一种基于蚁群算法的主题爬虫搜索策略

陈永彬,张 琢,张 添

(东北师范大学 理想信息技术研究院,吉林 长春 130024)

摘 要:针对目前主题爬虫采用"启发式"搜索策略出现的"近视"缺点,提出了一种基于蚁群算法的主题爬虫搜索策略。该方法将蚁群算法引入到主题爬虫的搜索策略中,并对蚁群算法中信息素的更新计算进行了改进,使其具有一定的自适应性。通过与其他搜索策略的比较实验,结果表明该算法能够更好地提高爬虫的全局搜索能力。

关键词:主题爬虫;蚁群算法;搜索策略;信息素

中图分类号:TP301 **文献标识码**:A **文章编号**:1674-7720(2011)01-0053-04

A searching stategy in topic crawler using ant colony algorithm

Chen Yongbin, Zhang Zhuo, Zhang Tian

(Institute of Ideal Information and Technology, Northeast Normal University, Changchun 130024, China)

Abstract: This text advances a searching stategy in topic crawler using ant colony algorithm, in accordance with the defect of "near-sighted" for the topic crawler using heuristic searching strategy. Ant colony algorithm is introduced to the topic crawler searching strategies and updated the computational method of pheromone, which has certain adaptability. Through the comparison with other searching strategies, the test results indicate that this algorithm has more excellent ability in searching the whole best solution.

Key words: topic crawler; ant colony algorithm; searching stategy; pheromone

主题网络爬虫是根据一定的网页分析算法,过滤与主题无关的链接,保留主题相关的链接并将其放入待抓取的超链接队列中,然后根据一定的搜索策略从队列中选择下一步要抓取的网页链接,并重复上述过程,直到达到系统的某一条件时停止。所有网络爬虫抓取的网页将会被系统存储,进行一定的分析、过滤,并建立索引[1]。相对于通用爬虫,主题爬虫搜索的内容只限于特定主题或专门领域,因而被通用网络爬虫广泛采用的基于广度或深度优先算法已不再适用。目前,主题网络爬虫通常采用启发式搜索策略,每次选择"最有价值"链接进行优先访问,但是这类策略容易过早陷入 Web 搜索空间中局部最优子空间的陷阱,缺乏全局性,从而导致整体回报率不高[2]。

蚁群算法不仅能够智能搜索和全局优化,而且还具有鲁棒性、正反馈、分布式计算、易于与其他算法结合等特点。利用正反馈原理,可以加快进化过程。分布式计算使该算法易于并行实现,个体之间不断进行信息交流和传递,有利于找到较好的解,不容易陷入局部最优。易与多种启发式算法结合,可改善算法的性能。稳健性强,故在基本蚁群算法模型的基础上进行修改,便可用于其他问题。

结合蚁群算法,本文针对主题爬虫搜索策略上的不足,提出了一种基于蚁群算法的主题爬虫搜索策略。由于对蚁群算法进行了改进,所以提出的算法还具有一定的自适应性。

1 蚁群算法模型

蚁群算法是群集智能体现的一个典型例子,该算法是意大利学者 Marco Dorigo[3]等人在 1991 年受蚂蚁觅食行为的启发而提出的。

蚁群算法借鉴和吸收了现实世界中蚁群的行为特征:蚂蚁属于群居昆虫,个体行为极其简单,而群体行为却很复杂。相互协作的一群蚂蚁很容易找到从蚁巢到食物源的最短路径。此外,蚂蚁还能够适应环境的变化,例如在蚁群的运动路线突然出现障碍物时,它们能够很快地重新找到最优路径。蚂蚁个体之间在觅食过程中通过信息素来进行信息传递,信息素随着时间的推移会逐渐

挥发。蚂蚁在觅食过程中能够感知信息素的存在及其强度，并以此来指导自己的运动方向，倾向于朝着信息素强度高的方向移动，即选择该路径的概率与当时这条路径上信息素强度成正比。信息素强度越高的路径，选择它的蚂蚁就越多，则在该路径上留下的信息素的强度就更大，而强度大的信息素又吸引更多的蚂蚁，从而形成一种正反馈。通过这种反馈，使得大部分蚂蚁都会走这个最佳路径。

正反馈的副作用就是当许多蚂蚁都选中同一条路径时，该路径中的信息素量会迅速增大，从而使得多只蚂蚁集中到某一条路径上，造成一种堵塞和停滞现象，表现在使用蚁群算法解决问题时就容易导致早熟和局部收敛。

2 基于蚁群算法的搜索策略

2.1 算法思想

本文提出了一种基于蚁群算法的主题爬虫搜索策略，其基本思想是：在 Web 页面中存在超文本页面 w_i 和 w_j，如果 w_i 中有一个链接指向 w_j，那么处于 w_i 的蚂蚁自身将根据一定的条件决定是否从 w_i 移动到 w_j。每个链接序列代表了一个可能的蚂蚁移动路线。蚂蚁个体之间在移动过程中通过信息素来进行信息传递。信息素在蚂蚁爬行过程中会随着时间的推移逐渐挥发。蚂蚁在页面之间的爬行被分为多个循环周期，在每个周期中，一个蚂蚁在 Web 页面间进行一系列的移动，直到探寻到目标资源并返回到源点为止。每完成一次爬行周期，蚁群对各路线上的信息素量进行更新。为解决蚁群算法的“早熟”和“局部收敛”问题，本文借鉴了参考文献[4]中动态自适应的调整信息素的思想。

假设 V 代表全体页面集合，E 代表由链接构成的路径集合，则 Web 页面（链接）构成有向图 $G=\{V,E\}$。因为蚂蚁在选择下一个 Web 页面时必须考虑其主题相关度，所以有向图 G 中页面 Pk 的主题相关度值可以参考 PageRank 算法公式。

为方便表述，作如下定义[5]：

定义 1 $R(Pk)$为页面的主题相关度值为：

$$R(Pk)=(1-d)+d(\sum_{t=1}^{k}R(Ti)/ci) \tag{1}$$

d 为调节因子（一般取值 0.8），Ti 为链接到 Pk 的页面集合，ci 为 Ti 的链接出度，即从 Ti 指向其他页面的链接数目。

定义 2 d_{ij} 表示页面 w_i 与 w_j 之间的距离，由下述公式计算得到：

$$d_{ij}=\frac{C}{\sum_{t=1}^{k}R(Pi)} \tag{2}$$

C 为一协调因子常量，$Pi(i=1,2,\cdots,n)$为由页面 i 转移到 j 所经过的页面集合。

定义 3 $\Delta\tau_{ij}$ 表示本次循环中留在路径 $e(i,j)$上的信息素量。$\Delta\tau^{k}_{ij}$ 表示第 k 只蚂蚁在本次循环中留在路径 $e(i,j)$上的信息素量。

$$\Delta\tau^{k}_{ij}=\begin{cases}\dfrac{Q}{L_k}, & \text{若第 } k \text{ 只蚂蚁本次循环经过}\\ 0, & \text{其他}\end{cases} \tag{3}$$

其中 Q 是常数，表示第 k 只蚂蚁在本次循环中所走过路径的长度，它可表示为：

$$L_k=\sum_{i=1,j=1}^{n}d_{ij} \tag{4}$$

式中，n 表示蚂蚁 k 在本次循环中所漫游的 Web 节点数目。

2.2 算法模型

设 Web 页面节点个数为 n，蚂蚁个数为 m，则 $b_{i(t)}$表示 t 时刻位于页面 w_i 的蚂蚁个数，则有 $m=\sum_{i=1}^{n}b_{i(t)}$；$\tau_{ij}(t)$表示 t 时刻在边上 $e(i,j)$残留的信息素量。初始时刻，在各条路径上的信息素量相等，设 $\tau_{ij}(0)=c$（c 为常数，通常取为 0）。蚂蚁 $k(i=1,2,\cdots,m)$在运动过程中，根据各条路径上的信息素强度来决定下一步所行进的路径，t 时刻蚂蚁 k 由位置 i 转移到 j 的概率 $p^{k}_{ij}(t)$由下式表示：

$$p^{k}_{ij}(t)=\begin{cases}\dfrac{[\tau_{ij}(t)]^{\alpha}\cdot[\eta_{ij}]^{\beta}}{\sum_{l(i,j)\in E}[\tau_l(t)]^{\alpha}\cdot[\eta_l]^{\beta}}, & j\in V\\ 0, & \text{其他}\end{cases} \tag{5}$$

其中，$l(i,j)\in E$表示给定页面 Pi 中存在一条从 w_i 到 w_j 的链接。为了避免蚁群出现环路爬行并约束蚂蚁进行页面递增探索，每个蚂蚁存储一个禁忌表 $tabu$，以记录被访问的链接。如果 w_i 属于 $tabu$，则从 w_i 到 w_j 的路径概率值 $p^{k}_{ij}(t)$为 0，从而禁止蚂蚁 k 探索该链接。在每个周期的结束，$tabu$ 表将被清空。式中参数 η_{ij}、α 和 β 将在 2.4 小节详细探讨。

信息素强度 $\tau_{ij}(t)$随时间的失衡会逐渐消逝。设信息素保留系数为 ρ（$0<\rho<1$），它体现了信息素强度的持久性；而 $1-\rho$ 表示信息素的消逝程度。当达到每周期限制的移动次数时，各路径上的信息素量要根据式(6)进行更新：

$$\tau_{ij}(t+1)=\rho\tau_{ij}(t)+\Delta\tau_{ij},\ \Delta\tau_{ij}=\sum_{k=1}^{m}\Delta\tau^{k}_{ij} \tag{6}$$

但是为了避免蚁群算法的“早熟”和“局部收敛”问题，本文根据参考文献[6]对各路径上信息素量采用如式(7)、式(8)作出调整：

$$\tau_{ij}(t+1)=\rho^{1+\varphi(n)}\cdot\tau_{ij}(t)+\Delta\tau_{ij},\ \text{当 } \tau>\tau_{\max} \text{ 时} \tag{7}$$

$$\tau_{ij}(t+1)=\rho^{1-\varphi(n)}\cdot\tau_{ij}(t)+\Delta\tau_{ij},\ \text{当 } \tau<\tau_{\min} \text{ 时} \tag{8}$$

其中 $\varphi(n)$是一个与收敛次数 n 成正比的函数，收敛次数 n 越多，$\varphi(n)$的取值越大：

$$\varphi(n)=n/c \tag{9}$$

其中 c 为常数。这样，根据解的分布情况自适应地

进行信息素量的更新，从而动态地调整各路径上的信息素量强度，使蚂蚁既不过分集中也不过分分散，从而避免了早熟和局部收敛，提高全局搜索能力[5]。

2.3 算法流程

提出的基于蚁群算法的爬虫搜索策略执行过程如下：

(1)给定初始超链接节点，得到它属下的其他链接，构成蚂蚁爬行 Web 节点集。

(2)将 m 个蚂蚁均置于 n 个不同节点，初始化 η_{ij}、α、β 和 c 等控制参数。

(3)对每条 $e\in E(i,j)$ 路径设置初始浓度：

For k:=1 to m do

$\tau(i,j)=c$；$\Delta\tau_{ij}=0$； //c 为一常量

(4)置 $s=1$(s 为 $tabu$ 表索引)

For k:=1 to m do

将第 k 个蚂蚁的起始 Web 节点 $\nu(0)$ 置入 $tabu$。

(5)重复直到 $tabu$ 表为空，重复 $n-1$ 次。

置 s:=$s+1$

For k:=1 to m do

根据 $p_{ij}^{k}(t)$ 选择下一个探索节点 $\nu(j)$；在时刻 t，第 k 个蚂蚁在节点 $\nu(i)=tabu_k(s-1)$。

将第 k 个蚂蚁移动节点 $\nu(j)$ 插入 $tabu$。

(6)For k:=1 to m do

计算第 k 个蚂蚁所行进路径长度 L_k，并求出最优长度的平均值 $Avel$。

(7)If 平均值 $Avel$，与前一次 $Avel$ 计算值相同，则：

$n=n+1$；

else $n=1$ $\varphi(n)=n/c$

(8)对每条边上的信息素量值，按式(7)、式(8)进行更新。

(9)一个周期结束，输出最佳路径。

2.4 算法参数分析

在蚁群算法的实现过程中，多个参数需要初始化设定。由蚁群算法的原理可知，不同参数的选择能够对蚁群算法的性能产生至关重要的影响[5]。目前对蚁群算法中参数的确定还没有严格的理论基础，所以以上诸式中出现的参数 η_{ij}、α、β 和 ρ 通常用试验方法来确定其最优组合。η_{ij} 表示由城市 i 转移到城市 j 的期望程度，可根据某种启发算法而定，例如可以取 $\eta_{ij}=1/d_{ij}$。α 表示蚂蚁在行进过程中所积累的信息素对它选择路径所起的作用程度。β 是一个表示信息素重要程度的参数。信息激素的保留系数为 $\rho(0<\rho<1)$，它体现了信息素强度的持久性，而 $1-\rho$ 则表示信息素的消逝程度。

参考文献[6]通过大量的实验数值分析表明，当满足 $0.01\leqslant\alpha\leqslant0.3$、$3\leqslant\beta\leqslant6$、$0.1\leqslant\rho\leqslant0.3$ 时，算法总体上有较好性能，达到的最优解与全局最优较接近，同时，所需的迭代次数也较少，不易陷入局部最优而导致算法停滞。

3 实验

3.1 实验说明

为了验证基于蚁群算法的主题爬虫搜索策略比传统的广度优先算法和基于最佳优先搜索策略具有更好的全局搜索能力和自适应性，本文在 Nutch 爬虫的基础上构建了一个主题爬虫。Nutch 爬虫具有可扩展和定制性。通过定义一个 ACOCrawler 插件来抓取特定主题的网页[8]。实现以“物理教学资源”为主题，选取了国内三个教育网站为种子集(如表 1 所示)，算法参变量设定如表 2 所示。

表 1 网站种子集

网站名称	网址
中国教学资源网	http://www.cnier.com/
学资源网	http://www.phynet.cn/
中小学资源网	http://www.i3721.com/

表 2 算法参变量设定值

参变量	初始值
α	0.2
β	5
ρ	0.3

3.2 结果分析

系统运行 12 个小时，共抓取 3 360 000 个网页及资源。为了便于比较，分别对基于广度优先算法和最佳优先搜索算法的搜索结果进行测试，统计三种搜索算法实现的爬虫所搜索的关于物理教学资源的网页及资源数，采用“相对回报率”来评价爬虫的性能。相对回报率 R 的计算公式为：

$$R=\frac{t\text{ 时刻已发现相关主题页面比例}}{t\text{ 时刻已搜索页面比例}}$$

通过计算，可以得到三种算法性能比较图，如图 1 所示。

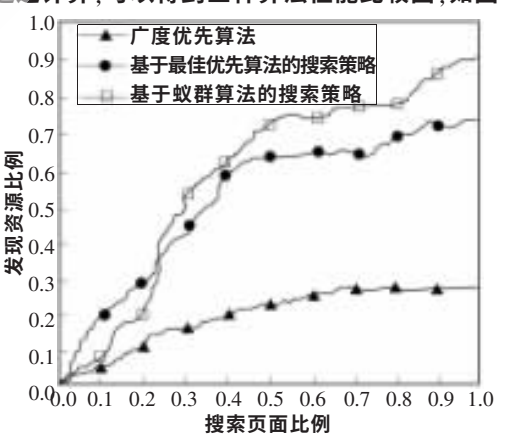

图 1 三种搜索算法性能比较图

由图 1 可以看出，在三种搜索策略中，广度优先算法的性能低于其他两种“启发式”算法。这两种搜索策略在访问了 50%的页面后，已经找到了 70%以上的相关物理资源，这表明基于“启发式”搜索策略具有优越性。

基于蚁群算法的搜索策略性能比较显著，除了在搜索初期其发现能力略低于基于最佳优先策略的搜索算法外，在其后的搜索中，新算法的性能明显高于基于最佳优先策略的搜索算法。其原因在于，基于蚁群算法的搜索策略采用了一种最优选择机制，一旦蚁群发现有好的全局最优个体，动态地更新路径上的信息素，作为最优选择路径，从而避免了局部最优，因而整体回报率较高。

本文针对现有主题爬虫所采用搜索策略出现的一些问题，将蚁群搜索模型引入主题爬虫搜索策略。实验结果表明，基于蚁群算法的搜索策略与基于广度优先搜索策略和基于最佳优先搜索策略相比，其在主题相关性上有比较明显的优势。通过对蚁群算法进行改进，能够动态地调整信息素，从而也能够较好地解决局部最优问题，提高了全局搜索的能力。但由于蚁群算法本身的一些缺陷，使得主题爬虫在搜索效率上还有待提高，这是下一步要做的工作。

参考文献

[1] 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(10)：26-29.

[2] 李学勇，田立军，谭义红，等.一种基于非贪婪策略的网络蜘蛛搜索算法[J].计算机与自动化，2004，23(2).

[3] DORIGO M，MANIEZZO V，COLORNI A. The ant system：optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems，Man and Cybernetics—Part B，1996，26(1)：29-41.

[4] 李开荣，陈宏建，陈崚.一种动态自适应蚁群算法[J].计算机与自动化，2004，40(29)：149-152.

[5] 陶剑文.基于蚁群计算的自适应Web检索算法设计[J].计算机工程与应用，2007(15)：163-165.

[6] 蒋玲艳，张军，钟树鸿.蚁群算法的参数分析[J].计算机工程与应用，2006(13)：31-35.

[7] MENCZER F，PANT G，SRINIVASAN N P. Topical Web crawler：evaluating adaptive algorithms[J]. ACM Transactions on Internet Technology，2004(4)：378-419.

[8] 荣光，张化祥.一种DeepWeb爬虫的设计与实现[J].计算机与现代化，2009(3)：32-34.

(收稿日期：2010-09-06)

作者简介：

陈永彬，男，1984年生，硕士研究生，主要研究方向：信息检索，数据挖掘。

张琢，女，1968年生，教授，主要研究方向：信息检索，数据挖掘。