基于蚁群计算的自适应 Web 检索算法设计

陶剑文
TAO Jian-wen

浙江工商职业技术学院 计算机应用研究所,浙江 宁波 315012
Computer Application Institute,Zhejiang Business Technology Institute,Ningbo,Zhejiang 315012,China

TAO Jian-wen.Algorithm design for adaptive Web retrieval system based on ant system computation.Computer Engineering and Applications,2007,43(15):163-165.

Abstract: This paper presents an adaptive and scalable Web search system,based on a multi-agent reactive architecture,which draws inspiration from biological researches on the ant foraging behavior.Its target is to search autonomous information on particular topics,in huge hyper textual collections,such as the Web,exploiting the outstanding properties of the agent architectures.The algorithm has been proven to be robust against environmental alterations and adaptive to user's information need changes,discovering valuable evaluation results from standard Web collections.
Key words: intelligent agent;information retrieval;similarity measure;multi-agent system

摘 要:受蚁群觅食行为仿生研究和蚁群系统模型理论所启发,提出了一种基于蚁群计算模型的分布、协作多主体(multi-agent)反应架构的自适应、可伸缩的 Web 搜索系统模型(MASAIR),其由大量智能主体组成,利用智能主体架构的优异特性,旨在从巨型超文档集合(Web)中自治地搜索特定主题的信息,从而为用户提供迅捷的信息检索服务。详细描述了 MASAIR 的计算模型及其算法,通过对标准 Web 文档集的检索仿真实验结果显示:该架构具有对环境改变的鲁棒性和对用户信息需求变更的自适应性。
关键词:智能主体;信息检索;相似性测量;多主体系统
文章编号:1002-8331(2007)15-0163-03 **文献标识码**:A **中图分类号**:TP393.09

1 引言

由于 Web 所具有的某些特性:巨型、复杂、高动态,在 Web 上有效地进行感兴趣信息的检索是一个很难达到的目标[2]。另外,由于 Web 站点镜像和别名的出现,导致某些 Web 资源重复多次呈现。更为糟糕的是,Web 还包含大量未知的"黑洞"(dark matter)影响(如 Web 页面访问限制、链接结构缺乏互连性、Web 页面自动生成器的存在等)。上述因素均为传统的信息检索(Information Retrieval,IR)技术带来严重的负面影响,致使传统的 IR 系统存在以下几个突出问题:

(1)在查询结果中缺少自适应性;

(2)查询库不具备高 Web 空间覆盖率;

(3)由于网站镜像和别名的影响,IR 搜索资源可能重复出现多次。

随着人工智能(AI)技术的发展,新的理论、架构与智能主体(agent)应用为解决上述问题提供了新的思路。由于 agent 的自治性、反应性、前摄性和社会性[3],agent 技术在当前信息系统架构中得以广泛应用。蚁群算法是由意大利学者 M.Dorigo 等人首先提出的一种新型的模拟进化算法,初步的研究已经表明该算法具有许多优良的性质[1]。本文提出一种基于蚁群计算模型的分布、协作多主体(multi-agent)反应架构的自适应、可伸缩的 Web 搜索系统模型(MASAIR)。本系统由多个相互协作的 agent 组成,这些 agent 将自治地适应用户与执行环境的变化,从而使本系统免受整体崩溃的威胁。通过蚁群计算算法的优化,本系统具有以下几个方面的优点:

(1)由于本架构固有的自治与反应特性,系统对环境改变具有一定的鲁棒性,对用户信息需求的改变也具有一定的自适应性;

(2)由于蚁元 agent 的反应性与协作性,MASAIR 将以最短的时间内定位到所需的 Web 资源,减少了 Web 空间的探索时间;

(3)由于蚁元 agent 的自适应性,提高了 MASAIR 信息检索的质量。

2 蚁群计算模型

虽然单个蚂蚁的行为极为简单,但由单个简单的个体组成的群体却表现出神奇的行为。蚁群协作觅食是一种社会性的昆虫行为模型,该模型由生物学家和行为学家创建,其旨在说明盲目的动物如何从巢穴到食物源(或反之)发现最短的行程路线[1]。对于这个现象的解释为:由于蚁群能够在其所移动的路线上释放外激素,一路上其留下外激素轨迹(trail),其强度与食物的品质和数量成正比,其他蚁群遇到该轨迹便会循迹移动(但也会有一定的走失率,走失率与轨迹的强度成反比),从而使得该轨迹被后来的蚁群所释放的激素所加强。因此,蚁群的集体行为便表现出一种信息正反馈现象:某路径上走过的蚂蚁

基金项目:浙江省教育厅科研项目(the Research Project of Department of Education of Zhejiang Province of China under Grant No.20040120)。
作者简介:陶剑文(1973-),男,教研室主任,讲师,主要研究方向:分布式智能计算、数据挖掘等。

越多,则后来者选择该路径的概率越大,蚂蚁个体之间就是通过这种信息的交流达到搜索食物的目的。随着多个周期正反馈的蚁群移动,所有盲目的蚁群将会收敛到某个最短路径[6]。

3 算法介绍

3.1 MASAIR 算法设计

MASAIR 是基于蚁群计算模型的多 agent 架构,其由一群带有反应能力的智能主体(agent)构成,这些主体生活在一个超文本的环境中,在该环境中进行资源与用户信息需求的相似性测量是可能的[5],每个主体与一个虚拟蚁元(ant)相对应。给定超文本资源 *urli* 与 *urlj*,假设在 *urli* 中有一个链接指向 *urlj*,则处于 *urli* 的蚁元自身将有机会从 *urli* 移动到 *urlj*。一个链接序列(或称 *url* 对)代表了一个可能的主体移动路线,在每次页面探索结束,主体在路线上释放外激素(pheromone)轨迹。系统的执行被分为多个循环周期,在每一个周期中,一个蚁元在超文本资源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,在周期的结尾,蚁群更新探索路线的外激素的强度值。外激素轨迹的意义在于:允许蚁群利用本地有限的知识对环境和蚁群行为作出较好的本地决策,蚁群利用这些轨迹来彼此交流探索结果。一个蚁元能够发现某资源越有兴趣,其释放在探索路线上的外激素数就越多。只要某路线承载了相关资源,路线上的外激素轨迹将被加强,同时,受吸引的蚁元数将会增多,从而形成一个正反馈环路。

本文所提出的 MASAIR 系统算法主要包括两个部分:蚁群探索算法;路径轨迹更新算法。为了表述方便,本文拟作如下定义:

定义 1 链接一致性。是指含有某个主题信息的页面 *urli*,其所链接的页面 *urlj* 将包含与 *urli* 相同(或相似)的主题信息[4]。

定义 2 访问一致性(Visit Coherence)。指的是这样一种现象,即如果某用户正在访问某个其感兴趣的主题页面,其也将有可能访问该页面所链接的页面。

定义 3 Web 信息页面(链接)有向图 $G=(V,E)$,V 代表超文本页面集合,E 代表由链接构成的路径集合。

定义 4 基于定义 1 与定义 2,G 中页面 Pk 的搜索权重值(PageRank)[2]计算公式为:

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

d 指调节因子($0.8<d<0.9$),Ti 指链接到 Pk 的页面集合,ci 指 Ti 的链接扇出度,即从 Ti 指向其它页面的链接数。

定义 5 页面 *urli* 与 *urlj* 间的距离 *dij*,由下述公式计算得到:

$$dij=\frac{C}{\sum_{i=1}^{k}R(Pi)} \tag{2}$$

C 为一协调因子常量,$Pi(i=1,2,\cdots,n)$指由页面 i 到页面 j 所经历的所有页面集合。

定义 6 用户查询 Q 与文档 D_d 的相似性测量函数为:

$$M(Q,D_d)=\sum_{t\in Q}\frac{tf_{q,t}*idf_t}{\sqrt{\sum_{t\in Q}(tf_{q,t}*idf_t)^2}}*\frac{tf_{d,t}*idf_t}{\sqrt{L_d}} \tag{3}$$

其中 L_d 是文档 d 的长度,该值可通过计算索引词条的数目获得,$f_{d,t}$ 表示术语 t 在文档 d 中出现的次数,通常称为文档内(within-document)频率,$tf_{d,t}$ 指术语 t 在文档 d 中的频率,术语 t 反转文档频率 idf_t 用于反向调节 t 的等级(rank),$tf_{d,t}$ 与 idf_t 的计算公式分别为:

$$tf_{d,t}=\sqrt{f_{d,t}} \tag{4}$$

$$idf_t=1+\log_e\frac{N}{f_t+1} \tag{5}$$

式(5)中 N 和 f_t 分别表示文档集中文档数和包含术语 t 的文档频率。

定义 7 $bi(t)$指在时间 t 位于页面 *urli* 的蚁元数,$m=\sum_{i=1}^{n}bi(t)$指所有蚁元数。

定义 8 链接转移概率。给定页面 p,$l(i,j)\in E$ 指示 p 中存在一条从 *urli* 到 *urlj* 的链接,则从链接 *urli* 到 *urlj* 的转移概率值为 $p_{ij}(t)$。设 $\tau_{ij}(t)$为对应于 *urli* 和 *urlj* 间外激素轨迹,则在时刻 t,$p_{ij}(t)$的计算公式为:

$$p_{ij}(t)=\frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{l:(i,j)\in E}[\tau_l(t)]^{\alpha}[\eta_l]^{\beta}} \tag{6}$$

式(6)中,η_{ij} 指 $E(i,j)$路径的显见度(Visibility),其可相对于页面 i 到页面 j 的距离 *dij* 计算得到,距离越短,显见度值越大,蚁元趋向概率便大。α 与 β 分别指蚁元在移动过程中所积累的信息及启发式因子在蚁元路径选择中所起的不同作用,其可以通过实验逐步协调确定。

3.2 蚁群探索算法

在系统初始时刻,设定 $\tau_{ij}(0)=C$(C 为常数),各条路径上信息量相等,蚁元 $k(k=1,2,\cdots,m)$在移动过程中,根据路径 $E(i,j)$上的信息量决定转移方向。在时刻 t,某个定位在资源 p 的蚁元抽取 p 中包含的每个连接 *urli* 与 *urlj* 的链接的转移概率值 $p_{ij}(t)$,蚁元根据该值来决定下步探索路径。为了避免蚁群出现环路爬行并约束蚁元进行页面递增探索,每个蚁元存储一个 tabu 列表以存储被访问的 *urls*,如果 *urlj* 属于 tabu,则从 *urli* 到 *urlj* 的路径相关概率值为 0,从而禁止蚁元 k 探索 *urlj*。在每个周期的结束,tabu 列表将被清空。

MASAIR 系统蚁群探索算法描述为:

1.初始化
将 m 个蚁元放置在 n 个不同节点
Set $\alpha:=1$ {α 值可通过实验调节}
Set $\beta:=5$ {β 值可通过实验调节}
2. Set $s:=1$ {s 为 tabu 列表索引}
For $k:=1$ to m do
将第 k 个蚁元的起始 Web 节点 $V(0)$置入 $tabu_k(s)$
3.Repeat until *tabu*=full {本过程重复($n-1$)次}
Set $s:=s+1$
For $k:=1$ to m do
根据 $p_{ij}(t)$选择下一个探索节点 $V(j)$
{在时刻 t,第 k 个蚁元在节点 $V(i)=tabu_k(s-1)$}
第 k 个蚁元移动到节点 $V(j)$
Insert $V(j)$ into $tabu_k(s)$
4. For $k:=1$ to m do
将第 k 个蚁元从 $tabu_k(n)$移动到 $tabuk(1)$
计算第 k 个蚁元探索的路径长度 L_k
更新第 k 个蚁元发现的最短路径

3.3 路径轨迹更新算法

当达到每周期的移动限制数时,蚁群便启动轨迹更新进

程。第 k 个蚁元的外激素变化 $\Delta\tau^k$ 与被访问的资源评分均值相对应:

$$\Delta\tau^k=\frac{\sum_{j=1}^{|P^{(k)}|}score(Q,P^{(k)}[j])}{|P^{(k)}|} \quad (7)$$

式(7)中 $P^{(k)}$指被第 k 个蚁元访问的排序页面集;$P^{(k)}[j]$是 $P^{(k)}$的第 j 个元素;$score(Q,p)$是一计算函数:对于每个页面 p,根据当前的信息需求 Q 计算出相似性度量值 $a\in[0,1]$,a=1 代表具有最高相似性。$score(Q,p)$可通过 $M(Q,D_d)$相对应计算得到。

上述更新过程随着 t 值的更新而完成。在 $t+n$ 时刻,从 *urli* 到 *urlj* 的 $E(i,j)$路径轨迹受蚁元外激素更新进程的影响,其通过 $\Delta\tau_{ij}^k$的值计算得到:

$$\tau_{ij}(t+n)=\rho\cdot\tau_{ij}(t)+\sum_{k=1}^{m}\Delta\tau_{ij}^k \quad (8)$$

$$\Delta\tau_{ij}^k=\frac{Q}{L_k} \quad (9)$$

式(8)中 ρ 指轨迹蒸发(evaporation)系数,其必须设置为一个小于 1($0<\rho<1$)的正数以免由于重复的正反馈产生的无限制信息聚集,在执行的开始,所有 $\tau_{ij}(0)$值被设定为一个较小的恒定值 $\tau0$($|\tau0|<\varepsilon$)。式(9)中 Q 为一常数,L_k 表示蚁元 k 在一次循环中所探索的路径长度,初始时刻,$\Delta\tau_{ij}$=0。

在周期的结尾,已经发现最有兴趣路线的 $N(N\leqslant m)$个蚁元的执行外激素更新进程,而其它($m-N$)的探测结果被遗弃。在每个周期, 外激素轨迹根据所访问的资源评分均值被更新,该技术确保了系统对环境改变的鲁棒性和对用户信息需求变更的适应性。例如,如果 $\Delta\tau_{ij}^k$的一个负值变量发生在周期 t(如由于查询精化或页面改变),$\tau_{ij}(t+1)$接收一个负的反馈,因此,几乎没有蚁群会被吸引,$\Delta\tau_{ij}^k$在 $t+1$ 周期的增量依旧进一步减少等等。换句话说,环境中的每一个变化都会促使一个修改全局系统行为的反馈信息,以确保系统自身对新环境的适应。

MASAIR 系统路径轨迹更新算法描述为:

1.初始化
Set t:=0 {t 为时间计数器}
Set NC:=0 {NC 为周期计数器}
Set ρ:=ε {$0<\varepsilon<1$}
For all $e\in E(i,j)$
Set $\tau_{ij}(0)$:=C {轨迹初始浓度}
Set $\Delta\tau_{ij}$:=0
2.For all $e\in E(i,j)$
For k:=1 to m do
$\Delta\tau_{ij}^k:=\frac{Q}{L_k}$
$\Delta\tau_{ij}:=\rho\cdot\Delta\tau_{ij}+\Delta\tau_{ij}^k$
3. For all $e\in E(i,j)$
依据式(8)计算 $\tau_{ij}(t+n)$
Set t:=$t+n$
Set NC:=NC+1
For all $e\in E(i,j)$
Set $\Delta\tau_{ij}$:=0
4. If ($NC<NC_{MAX}$) and (stop<>true) then
清空所有 tabu 列表
Goto step 2
else
打印最短路径
Stop

4 系统仿真实验

4.1 仿真环境介绍

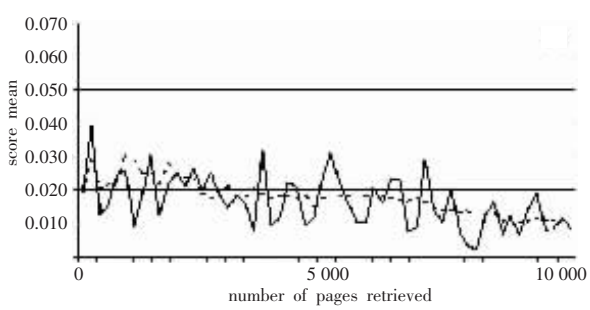
MASAIR 系统模型原型实现基于 JADE 平台,JADE 是一款由 JAVA 语言开发的跨平台的移动 agent 开发环境, 其提供了移动 agent 执行的宿主环境[5]。移动 agent 群间通信采用 KQML(Knowledge Query and Manipulation Language)语言实现。Web 资源的抽取、分析、索引与搜索核心功能基于 Lucene 系统[7]。抽取页面均取自 ODP(www.dmoz.org)。向量空间模型被用于对每一个文档进行一个相似性测量,该测量结果用以评估文档与查询(即用户信息需求)间的亲近程度。起始 Web 页集 S 由一些从符合用户查询的文档集中抽取的随机 *urls* 组成,在访问了 10 000 张页面后,系统停止爬行。系统仿真主要参变量设定如表 1 所示。

表 1 系统主要参变量设定情况

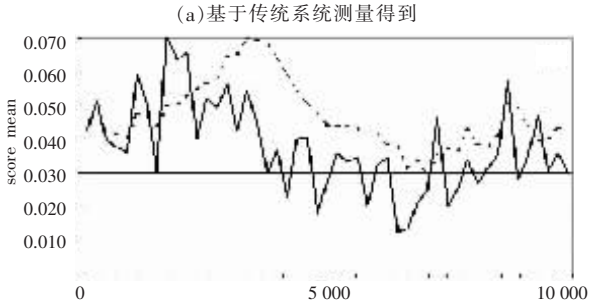
参变量	设定值
α	1
β	5
ρ	0.5
Q	100
m	30

4.2 实验结果分析

如文献[2]中所指出的,对智能爬行系统评价来说,相关页面的获取速率是最为重要的。为此, 每一次测试探索(exploration)中,最近分析的资源的评分均值(score mean)将被评测。通过传统爬行系统和本文推荐的搜索系统分别对最近检索的 100 张和 1 000 张页面的评分均值测量得到如图 1 所示的相关页面的获取速率图。从图 1 所显示的结果推断看,相较于传统



(a)基于传统系统测量得到



(b)本文系统测量获得

图 1 相关页面的检索速率

(下转 169 页)