基于遗传算法的主题爬虫

张海亮，袁道华
（四川大学 计算机学院，四川 成都 610065）

摘 要：针对目前主题网络爬虫搜索策略难以在全局范围内找到最优解，通过对遗传算法的分析与研究，文中设计了一个基于遗传算法的主题爬虫方案。引入了结合文本内容的 PageRank 算法；采用向量空间模型算法计算网页主题相关度；采取网页链接结构与主题相关度来评判网页的重要性；依据网页重要性选择爬行中的遗传因子；设置适应度函数筛选与主题相关的网页。与普通的主题爬虫比较，该策略能够获取大量主题相关度高的网页信息，能够提高获取的网页的重要性，能够满足用户对所需主题网页的检索需求，并在一定程度上解决了上述问题。

关键词：遗传算法；爬虫；主题爬虫；主题相关度；网页重要性
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1673-629X(2012)08-0048-05

Focused Crawling Based on Genetic Algorithms

ZHANG Hai-liang, YUAN Dao-hua
(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: Optimized solution can't be found in the global scope based on the present searching strategy of focused crawler. A focused crawler method based on genetic algorithm is proposed through the analysis and study of genetic algorithm. This method introduces the PageRank algorithm combined with text contents, computes the page topic similarity with vector space model algorithm, and judges the importance of web page according to web link structure and topic similarity. At the same time, the genetic factors are selected on basis of the importance of web page. The system sets fitness function to select pages relevant with topic. Compared to focused crawler, the topic crawler based on genetic algorithms could obtain the web pages which have strong correlation with subjects, and improve the importance of access web pages, and satisfy user's demand for searching topic webs they're interested in. So in a certain extent, the above problems are solved.

Key words: genetic algorithm; crawler; focused crawler; topic similarity; web importance

0 引 言

随着互联网信息的剧增，用户越来越依赖搜索引擎，而传统的通用搜索引擎查询的数据是海量无序的且查询的深度不够，而用户最大的体验就是查询不到自己想要的信息。为了克服通用搜索引擎不能快速、准确地查找有用的信息，并满足用户的需求体验，于是主题搜索引擎应运而生。

主题搜索引擎是信息技术在大规模文本集合上检索与主题相关信息的实际应用。主题爬虫对于主题搜索引擎发现和抓取文档具有首要的责任，主题爬虫设计的优秀与否决定着主题搜索引擎的好坏。主题网络爬虫采用分类技术来限制访问的网页是关于同一个主题的。

收稿日期：2011-12-24；**修回日期**：2012-03-27

作者简介：张海亮(1987-)，男，四川成都人，硕士，研究方向为分布式并行处理与网络计算；袁道华，教授，硕士生导师，研究方向为分布式并行处理与网络计算。

1 主题爬虫

自动地发现并下载网页称为爬取，下载网页的程序称为网络爬虫[1]（Crawler 或 Spider 程序），是搜索引擎的重要组成部分。

主题网络爬虫[2,3]的工作流程主要有三个步骤：

第一，从请求队列中取出 URL 地址，分析 URL 与主题的相关性，如果该 URL 与主题相关且以前没有遇到过，则将其放入待抓取的 URL 队列中。

第二，按照某种排序算法对上述 URL 队列进行排序，使重要的页面置于队列前端，并重复上述过程，直到满足系统设计的停止条件。

第三，存储与主题相关网页，进行一定的分析、过滤，并建立索引，并对后续的抓取过程进行反馈和指导。

爬虫开始的时候，需要给爬虫输送一个 URL 列表，在实际应用中该列表是一个特定话题的多个权威页面的集合，于是这个列表中的 URL 地址便是爬虫的起始位置，爬虫从这些 URL 出发，不断地发现新的

URL,然后再根据预先设计搜索策略爬行这些新的URL,最后得到与主题相关的网页。一般的爬虫都自己建立 DNS 缓冲,目的是加快 URL 解析成 IP 地址的速度[4]。爬虫的运行过程如图 1 所示。

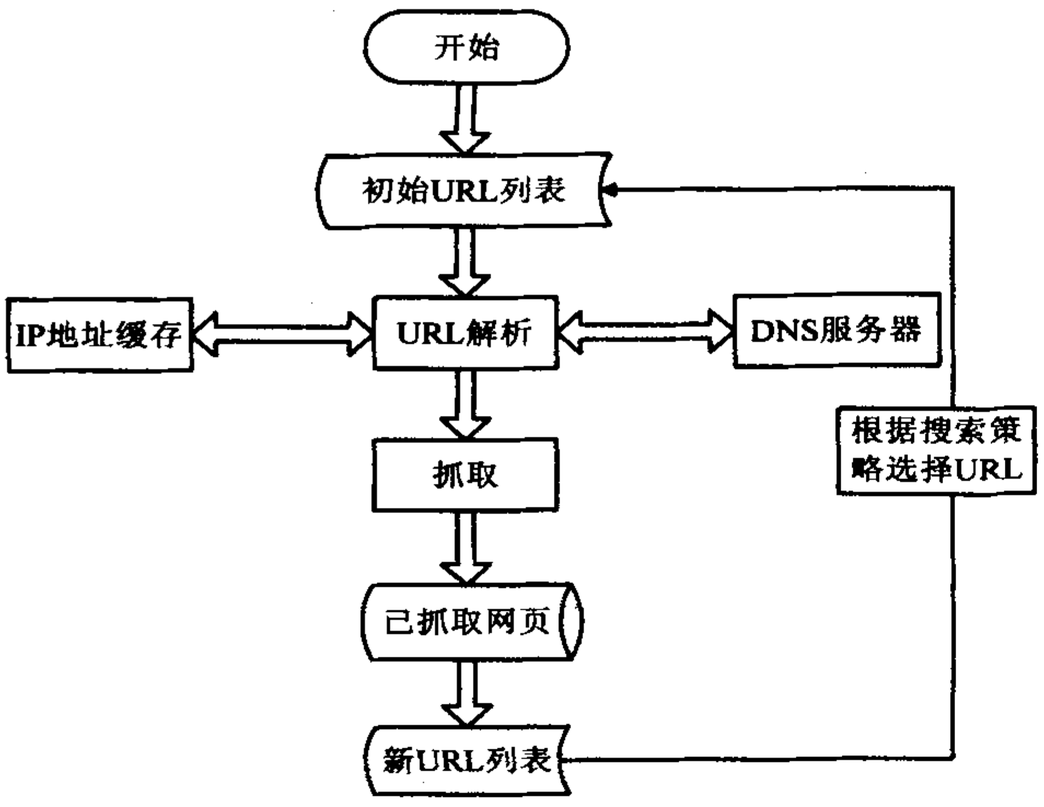

图 1　主题爬虫运行过程

2　遗传算法的基本思想

遗传算法[5] GA(Genetic Algorithms)是一种模拟自然界生物进化过程的计算模型,用于解决最优化的搜索算法。GA 克服了传统算法的一些缺点,对一些复杂的、难于数学建模的问题有着明显的优势。它体现了适者生存、优胜劣汰的进化原则,通过复制、交叉、变异等操作不断产生新的个体,并逐步淘汰适应度函数值低的个体,使群体不断进化,同时以全局并行搜索技术来搜索优化群体中最优个体,以求得到满足要求的解。

主要步骤如下 :

① 处理用户提交的 URL 集,抽取关键词;

②选中输入集的所有链接,并获相应的 WWW 表示,结果集作为第一代;

③对第一代所有元素计算其适应度;

④不断复制、交叉、变异操作,并从上代进入到下一代[6]。

其中,选择、交叉和变异三个主要遗传算子构成了遗传算法的遗传操作。

在主题爬虫中应用遗传算法可以在如下方面提高搜索的性能[7]:

1)内在启发式随机搜索:指导主题爬虫的搜索方向;

2)内在隐并行性:不容易陷入局部最优;

3)更好的全局寻优能力:提高主题爬虫的全局搜索能力;

4)渐进式优化:利用选择、交叉和变异等操作,通过不断遗传,选择最优集;

5)扩展性强:能够与其他技术混合使用在主题爬虫的设计中,如组合优化、机器学习和自适应控制。

3　基于遗传算法的搜索策略

3.1　设计思想

由于因特网上的信息是按照主题相关来分类,相关或重要的网页相互链接,所以基于网页链接结构评价的搜索策略[8,2]以此为基础来分析,但易出现“主题漂移”问题;基于内容评价的搜索策略[2]采用向量模型算法,却忽略了网页之间的链接关系。因此通过对以上两种策略的分析,再结合遗传算法的特性,决定采用如下策略:基于从优质个体链接出去的 URL 可能是优质个体和链接目标是优质个体的 URL 也可能是优质个体的原则,选出与主题相关度高的个体(URL)作为初始的种子集;交叉操作对父代个体进行基因变换,产生新的个体,再从中选择优异个体;通过变异操作,扩大搜索范围(URL 集)。

3.2　系统结构

通过对普通主题的研究分析,基于遗传算法的主题爬虫搜索策略在以下几方面对传统的设计进行了扩展:在海量网页的处理过程中加入选择操作,通过对结合文本内容的 PageRank 算法计算得到的网页重要性值排序来选择与主题相关的网页,在一定程度上解决了信息检索中出现的主题漂移问题;在每代种子挑选过程中增加变异和交叉操作,增加相关度较低的页面被爬行的机会并扩大相关网页的爬行范围,形成一个能反映主题特征的种子集合,其实现流程如图 2 所示。

3.3　主题的选择

用户在使用搜索引擎时是通过相关主题来检索他们想要的信息,从这个角度主题可以描述成某一事物或某一信息的若干特征的集合。实际中,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[9]。主题的确立可以采用手工设置关键词集的方法,该方法易于实现,依赖于专家的经验[10]。

3.4　初始种子集合的确立

初始种子集合的确立有三种常用的方法,第一是人工指定;第二是自动生成;第三是前两种方式的混合模式。由于爬虫是从初始种子集合中的 URL 开始,并解析这些链接,根据网页的重要性来选择进入 URL 的顺序,所以初始种子集合的选择将直接影响主题页面的采集以及采集工作的效率。另外,由于主题爬虫只搜索跟主题相关的 URL,搜索的范围局限于万维网中很小的一部分,所以更加要求初始种子集合是主题相关的。鉴于上述理由,文中采用混合模式,利用元搜索

引擎获取,在领域专家的指导下并结合人的经验共同进行筛选,选出与主题相关的 URL 作为初始种子集合[11]。

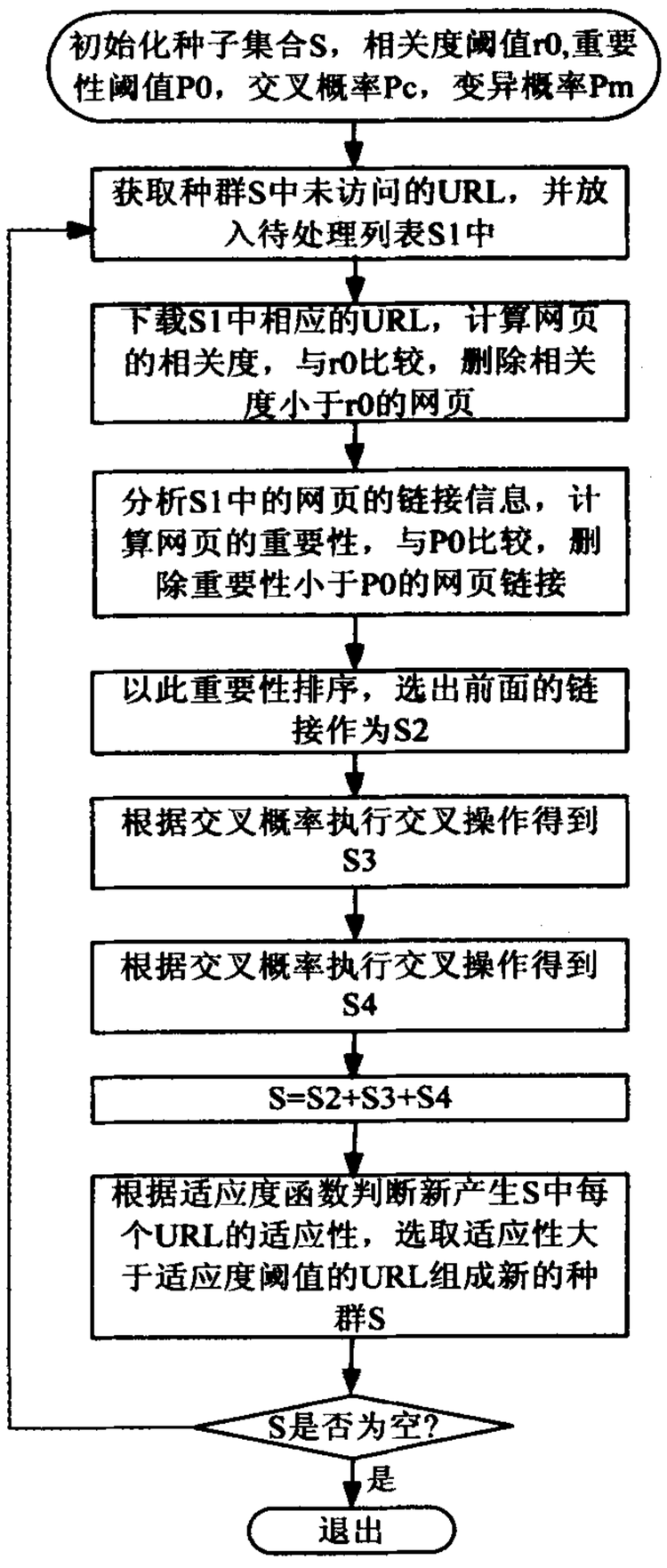

图 2 基于遗传算法的主题爬虫流程

3.5 主题相关度

计算主题相关度[12]采用向量空间模型算法。向量模型不判断文档与查询是否相关,而是根据文档与查询的相似度对文档进行排序。

选取文档中的词为向量空间的一个向量,由这些词作为向量的维数表示文档。文档和关键词都被假设是一个 n 维向量空间的一部分,其中 n 是关键词的个数,每一维分量的大小为每个关键词的权值 w_i,主题用向量则表示为:$\boldsymbol{\alpha}=(a_1,a_2,\cdots,a_n)$,$a_i=w_i$,$i=1,2,\cdots,n$。

采用绝对词频对页面进行统计分析,在实际计算中,让频率出现最高的关键词作为基准,设其频率 $x_i=1$,通过频率之比,计算出其他关键词的相对频率 x_i,则该页面对应向量的每一维分量为 x_iw_i。

页面主题用向量则表示为:

$$\boldsymbol{\beta}=(x_1w_1,x_2w_2,\cdots,x_nw_n),i=1,2,\cdots,n$$

用两个向量夹角的余弦表示页面的主题相关度:

$$\begin{aligned}\cos<\boldsymbol{\alpha},\boldsymbol{\beta}>&=\frac{(\boldsymbol{\alpha},\boldsymbol{\beta})}{|\boldsymbol{\alpha}||\boldsymbol{\beta}|}\\&=\frac{((a_1,a_2,\cdots,a_n),(x_1w_1,x_2w_2,\cdots,x_nw_n))}{|(a_1,a_2,\cdots,a_n)||(x_1w_1,x_2w_2,\cdots,x_nw_n)|}\\&=\frac{x_1w_1^2+x_2w_2^2+\cdots+x_nw_n^2}{\sqrt{w_1^2+w_2^2+\cdots+w_n^2}\sqrt{x_1^2w_1^2+x_2^2w_2^2+\cdots+x_n^2w_n^2}}\end{aligned}\tag{1}$$

阈值 r_0 是一个常量,当 $\cos<\boldsymbol{\alpha},\boldsymbol{\beta}>\geqslant r_0$ 时,该页面和主题相关。

3.6 网页重要性

能否把与用户检索需求相关的高质量文档放在排序结果最靠前的位置,是衡量搜索引擎性能的关键技术之一,因此网页的排序显得尤为重要。考虑到网页与主题的相关性以及网页之间的链接性,可以综合考虑主题相关度和链接分析两个关键因素。

其中链接分析采用结合文本内容的 PageRank 算法。PageRank[13,14]的基本思想是试图为所有网页赋予一个量化的价值度,每个网页被量化的价值通过一种递归的方式来定义,由所有链接到该网页的价值程度决定。具体地,基于网页之间的链接信息,设 F_i 是页面 i 的链出页面集,B_i 是页面 j 的链入页面集。则在任意时间点,爬虫位于页面 j 的概率 $P(j)$ 见式(2):

$$P(j)=(1-\beta)+\beta\sum_{i\in B_j}\frac{P(i)}{|F_i|}\tag{2}$$

其中,$0<\beta<1$,通常取值为 0.85。该算法认为一个网页被多次引用,则它可能是很重要的;一个网页虽然没有被多次引用,但是被重要的网页引用,则它也可能是很重要的,这就是权威(Authoritative)网页,对每个网页计算它的权威值,就可以对网页进行排序,从而找到最重要的权威网页[14]。

为了能够将页面的权威性与主题相关性结合起来,在网页间传递 PR 时考虑页面的主题相关性,相关度大的链出页面的 PR 就相对大一些[15]。改进后的 PageRank 见式(3):

$$P_t(j)=(1-\beta)P_t'(j)+\beta\sum_{i\in B_j}P_t(i)P_t(i\to j)\tag{3}$$

其中,t 为待搜索的主题,$P(i)$ 可由(2)式计算出来,$P_t(i\to j)$ 是当爬虫位于网页 i 时选择链出页面 j 的概率。$P_t'(j)$ 表示爬虫不选择出链而是直接跳转到页面 j 的概率,它们都和主题相关度有关。设 W 为所有网页的集合,$R_t(k)$ 为页面 k 关于主题 t 的相关性:

$$P_t(i\to j)=\frac{R_t(j)}{\sum_{k\in F_i}R_t(k)}=\frac{\cos(\alpha_t,\beta_j)}{\sum_{k\in F_i}\cos(\alpha_t,\beta_k)}\tag{4}$$

$$P_t'(j)=\frac{R_t(j)}{\sum_{k\in W}R_t(k)}=\frac{\cos(\alpha_t,\beta_j)}{\sum_{k\in W}\cos(\alpha_t,\beta_k)}\tag{5}$$

其中,$\cos < \alpha, \beta >$ 由式(1)可得。由式(4)与式(5)可得,结合了主题的 PageRank 按照链出页面的主题相关度来传递 PageRank 值。这样,同一页面的链出页面集里和主题相关性大的就能获得此页面的父页面较多的 PageRank 值。

3.7　适应度评价函数

个体适应度计算公式为:$\mathrm{Fit}(\mathrm{link}_i) = r_{pi} + r_{li}k$。其中$\mathrm{Fit}(\mathrm{link}_i)$代表第 i 个 URL 的主题相关度。r_{pi} 代表 link_i 对应父网页的主题相关度。r_{li} 反映了链接提示信息的主题相关度。k 是因子,取 100。

3.8　遗传操作

3.8.1　选择操作

在每一代种群中,对个体的适应度值进行排序,选择适应度值高的个体遗传到下一代群体中,淘汰适应度值低的个体。具体实现方法为:

①淘汰已搜索过的 URL;

②删除重复的 URL,并合并链接提示信息;

③计算链接对应 URL 的适应度值;

④根据适应度值,对个体的适应度值进行排序,选取适应度值大于适应度阈值的个体,构成新集合 S,然后进行下一轮的遗传操作[16]。

3.8.2　交叉操作

交叉操作把两个适应度值高的个体上的基因重组,产生两个新的个体,新的个体将进入新一轮的遗传操作。

第一,分析集合 $S1$ 中所有 URL 对应的网页,解析每个网页所包含的链接及链接提示信息,并计算每个网页的重要性;

第二,依照网页的重要性进行降序排序;

第三,若设交叉概率为 P_c,则选出前 $m \times P_c$ 个 URL 进行交叉操作,交叉结果集为 $S3$。

3.8.3　变异操作

变异操作通过个体基因突变的方式,产生新的个体,从而增加相关度较低的页面被爬行的机会,扩大相关网页的爬行范围,能够防止局部最优化。假设种子集合有 m 个 URL,由选择操作可知当前网页的重要性,按照重要性进行降序排序。设变异概率为 P_m,选出前 $m \times P_m$ 个 URL 进行变异,得到集合记为 $S4$。令 $S = S2 + S3 + S4$,则将 S 集合中对应的 URL 作为下一代种子。其中,交叉概率 P_c 与变异概率 P_m 之和为 1。

4　实验设计与数据分析

软硬件环境:Windows XP 系统,CPU:Pentium E6500,内存 2G,硬盘容量为 500GB,系统用 Myeclipse 开发,语言为 Java。

实验参数:线程数=10,起始种子=5,$r_0 = 0.0002$,变异概率 $P_m = 0.1$,交叉概率 $P_c = 0.9$,$P_0 = 2$,适应度阈值取 $r_0 = 0.0002$。

以“奥运”和“世博”为主题,从百度及 Google 上分别搜索获取的 5 个跟主题相关的 URL 作为初始 URL 集合。图 3 是以“奥运”为主题的基于两种算法的爬虫爬行网页的相关性的对比图,图 4 是以“世博”为主题的基于两种算法的爬虫爬行网页的相关性的对比图,均设定搜索深度为 2(设为较小值,防止搜索规模过大)。

图 3 与图 4 表明,基于遗传算法的主题爬虫策略与普通的爬出策略相比,爬取网页的主题相关性高;随着爬行网页的增多,基于遗传算法的主题爬虫策略,爬取网页的主题相关性趋于稳定,而普通的主题爬虫策略,爬取网页的主题相关性呈下降趋势。原因在于:在计算网页的重要性时,基于遗传算法的主题爬虫策略综合考虑了主题相关度和链接分析两个关键因素。

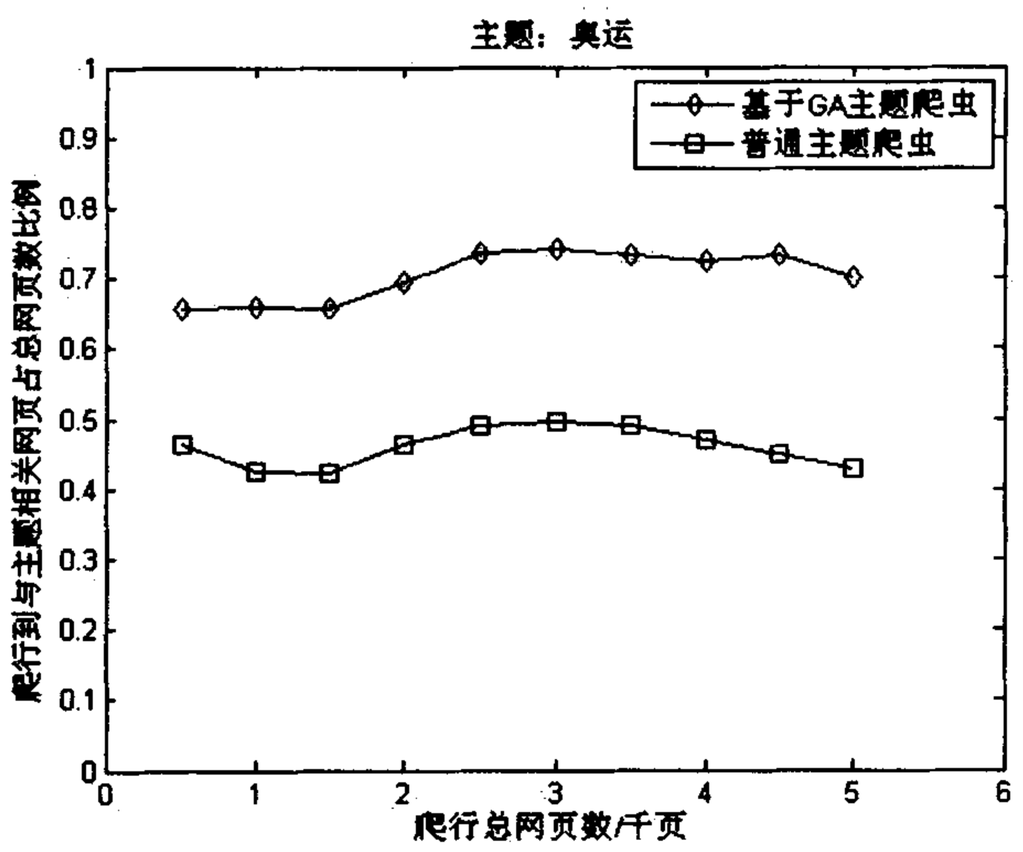

图 3　主题为“奥运”的爬行到网页相关性对比图

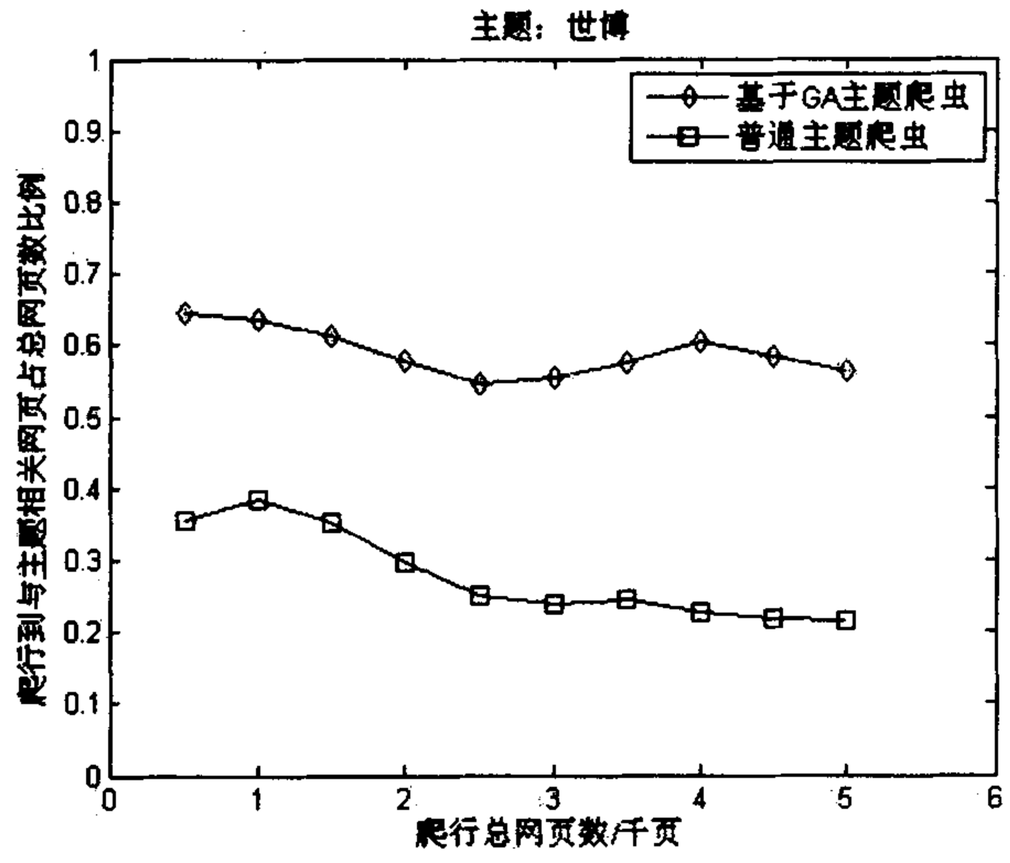

图 4　主题为“世博”的爬行到网页相关性对比图

表 1 是基于两种搜索策略在不同爬行深度的情况下对所抓取的网页的相关性进行的比较,从表 1 可以看出,在爬行深度不同而其他条件相同的情况下,随着

深度的增加,与主题相关的网页数占总网页数的比例有所下降,其原因在于:搜索的路径越深,一方面抓取到的网页的适应度越低,另一方面适应度低的网页的数量却越多。从表 1 还可以得出另一个结论,即基于遗传算法的主题爬虫爬行到与主题相关的网页占总网页数比例下降的幅度比普通主题爬虫爬行到与主题相关的网页占总网页数比例下降的幅度小。原因在于:通过设置变异算子,增加相关度较低的页面被爬行的机会,并扩大爬行与主题相关的网页的范围;在计算网页重要性时引进结合文本内容的 PageRank 算法,考虑了网页与网页之间的链接关系。

表 1 不同爬行深度下两种爬虫抓取网页的相关性的比较

爬行的深度	基于不同算法的爬虫	爬行到与主题相关的网页占总网页数比例
Depth:2	普通主题爬虫	40.16%
	基于 GA 主题爬虫	67.63%
Depth:3	普通主题爬虫	36.88%
	基于 GA 主题爬虫	65.32%

实验中还发现,网络的状况会影响到爬虫的爬行速度与结果。网络状况较好的情况,不但爬行速度快,而且爬行的网页的相关度高于同等情况下网络状况较差时爬取的网页的相关度。

5 结束语

分析结果表明:基于遗传算法的爬行策略能够有效地加快抓取网页的速度、扩大搜索范围以及提高搜索精度。但是对于如何便捷、有效地定义有实际意义的主题,如何在搜集网页信息时准确、快速地判定页面与主题是否相关以及如何提高系统搜索的完全度等,将是下一步要做的工作。

参考文献:

[1] Shokouhi M, Chubak P, Raeesy Z. Enhancing Focused Crawling with Genetic Algorithms [C]//International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II. [s. l.]: [s. n.], 2005: 503-508.

[2] 刘金红,陆余良. 主题网络爬虫研究综述[J]. 计算机应用研究,2007,24(10):26-29.

[3] 刘汉兴,刘财兴. 主题爬虫的搜索策略研究[J]. 计算机工程与设计,2008,29(12):3160-3162.

[4] 袁津生,李 群,蔡 岳. 搜索引擎原理与实践[M]. 北京:北京邮电大学出版社,2008.

[5] 赵大明,鱼 滨. 基于遗传算法的专业搜索引擎[J]. 计算机工程,2009,35(21):192-194.

[6] 丁永生. 计算智能-理论、技术与应用[M]. 北京:科学出版社,2004.

[7] 刘国靖,康 丽,罗长寿. 基于遗传算法的主题爬虫策略[J]. 计算机应用,2007,27(12):172-174.

[8] 刘 林,汪 涛,樊孝忠. 主题爬虫的解决方案[J]. 华南理工大学学报(自然科学版),2004,32(S1):137-140.

[9] 王峰松. 新一代智能搜索引擎-网典[J]. 网络世界,1999,13(2):12-21.

[10] 方加沛,黄 战. 基于单类别文档分类的主题爬虫[J]. 计算机工程与应用,2010,46(16):63-66.

[11] 关慧芬,师 军,马继红. 基于遗传算法的主题爬行技术研究[J]. 计算机与数字工程,2008,36(10):50-53.

[12] 袁 浩,黄烟波. 网页标题分析对主题爬虫的改进[J]. 计算机技术与发展,2009,19(6):22-24.

[13] Arasu A, Novak J, Tomkins A, et al. PageRank Computation and the Structure of the Web: Experiments and Algorithms [R]. [s. l.]: IBM Almaden Research Center, 2001.

[14] 冯振明. Google 核心-PageRank 算法探讨[J]. 计算机技术与发展,2006,16(7):82-84.

[15] Richardson M, Domingos P. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank [C]//Advances in Neural Information Processing System. [s. l.]: [s. n.], 2002: 673-680.

[16] 曹道友,程家兴. 基于改进的选择算子和交叉算子的遗传算法[J]. 计算机技术与发展,2010,20(2):44-47.

(上接第 47 页)

Distributed Systems, 2005, 16(11): 1078-1091.

[6] 王新生,梁 平,张云超,等. 结构化 P2P 路由协议的改进[J]. 计算机工程,2010,36(10):105-107.

[7] 林雅榕,候整风. 对哈希算法 SHA-1 的分析和改进[J]. 计算机技术与发展,2006,16(3):124-126.

[8] 徐 乾. 对等网中 chord 协议及算法的研究改进[D]. 哈尔滨:哈尔滨工业大学,2007.

[9] 张 浩,金 海,聂江武,等. Dual-chord:一种更加有效的分布式哈希表[J]. 微型计算机系统,2006,27(8):1450-1454.

[10] 张 震,王晓明. 对等网中 Chord 资源查找算法研究[J]. 计算机工程与应用,2006,42(11):147-152.

[11] 刘晓峰,吴亚娟,钟乐海. chord 路由表结构的改进与优化[J]. 计算机工程,2007,33(21):102-104.

[12] Karger D, Lehman E, Leighton T, et al. Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web[C]//Proceedings of the 29th Annual ACM Symposium on Theory of Computing[C]. [s. l.]: [s. n.], 1997.

基于 HKS1553BCRT 芯片的1553B总线通信软件设计

作者： 王绮卉， 田泽， 赵彬
作者单位： 中国航空计算技术研究所,陕西西安710119
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2012(8)

参考文献(16条)

1. Shokouhi M;Chubak P;Raeesy Z Enhancing Focused Crawling with Genetic Algorithms 2005
2. 刘金红;陆余良 主题网络爬虫研究综述[期刊论文]-计算机应用研究 2007(10)
3. 刘汉兴;刘财兴 主题爬虫的搜索策略研究[期刊论文]-计算机工程与设计 2008(12)
4. 袁津生;李群;蔡岳 搜索引擎原理与实践 2008
5. 赵大明;鱼滨 基于遗传算法的专业搜索引擎 2009(21)
6. 丁永生 计算智能-理论、技术与应用 2004
7. 刘国靖;康丽;罗长寿 基于遗传算法的主题爬虫策略 2007(12)
8. 刘林;汪涛;樊孝忠 主题爬虫的解决方案[期刊论文]-华南理工大学学报（自然科学版） 2004(z1)
9. 王峰松 新一代智能搜索引擎-网典 1999(02)
10. 方加沛;黄战 基于单类别文档分类的主题爬虫[期刊论文]-计算机工程与应用 2010(16)
11. 关慧芬;师军;马继红 基于遗传算法的主题爬行技术研究[期刊论文]-计算机与数字工程 2008(10)
12. 袁浩;黄烟波 网页标题分析对主题爬虫的改进[期刊论文]-计算机技术与发展 2009(06)
13. Arasu A;Novak J;Tomkins A PageRank Computation and the Structure of the Web:Experiments and Algorithms 2001
14. 冯振明 Google核心-PageRank算法探讨 2006(07)
15. Richardson M;Domingos P The Intelligent Surfer:Probabilistic Combination of Link and Content Information in PageRank 2002
16. 曹道友;程家兴 基于改进的选择算子和交叉算子的遗传算法[期刊论文]-计算机技术与发展 2010(02)