自适应遗传算法在主题爬虫搜索策略中的应用研究

荆文鹏 王育坚 董伟伟
(北京联合大学信息学院 北京 100101)

摘 要 如何提高爬虫覆盖率和准确率是主题爬虫的研究热点之一。目前大多采用最佳优先搜索策略,针对该类主题爬虫易陷入局部最优的不足,设计结合遗传算法的主题爬虫搜索策略,并设计动态适应度函数和遗传算子使得爬虫具有一定的自适应性。与其他搜索策略以及结合非自适应遗传算法的搜索策略进行了比较,结果表明该算法能够在一定程度上提高爬虫性能。

关键词 主题爬虫,重要度,遗传算法,遗传算子,适应度函数
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.8.051

Research on Adaptive Genetic Algorithm in Application of Focused Crawler Search Strategy

JING Wen-peng WANG Yu-jian DONG Wei-wei
(College of Information Technology, Beijing Union University, Beijing 100101, China)

Abstract How to design the crawler search strategy to improve the crawler's coverage and accuracy has become a hot research point in the focused crawler. Mostly crawler uses best-first search algorithm. Based on the focused crawler which uses this search strategy will easily plunge into local optimum, we combined genetic algorithm with focused crawler search strategy. We set dynamic fitness function and genetic-operators to make the crawlers have certain adaptive searching adaptability. By comparing with those crawlers which use the other search strategy or which combine with traditional genetic algorithm search strategy, the experimental results show that this algorithm can partly improve the crawler search ability.

Keywords Focused crawler, Important degree, Genetic algorithm, Genetic operators, Fitness function

网络爬虫是一种自动搜索并下载互联网资源的程序或脚本,是搜索引擎的重要组成部分,主要负责将互联网上的资源下载到本地,在本地形成网页镜像备份。百度、Google采用的通用网络爬虫将全部网络资源下载到本地,这需要大量的存储空间、计算量和带宽,但该类爬虫对于一些主题针对性较强或特定领域的搜索引擎并不合适。主题爬虫在通用网络爬虫模型的基础上加入网页分析、链接分析部分,过滤掉无关网页,尽量减少无关页面在本地生成镜像备份,因此在一定程度上减少了资源的消耗。图1示出一个主题爬虫模型。该模型首先手动指定一部分URL作为种子集合,将这些种子URL放入等待队列等待下载,按照一定规则从队列中读取URL,并从Web下载对应页面。下载到本地的页面经过内容的相关度分析,直接舍弃小于相关度阈值的网页。对于大于相关度阈值的网页,一方面将其存储到本地,等待建立索引等后续处理;另一方面,将下载页面的URL放入完成队列。对于刚下载的页面,提取其中包含的链接进行分析,通过预测链接所指向的页面的相关度,将符合要求的链接加入到等待队列等待下载。循环此过程,直到等待队列为空,即表示进行了一轮完整的爬取。

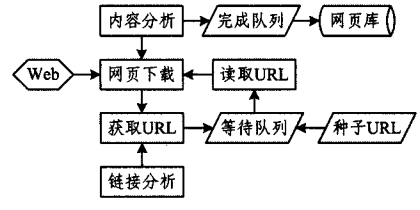

图1 主题爬虫模型

网页分析可分为基于网页内文本内容和基于网页内Web链接两种方式。针对网页内容的分析,国内外研究人员提出了布尔模型[1]、向量空间模型[2]、贝叶斯分类器算法[3]以及Shark-Search算法[4]等,针对网页内Web链接的分析,研究人员提出了PageRank算法[5]、HITS算法[6]等,并在基于内容分析和Web链接的基础上为了优化爬虫性能提出了二次或多次爬行算法[7]、基于概率模型的算法[8]等。针对现在同类型网站(例如微博、视频网站、SNS(Social Networking Services)网站等)的网页结构大多类似的情况,又提出了针对同类型网站的主题爬虫搜索策略[9]。

网页重要度用于决策按照何种规则从等待队列获取URL,从而决定爬虫的爬取方向。因此网页搜索策略的本质

到稿日期:2015-05-21 返修日期:2015-08-21 本文受国家自然科学基金项目:基于超图形XGML的图像半结构化研究(61271369)资助。
荆文鹏(1989—),男,硕士生,主要研究方向为网络爬虫、智能算法;王育坚(1963—),男,教授,主要研究方向为软件理论、图像处理,E-mail:xxtyujian@buu.edu.cn(通信作者);董伟伟(1988—),男,硕士生,主要研究方向为数字图像处理。

就是通过搜索最优解，确定在待下载页面队列中选择哪个页面进行下载，从而决定爬虫的搜索路径。目前大多爬虫采用最佳优先搜索策略，即每次都下载当前队列中重要度最高的网页，但由于最佳优先算法本身欠缺全局性，因此爬虫容易陷入局部最优。

遗传算法基于群体，从而在一定程度上提高了全局搜索性，避免了爬虫陷入局部最优；并且遗传算法的适应性比较强，往往可以结合多种其他策略对爬虫性能进行优化。基于以上研究和分析，并结合遗传算法，对目前搜索策略的不足加以改进，使得爬虫不易陷入局部最优，并且具有一定的自适应性。

1 遗传算法和爬取策略研究

1.1 遗传算法

遗传算法 GA(Genetic Algorithm)由 Holland 教授在 1975 年提出。遗传算法作为一种智能算法，模拟了自然生物的优胜劣汰，从而获得最优解，适用于多类问题，如函数优化、组合优化等。传统遗传算法的流程如图 2 所示，主要分为以下几个步骤：

(1)确定待解决问题的规模，即建立种群初始状态。

(2)按照一定的策略对个体进行编码，构建群体。

(3)根据适应度函数评估个体适应度，按照一定的规则淘汰适应度较低的群体。

(4)对剩余群体进行选择、交叉和变异 3 步遗传操作，从而获得适应度更高的群体，产生下一代。选择操作选择优胜个体，淘汰劣势个体，保证适应度较高的个体进入下一代遗传操作；交叉操作是指把两个父代个体的部分结构加以替换重组，从而生成新的个体，大大提高了遗传算法的搜索能力；变异操作则在一定程度上提高了种群多样性。

(5)反复进行步骤(3)和(4)，直到满足预设终止条件，从而获得最优解。

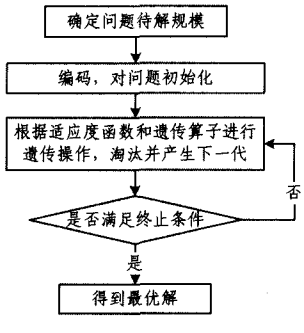

图 2 遗传算法流程图

1.2 PageRank 算法

网页重要度主要通过对等待队列中 URL 的分析，决定等待队列中的 URL 下载顺序，从而决定爬虫的爬行方向。网页重要度的代表算法有 PageRank 算法、HITS 算法等。PageRank[10]算法由 Google 创始人于 1997 年在构建搜索系统原型时提出，算法的基本思想为：第一，如果一个页面 P 接受到其他页面所指向的入链数目越多，则页面 P 越重要；第二，入链质量不同，若越是质量高的网页指向 P，则 P 页面越重要。

在计算页面 PageRank 值中，给每个页面赋以一个相同的初始 PageRank 值，每个页面将其 PageRank 值平均分配给本页面包含的所有出链所对应的页面上，每个页面对所有入链传递的 PageRank 值求和，得到该页面的真实 PageRank 值。任意一个页面 A 的 PageRank 值可以由式(1)计算得到。

$$PR(A)=\sum_{N\in P}\frac{PR(N)}{L(N)} \tag{1}$$

式中，P 为 A 页面所有入链对应的页面集合，$PR(N)$ 表示页面 N 的 PageRank，$L(N)$ 表示 N 页面所包含的出链数目。

可以看出，PageRank 算法是一种全局性算法，需要所有页面下载完成即页面集合 P 确定后，才能通过计算得到可靠的页面 PageRank 值，但这在实际中是不可能的，因此常采用非完全 Page-Rank算法来决定网页重要度。非完全 PageRank 算法的基本思想为：将已下载页面与等待队列中所有 URL 一起形成一个网页集合，在此集合内进行页面的 PageRank 值计算。由于不断有新的 URL 进入等待队列，每抓取到一个新的 URL 就进行一次 PageRank 值的计算，效率明显过低，因此采用这样的方式：1)在新抓取网页数目达到一个预设值 K 时，将所有已下载页面与等待队列中的 URL 形成一个新的集合，重新计算一次该集合内页面的 PageRank 值；2)新抓取网页数目尚未达到 K 时，有一些新加入的 URL 的重要度非常高，理应优先下载这些 URL。因此为每个新加入的 URL 设置一个临时 PageRank 值，将所有入链的 PageRank 值相加得到该网页的临时 PageRank 值，如果这个值比待抓取 URL 队列中已经计算得到的最大 PageRank 值大，那么优先下载这个 URL 对应的页面。

1.3 VSM 向量空间模型

VSM (Vector Space Model)算法[11]通过对文本的比较判断两个文本文档之间的相似度。计算方式如式(2)、式(3)所示。

$$Sim(s_i,s_j)=\frac{\sum_{k=1}^{n}(w_{i_k}\times w_{j_k})}{\sqrt{\sum_{k=1}^{n}(w_{i_k})^2\sum_{k=1}^{n}(w_{j_k})^2}} \tag{2}$$

$$w_{i_k}=\frac{tf(i_k,p)}{\sum_{m=1}^{n}if(i_m,p)}\times\log(\frac{N}{n_k}) \tag{3}$$

式(2)中，s_i，s_j 分别为用户给定的关键字和网页锚文本两个文本集合；w_{i_k}，w_{j_k} 分别表示这两个集合中第 k 个向量的权值。利用式(3)计算 w_{i_k}，w_{j_k}。式(3)是 TF×IDF 公式，TF 为词频(Term Frequency)，词频为词语 i_k 在页面 p 中出现的频率。IDF 为逆文档频率(Inverse Document Frequency)，N 表示 Web 上的网页总数，n_k 表示包含该词的网页总数+1(防止分母为 0)。显然，分母越大表示该词越常见，这些词越难代表某类主题，相应的权重就越小。利用 VSM 进行计算时，对文本的中文分词方式有很多种，选择不同的分词方式，计算结果也不尽相同。

2 主题网络爬虫设计

2.1 适应度函数的设计

适应度作为决定是否被淘汰的指标，若直接使用页面 PageRank 大小作为页面下载与否的标准，会存在几个问题。

由于互联网链接十分复杂，很多时候在某几个页面中会形成互相有链接指向的环形结构，导致这种结构中的页面只能接受传入的 PageRank 值，而不能将 PageRank 值传出，从而使得结构内页面的 PageRank 值越来越高。这个结果势必影响

选择过程，导致过早收敛，从某种程度上降低了爬虫的搜索能力。

针对这个问题，采取如下解决办法：网页通过出链传递PageRank值时，以一定概率加入一个其他网页，相当于虚拟一条出链指向其他网页，PageRank值可以通过这条虚拟出链向外传递，从而避免了陷入环形的链接陷阱。

PageRank算法计算网页重要度的过程中易出现主题漂移现象，即网页重要度很高，但与主题无关。为了避免出现主题漂移现象，对于链接附近的锚文本，利用VSM模型对URL所指向的页面进行主题相关度预测。另外，在此过程中统计网页中相关链接数目在全部链接中所占的比例，动态调整页面PageRank值，可以在一定程度上防止通过增加许多不相关链接来提高网页重要度的网页作弊行为。

网页随着时间变化不断地更新，入链数目会随时间快速增加，导致对应页面的重要度增加。适应度函数设计中引入时间调节因子，假设某个网页随时间变化，入链数目增加，并在24小时后达到峰值。

根据以上分析，提出如式(4)所示的适应度函数。

$$F(P)=\frac{t}{24}((1-k)PR(P)+Sim'(P)) \tag{4}$$

$$Sim'(P)=\frac{I(P)}{L(P)}\sum_{i=1}^{I}Sim(P) \tag{5}$$

式(4)中，$PR(P)$表示P页面的PageRank值；$Sim'(P)$表示链接锚文本与主题通过VSM模型计算得到的URL相关度预测值；k为存在虚拟出链从而跳转到其他页面的概率。式(4)中的$Sim'(P)$通过式(5)进行计算，$I(P)$表示P页面中预测结果为主题相关的链接数目，$L(P)$表示页面P包含的全部链接数目，$Sim(P)$为每个相关链接的相关度。

2.2 遗传算子设计

进行遗传操作(选择、交叉和变异操作)之前，需给定两个参数的值：交叉概率p_c以及变异概率p_m。这两个参数值可以根据实验结果进行调整，从而优化爬虫。为了在动态适应度函数(4)的基础上进一步提高爬虫的自适应性。p_c和p_m利用式(6)、式(7)进行计算，并采用动态调整的方式[12,13]。

$$p_c=\begin{cases}\frac{i}{I}(f_{\max}-f_v), & f_{\max}-f_v<kf_{\max}\\ \alpha, & f_{\max}-f_v>kf_{\max}\end{cases} \tag{6}$$

$$p_m=\begin{cases}(\frac{i}{I}+\frac{f_v}{f_{\max}})(f_v-f_{\min}), & f_v-f_{\min}>kf_{\min}\\ \beta, & f_v-f_{\min}<kf_{\min}\end{cases} \tag{7}$$

其中，α,β可通过实验进行调整(在$[0,1]$之间取值)，用来保证p_c,p_m保持在一定范围内；k一般取值0.9；$f_{\max},f_{\min},f_v$分别代表本次遗传操作中适应度的最大值、最小值和平均值；i表示当前进化次数；I表示总进化次数。

2.3 算法流程

URL由5个队列共同进行管理和维护，同一URL同时只能存在于以下队列中的一个。1)等待队列。新发现的URL被放入此队列。2)处理队列。爬虫对URL进行处理时，URL移动到此队列。URL被处理后，根据处理结果，移动到完成、错误、舍弃这3个队列中的一个。3)完成队列。正常下载后，对下载页面进行相关度计算，如果计算结果大于预设阈值，则URL移动到此队列。4)舍弃队列。正常下载后，相关度计算结果小于预设阈值，URL移动到此队列。5)错误队列。下载时发生错误，如网站发生变更、等待时间过长时，URL移动到此队列。

进行第一次抓取前，指定初始URL种子集。指定主题后使用如Google、百度等通用搜索引擎进行搜索，对搜索结果进行进一步精选后，将结果作为初始页面种子集合，记为S_0。

进行选择操作，将S_0中的URL写入到等待队列，爬虫依次从等待队列中取出URL所对应网页进行处理，计算对应网页的适应度，如果大于预设阈值，将该URL移动到完成队列。反复操作，直到等待队列为空，将此时完成队列中的URL记为集合S。

获取S中所有URL对应页面包含的子链接，对结果集过滤，删除重复和已下载的页面，并计算过滤后集合中URL所对应页面的适应度，将结果集记为S_1。

进行交叉操作，对集合S_1中的URL按照适应度降序排列，记S_1中URL数量为N_1，根据交叉概率p_c取得S_1中前p_c*N_1个URL，对结果集过滤，删除重复和已下载的页面，记为S_2。

进行变异操作，根据URL命名规则“协议://主机名[:端口号]/路径/[;可选参数][?查询]#片段”，获取S_2中URL所对应相同主机地址及一级目录的所有URL集合，计算该集合中URL数量N_2，并对此集合中URL所对应页面进行适应度计算，对结果降序排序，根据变异概率p_m取得S_1中前p_m*N_2个URL，结果集记为S_3。

进行并集操作即$S_1\cup S_2\cup S_3$，其结果将作为下一代遗传操作的初始集合。

重复以上操作，直到满足以下3个终止条件中任意一个为止：1)等待队列为空；2)新一代种群所有URL适应度低于阈值；3)已下载页面达到设置的最大下载页面数。另外，每次遗传操作之后都要进行网页过滤，即删除重复、无效和已下载页面。算法流程如图3所示。

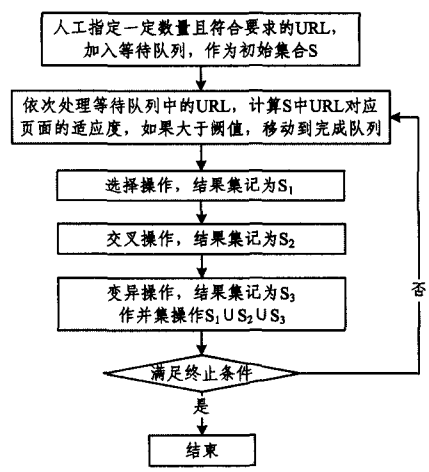

图3 爬虫算法流程图

3 实验结果

3.1 评价指标

准确率和覆盖率是两个评价主题网络爬虫优劣的重要指标。正确率A根据(8)式进行计算，覆盖率C根据式(9)进行计算。其中，R表示主题相关网页的数量，N表示所有已下载网页的数量，T表示互联网上所有主题相关网页的数量。

$$A=R/N \tag{8}$$

$$C=R/T \tag{9}$$

由于新的网页层出不穷，式(8)中分母不容易统计，因此一般直接采用分子来代表覆盖率，即在一定的准确率的基础上，爬行到的网页越多，爬虫性能越好。

3.2 实验设计及结果分析

实验在开源爬虫 Heritrix 的基础上定制完成。以“计算机”、“编程”、“软件”、“开源”、“程序”、“数据库”6 个词语构成主题文档，分别使用百度及 Google 搜索获取前 50 个主题相关的页面 URL，并从中精选出 20 个作为初始 URL 种子集合。实验中设置最大下载页面数目作为终止条件，在此基础上统计相关页面的数量。为了对比，实验采用另外 3 种不同的计算方式：广度遍历、结合 HITS 算法的搜索策略和结合非自适应遗传算法(GA)的搜索策略。在覆盖率相同的情况下对准确率作横向对比；对传统遗传算法和自适应的遗传算法作纵向对比。统计实验数据，并用比较直观的折线图对结果进行分析，实验结果如图 4、图 5 所示，图中 X 轴表示最大下载页面数，Y 轴表示相关页面数。

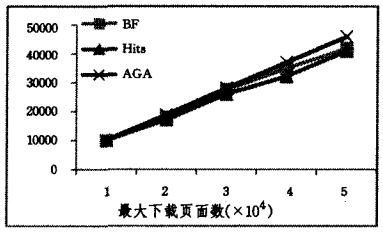

图 4 实验结果(横向比较)

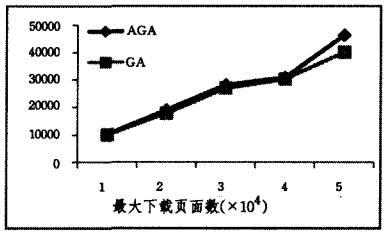

图 5 实验结果(纵向比较)

由图 4 可以得出，结合自适应 GA 的搜索策略在爬行网页数目较少时与其余两种搜索策略相比准确率基本相同，但随着爬行网页数目的增多，应用结合自适应 GA 搜索策略的主题爬虫不易陷入局部最优，能够搜索到较多的相关页面。由图 5 可以得出，随着爬行页面数目的增多，结合自适应 GA 搜索策略的主题爬虫的搜索能力相比结合非自适应 GA 的主题爬虫有明显提高。

结束语 主题爬虫爬行策略是主题爬虫的关键技术。遗传算法具有较强的全局优化能力，针对利用启发式搜索策略的主题爬虫容易陷入局部最优的情况，将遗传算法与主题爬虫搜索策略相结合。采用动态适应度函数和动态交叉、变异算子，使得页面淘汰及优选更为准确。实验结果表明，在搜索页面不断增多的情况下，该搜索策略有助于提高主题爬虫的准确率。遗传算法与其他算法有着很好的兼容性，今后将以与其他算法结合的主题爬虫搜索策略为方向进行进一步研究。

参 考 文 献

[1] Xian Xiao-ping. An algorithm based on a comprehensive improvement of PageRank algorithm[D]. Xi'an: Northwest University, 2010(in Chinese)
县小平. 搜索引擎 PageRank 算法研究[D]. 西安：西北大学，2010

[2] Zou Yong-bin, et al. Research on focused crawler based on Bayes classifier[J]. Application Research of Computers, 2009, 26(9): 3418-3420, 3439(in Chinese)
邹永斌，等. 基于贝叶斯分类器的主题爬虫研究[J]. 计算机应用研究，2009，26(9)：3418-3420，3439

[3] Luo Lin-bo, et al. Research on Topical Crawler of Shark-Search Algorithm and HITS Algorithm[J]. Computer Technology and Development, 2010, 20(11): 76-79(in Chinese)
罗林波，等. 基于 Shark-Search 和 Hits 算法的主题爬虫研究[J]. 计算机技术与发展，2010，20(11)：76-79

[4] Song Hai-yang, et al. A Novel Crawling Strategy of Focused Web Crawler[J]. Computer Application and Software, 2011, 28(11): 264-267, 293 (in Chinese)
宋海洋，等. 一种新的主题网络爬虫爬行策略[J]. 计算机应用与软件，2011，28(11)：264-267，293

[5] Wei Jing-jing, et al. Focused Crawler Based on Improved Algorithm of Web Content Similarity[J]. Computer and Modernization, 2011, 193(9): 1-4(in Chinese)
魏晶晶，等. 基于网页内容相似度改进算法的主题网络爬虫[J]. 计算机与现代化，2011，193(9)：1-4

[6] Bai Yu-zhao, et al. Research and implementation for focused crawler based on probabilistic model[J]. Computer Engineering & Science, 2013, 35(1): 160-165(in Chinese)
白玉昭，等. 基于概率模型的主题爬虫的研究和实现[J]. 计算机工程与科学，2013，35(1)：160-165

[7] Liu Zuo-da, et al. Focused Crawling Algorithm for BBS Information Retrieval [J]. Journal of Zhengzhou University (Natural Science Edition), 2010, 42(2): 22-25(in Chinese)
刘佐达，等. 一种面向 BBS 信息检索的主题网络爬虫算法[J]. 郑州大学学报(理学版)，2010，42(2)：22-25

[8] Deng Yue-gui. Heuristic Search in Network Crawler Application Analysis[J]. Software Guide, 2008(2): 80-82(in Chinese)
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(2)：80-82

[9] Salton G. Automatic Text Processing: The Transformation, Analysis, and Tetrieval of Information by Computer[M]. Addison-Wesley, Reading, Pennsylvania, 1989

[10] 玄光男，程润传. 遗传算法与工程设计[M]. 汪定伟，等译. 北京：科学出版社，2000

[11] Li Lu, Zhang Guo-yin, et al. Defence Industry Secrecy Examination and Certification Center Laboratory[J]. Computer Science, 2015, 42(2): 118-122(in Chinese)
李璐，张国印，等. 基于 SVM 的主题爬虫技术研究[J]. 计算机科学，2015，42(2)：118-122

[12] Li Dong, Pan Zhi-song. Research on Parallel Genetic Algorithms Based on MapReduce[J]. Computer Science, 2012, 39(7): 182-184, 204(in Chinese)
李东，潘志松. 一种适用于大规模变量的并行遗传算法研究[J]. 计算机科学，2012，39(7)：182-184，204

[13] Srinivas M, PatnaikI M. Adaptive Probabilities of Crososverand Mutationin Genetie Algorithm [J], IEEE Trans. on Systems. Manand Cybenreties, 1994(4): 656-667