基于模拟退火算法的主题爬虫

贺 晟[1],程家兴[1],蔡欣宝[2]
(1. 安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2. 苏州大学 智能信息处理及应用研究所,江苏 苏州 215006)

摘 要:主题爬虫是主题搜索引擎的基础与核心,主题爬行策略的好坏直接影响搜索结果。为了搜索到更多相关的网页,通过利用模拟退火机制选择下一步要访问的链接,使那些蕴含"综合价值"高的链接在搜索初期有机会被选中,同时利用"隧道技术"扩大相关网页的搜索范围。计算链接价值时,综合考虑了链接所在页面内容的价值和链接提示文字的价值,根据它们对链接价值的影响程度不同,分别赋予它们不同的权值。实验证明,该方法对提高网页覆盖率和准确率都有很好的效果。

关键词:模拟退火算法;隧道技术;召回率
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2009)12-0055-04

Focused Crawler Based on Simulated Anneal Algorithm

HE Sheng[1], CHENG Jia-xing[1], CAI Xin-bao[2]
(1. Ministry of Edu., Key Lab. of Intelligent Computing & Signal Processing, Anhui Univ., Hefei 230039, China;
2. Institute of Intelligent Information Processing and Application, Soochow University, Suzhou 215006, China)

Abstract: Focused crawler is the core and foundation of the topic-specific search engine. Special crawling strategy quality gives a direct impact on search results. In order to select more relevant pages, through using the simulated annealing mechanism to choose the next link to visit, makes those high "overall value" link have the opportunity to be selected early in the search, and using "Tunneling" to broaden the searching scope. When calculating the link value, consider the value of the page content and the link text synchronously, and according their different impact to the link value, gives them different weight. Experiments indicate that the method has a good effectiveness.

Key words: simulated anneal algorithm; tunneling; recall fraction

0 引 言

随着 Internet 的飞速发展,网络上的信息更是成指数级增长,人们在 Internet 上获取信息时,越来越难以离开搜索引擎的帮助。但目前传统方式的搜索引擎已经不能满足特定用户的需求,适应特定主题和个性化搜索引擎的主题爬虫便应运而生。主题爬虫所抓取的内容只限于特定的主题或专门领域,在搜索过程中无须对整个网络进行遍历,只需选择与主题相关的页面进行访问。主题爬行策略的目标就是保证使其尽可能多地爬行相关网页,尽可能少地爬行无关网页,,以提高主题信息的发现率与召回率。在主题搜索引擎中,网络爬虫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。在制定主题爬行策略时通常要考虑多种因素,如:待爬 URL 取舍策略、优先级排序策略、隧道技术、主题飘移策略等。

文中把模拟退火算法与"隧道技术[1]"结合。模拟退火算法在选择优化解方面具有"非贪婪性",在网络爬虫搜索过程中,每次除选择评价值最优的链接,还以一定概率有限度地接收评价值次优的链接,确保那些蕴含"综合价值"的链接在搜索初期有机会被选中[2]。"隧道技术"使搜索有机会穿过低相关区域进入高相关区域,当页面内容的相关度低于设定的阈值时,扩大主题范围,使更多的相关的链接加入到链接优先机队列,提高相关网页的召回率。同时考虑链接文字和链接所在页面内容的对链接价值的影响,克服了主题漂移。

1 相关主题爬虫分析

Breath-First 搜索算法(Pinkerton,1994)是一种

收稿日期:2009-04-14;**修回日期**:2009-07-28
基金项目:国家自然科学基金(60273043);安徽大学研究生创新基金(20073053)
作者简介:贺 晟 (1984-),女,安徽灵璧人,硕士研究生,研究方向为数据挖掘、智能计算;程家兴,教授,博士生导师,研究方向为智能计算、算法分析及设计及最优化方法。

无遗漏的广度优先搜索算法，优点是可以让爬虫并行处理，提高抓取速度。但用在主题爬虫中，准确率不高，造成了资源的浪费。

Best－First[3]搜索算法是对 Breath－First 算法的一种改进。其基本思想是构建一个 URL 链接列表，然后按照某种评价选择策略选择出最好的链接进行访问。它具有很大的贪婪性，容易过早地陷入 Web 搜索空间中局部最优子空间的陷阱。Best－First 算法只适用于小范围内主题相关页面的搜索，它只选择“立即回报[4]”价值高的链接，而没有考虑链接的“未来回报[4]”价值。

Fish－Search[5]搜索算法是对深度优先搜索算法的一种改进，在信息搜索过程中，相关网页包含的超链被赋于比不相关网页包含的超链更高的优先权值，插入到未被搜索的 URL 列表中。优点是模式简单、动态搜索，但由于只能使用简单的字符串匹配来分配孩子节点的相关度值，相关度值不够精准，同时也使爬行队列中链接优先级差别太小，网页之间的优先次第关系不明显。

限定搜索算法[6](Limited Memory Search)只保留待爬队列中相关度最高的前 N 个链，将第 $N+1$ 及其以后的链接 URL 作为低相关或不相关页面丢弃。该方法由于舍去了相关度较低的，减少了系统占用的缓冲空间，同时爬行覆盖范围限制在高相关度领域，爬行结果的主题相关度高，缺点是错过了经由低相关度页面发现高相关度页面的机会。

2 基于模拟退火算法的主题爬虫

2.1 模拟退火搜索算法的其本思想

模拟退火算法具有渐近收敛性，已在理论上被证明是一种以概率 1 收敛于全局最优解的全局优化算法。其算法的核心是在搜索最优解过程中，除了可以接受优化解外，还用一个随机接受准则有限度地接收恶化解，并且接受恶化解的概率逐渐趋近于 0，这使得算法有可能从局部最优中跳出，找到全局最优解。通常只要退火过程足够慢，算法寻找到全局最优解的概率趋近于 1。

Ester[1]将搜索穿过低相关区域进入高相关区域的技术称为隧道技术，基本思想是：当爬行器进入低相关网页区域时，扩大主题范围，而当爬行器重新进入正常区域时，恢复到原来定义的主题范围。

实现的方法主要有以下几种：

(1)主题词泛化，即当爬行器所处区域页面相关度低于设定的阈值时，则取主题词的上位类词，如用“花朵”来代替原来的主题词“玫瑰“，当爬行器所处区域页面相关度上升并超过阈值时，恢复初始指定的主题词，如把“花朵”恢复为“玫瑰”；

(2)表达词泛化，对于形如 $\Phi = A \cap B$ 的提问表达式，用 A 相关度 f_A 取代 $A \cap B$ 的相关度 $f_{A\cap B}$(如 $f_A < f_B < f_{A\cap B}$ 果)；

(3)调整权值，即通过增加关键词的权重来提高相关度，使本来相关度低于阈值的页面被作为“桥梁”页面下载，将爬行器引导到其后续相关链接页面。

模拟退火算法具有较强的局部搜索能力，但模拟退火算法对整个搜索空间的了解不多，不便于搜索过程进入最有希望的搜索区域[7]。这里把模拟退火选择策略与“隧道技术”结合，当搜索的页面相关度低于事先设定的阈值时，就扩大主题范围，计算页面中的链接的价值，再把大于阈值的链接加入到链接优先权队列中，扩大了搜索范围，使搜索进入到最有希望的搜索区域，在更好的空间里采集与主题相关的页面。

2.2 种子集生成

种子页面即主题爬行的起始页面。种子页面的选择将直接影响信息采集的质量以及采集工作的效率。为此，要求种子页面具有较高的主题相关性以及主题链接的中心度。

种子页面的生成，一般采用三种方法：①人工指定，即由专家给出相关的种子页面；②自动生成，用户指定部分关键词，提交给通用搜索引擎，从检索结果中抽取前 N 个页面作为种子页面；③混合模式，先用通用搜索引擎获得部分相关页面，然后再经过人工筛选、过滤、评价，形成一个能充分反映主题特征的种子页面集。

文中就是采用混合模式，利用元搜索引擎获得可能相关的 URL，除去重复的 URL，然后在领域专家的指导下进行人工选择。

2.3 页面内容的相关度计算

相关度计算分为三步：①确定一组带有权重的能够代表受限领域的关键词，用它来表示确定的主题；②页面的关键词提取；③计算相关度。把关键词的个数作为向量空间的维数，每个关键词的权重作为每维分量的大小，文中关键词的权重大小由关键词所在页面的不同位置决定。

根据 W3C 维护的 HTML4.0 标准以及对大量网页文本的分析，将 HTML 标签分为以下几组类别：

(1)TITLE、META、H1：描述标题、关键字，含有重要信息；

(2)H2、H3：描述二级、三级标题；

(3) H4、H5、H6、Strong：描述三级以下标题以及具有加粗效果的文字；

(4)P、TD、LI:描述正文信息;

(5)其他标签:描述非正文信息。

根据标签所描述的不同内容,给五组标签赋予不同的权重构成一个权重向量:$TW=(tw_1,tw_2,tw_3,tw_4,tw_5)$,规定 $tw_1>tw_2>tw_3>tw_4>tw_5$。

为了简化计算,缩短计算的运行时间,提高系统效率,采用改进后的算法计算页面内容相关度[8]。在对网页进行分析时,直接整理出特征词在标签中出现的次数,不用算出它们的频率比值。分别用$(a_1,a_2,\cdots,a_n)$、$(b_1,b_2,\cdots,b_n)$、$(c_1,c_2,\cdots,c_n)$、$(d_1,d_2,\cdots,d_n)$、$(e_1,e_2,\cdots,e_n)$来表示特征词在对应标签中的出现的次数。则相关度可以表示为:

$$Ti=(a_1+a_2+\cdots+a_n)\times tw_1+(b_1+b_2+\cdots+b_n)\times tw_2+(c_1+c_2+\cdots+c_n)\times tw_3+(d_1+d_2+\cdots+d_n)\times tw_4+(e_1+e_2+\cdots+e_n)\times tw_5 \quad (1)$$

设定阈值 r,当 Ti 大于或等于 r 时,认为网页与主题相关,当 Ti 小于 r 时,认为网页与主题无关。

2.4　链接的评价函数

对于链接价值,也同样需要根据其内容相关度进行确定。一方面,链接文字内容表明其价值。为了用户浏览的方便,通常情况下,链接的文字与其指向的页面内容是相关的。另一方面,链接所在页面的信息内容也对其价值有影响。链接指向的下一级网页往往是对当前网页有关问题的具体阐述或相关介绍[9]。如:

假设 L 是网页 P 指向网页 C 的一个链接,见图 1,网页 P 已经被下载并被解析,网页 C 为待下载页面,那么,基于 L、P 以及爬行主题 Q 等信息,在估算网页 C 潜在的主题相关度时,可以考虑的启发式策略包括:①网页 P 与 Q 的相关度;②链接 L 的提示文字与 Q 的相关度;③链接 L 的 URL 周围文字与 Q 的相关度;④链接 L 的兄弟链接与 Q 的相关度;⑤L 的上下文环境与其他已知相关网页的上下文环境的相似度等。基于计算简便,这里只考虑①②两个方面。

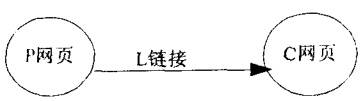


图 1　P 网页指向 C 网页

所以评价函数为:

$$E(\text{link}_i)=\varphi\times f_i+(1-\varphi)\times g_i \quad (2)$$

式中,φ— 权重因子,f_i— 网页内容评价值,g_i— 链接提示文字的评价值。由于链接文字通常是由为数不多的几个词或短语构成的短文本,正常情况下所包含的特征对链接目标内容的表征能力是较强的[9],为了提高计算性能,计算链接提示文字对链接的得分时使用高频特征库。

2.5　算法的执行过程

模拟退火算法在选择优化解方面具有“非贪婪性”,在网络蜘蛛搜索过程中,每次除选择“最优链接”外,还以一定概率有限度地接收“次优链接”,随着搜索过程的进行,逐渐使选择次优链接的概率趋近于 0。这使得那些“蕴涵”较高“综合价值”的链接在搜索初期有机会被选中,使网络爬虫有可能跳出局部最优空间的陷阱,寻找到最优的行动选择序列。另外当页面的内容相关度低于设定的阈值,扩大主题范围,使搜索始终控制在一个较大的范围,提高了链接于相似度不高的页面之后的页面被搜索的机会。

对算法执行过程描述如下:

(1) 初始化控制参数:初始化链接优先权队列 S,页面相关度阈值 r_0,链接阈值 r_1,退火初始温度 T,温度冷却参数 $0\leqslant C\leqslant 1$,L 为某一温度下达到温度平衡状态的循环次数,设定计数变量 $t\leftarrow 0$。

(2) 链接优先权队列为空,转入(9);否则,队头链接出队,请求并下载该链接指向的页面。

(3) 如果页面主题相关度大于 r_0(用式(1)计算),保存以备索引,转入(4);否则转入(5)。

(4) 提取页面中所有 URL② 提取链接文本和链接所在的页面的文本信息,计算所有未被访问的链接的链接价值。转入(6)。

(5) 提取页面中所有 URL② 提取链接文本和链接所在的页面的文本信息,扩大主题范围,计算所有未被访问的链接的链接价值。

(6) 把链接相关度大于 r_1 的链接插入到链接队列 S 中。重排优先权队列。

(7) 根据模拟退火机制选择链接:① 从链接优先权队列中,按最好优先策略选出价值最大的链接 current,再以等概率任选一链接 next;② 若 $\exp[(\text{value(next)}-\text{value(current)})/T]>\text{random}$,则接收 next 为下一步将访问的链接;否则接收 current 作为下一步将访问的链接。其中,random 为[0,1]上的随机数;③ 链接优先权队列重排,使得由 ② 选中的链接置于队头。重排优先权队列。

(8) 若 $t>L$,则调整退火温度 $T\leftarrow C\times T$,$t\leftarrow 0$。

(9) $t\leftarrow t+1$,转入(2)。

(10) 结束。

基本流程图如图 2 所示。

其中,温度冷却参数 C 控制温度的冷却速度,C 值越大,温度冷却的越快,反之则越慢。步骤(5) 当页面的相并度低于设定的阈值,就扩大主题范围,再计算链接价值。步骤(7) 可以看出,算法除接收最优链接外,还以概率 $\exp[(\text{value(next)}-\text{value(current)})$

/T]接收次优链接，因而是一种“非贪婪”的链接选择策略。其中的value函数用链接评价值函数式(2)计算。步骤(8)用于调整退火温度，容易证明，当 T 趋近于0时，接收次优链接的概率也趋近于0。

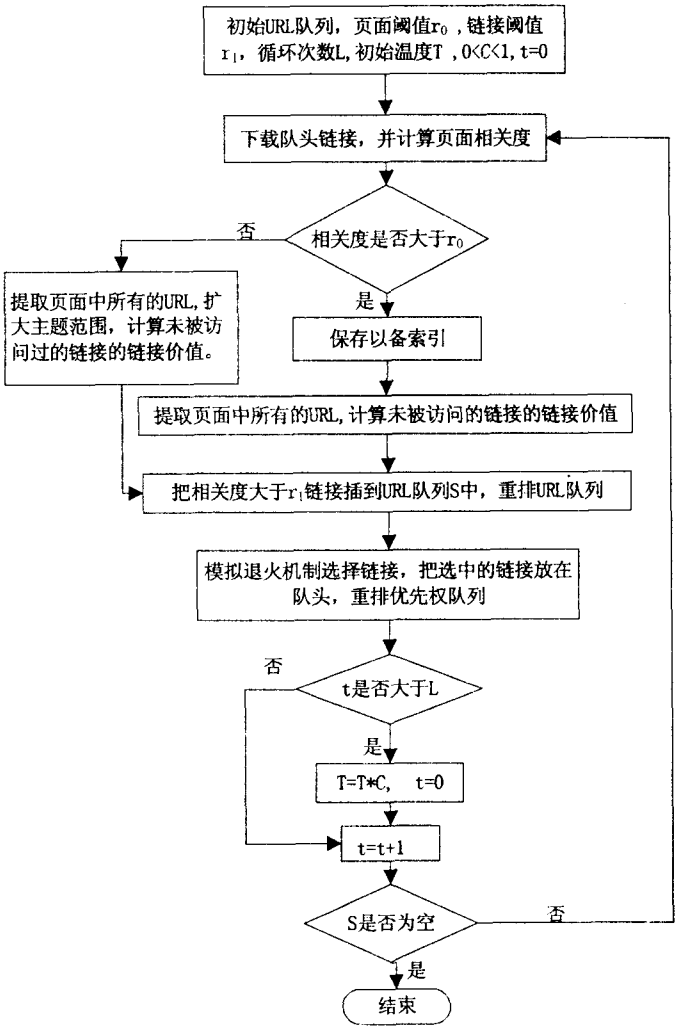


图2 算法执行的其本流程

3 实验结果分析

实验：以“化妆品”为主题，线程数= 200（要求在网络环境较好的情况下），初始种子= 100（都是经过人工选择的较好的种子），链接价值阈值 r =10（采用改进后的计算方法）。机器配置：P4 3.0G CPU，内存1G，硬盘容量为160GB。为了验证文中提出的搜索策略，使其与OPS（最佳优先搜索策略）、BFS（广度优先搜索策略）性能进行比较。

图3显示了三种不同的Spider在搜索同一主题时的性能，从图中可以看出，搜索初期OPS算法下载的相关网页数占整个下载网页数的比例明显高于BFS和SA1，随着搜索的进行，文中提出的方法优势进一步明显，下载的相关网页数所占的比例高于OPS和BFS。从图中可以看出BFS随着下载网页数的增多，相关的网页数增辐越来越小，而且很快相关网页数就不再增多。而文中提出的方法由于考虑了“综合回报值”，相关页面数不断增长，召回率明显提高，同时在计算链接价值时，综合考虑了链接所在页面的内容和链接文本的价值，根据它们对链接价值的影响程度不同，分别赋予它们不同的权值，使得所搜索的与主题相关的页面更准确。

4 结束语

通过实验可以看出，文中提出的方法具有非贪焚性，既考虑了“立即回报值”，也考虑了“未来回报值”，有效地提高了相关网页的召回率。

据统计Deep Web中隐含的信息量一般是Surface Web的400～500倍，它的信息来源于后台数据库，对于传统的搜索引擎来说，这部分页面并不能被索引到，所以下一步工作主要对Deep Web数据源的发现方法进行研究。

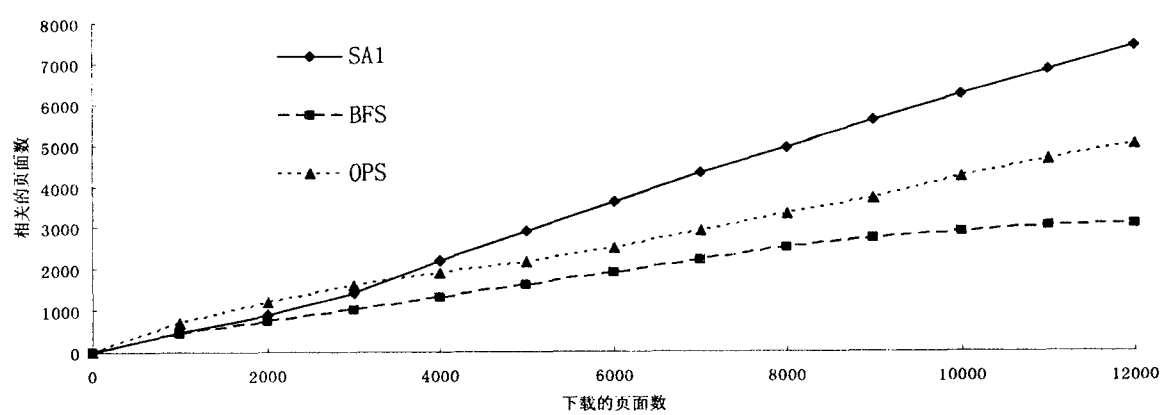
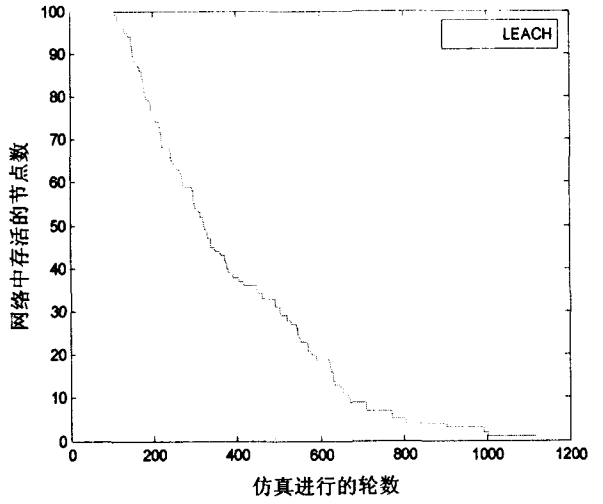


图3 三种算法搜索性能比较

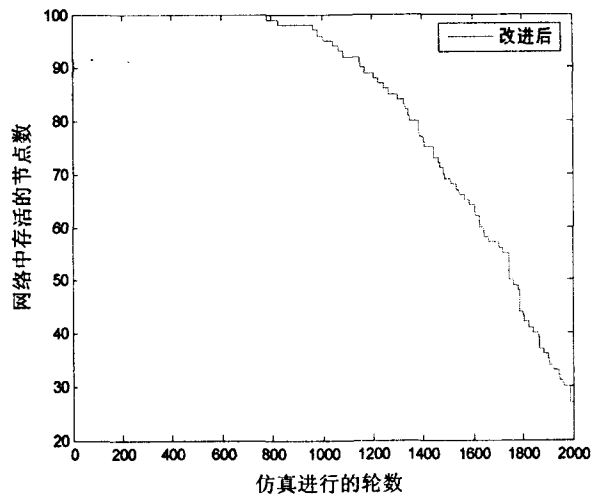
（下转第62页）

参考文献：

[1] Ester M, Gross M, Kriegel H P. Focused Web crawling: a



(a)



(b)

图 3 LEACH 算法和改进后算法比较

5 结束语

该文分析了无线传感器网络的体系结构和特点，重点对无线传感器网络中的路由问题进行了研究。在分析了基于 ZigBee 协议的分簇路由算法 Cluster - tree 的基础上，对该算法提出了改进。改进之后的算法利用 ZigBee 节点的深度信息来简化路由的过程。同时考虑了节点的能量均衡利用的问题。仿真实验的结果证明，利用深度信息改进后的路由算法，确实能有效延长网络的生存期，同时也就提高了能量的利用率。

参考文献：

[1] 马祖长，孙怡宁，梅 涛. 无线传感器网络综述[J]. 通信学报，2004，25(4)：114 - 124.

[2] Estrin D. Wireless Sensor Networks tutorial part IV: Sensor network protocols[C]//Proceedings of the ACM Mobile Computing and Networking (MobiCom). Atlanta, GA: [s. n.], 2002.

[3] 杨菊英，吕光宏. 无线传感器网络分层路由协议研究[J]. 计算机技术与发展，2008，18(6)：115 - 118.

[4] Heinzelman W, Chandrakasan A, Balakrishnan H. Energy efficient communication protocol for wireless microsensor networks[C]//Proceedings of the 33rd Annual Hawaii International Conference on System Sciences. Mani, HI: [s. n.], 2000：1 - 10.

[5] ZigBee Specification, ZigBee 联盟白皮书[EB/OL]. 2004. http://www.zigbee.org/.

[6] 刘元安，唐碧华，胡月梅. Ad hoc 网络中的路由算法[J]. 北京邮电大学学报，2004，27(2)：1 - 7.

[7] 吴小兵，陈贵海. 无线传感器网络中节点非均匀分布的能量空洞问题[J]. 计算机学报，2008，31(2)：253 - 261.

[8] 黄海平，王汝传，孙力娟，等. 基于父亲树的无线传感器网络路由协议[J]. 计算机技术与发展，2008，18(8)：4 - 7.

[9] 李方敏，刘新华，徐文君，等. 无线传感器网络的链路稳定成簇与功率控制算法[J]. 计算机学报，2008，31(6)：965 - 978.

(上接第 58 页)

generic framework for specifying the user Interest and for adaptive crawling strategies[C]//In: Proceedings of 27th International Conference on Very Large Database (VLDB'01). Roma: Springer - Verlag, 2001：527 - 534.

[2] 邓岳贵. 启发式搜索在网络爬虫中应用的分析[J]. 软件导刊，2008，7(2)：80 - 82.

[3] Cho J, Garcia - Molina H, Page L. Efficient crawling through URL ordering[J]. Computers Networks and ISDN Systems, 1998, 30: 161 - 172.

[4] 林海霞，原福永. 一种改进的主题网络蜘蛛搜索算法[J]. 计算机工程与应用，2007，43(10)：174 - 176.

[5] DeBra P, Post P. Information retrieval in the World - Wide web: making client - based searching feasible[J]. Computer Networks and ISDN Systems, 1995, 27(2)：183 - 192.

[6] 李春旺. WEB 信息主题采集技术研究[J]. 图书情报工作，2005，49(4)：76 - 80.

[7] 王知人，章 胤. 一种改进的模拟退火算法[J]. 高等学校计算数学学报，2006，28(1)：15 - 19.

[8] 郑国良，叶飞跃，张 滨. 基于网页内容和链接价值的相关度方法的实现[J]. 计算机工程与设计，2008，29(23)：6020 - 6022.

[9] 黄 旭，朱艳琴，罗喜召. 基于内容评价的爬虫搜索策略研究[J]. 微电子学与计算机，2008，25(11)：25 - 28.