(4)
13. [张春菊](#), [张雪英](#), [朱少楠](#), [徐希涛](#) 基于网络爬虫的地名数据库维护方法[期刊论文]-[地球信息科学学报](#) 2011(4)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_jsjgcykx200803002.aspx