DOI:10.11721/cqnuj20150223

基于查询代理和广义蚁群算法的 P2P 资源搜索*

王文豪[1], 陈晓兵[1], 蒋道霞[2]
(1.淮阴工学院 计算机工程学院, 江苏 淮安 223003; 2. 江苏财经职业技术学院 机械电子与信息工程学院, 江苏 淮安 223003)

摘要:随着 Intranet 迅速发展,私有网络中包含了大量的资源,如何将 P2P 网络延伸到私有网络中,并对其中的资源进行整合具有重要的研究价值。在分析现有 P2P 网络资源搜索方法的基础上,提出了一种基于查询代理和广义蚁群算法的资源搜索模型。即在私有网络内部使用查询代理完成资源的查找,在公有网络中使用广义蚁群算法进行资源搜索,并对广义蚁群优化算法进行改进,将节点可信度融入到算法优化条件中。实验表明:该算法能够提高搜索效率和命中率,能够有效地孤立网络中的虚假节点,对提高网络的服务质量具有重要意义。

关键词:P2P;资源搜索;查询代理;广义蚁群优化算法
中图分类号:TP393 **文献标志码**:A **文章编号**:1672-6693(2015)02-0117-06

P2P 网络是一个无中心的分布式网络,网络中的每个节点都处于完全平等的地位,既是服务提供者也是服务接受者,资源分布在各个对等节点上,每个节点可以随意地加入或者离开网络,这使得 P2P 网络具有很好的可扩展性、健壮性和负载均衡性[1],在文件共享方面得到了广泛的应用。

搜索是 P2P 网络必须提供的基本功能,P2P 的搜索是在一个分布式的网络中进行信息检索,随着 P2P 网络规模的不断扩大和 P2P 网络自身的开放、匿名和高度动态的特性[2],使得 P2P 搜索相对于传统的信息检索方法有很多不同之处,如何在这庞大的网络上快速地搜索到所需要的资源已成为人们日益关注的问题,本文在分析 P2P 现有资源搜索算法的基础上,提出了一种基于查询代理和广义蚁群算法的资源搜索方法,该方法不仅能高效地搜索到所需要的资源,而且还能对私有网络内部的资源进行搜索,实现对私有网络资源的整合。

1 传统的 P2P 资源搜索算法

资源搜索技术是 P2P 技术研究的重点,针对不同的 P2P 网络,研究者和学者们提出了很多 P2P 搜索的方法。

1) 集中式 P2P 网络搜索:这种搜索技术依赖于中央索引服务器,所有的节点都通过中央服务器来完成资源的定位通讯信息。查询效率高,可维护性好,但对中央索引服务器的处理能力和带宽的要求很高,随着网络规模的不断扩大,维护成本会变高。

2) 非结构化 P2P 网络搜索:这种搜索技术中最典型的方法是洪泛,通过将查询消息反复地转发给邻居节点进行资源搜索,直到找到资源或者洪泛次数达到预先确定的最大值为止。该方法能有效地避免集中式搜索存在的问题,但在网络中产生大量冗余的查询包,对网络带宽的消耗较大[3]。为此人们又提出了迭代洪泛方法,虽然减少了查询的节点数目,但也额外地增加了节点处理负担。

3) 结构化 P2P 网络搜索:这种搜索技术目前主要是基于分布式哈希表(DHT),根据资源的标识快速定位到相应节点。由于 DHT 维护开销较大,人们又提出了一些改进方法:文献[4]提出 SWOP(Small word overlay protocol)算法,通过聚类方法来提高查找性能;文献[5]提出基于协作文件的文件共享系统,通过使用 Chord 分布式查找协议来定位相应文件。但这类算法对模糊匹配支持程度较低。

4) 混合式 P2P 搜索:这种搜索技术综合了集中式快速查找和非结构化 P2P 网络搜索的优点,将网络中的节

* **收稿日期**:2013-10-20 **修回日期**:2014-05-21 **网络出版时间**:2015-01-22 11:56
资助项目:江苏省科技支撑计划(No. BE2012112);江苏省淮安市科技支撑计划(No. HAG2013068)
作者简介:王文豪,男,副教授,研究方向为智能计算、P2P 计算,E-mail: wangwenhao1407@163.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20150122.1156.014.html

点分为普通节点和超级节点两类，当普通节点需要查找文件时，直接从它连接的超级节点中查找，如果找到了相应文件，则直接根据所存储机器的IP地址建立连接，如果没有找到，超级节点则将这个查询请求转发给与它连接的其他超级节点。这种方法可扩展性较好，较易管理，但对超级节点依赖性大。

2 查询代理

2.1 资源搜索模型

本文的资源搜索模型如图1所示：私有网络被抽象为公有网络中的一个普通节点，对私有网络内部资源的查询由查询代理完成，模型中的服务器主要用于管理用户信息和文件的信任度信息以及辅助穿越NAT设备，并不指示用户物理地址。其工作原理如下。

当公有网络中的节点要搜索资源时，例如图1中用户A，会向服务器发送握手消息，服务器收到该消息后，向私有网络中的主超级节点发送资源查询打洞命令，该命令中包含用户A的IP地址和端口号，私有网络主超级节点收到该打洞命令后，先提取其中的IP地址和端口号，然后再向该地址发送一个数据包，这样，私有网络的NAT设备就被成功地穿越。用户A就可以将查询消息交给私有网络中的主超级节点代理查询。

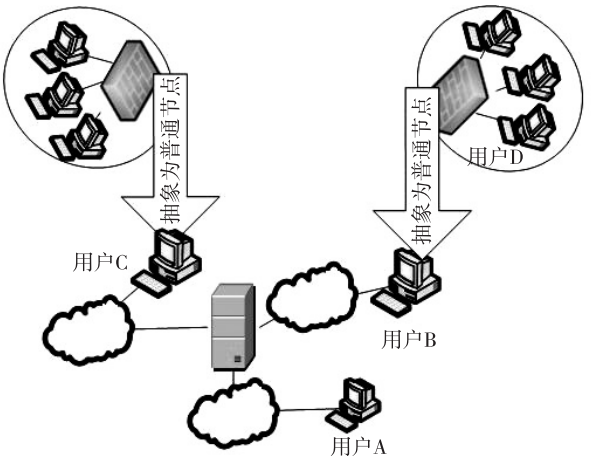


图1 资源的搜索模型

当私有网络中的节点需要搜索资源时，例如图1中用户D需要搜索资源，就相当于由该私有网络抽象而来的节点B需要搜索资源。对于这种搜索，可以采用转移查询消息的方法进行搜索。如果用户D所发起的查询消息能到达公有网络中节点A，并且节点A与用户D之间有较好的网络环境资源，那么就可以用节点A代替用户D发起资源搜索命令；如果用户D所发起的查询消息不能到达公有网络中任何节点，那么可以借助于公有网络中的服务器，利用STUN协议穿透NAT设备，进行私有网络之间的资源搜索。

2.2 查询代理设置

本搜索模型将私有网络抽象为公有网络中的一个普通节点，公有网络中的查询消息进入私有网络时，就不继续公有网络中的资源搜索方法，而是按照私有网络中的特点，将查询消息交给私有网络中的主超级节点代理查询。

本文的资源搜索分为两种情况：私有网络和公有网络，在公有网络中采用蚁群算法进行资源搜索，在私有网络中使用查询代理进行资源搜索。当有公有网络中的查询消息到达私有网络时，查询消息被转交给私有网络中的主超级节点代理查询，由主超级节点负责整个查询。在整个查询过程中，各局域网的超级节点被设置为子查询代理，负责查询本网络内的资源信息，私有网络的查询代理只是负责接受从公有网络中进入私有网络中的查询消息并搜集各个子查询代理所查询到的资源信息。

3 蚁群算法

3.1 蚁群算法原理及广义蚁群算法数学模型

蚁群算法最初是由意大利学者M. Dorigo等人在蚂蚁觅食行为启发下提出的一种群体智能优化算法[6-7]。研究发现，蚂蚁是在所经过的路上留下一种具有挥发性的物质——“信息素”来进行信息传递。随后的蚂蚁不仅可以检测出该物质的存在和强度，而且可根据信息素的强度来指导自己对前进方向的选择。这样，当某路径上走过的蚂蚁越多，留下的“信息素”就越多，随后的蚂蚁选择该路径的概率就越大。这就构成了蚂蚁群体行为表现出的一种信息正反馈现象。通过这种方式，蚂蚁很容易找到从蚁巢到食物源的最短路径。

广义蚁群优化算法[8-10](Generalized ant colony optimization algorithm, GACO)的数学模型为

$$p_{ij}^{k}(t_m)=\begin{cases}\dfrac{F(\tau_{ij}(t_m))}{\sum\limits_{s\subset allowed_k}F(\tau_{is}(t_m))}, & \text{若 } j\in allowed_k\\ 0, & \text{其他}\end{cases},\tag{1}$$

其中 $p_{ij}^{k}(t_m)$ 表示第 k 只蚂蚁在节点 i 处选择节点 j 的概率；$F(\tau)$ 是关于变量 τ 的严格单调递增函数，即 $\dfrac{\mathrm{d}F}{\mathrm{d}\tau}>0$ 且 $F(\tau)>0$；$allowed_k$ 表示蚂蚁 k 下一步可以选择的城市。

其信息素的更新可以按照如下的策略进行：

$$\forall(i,j):\tau_{ij}(t_m)\leftarrow\delta(t_m)\cdot\tau_{ij}(t_{m-1}),\tag{2}$$

$$\text{if } f(s_{t_m})<f(\bar{s}),\ \text{then } \bar{s}\leftarrow s_{t_m},\tag{3}$$

$$\forall(i,j)\in\bar{s}:\tau_{ij}(t_m)\leftarrow\tau_{ij}(t_m)+\mu(t_m),\tag{4}$$

上式中，$\delta(t_m)$ 和 $\mu(t_m)$ 是关于 t 的连续函数，$\tau_{ij}(t_m)$ 是 t_m 时刻从节点 i 到节点 j 路径上的信息素浓度，$\bar{s}$ 表示到目前为止所求解问题的最优解，s_{t_m} 表示第 t_m 次迭代的最优解，$f(s_{t_m})$ 和 $f(\bar{s})$ 分别表示 s_{t_m} 和 $\bar{s}$ 所对应的路径的代价函数的值。$\delta(t_m)$ 为信息素挥发率函数，它是关于 t_m 的严格单调递减函数，即 $\dfrac{\mathrm{d}\delta(t)}{\mathrm{d}t}<0$。$\delta(t_m)$ 的通解形式为

$$\delta(t)=\frac{\gamma_0\mathrm{e}^{-\lambda t}}{1+\gamma_0\mathrm{e}^{-\lambda t}},\tag{5}$$

其中 λ 为正常数，设 $\delta(t)$ 的初始值为 δ_0 且在 0 和 1 之间，则有 $\gamma_0=\dfrac{\delta_0}{1-\delta_0}$，很显然 $\delta_t\in(0,1)$。

$\mu(t_m)$ 为信息素的全局函数。它是关于 t_m 的具有上下界严格单调增函数，即 $0<A<\mu(t)<B<+\infty$ 且 $\dfrac{\mathrm{d}\mu(t)}{\mathrm{d}t}>0$。$\mu(t_m)$ 的通解形式为

$$\mu(t)=\frac{A+B\eta_0\mathrm{e}^{\kappa(B-A)t}}{1+\eta_0\mathrm{e}^{\kappa(B-A)t}},\tag{6}$$

其中 κ 为常数，设 $\mu(t)$ 的初始值为 $\mu_0(\mu_0\in(A,B))$，则有 $\eta_0=\dfrac{\mu_0-A}{B-\mu_0}$。

3.2 广义蚁群算法在 P2P 资源搜索中的应用

P2P 网络模型可以用一个无向图 $G(V,E)$ 表示，其中 V 为网络中所有节点的集合，E 为 P2P 网络中的链路集合。寻找最佳资源的路径问题可以转换为在无向图 G 中找到一条满足约束条件要求，并且费用最小的最佳路径[9]。这恰好适合蚁群算法求解问题的特征，因此可以运用蚁群算法来解决。

文献[9]对广义蚁群算法在 P2P 网络中的应用做了详细介绍。其最优化条件为

$$\min\ y(e)=u_1cost(e)+u_2bandwidth(e)+u_3(delay(e)+delay(v))$$

$$\text{s.t.}\begin{cases}\sum\limits_{e\in p}cost(e)\leqslant C\\ \forall e\in E,\ bandwidth(e)\geqslant B\\ \sum\limits_{e\in p}delay(e)+\sum\limits_{v\in p}delay(v)\leqslant D\\ cpu\_step(v_{dest})\geqslant S,\quad \text{其中 } v_{dest}\text{ 是目标节点}\end{cases},\tag{7}$$

其中 u_1、u_2 和 u_3 分别表示各因素所占的权重；B、C、D、S 分别代表要求满足的带宽、费用、时延、CPU 剩余利用率；$cost(e)$ 为费用函数；$bandwidth(e)$ 是链路带宽函数，与链路带宽成反比；$delay(e)$ 和 $delay(v)$ 分别代表链路和节点的时延函数。

在 P2P 网络中，上述最优化条件是存在问题的。应用上述优化条件虽然可以找到最优解，但并不能说明它就是综合最优的。例如，如果用户运用该方法搜索到一个资源，下载后却发现这个资源与文件名不符合，那么这时候就不能说这个资源是最优的。这必然会影响用户对系统的满意度，同时也会导致系统资源的白白浪费以及不可信资源在整个网络中的传播。

针对这个问题，本文对文献[9]中的优化条件进行改进，将节点（用户）的信任度融入到最优化条件中去，并

称这种算法为TGACO(Trust-GACO)。于是公式(7)可被改写为公式(8)。

$$\min\ y(e)=u_1 cost(e)+u_2 bandwidth(e)+u_3(delay(e)+delay(v))+u_4\frac{1}{T_U}$$

$$\text{s.t.}\begin{cases}T_U\geqslant T\\ \sum\limits_{e\in p}cost(e)\leqslant C\\ \forall e\in E,\ bandwidth(e)\geqslant B\\ \sum\limits_{e\in p}delay(e)+\sum\limits_{v\in p}delay(v)\leqslant D\\ cpu\_step(v_{dest})\geqslant S,\quad \text{其中 } v_{dest}\text{ 是目标节点}\end{cases},\tag{8}$$

其中 u_4 为节点的信任度的权重因子,T_U 为节点(用户)的信任度。用户的信任度与其发布资源质量密切相关,如果用户发布虚假的、恶意的或病毒等信息,则其信任值必然较低,那么蚂蚁在搜索资源时应尽量回避这类用户的资源;相反,如果用户提供的资源都是货真价实的高质量资源,则其信任值必然很高,那么蚂蚁在搜索时应该在这类用户发布的资源中进行搜索,这样有利于提高搜索的成功率。(8)式体现了这种思想,用户的信任度越低,即 T_U 值越小,在相同条件下,蚂蚁搜索该资源时综合费用 $y(e)$ 值就越大,因而该资源也就不是最优资源;同样,用户的信任度越高,$y(e)$ 值就越小,即蚂蚁搜到该资源时所花费的费用越少,该资源越优秀。

设蚂蚁的数量为 m,则综合最优资源是 m 只蚂蚁在搜索中所花费用最小的那个目标节点的资源,用公式表示为

$$y_{best}=\min\{\min y_i(e)\mid 1\leqslant i\leqslant m\}。\tag{9}$$

信息素的更新按照公式(2)~(4)进行,因此TGACD算法查找最佳资源节点的具体步骤如下:

步骤1:初始化各个节点的CPU剩余利用率和时延大小,初始化各条边上的费用和带宽大小与信息素 $\tau_{ij}(0)=\tau_0$,初始化各节点用户的信任度。并将 m 只蚂蚁放在源节点 S 处,每只蚂蚁生命周期为 C,即蚂蚁最多访问 C 个节点。

步骤2:删除不满足约束 $bandwidth(e)\geqslant B$ 的边。

步骤3:删除用户信任度 $T_U<T$ 的节点,以及与该节点相邻的边。

步骤4:设置每只蚂蚁的禁忌列表 $tabu_k$,源节点 S 放入 $tabu_k$ 列表中,for $k=1$ to m $tabu_k[k]=S$。

步骤5:重复以下步骤直到蚂蚁找到资源节点或者其生命周期结束。

1) 蚂蚁k根据公式(1)选择路径,并分别计算链路延时 $\sum\limits_{e\in p}delay(e)+\sum\limits_{v\in p}delay(v)$ 和节点代价 $\sum\limits_{e\in p}cost(e)\leqslant C$。

2) 如果 $\sum\limits_{e\in p}delay(e)+\sum\limits_{v\in p}delay(v)\leqslant D$ 和 $\sum\limits_{e\in p}cost(e)\leqslant C$,则将该节点 S' 放入 $tabu$ 表中,并将蚂蚁移动到该处;否则转1)。

3) 查询节点 S' 处的资源列表,如果找到所需要的资源,则计算该节点CPU的剩余利用率 $cpu\_step(v_{dest})$。

4)如果 $cpu\_step(v_{dest})\geqslant s$,则停止搜索,否则继续搜索后继节点。

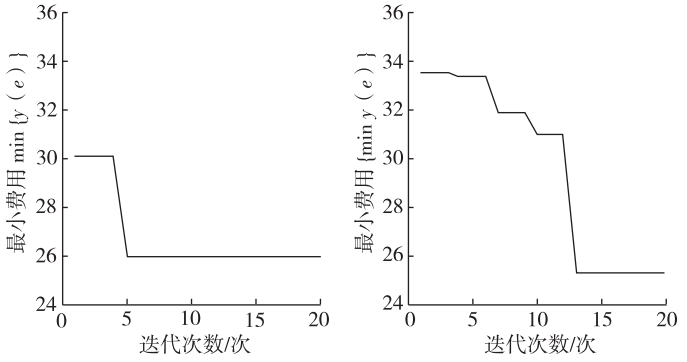
步骤6:将蚂蚁放回源节点处 S,依据公式(2) ~(4)更新信息素浓度;计算综合费用 $y(p)$,如果 $y(p)$ 小于当前最优解则更新最优解集 P_{best},并将其他的解放入备用解集 P_{bak}。

步骤7:$i=i+1$,若 $i<T$(迭代次数),则转到步骤4,否则停止算法,输出最优解和备用解。

4 仿真实验与分析

实验1 确定广义蚁群算法的参数。

仿真实验的方法:广义蚁群算法中需要确定的参数有公式(5)中的 δ_0、λ,公式(6)中的 μ_0、A、B、κ。为了确定这些参数,用Matlab编写蚁群算法程序预先假定这些参数的值,反复运算观察实验结果,所使用的网络的拓扑结构图为[2, 5; 3, 3; 3, 7; 4, 5; 5, 4; 6, 2; 6, 6; 7, 5; 8, 3; 8, 7; 9, 6],最大迭代次数为20,人工蚂蚁的数目为11,表1是其中的两组实验参数,图2中的(a)和(b)分别是其对应的仿真结果。



(a)　　(b)

图 2　迭代次数与收敛速度

表 1　广义蚁群算法两组参数

参数	组 1	组 2
A	10	10
B	20	20
κ	1.5	1.5
λ	2	2
δ_0	0.8	0.4
μ_0	11	2

从上述的实验结果可以看出，图 2(a)所对应的参数使得广义蚁群算法收敛速度较快且稳定，而图 2(b)所对应的参数使得广义蚁群算法收敛速度要慢些。因此在下面的 P2PSim 实验中的，选用组 1 参数来进行广义蚁群算法实验。

实验 2　本文 TGACO 算法和文献[9]中的 GACO 算法比较。

实验采用 P2Psim 仿真器进行模拟，参数设置：网络节点数目为 10 000，其中，5%的节点含有所需的资源，2%的节点用户的信任度不符合 TGACO 要求，迭代次数 $T=5$，蚂蚁的生命周期 $C=10$，TGACO 算法参数使用实验 1 中确定的参数和 $u_1=0.4$、$u_2=0.3$、$u_3=0.1$、$u_4=0.2$。GACO 的参数的选取参照了文献[9]中对参数选择的结论：选择 $\alpha=1$、$\beta=5$、$\rho=0.8$，$u_1=0.5$、$u_2=0.3$、$u_3=0.2$。重复实验 50 次，统计其平均综合费用如表 2 所示。图 3 是搜索次数与所得到的虚假文件数目的比较结果。

表 2　TGACO 和 GACO 平均综合费用比较

网络节点数/个	GACO 算法平均综合费用 $y_{GACO}(e)$	TGACO 算法平均综合费用 $y_{TGACO}(e)$
100	458.2	462.7
500	420.6	427.3
1 000	379.5	390.1
5 000	321.9	340.7
10 000	312.7	321.5
20 000	307.9	314.7

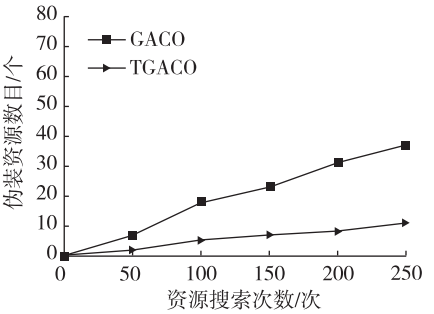


图 3　TGACO 和 GACO 搜索次数与所得到的假文件数目比较

从表 2 可以看出，在最优化条件中加入了用户信任度的广义蚁群算法(TGACO)的平均综合费用比一般广义蚁群算法(GACO)的平均综合费用有所上升，但从图 3 可以看出 TGACO 算法可以明显地减少搜索到的虚假文件的数量，从另一角度提高了算法的搜索效率。因此使用该方法能有效的隔离 P2P 网络中的提供虚假信息的节点，从而提高网络服务质量。

5 结语

随着私有网络的规模不断扩大，其内部包含的资源越来越多，如何将 P2P 技术延伸到私有网络中，整合其资源具有重要的意义。本文针对此问题提出了一种基于查询代理和广义蚁群算法的资源搜索模型，详细地阐述其相关原理。实验仿真证明了该算法可以得到局部的最优资源。

参考文献：

[1] 黎梨苗，陈志刚，桂劲松，等. 基于优先权的 P2P 网络信任模型[J]. 计算机工程，2013，39(5)：148-152.
Li L M，Chen Z G，Gui J S，et al. P2P network trust model based on priority[J]. Computer Engineering，2013，39(5)：148-152.

[2] 陈国强，苏静. P2P 文件共享系统中的一种双向并发信任机制[J]. 微电子学与计算机，2011，28(2)：120-123.
Chen G Q，Su J. A Bidirectional parallel trust mechanism

for P2P file—sharing systems[J]. Microelectronics & Computer,2011,28(2):120-123.

[3] 殷嘉乐.基于预算机制的非结构化网络分段搜索策略[J].电子设计工程,2013,21(24):124-130.
Yin J L. Divisional searching strategy based on budget in unstructured P2P network[J]. Electronic Design Engineering,2013,21(24):124-130.

[4] Ken Y,Hui K,John C,et al. Small world overlay P2P networks[J]. IEEE Quality of Service,2004(7):201-210.

[5] 盛明超.P2P网络的延伸—对私有网络中资源的整合[D].南京:南京邮电大学,2009.
Sheng M C. An extension of P2P network—integration of resources in private network[D]. Nanjing:Nanjing University of Posts and Telecommunications,2009.

[6] 杨清平,邓小清,蒲国林,等.网格资源的组织与发现研究[J].重庆师范大学学报:自然科学版,2008,25(4):70-73.
Yang Q P,Deng X Q,Pu G L,et al. Research into organization and discovery of grid resource[J]. Journal of Chongqing Normal University:Natural Science,2008,25(4):70-73.

[7] 黄毅然,钟诚,李智,等.融合蚁群算法和路由侦听的移动P2P搜索[J].小型微型计算机系统,2011,32(8):1515-1520.
Huang Y R, Zhong C, Li Z, et al. Mobile peer-to-peer searching using Ant Algorithm and routing detection[J]. Journal of Chinese Computer System,2011,32(8):1515-1520.

[8] 张代远.一类新型改进的广义蚁群优化算法[J].计算机技术与发展,2012,22(6):39-44.
Zhang D Y. A new improved generalized Ant Colony Optimization Algorithm[J]. Computer Technology and Development,2012,22(6):39-44.

[9] 朱骏,潘理,李建华.基于蚁群算法的P2P网络资源发现算法[J].信息安全与通信保密,2007,2:166-168.
Zhu J,Pan L,Li J H. An Ant Colony Algorithm based P2P network resource detection algorithm[J]. Information Security and Communications Privacy,2007,2:166-168.

[10] Bontoux B, Feillet D. Ant Colony Optimization for the traveling purchaser problem[J]. Computer & Operations Research,2008,35:628-637.

Search Resources in P2P Network Based on Query Agent and Generalized Ant Colony Algorithm

WANG Wenhao[1], CHEN Xiaobing[1], JIANG Daoxia[2]

(1. Faculty of Computer Engineering, Huaiyin Institute of Technology, Huaian Jiangsu 223003;
2. Faculty of Mech-E&Info Engineering, Jiangsu Vocational and Technical College of Finance and Economics, Huaian Jiangsu 223003, China)

Abstract: With the rapid development of Intranet, there are a lot of resources in private network, how to extend the P2P network to private network, and conduct the internal resources of private networks becomes a much valuable research subject. On the basis of analysis of the existing P2P network resource searching method, a P2P resource searching method based on query broker and generalized ant colony algorithm is introduced. The query broker is used to search resources in private network, the generalized ant colony algorithm is used to search resources in the public network, In addition the generalized ant colony algorithm is improved by integrate the credibility of the nodes into the optimization conditions of this algorithm. Experiments show that: this method can improve the searching efficiency and accuracy, and can effectively isolate the false nodes in the network. It is important significance to improve the quality of network service.

Key words: P2P; resource searching; query broker; generalized Ant Colony Optimization Algorithm

(责任编辑 游中胜)