基于蚁群算法的非结构化 P2P 网络搜索的改进

詹晓亮,余 强,熊 健
(西华大学 数学与计算机学院,四川 成都 610039)

摘 要:针对非结构化的对等网络一般以广播方式作为其搜索的基本策略而引发较大的网络流量和盲目性这一问题,引入人工智能领域的蚁群算法,利用蚂蚁信息素的多样性和正反馈机制,有效地指导节点选择查询,以便更快地找到查询结果。仿真结果表明,该算法有效地减少了查询带来的网络流量和盲目性,提高了查找的成功率。

关键词:对等网络;蚁群算法;信息素;非结构化 P2P;查询
中图分类号:TP393.02 文献标识码:A

Improved searching algorithm for unstructured P2P network based on ant colony optimization

ZHAN Xiao Liang, YU Qiang, XIONG Jian
(School of Mathematics and Computer Engineering, Xihua University, Chengdu 610039, China)

Abstract: Ant colony optimization (ACO) in artificial intelligence is introduced. This mechanism directs the query routing effectively according to the diversity and the positive feedback principle of the ant pheromone. In the way the peer reduced the blind search..Simulation results show that the new algorithm reduces the network traffic and blindness greatly and improves the searching success ratio.

Key words: peer-to-peer network; ant colony optimization algorithms ; pheromone; unstructured P2P; query

P2P(Peer to Peer)即对等计算或对等网络,通过直接交换共享计算机资源和服务。在 P2P 网络环境中,成千上万台彼此连接的计算机都处于对等地位,整个网络不依赖专用的中心服务器。网络中的每台计算机既能充当网络服务的请求者,又能对其他计算机的请求做出响应,提供资源与服务。

按照数据信息在网络中的存储方式,可以将 P2P 系统分为两大类:结构化 P2P 系统,以 Chord[1]、CAN 等为代表,在结构化网络中每个节点存储的信息与网络拓扑结构有关;非结构化 P2P 系统,以 Gnutella[2]、Freenet[3]为代表,非结构化网络与网络拓扑无关,其节点可任意存储信息。以上所说的这些网络主要是通过直接交换来共享计算机资源,因此搜索技术成为 P2P 网络的关键技术。

目前,大多数应用系统是非结构化拓扑结构,这种结构的覆盖网络一般采用基于完全随机图的搜索算法,容错性好、支持复杂查询、受节点频繁加入和退出系统的影响小。随着联网节点不断增多,以这种方式定位资源将造成网络流量急剧增加,导致网络带宽消耗大,并由此产生可扩展性差等问题。所以不能采用传统路由扩散的方法解决搜索信息的定位问题;同时,如何从大量分散的对等点中快速找到所需要的资源也是个巨大的挑战。本文对无结构对等网络现有的算法进行研究,在此基础上引入人工智能领域的蚁群算法[4],利用蚂蚁信息素的多样性和正反馈机制来指导节点的选择查询,以达到减少查询所带来的网络流量和盲目性,提高查询成功的概率。

1 常见的搜索算法与缺点

根据对邻居转发方式的不同,可将非结构化对等网络的搜索算法分成两类:泛洪搜索算法与漫游类搜索算法。

(1) 泛洪搜索算法一般有 Gnutella 算法、Iterative-Deepening/Expanding-Ring 算法[5]、改进的 BFS 算法[6]等。目前,在运作的大型的 P2P 网络都采用类似于 Gnutella 网络的模式,Gnutella 实现了对搜索规模的简单控制,而对搜索消息传播的控制是通过 TTL 的减值来实现的。在 Gnutella 网络中,节点检查收到的 Query 消息或者 Ping

消息的 TTL 值,如果消息的 TTL 值不等于零,则将该消息的 TTL 减 1,然后将该消息转发给节点的所有邻居节点;如果消息的 TTL 值等于零,则不再转发该消息。但在 Gnutella 算法中,TTL 的设置是一个两难的问题:TTL 值设置过大,很可能造成大量无谓的搜索开销;而 TTL 值设置的过小,则搜索不易满足,使得搜索需要重复发起。因此,LV 等研究人员提出了 Expanding-Ring 算法,接着 Yang 等研究人员又提出了 Iterative-Deepening 算法,这两个算法本质相似,算法的前半部分完全与 Gnutella 算法一样,主要区别在于算法的后半部分。在该策略中,查询节点先初始化一个较小的生命周期 TTL 值,如果成功查询到生命周期正好降到 0 时,所有查询消息都暂时休眠,等待搜索结果返回查询发起节点。查询发起节点收到所有结果之后,判断该次搜索是否已满足查询结果,如果已满足,则放弃继续搜索;若不满足,则逐步增加 TTL 值,直到 TTL 值达到预定的最大值或成功查询到目标数据。这种策略在某种程度上降低了网络流量,但依然会延长找到目标数据的时间。由此,有人提出另外一种泛洪算法:改进的 BFS 算法,该策略是所有节点随机抽取一定比例的相邻节点传递查询信息包 Query,而不是像原始泛洪搜索算法那样把查询信息包转发给所有的邻居节点。选择的邻节点的数目占总邻居数的比例是这种机制的参数。虽然它减少了平均消息量的产生,但是它仍然需要与大量的节点相关联,而且有一定的随机性。

(2)漫游类搜索算法最常见的是 K-Random-Walker 算法[7],该策略是研究人员在原来 Expanding-Ring 算法的基础上构想了这种叫做 K 步漫游行走器的算法,节点将 K 个请求消息转发给在其相邻节点中随机选取的 K 个节点,然后这 K 个节点将请求消息随机地向它的一个相邻节点进行转发,依次类推。直到搜索成功或者 TTL 值为零。这种策略的最大优点是大大减少了消息的产生数量,但缺点是难以优化 K 值。K 值小则冗余开销小,但用户时延大;如 K 值大则可以获得低时延,但导致冗余开销增大,并使查询发起者附近节点重复收到查询消息。对 K 的优化需预先了解 P2P 网络全局信息,但无结构 P2P 网络中单个节点难以做到。

2 基于蚁群算法的非结构化 P2P 网络搜索的改进

前述搜索算法,不是导致查询产生大量的网络流量而对网络造成很重的负担,就是导致搜索结果的可靠性受到影响。因此,将基于人工智能的蚁群算法应用到非结构化 P2P 网络的搜索中具有十分重要的意义,基于此提出了一种基于蚁群算法的非结构化 P2P 网络搜索的改进。该策略是利用蚁群算法的两种机制:多样性和正反馈机制。多样性保证了蚂蚁在觅食的时候不致走进死胡同而无限循环;正反馈机制则保证了相对优良的信息能够被保存下来,从而建立它们的对应关系表,利用该路由表来指导资源搜索。仿真实验表明,该搜索算法有效地减少了查询带来的网络流量和盲目性,提高了查找成功率。

2.1 算法描述

在基于蚁群算法的非结构化 P2P 网络搜索机制中,整个 P2P 系统可以视为一个由很多蚁巢以及连接这些蚁巢的通路构成的网络。每个蚁巢相当于 P2P 网络中的一个对等节点,存放了一定量的 P2P 文件资源,查询消息包可以看作是蚂蚁,搜索的目标视为食物,存在搜索目标的节点就是食物源。当某节点发起查询请求,相当于蚁巢派出蚂蚁寻找食物,当一只蚂蚁找到食物以后,它会在环境释放一种信息素,吸引其他的蚂蚁过来,这样越来越多的蚂蚁会找到食物,每一只蚂蚁每一步的行动是,根据一定的依据选择下一个它还没有访问的节点,同时在完成一步(从一个节点到达另外一个节点)或者一个循环(完成对所有 n 个节点的访问)后,更新所有路径上的残留信息浓度,由此,网络中的节点都维护一张信息素表,来保留本节点的信息和信息素浓度以作为下一跳节点的依据。用户通过节点查找关键词 Key 时,先检查自己的信息素表,如果有,则返回给用户,并更新信息素表;如没有,先从信息表中找相似主题,如找到与主题相似度达到一定值的时候,查询蚂蚁再根据本节点的信息素浓度作出选择。浓度高,就作为下一跳节点,如果下一节点没有要找的食物,就将消息包中的 TTL(存活时间)值减 1。某一节点在收到消息包后,如果发现 TTL 值为 0,就停止转发消息包。如果消息包在 TTL 减为 0 之前到达了拥有食物的目标节点,就会返回一个命中消息包。这可以看成是找到食物的蚂蚁沿原路返回源节点,沿途释放信息素,并且修改节点的信息素表。

2.2 路由表的建立

每个蚁巢需要维护 3 张表:本地资源存储表、关键字信息存储表和信息素表,如表 1 所示。本地资源存储表负责本地资源的维护,它包括节点的 ID 和 IP 信息,这些信息可以用来开辟新的连接;关键字信息存储表存储该节点以前访问的相应的请求信息,它可以对外提供资源信息的查询。信息素表存储了通往邻居节点上关键词信息素浓度值。

当节点新加入时,需要初始化路由表,信息素浓度初始值都为常量 0,当用户通过源节点进行关键词 Key 查询,先检查本地的路由表是否有满足条件的资源,若有则将其返回给用户,并修改路由表;若没有,则根据主题相似函数检查是否有相关主题。2 个主题的相似度用式(1)夹角相似公式[8-9]计算:

$$S(\boldsymbol{V},\boldsymbol{U})=\cos(\boldsymbol{V},\boldsymbol{U})=\frac{\sum_{k=1}^{m} a^{(v)}(t_k)\cdot a^{(u)}(t_k)}{\sqrt{\sum_{k=1}^{m} a^{(v)}(t_k)^2}\cdot\sqrt{\sum_{k=1}^{m} a^{(u)}(t_k)^2}} \quad (1)$$

每个主题 D 可以表示为一个范化特征矢量 $\boldsymbol{\Psi}(D)=[\alpha$

表 1 路由表信息

节点信息	关键词	信息素浓度值
[ID$_1$, IP]	Keyword$_1$	m$_1$
	…	…
	Keyword$_n$	m$_n$
[ID$_2$, IP]	Keyword$_1$	m$_1$
	…	
	Keyword$_n$	m$_n$
…	…	…

$(\Psi)(t_1),\ldots,\alpha(\Psi)(t_i),\ldots,\alpha(\Psi)(t_m)]$。$\alpha(t_i)$表示词条 i 权重要程度的权值。假设两个查询的矢量分别是 $\boldsymbol{U}$ 和 $\boldsymbol{V}$，夹角越小说明查询的相似度越高。其决策规则为：若 $S(\boldsymbol{V}, \boldsymbol{U}) \geqslant \zeta$，则 $\boldsymbol{U}$，$\boldsymbol{V}$ 相似；否则不相似。相似度阈值的确定公式为：

$$\frac{\sum_{t=1}^{T}\sigma[s,1]\zeta[s,1]}{\sum_{t=1}^{T}\sigma[s,1]} \tag{2}$$

其中 $\sigma[s,1]$表示阈值的自信度，$\zeta[s,1]$表示划分的相似度阈值。

如果相似度一样，再根据信息素浓度值的高低来判断下一跳。当用户在返回命中消息包时，沿途节点的信息素表更新规则如下：

$$m=(1-\rho)\times m+\Delta m \tag{3}$$

其中 ρ 是挥发系数，通常设置 $\rho < 1$，用来避免信息素的无限增加。Δm 为信息素的增量，这里根据下式得出：

$$\Delta m = W\times(\mu^{S_{ij}}) \tag{4}$$

W 是增加的信息素原始总量，为常数。μ 为节点浓度差，是一介于 0 和 1 之间的调节因子。S_{ij} 为目标节点 i 到节点 j 的跳数，这样就保证了在每次增加信息素浓度时，离目标节点越近的节点增量越多，其增量随着当前节点与目标节点距离的增大而递减。每隔一段时间，信息素会自动更新，形成信息素的挥发现象，这时 Δm 取值为 0 。

2.3 路由表的更新与维护

为了避免残留信息过多引起的残留信息淹没启发信息的问题，在每一只蚂蚁完成对节点的访问后，并返回消息包时，必须对残留信息进行更新处理，又因为缓存是有限的，因此每个节点根据信息量大小来管理缓存空间，路由信息表的更新与维护的算法框图如图 1 所示。

3 仿真与实验

本文通过实验对基于蚁群算法的非结构化 P2P 网络搜索的改进算法的性能进行评估，考察其搜索成功率、响应率(平均每次搜索产生的响应结果数) 、产生的冗余消息以及整体性能(搜索效率) ，并与 2 个典型的洪泛算法和随机漫步算法进行性能比较。

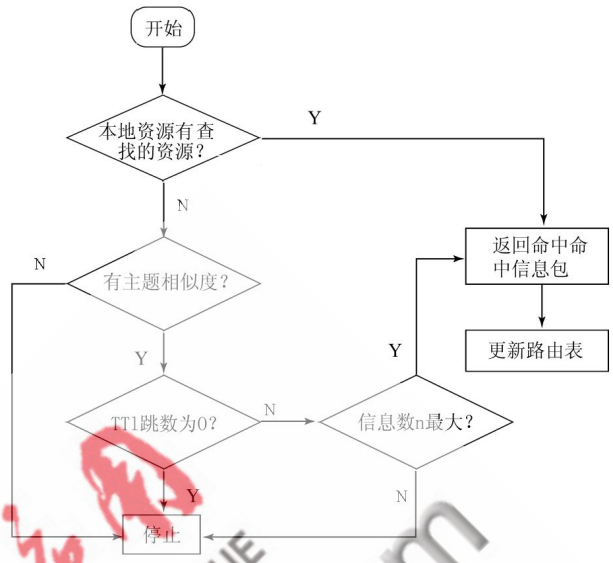

图 1 路由信息表的更新与维护的算法框图

所有实验在一台 Core(TM)2 Duo 2.00 GHz 处理器、2GB内存、Windows XP 操作系统的 PC 机上完成。用Peer-Sim[10]模拟 P2P 网络环境，网络拓扑是基于 BRITE[11]产生的 10000 个节点的 power law 模型。其中最大节点度数为 315、最小节点度数为 4，平均节点度数为 8，以上三种搜索算法采用 Java 语言编程实现并在 PeerSim 上运行。

实验中采用 100 个对象，其拷贝分布和查询分布服从 Zipf-like 分布，且分别为 Zipf-like(α=1)和 Zipf-like(α= 0.9)。这一设置符合参考文献[12]的调查结果：流行对象的拷贝数量占所有拷贝数量的 50% ，对它们的查询次数占所有查询次数的 60%。另外，随机选择 10%的网络节点(即1 000个)进行查询，每个节点进行约 2600 次查询。

当搜索深度 TTL 达到 10 跳时，仿真实验结果如 2 所示。从图 2 中可以看到本文中改进的算法搜索成功率明显高于洪泛和随机漫步算法。另外从图 3 看出在响应率方面，基于蚁群算法的非结构化 P2P 网络搜索的改进算法比洪泛和随机漫步算法提高了近一个数量级。其次，图 4 用消息冗余度(即冗余消息数量占总消息数量的百分比) 来比较各种算法在搜索过程中产生的冗余消息。图 5 比较了各种算法的搜索效率。随着搜索深度的递增，基于蚁群算法的非结构化 P2P 网络搜索的改进算法不仅能够有效地减少冗余消息，而且在效率方面都比洪泛和随机漫步算法高。

本文基于蚁群算法的非结构化 P2P 网络搜索的改进算法，在一定程度上减少了节点的信息处理量，减少了查询带来的网络流量和盲目性，对网络拥塞控制起到了一定的作用，使 P2P 网络的性能得到了进一步的提高。如何动态的确定 TTL 参数的选取，更好地提高 P2P 网络的搜索性能是下一步研究的重点。

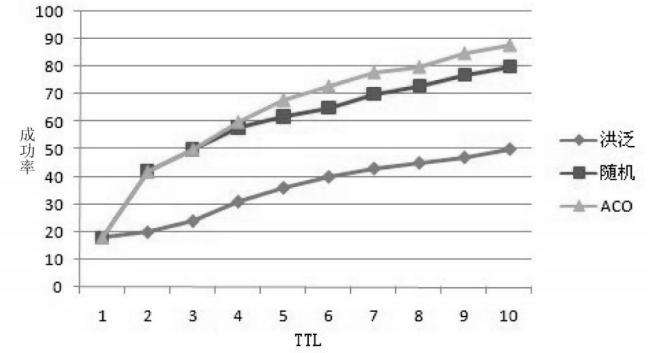

图 2 成功率与 TTL

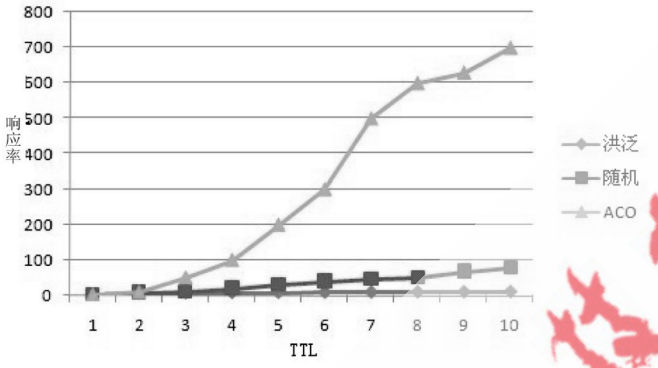

图 3 响应率与 TTL

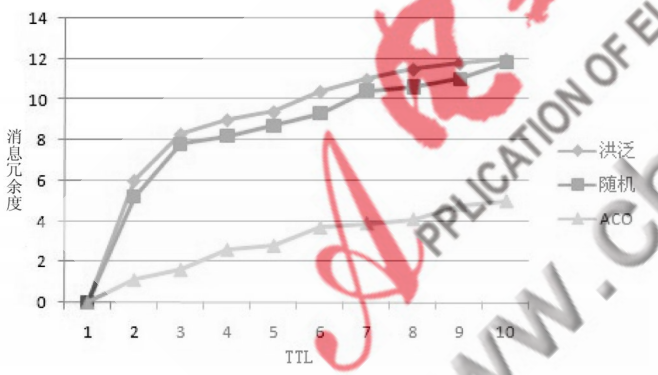

图 4 消息冗余度与 TTL

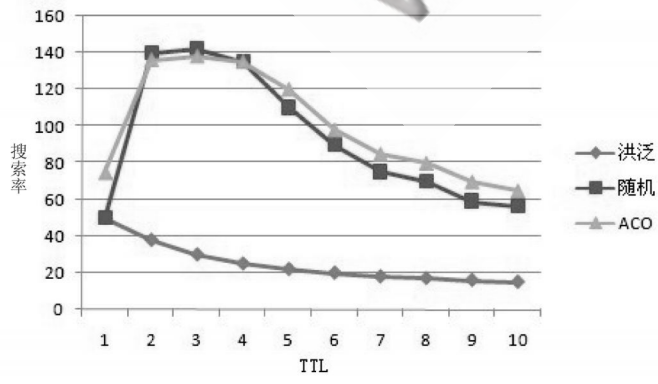

图 5 搜索效率与 TTL

参考文献

[1] STOICA M R, KARGER D C. A scalable peer-to-peer loopup service for internet applications [C].ACM SIGCOM M.san Diego, USA,2001.

[2] RIPEANU M. Peer-to-peer architecture case study: gnutella network [C]. IEEE International Conference on Peer-to-peer Computer(P2P2001).Linkoping,Sweden,2001.

[3] CLARKE I,SANDBERG O,WILEY B.Freenet:a distributed anonymous information storage and retrieval system .The ICSCI Workshop on Design Issues in Anonymity and Unobservability . Berkeley,California, USA,2002.

[4] COLORNI A , DORIGO M , MANIEZZO V. An investigation of som properties of an ant algorithm [C].In : Proc. of the Parallel Problem Solving from Nature Conference (PPSN'92) . Brussels , Belgium: Elsevier Publishing , 1992 :509 - 520.

[5] YANG B, GARCIA H M. Improving search in Peer-to-Peer Networks[C]. ICDCS, 2002.

[6] KALOGERAKI V, GUNOPULOS D, ZEINALIPOUR Y D. A local search mechanism for Peer-to-Peer Networks [C]. Proc. of the 11th International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2002:300-307.

[7] LV Q,CAO P,CPJEN E,et al. Search and replication in unstructured peer-to—peer networks[C].New York:Proceedings of 16th ACM International Conference on supercomputing(ICS'02), 2002.

[8] ROCCHIO J. Relevance feedback in information retrieval [A]. Salton G. SMART Retrieval Sys: Experiments in AutomaticDoc Proc[C].NJ,USA:Prentice- Hall,1971:313~323

[9] SALTON G,WONG A,YANG C. A vector space model forauto maticindexing[J]. Commu of ACM,1995(18):613-620

[10] PeerSim P2P simulator [EB/OL].[2007201218]. http: //www. peersim.sourceforge. net /.

[11] BR ITE: Boston University rep resentative Internet topology generator[EB/OL].[2007201218].http: //www.cs.bu.edu /brite /.

[12] CHU J, LANONTE K, LEV INE B N. Availability and locality measurements of peer2to2peer file systems [C]. F IRO IU V, ZHANG Zhi-li. Proc of the SPIE Vol. 4868, Scalability and Traffic Control in IP Networks Ⅱ. Boston: SP IE, 2002: 310-321.

(收稿日期:2009-09-29)

作者简介:

詹晓亮,男,1982 年生,硕士研究生,主要研究方向:对等网络,分布式计算;

余强,男,1973 年生,副教授、博士,主要研究方向:对等网络,分布式计算;

熊健,男,1986 年生,硕士研究生,主要研究方向:对等网络。