

文章编号:1001-9081(2009)11-2881-03

非结构对等网络中一种有效的资源搜索策略

安世全,高 涛,丁进标

(重庆邮电大学 计算机科学与技术学院,重庆 400065)

(gtahwan@163.com)

摘 要:针对非结构化对等网络中资源搜索算法效率不高、搜索过程中产生的冗余消息数过大而造成的网络带宽消耗及网络拥塞等状况,提出一种基于路由搜索机制的改进算法。该算法利用邻节点之间的关系,生成邻节点的转发路由表。实验证明,该算法有效抑制了网络中冗余搜索消息数量,减小了网络带宽的消耗,有效避开了搭便车节点,从而提高了搜索效率。

关键词:对等网;资源搜索;路由表;邻接关系

中图分类号: TP301.5 **文献标志码:** A

New effective resources search strategy in unstructured P2P network

AN Shi-quan, GAO Tao, DING Jin-biao

(College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Concerning the low efficiency of resources search and the dramatic consumption of bandwidth when searching in unstructured P2P network, the authors put forward an improved algorithm based on routing mechanism. The algorithm generated neighboring nodes' routing tables by analyzing the neighboring relations. Simulation results show that the proposed algorithm effectively reduces the network traffic and consumption of bandwidth and keeps away from selfish nodes. Meanwhile, it improves the efficiency of resources search.

Key words: Peer-to-Peer (P2P); resources search; routing table; neighboring relation

0 引言

对等网络(Peer-to-Peer, P2P)中每个节点地位对等,既充当服务器,为其他节点提供服务,同时也为客户机,享用其他节点提供的服务。P2P网络从结构上分为无结构化P2P和结构化P2P。结构化P2P网络建立在分布式哈希表(Distributed Hash Table, DHT)上,在给定资源索引下的情况下,能够在 $O(\log N)$ 跳之内定位到目的节点,资源定位快,但需要以很高的代价维护既定拓扑,而且节点的加入和退出所引起的网络波动较大,不能很好地适应高度动态的P2P环境。无结构化P2P资源的搜索和定位通过消息扩散来实现,容错性好,支持复杂查询,受节点频繁加入和退出的影响小,但搜索数据几乎是随机搜索,容易造成网络流量急剧增加,导致网络拥塞。因此,无结构化P2P的一个核心问题就是如何进行资源的快速搜索,同时降低网络带宽消耗,保证系统可扩展性和稳定性。本文主要讨论无结构P2P网络环境下的搜索策略。

Gnutella是非结构化网络的典型代表,其设计思想简单,资源搜索采用洪泛搜索机制,每个节点在查找过程中将接收到的搜索请求信息转发给所有的邻居节点。随着对等网络规模的不断增大,洪泛机制的弊端越来越明显:搜索过程中产生大量冗余的消息,严重吞噬带宽而使得系统极不可扩展。文献[1]中提出的随机漫步(Random Walk, RW)算法的基本思想是:查询节点将 K 个查询请求转发给在其相邻节点中随机选取的 K 个节点,然后这 K 个节点将查询请求随机地向它的一个相邻节点进行转发,以此类推。这些请求消息叫作漫步

者,漫步者在搜索成功或者生存时间(Time To Live, TTL) t_{TTL} 为0时结束。相比洪泛方法,随机漫步算法能大量减少消息冗余,最坏的情况下所产生的最大消息数为 $K \cdot t_{TTL}$,但其搜索成功率随着网络拓扑结构和 K 的选取有很大的变化。文献[2]中指出Gnutella具有Power-law特性,并依据这种特性提出了将搜索请求转发到度数较高的节点,度数较高的节点指向更多的节点,包含了较多的信息。文献[3]中给出了一种前向学习的方法,在这种方法中建议每个节点包含的元数据可以给出哪些节点可能含有应答信息的隐含信息,依据此隐含信息可以选择节点转发请求信息。文献[4]中的研究结果表明,在Gnutella中,有42%的节点不共享任何资源,而10%的节点提供了大部分的资源。文献[5]中提出的基于扩散路由机制的改进算法(Improved algorithm Based Spread Routing, ISR)在一定程度上控制了节点转发的搜索消息数量,但其未考虑到对等网络中大量存在的搭便车的自私节点及节点负载情况,使搜索的效率和可靠性大打折扣。

由分析得,为提高无结构化对等网络中资源搜索效率,在保持节点负载平衡的情况下,要使搜索消息尽量流向度数较高的节点,避开自私节点,同时降低搜索消息的冗余度和减小消息数据包的大小。因此,本文提出了一种基于路由搜索机制的改进算法,较之洪泛算法和随机漫步算法在搜索消息控制及搜索效率方面均有较大的改进。

1 算法描述

1.1 转发路由表的引入

通过研究Gnutella网络中节点的聚集性和幂率特性,分

收稿日期:2009-05-14;修回日期:2009-07-03。

基金项目:高校教师专业发展研究项目(06AIIJ0180031);中国高等教育学会教育科学“十一五”规划重点研究课题(06AIIJ0180024)。

作者简介:安世全(1962-),男,甘肃天水人,教授,主要研究方向:非线性分析、系统科学、应用数学;高涛(1984-),男,湖北荆州人,硕士研究生,主要研究方向:信息与计算理论;丁进标(1980-),男,山东枣庄人,硕士研究生,主要研究方向:数据挖掘。

析节点的邻接关系,发现一个节点的直接前驱节点利用其邻节点集合能够得知该节点的消息转发路由。节点 n 的搜索消息转发路由表由它的前驱节点生成,通过附加在搜索消息数据包中传递给节点 n 。网络中各节点存储各自的邻居节点集合 $Neighbor(n)$ 以及每个邻节点的邻节点集合 $Connection(n, i)$ [6],即每个节点必须了解其邻居节点的邻接关系信息,即每个节点都知道谁与它的邻居节点直接相连。这些信息相对比较稳定,只有当网络中有新的节点加入或者有节点离开时才需要更新,而这种更新可以通过在相邻节点之间周期性交换邻居连接关系查询包来触发。有关定义如下。

定义 1 集合 $Neighbor(n)$ 表示节点 n 的所有邻接点集合。

定义 2 设当前节点为 n , 节点 i 是节点 n 的一个邻节点,集合 $Connection(n, i)$ 表示节点 n 的邻节点 i 的所有邻节点的集合。

定义 3 如果节点 j 从节点 i 接收到搜索消息数据包,节点 k 是节点 j 的待转发节点,则称节点 i 是节点 j 的直接前驱节点,节点 k 是节点 j 的直接后继节点。

设当前节点为 n , 节点 i 是节点 n 的直接后继节点,节点 i 的搜索消息转发路由表 $Forward(i) = Connection(n, i) - (Neighbor(n) \cap Neighbor(i)) - n$, 即节点 i 的转发路由表为从其所有邻节点中去掉与直接前驱节点 n 的共同邻节点的剩余邻节点集合,另外还要去掉节点 n 。这样处理的目的是防止搜索消息回流和侧向流动,使得消息始终向外扩散。若将搜索起始点看作制高点,则这种搜索消息扩散机制有类似于水流特性。

1.2 对节点负载及自私节点的控制

研究表明,Gnutella 网络具有很强的聚集性。所谓聚集性是指网络中的一些节点互相连接聚集在一起。具有聚集性的节点具有很高的连通度,又由于高连通度的节点与其他节点联系更为频繁而显示出更高的稳定性和更丰富的资源,通过它查找到待查资源概率较高。因此,若控制搜索消息流向高连通度的节点,搜索效率和系统的稳定性将会大大提高。在 P2P 这样一个高度动态和异构的网络中,各个节点都有计算和带宽的限制。节点接受查询请求,执行查询处理,返回查询结果需要消耗一定的计算资源和网络带宽资源。本文引入节点负载^[7]和节点负载率的概念。

定义 4 节点负载。节点在单位时间内接受的查询请求数量,称之为节点负载。

定义 5 节点负载率。令节点 n 的能力为 c_n , 负载为 l_n , 那么 n 的负载率为 $u_n = l_n/c_n$ 。令节点 n 的负载率阈值为 t_n , 当 $u_n \geq t_n$ 时,称节点 n 过载,否则称节点 n 欠载。

1.3 算法分析

在算法中,每个节点维护着一个如表 1 所示的路由信息表,表中存储邻节点的相关信息,表中各项按照节点连通度从大到小顺序排列。搜索过程中,控制搜索消息流向连通度高且处于欠载状态的节点。若连通度最高的节点处于过载状态,则选择连通度次高的节点,以此类推。另外,考虑到网络中存在搭便车节点的情况,用 Share 关键字来表示节点是否共享资源。若节点不共享资源,则在形成转发路由信息表的时候去掉该节点,从而使搜索消息不流向该节点。算法描述如下所示。

1) 初始化搜索路由信息表,表项数为 $items, items[i]$ 表示表中第 i 项。

2) 选取表中第一项开始处理搜索请求,描述如下:

```
for(int i = 0; i <= items; i++) {
    if (! Tol && Share) //判断邻节点是否共享资源及是否过载
    {生成 Forward(i);
    if (Forward(i) <> null)
    {将搜索消息转发到邻节点 i;
    if (success) //搜索到请求资源
    {停止搜索并返回已找到请求资源信息;
    }
    }
```

下面以一个含有 11 个节点的 Gnutella 网络来分析算法执行过程。图 1 为洪泛算法转发搜索消息数据包过程。从节点 A 开始搜索,搜索过程中总共产生了 28 个搜索消息数据包,图 2 为改进算法,所产生的数据包数为 7 个。由表 2 可以得出,节点 H 处于过载状态,搜索消息不转发给 H,节点 I 和节点 L 为不共享资源的搭便车节点,搜索消息路由扩散时避开了这两个节点。显然,采用改进算法在 Gnutella 网络中所产生的冗余搜索消息数据包数量大大减少,随着网络中节点数量的增加,节点聚集性的增强,搜索效率和性能行将会大大提高。

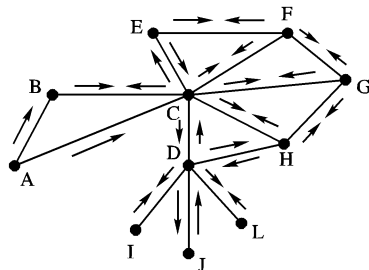


图 1 洪泛算法转发搜索消息数据包示意图

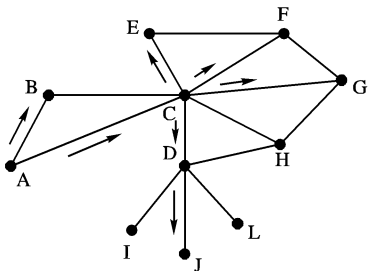


图 2 改进算法转发搜索消息数据包示意图

表 1 节点搜索路由信息表

连通度	邻节点集	邻节点的邻节点集	负载状态	共享标志
Degree	Neighbor(n)	Connection(n, i)	Tol	Share

表 2 节点共享及负载状态信息

节点	是否过载	是否共享资源	节点	是否过载	是否共享资源
A	否	是	G	否	是
B	否	是	H	是	是
C	否	是	I	否	否
D	否	是	J	否	是
E	否	是	L	否	否
F	否	是			

2 实验分析

为了验证算法的搜索效率,选择 GnuSim 仿真器进行系统模拟和实验数据分析,并将结果与 Flooding 算法及 RW 算

法进行对比。GnuSim 是一个通用 Gnutella 和非结构化 P2P 网络仿真器,用于构造 Gnutella 和非结构化 P2P 网络模型。该仿真器的目的是验证在非结构化 P2P 网络中使用的各种模式,并评估其性能和价值。

由于在 Gnutella 网络中存在大量查询消息,但并不是所有查询消息都能找到所需的资源^[8],因此,可以通过考查在一定查询消息数的情况下找到的资源数量来衡量搜索机制的有效性。搜索效率定义如下:

$$E_i = \sum_{k=1}^{n_i} S_k / \sum_{k=1}^{n_i} M_k \quad (6)$$

$$SE(\mu) = \frac{1}{\mu} \sum_{j=1}^{\mu} E_j = \frac{1}{\mu} \sum_{j=1}^{\mu} \left(\sum_{k=1}^{n_i} S_k / \sum_{k=1}^{n_i} M_k \right) \quad (7)$$

其中: E_i 表示第 i 组实验的搜索效率; n_i 表示第 i 组实验的搜索次数; S_k 表示第 i 组实验时第 k 次搜索找到的资源数; M_k 表示第 i 组实验时第 k 次搜索发出的搜索消息数; $SE(\mu)$ 为 μ 组实验的搜索效率的平均值。

平均搜索消息数据包比较如图 3 所示,搜索效率如图 4 所示。

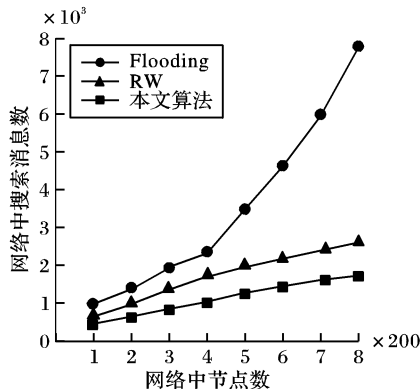


图3 网络中搜索消息数比较

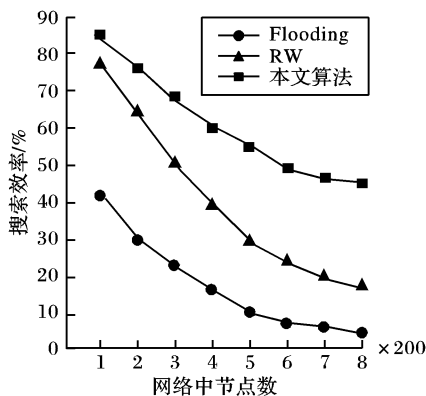


图4 搜索效率比较

为验证算法在搜索过程中搜索消息数据包对网络中搭便车节点的转发控制情况,设定网络中自私节点比例为 30%,在实验过程中记录在两种情况下网络中搜索消息数据包的数量,如图 5 所示。未考虑 Share 状态的算法在转发搜索消息数据包时存在盲目性,效率低下;而考虑 Share 状态的算法则有效避开了网络中的自私节点,使得转发消息数据包的目的性非常明显,因而提高了搜索效率。

在搜索过程中,若某些搜索频度很高的关键字所在的节点(称作热点)长期处于欠载状态而不加以控制的话,会严重影像搜索响应时间,降低搜索效率。在实验中,对节点 D 的负载情况进行采样来说明本文算法对节点负载状态的控制。

在模拟网络中,每隔 1 s 对节点 D 所含的高频度关键字“key”发起搜索,采样间隔时间为 1 min,连续采样 30 min 节点 D 的负载情况如图 6 所示。从图上可以看出节点 D 的负载在 25 左右,且一旦超过 25 会在很短时间内调整至 25 以内。

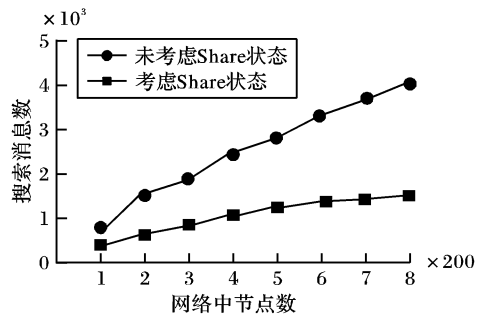


图5 本文算法中对 Share 状态考虑情况的比较

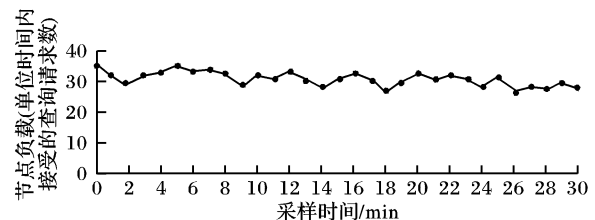


图6 节点 D 在 30 min 内负载的采样结果

3 结语

本文通过研究 Gnutella 网络中节点的聚集性和幂率特性,分析节点的邻接关系,提出的资源搜索算法,显著提高了无结构化对等网络中资源搜索效率,降低了网络带宽的消耗。在网络中存在搭便车的自私节点的情况下,降低了搜索消息的冗余度,提高了搜索效率,从理论上以及模拟实验均验证了算法的有效性。

参考文献:

- [1] LV QIN, CAO PEI, COHEN E, *et al.* Search and replication in unstructured peer-to-peer networks[C]// Proceedings of the 16th International Conference on Supercomputing. New York: ACM Press, 2002: 84-95.
- [2] ADAMIC L A, LUKOSE R M, PUNIYANI A R, *et al.* Search in power law networks[J]. Physical Review E, 2001, 64(4):046135-046143.
- [3] JOSEPH S. NeuroGrid: Semantically routing queries in peer-to-peer networks[C]// NETWORKING 2002 Workshops on Web Engineering and peer-to-peer Computing, LNCS 2376. Berlin: Springer-Verlag, 2002: 202-214.
- [4] ASVANUND A, BAGLA S, KAPADIA M H. Intelligent club management in peer-to-peer networks[EB/OL]. [2006-06-10]. <http://www.sims.berkeley.edu:8000/research/conferences/p2pecon/papers/s6-asvanund.pdf>.
- [5] GUO YU-TANG, LV WAN-LI, LUO BIN. Improved resource discovery algorithm on Gnutella based on P2P networks[C]// CCC 2007: Proceedings of the Chinese Control Conference. Washington, DC: IEEE Press, 2007: 599-602.
- [6] 黄道颖,陈新,张安琳,等. P2P 网络 Gnutella 模型中搜索消息的路由机制及改进研究[J]. 计算机工程与应用, 2003, 39(25): 13-15.
- [7] 刘德辉,周宁,尹刚,等. QFMA: 一种支持负载均衡的多属性资源定位方法[J]. 计算机学报, 2008, 31(8): 1376-1382.
- [8] 袁静波,石鸿伟,丁顺利. 非结构化 P2P 网络搜索算法的研究与改进[J]. 计算机工程, 2008, 34(22): 109-111.