异构P2P网络的分布式查询协议

任 超，李战怀，张 英

(西北工业大学计算机学院 西安 710072)

【摘要】从P2P网络节点的异构性出发，考虑了节点物理位置，提出了一种分层的混合路由查询协议。该协议利用时间戳在局部实现了节点逻辑位置和物理位置的统一，并引入超级节点解决了节点负载失衡和性能瓶颈的问题。在仿真平台P2PSim上的性能测试表明，该协议在大规模P2P网络中优于Kademlia、Chord、Tapestry。

关 键 词 分布式哈希表； 异构； 对等网； 查询； 仿真； 覆盖网络

中图分类号 TP311.11 文献标识码 A

Distributed Query Protocol on Heterogeneous P2P Overlay Networks

REN Chao, LI Zhan-huai, and ZHANG Ying

(School of Computer Science, Northwestern Polytechnical University Xi'an 710072)

Abstract Considering the heterogeneity and physical location of nodes, this paper represents a layered mixed query routing protocol. It uses time-stamp to achieve the agreement of physical location and logical position, and introduces super-node to cope with load imbalance between nodes and performance bottlenecks. Simulation test on P2PSim shows that this solution is superior to other protocols, such as Kademlia, Chord, and Tapestry, on large-scale peer-to-peer (P2P) networks.

Key words distributed hash table; heterogeneity; P2P; query; simulation; overlay networks

对等网(peer to peer，P2P)是在物理网络之上形成的一种覆盖网络，以提供资源发现和定位的服务。P2P网络是一种逻辑网络，一般分为两种，一是基于泛洪(flooding)算法的非结构化网络，二是基于分布式哈希表(DHT)的结构化网络。由于泛洪算法极易引起网络风暴，所以，目前的研究热点主要是以文献[1]等为代表的基于(DHT)的结构化网络。

当前流行的基于DHT的结构化网络协议均有如下假设：所有节点的能力是相当的；资源在覆盖空间中的分布是随机的、均匀的；所有节点都可以和其他任何节点进行连接。事实上，这些假设与客观情况是有出入的。对等网指的是地位上的平等而不是能力的相同。在实际情况中，各节点的计算能力、通信带宽和活动时间的不同[2-4]，将会导致节点负载失衡，造成网络的性能瓶颈。

针对P2P网络的异构性，同时考虑节点位置信息，本文提出一种新的P2P网络路由模型RCDHT。在RCDHT模型中引入了超级节点的概念，提出了分层和混合路由查找思路。在P2PSim仿真平台上对RCDHT的测试结果表明，RCDHT在查询耗时、逻辑跳数以及查询成功率等性能指标上都有了一定的提升。

1 P2P网络路由模型存在的问题

1.1 节点异构性

对等网中的“对等”指的是节点地位上的平等，而不是能力上的平等。事实上，P2P网络中的各节点存在普遍的差异性，以往的P2P网络模型均没有考虑这些节点之间的差异。

Chord[5]网络中节点的负载与能力分布如图1所示。

从图中可以看出，大部分节点的负载低于100，只有少数节点负载高于500；处理能力弱的节点对此负载压力较大，而处理能力强的节点对此负载压力较小，这会导致对等节点“负载失衡”。

带宽是影响网络性能的重要因素。同样数量的请求/应答，对于带宽大的节点，可能可以轻松应对，但造成带宽浪费；而对于带宽较小的节点，造成网络堵塞。事实上，实际的网络节点的带宽是极端异

收稿日期：2008－10－29；修回日期：2008－12－03
基金项目：国家自然科学基金(60873196)
作者简介：任超(1982－)，男，博士生，主要从事网络安全、数据挖掘方面的研究.

构的。平等看待各节点带宽的算法也会造成网络性能的瓶颈[6-8]。

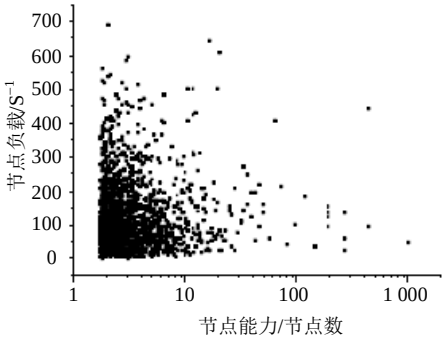

图1　Chord网络中节点的负载与能力分布

另外，节点活动时间也有很大差异。一项对BitTerrent系统的测试表明，只有17%的节点在下载完毕后仍然在线1 h以上，3.1%的节点仍然在线10 h以上，0.34%的节点在线100 h以上[3]，如图2所示。

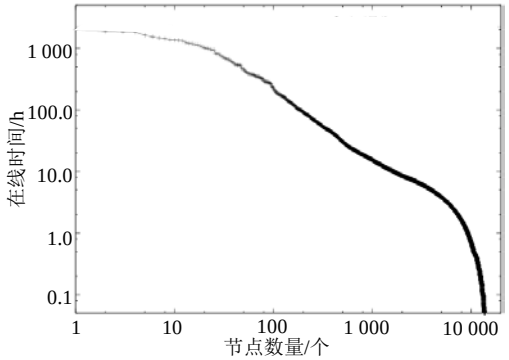

图2　BT中节点下载完成后在线时间分布

1.2　节点位置信息对查询速度的影响

基于DHT的P2P路由算法将各节点ID和资源key映射到同一命名空间，共享资源存储在距离其映射序列最近的节点上。不同的路由算法只是采用不同的探测方法，在映射空间中逐步逼近资源所存储的节点，最终获取资源[9-11]。这种路由机制的着眼点是节点的逻辑距离，如在Kademlia中使用XOR度量节点的逻辑距离，而在Chord中也使用了前趋、后继的概念。各种路由算法在某种程度追求的是逻辑上的高效，如平均跳数最少、时间复杂度最低。但是，这种没有考虑节点实际物理位置的机制可能给系统的性能带来一些困扰。尽管在路由算法中源节点到目的节点的跳数能有效地控制在某一范围内，但却不能保证每一跳的合理性：两节点逻辑距离很近、但物理距离很远的情况普遍存在，由此而引起的通信延迟必然影响到查询的速度。

综上所述，完全同等的对待所有节点，将影响系统的性能。

2　分层的P2P网络模型

针对P2P路由模型存在的问题，考虑网络节点的异构性、节点的物理位置等因素，本文提出了一种新的覆盖网络路由模型——RCDHT。该模型对现有的模型做了如下改进。

(1) 由于节点的异构性，从节点的计算能力、存储能力、带宽、活动时间等因素量化出“节点综合性能”，综合性能高的节点将承担更多的访问负载、存储和提供更多的信息，同时也会在更多的节点上注册自己，让更多的节点有机会查询自己。

(2) 将整个P2P网络分为主干网和子网两层。主干网的结构在Kademlia的基础上进行改进，而子网采用中心化拓扑结构。主干网的每个节点都是一个子网的集合点，子网里的非集合点叫做边缘节点，它们只能在子网内，通过集合点和其他节点进行通信。边缘节点要与子网外的节点进行通信，必须通过集合点转发请求。集合点将由综合性能强的节点来担任，以发挥其性能优势，在子网内实行中心化拓扑常用的索引查询以加快局部查询速度。

(3)　针对节点的物理位置信息，引入“时间戳”概念，在分配节点ID时考虑其物理位置信息。让物理位置较近的节点其逻辑距离也较近，使得每一跳的时间花销趋于平衡，也可以使同一子网内的节点尽可能地在同一个物理区域。

RCDHT节点覆盖空间如图3所示。

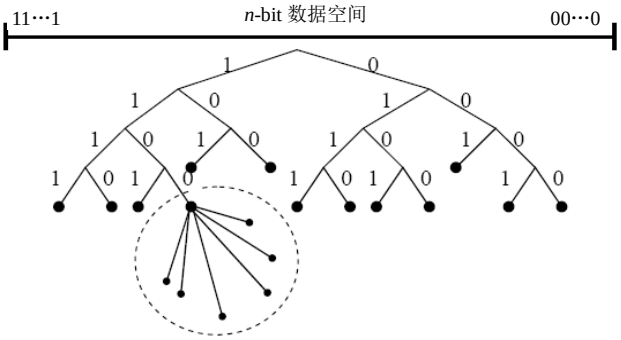

图3　RCDHT节点覆盖空间

2.1　路由模型

RCDHT模型将网络分为主干网和子网，如图4所示。主干网的逻辑拓扑与Kademlia类似，每个节点是二叉树的一个叶子。主干网的节点(即子网的集合点)存在于一个n-bit的覆盖空间内，其中$n=\lceil \log_2 m! \rceil$，$m$为landmark的个数[12]。对于这些节点，定义其逻辑距离为节点ID号的XOR运算[1]。

子网是采用中心化拓扑结构，集合点相当于充当其索引服务器。子网内各节点的覆盖空间以集合

点的ID开头，后面跟上自己*f*-bit的ID以示区别。因此RCDHT模型中节点的ID组成如图4所示。

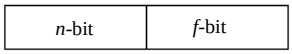

图4 RCDHT中nodeID

RCDHT网络中存在*m*个公开的全局landmark节点，节点ID的分配通过计算节点和这*m*个全局landmark节点的传输延时来实现(而传输延时正是物理位置信息的体现)，然后，再在通过XOR运算得到的节点逻辑距离较近的节点，它们的物理距离也相对较近。

在子网中，边缘节点通过集合点只需一步就可以查到网内的资源和路由信息，所以在分配ID时可以根据实际情况选取算法，如可以根据综合性能排序，或者根据物理距离分配，也可以随机生成。

2.2 路由表

RCDHT网络的路由表分为两个层次，一是主干网通信使用的路由表，即改进后的*k*桶；二是路由索引表，用来在子网内进行定位。

(1) 主干网的路由表

主干网中的路由表是改进后的*k*桶列表，如表1所示。

表1 改进的*k*桶

I 值	距离范围	存放的信息
0	$[2^0,2^1)$	(IP address, UDP port, Node ID, Distance) ……
1	$[2^1,2^2)$	(IP address, UDP port, Node ID, Distance) ……
⋮	⋮	⋮
i	$[2^i,2^{i+1})$	(IP address, UDP port, Node ID, Distance) ……

在*k*桶的信息项中添加了一个Distance字段，用来记录该节点与被记录节点的物理距离。物理距离是通过时间戳产生的一个值，并不是真正的物理距离，但它能从一定程度上反映了真实物理距离。

更新机制和Kademlia协议的*k*桶相同，只是在更新时同时将Distance字段进行更新，因为它会随着网络的带宽等因素变化而改变，对它的更新是非常必要的。对于*k*桶的这些优化提高了高性能节点的利用率，也给路由算法的优化做了准备工作。

(2) 子网的路由表

第二层路由表是子网内的路由索引表，这个表由集合点来维护，每个边缘节点加入到这个子网时，就在集合点上注册自己的信息。子网路由索引表如表2所示。

表2 子网路由表

节点ID	存放的信息
***** 00001	(IP address, UDP port, Node ID)
***** 00010	(IP address, UDP port, Node ID)
⋮	⋮
***** 10110	(IP address, UDP port, Node ID)

集合点定期对索引表内的节点做查询操作以判断其是否异常退出，异常退出节点的信息应该予以删除。

2.3 路由算法

(1) RPC操作

RCDHT网络的路由算法分为两个层次，第一层是子网内部的通信，第二层是子网间节点通过主干网的通信。本文首先要定义几个RPC，分别是PING、STORE、FIND_NODE、FIND_VALUE(和Kademlia中定义的基本相同)，以及REGISTER/LOGOUT，用来在子网中进行边缘节点的注册与注销。各操作的含义如表3所示。

表3 RPC操作表

RPC名	执行的操作
PING	探测一个节点是否在线
STORE	通知节点存储一个(key, valu e)对以便将来查询
FIND_NODE	对于主干网，参考文献[1]；对于子网则直接查找集合点路由索引表，如果查找失败，则由集合点向主干网发送查询请求
FIND_VALUE	对于主干网，参考文献[1]；对于子网则直接到集合点上查找，如果查找失败则由集合点再向主干网发送该请求。
REGISTER/LOGOUT	边缘节点向集合点注册/注销自己

当节点加入某子网时，就向该子网的集合点进行REGISTER操作，即在集合点的路由索引表上注册自己的信息(IP address、UDP port、node ID)供其他节点查询；当节点离开子网时，则向集合点注销(LOGOUT)，集合点将删除它在路由索引表上的记录。

RCDHT网络中节点查询的流程如下：

```
//to find a node with an ID
if ID belong to Subnet
    FIND_NODE-inSubnet(ID);
```

```
        if ID exist
            return (IP address, UDP port, Node ID);
        return err;
    else // ID belong to backbone
        FIND_NODE-inBackbone (ID);
        update k-buckets
        FIND_NODE-inSubnet(ID);
    if ID exist
            return (IP address, UDP port, Node ID);
        return err;
```

(2) 节点综合性能评估

考虑到综合性能包括许多方面，比如计算机的CPU计算能力、主存的物理大小、网络的带宽、活动时间等，本文提出模型$C=\omega f(B,T_\alpha,U,M)$来计算节点综合性能，其中$C$为本文要得到的综合性能指标的量化值；$\omega$是常数；$B$为带宽；$T_\alpha$为活动时间；$U$为计算能力；$M$为存储能力。

$f(B,T_\alpha,U,M)$根据节点各方面的性能进行计算以得出一个数值，这个数值越大则表示综合性能越高。根据实践经验，参数中以网络带宽最为重要，这两个参数在计算中应占有较大的比重。

(3) ID分配算法

节点加入的时候充分考虑它的物理距离，将它放到和它距离较近的子网中去，这样使得逻辑距离较近的节点其物理距离也较近。

假设在整个网络中存在了m个公开的全局landmark节点，新加入的节点u通过时间戳可以测量出自身和各landmark节点的延时，并得到一个对各landmark节点延时的排序。通过排序的结果就可以得到一个ID，即节点u应该加入的子网的集合点ID(或者它就可以作为该集合点)。对应方式为ID number=f(landmark$^\tau$)，其中landmark$^\tau$为一种landmark排序结果。f为Hash函数，将一种排序结果映射到一个ID。

(4) 对k桶的改进

根据节点的综合性能在初始化时动态地决定k桶的k值，即每个k桶存放的信息项个数。在主干网上查询的时候，接收到查询的k桶都会在每一步返回α个信息给消息发送方。α也由该节点k桶初始化的时候由综合性能决定，综合性能较高的节点的α值也较高。同时，在每一步返回α个信息的时候，都选择Distance最小的α个信息项。

(5) 结合点与继承节点

节点在成为集合点时，在子网内选出综合性能最高的r个边缘节点作为继承节点。子网内的边缘节点与集合点通信时得到继承节点列表。选出的继承节点将从集合点处得到路由索引表，同时也得到作为主干网节点的k桶等信息，并与集合点同步更新。

集合点退出时，需通知继承节点就任，并告知所有的边缘节点。集合点在异常退出的时候，边缘节点无法找到集合点，但曾经成功访问过集合点的节点已经获得了继承节点的信息，这时它向继承节点发送查询请求。当继承节点接收到查询时，说明原集合点已经离线，它可以自动就任。

新就任的集合点要在索引表中删除自己(因为它不再需要子网中的ID号)。上任后它就可以选择自己的继承节点，并开始前任的所有的工作。

这种继承机制有很强的自组织能力。新加入的节点如果综合能力很强，它虽然一开始不一定能成为集合点，但它可能很快就能成为继承节点，并最终成为集合点，使网络性能提升。

2.4　节点的加入与离开

由于RCDHT网络分为两层，并且引入了节点的综合性能评估机制，所以节点的加入和离开相对复杂。当节点u加入RCDHT网络时，首先假设它已经知道了m个landmark节点的信息以及一个已经存在于网络中的实际节点w的信息。

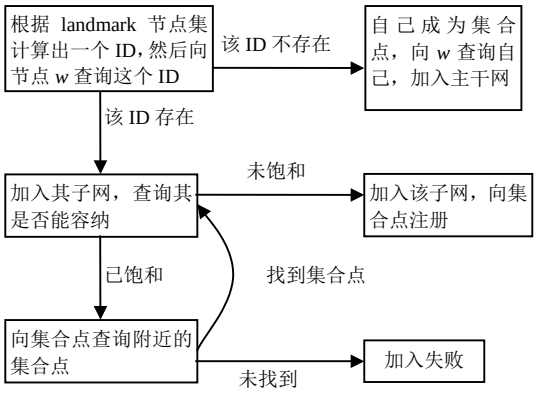

图5　节点加入流程图

节点的退出则要看其是边缘节点还是集合点。边缘节点退出时只要在集合点上注销即可，集合点退出时，对主干网来说不需要做任何工作，这符合Kadmlia协议的要求。对子网来说，它需要通知其继承节点就任，并告之于所有边缘节点。

3　性能评价

本文在P2PSim平台上对RCDHT的性能进行了仿真，在查询耗时、平均查询跳数以及查询成功率上与Kademlia、Chord、Tapestry进行了比较。仿真模型采用了E2EGraph网络拓扑结构，用King[13]数据

集作为输入，仿真的节点个数为1 024个。

为了体现数据丰度，分别选取了10%、50%和90%分位的值画图，并以虚线表示“平均值”，实线表示50%分位数值，如图6～8所示。

图6表明，RCDHT相比Kademlia的平均查询耗时有所下降，特别是10%分位值比Kademlia有明显降低，这说明了在子网中的中心化索引算法对加快查询速度是有效的，而90%分位线与Kademlia持平，说明“分层”的设计思路给系统带来一些性能损失。

图7表明，子网中心化索引的“一跳式”解决方案使得RCDHT相比Kademlia的平均跳数有所下降。

图8表明，由于RCDHT在分配nodeID时考虑了位置信息，使得物理网络延时与逻辑网络延时的比值进一步降低。

图9标明，RCDHT对Kademlia的改进，并没有影响查询的成功率。

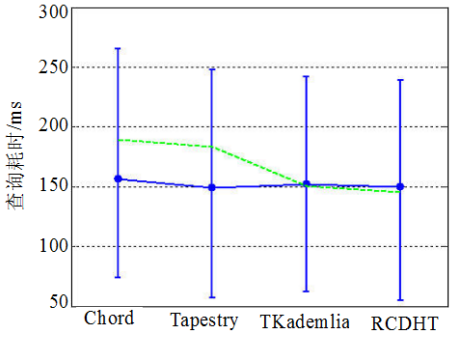

图6 查询耗时对比图

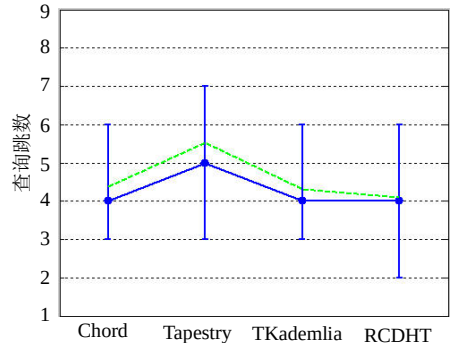

图7 跳数对比图

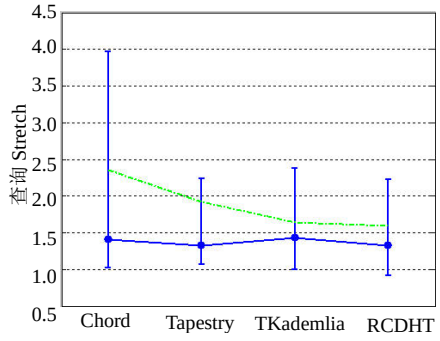

图8 延时对比图

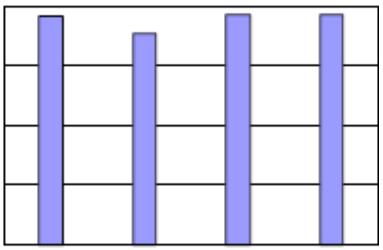

图9 查询成功率对比

4 总 结

本文总结了结构化P2P覆盖网络技术的特点和存在的缺陷，通过对主要的P2P协议的研究，提出了一种分层的、分布式覆盖网络查询协议。文中给出了该协议的覆盖空间、路由表、路由算法和节点加入与退出的详细描述，最后在P2PSim仿真平台上对其进行了仿真，并与Kademlia、Chord、Tapestry进行了对比。仿真数据表明，该协议在查询耗时、查询跳数以及查询成功率等性能上都较Kademlia、Chord、Tapestry有一定的提升。

参 考 文 献

[1] MAYMOUNKOV P, MAZIERES D. Kademlia: a peer-to-peer information system based on the XOR metric[J]. Peer-to-Peer Systems, 2002, 2429: 53-65.

[2] SAROIU S, GUMMADI P K, GRIBBLE S D. A measurement study of peer-to-peer file sharing systems[C]// The International Society for Optical Engineering. San Jose, CA, United States: [s.n.], 2002, 156-170.

[3] POUWELSE J A, GARBACKI P, EPEMA D H J, et al. A measurement study of the bit torrent Peer-to-Peer file-sharing system[C]//Proceedings of the Multimedia Computing and Networking (MMCN). San Jose, California, USA: SPIE, 2002: 281-297.

[4] 吴增德. 异构环境下结构化对等网络路由算法的研究[D]. 上海: 上海交通大学, 2003.
WU Zeng-de. Study on the structured peer-to-peer routing algorithm for heterogeneous environment[D]. Shanghai: Shanghai Jiaotong University, 2003.

[5] STOICA I, MORRIS R, LIBEN-NOWELL D, et al. Chord: a scalable peer-to-peer lookup protocol for Internet applications[J]. IEEE/ACM Transactions on Networking, 2003, 11(1): 17-32.

[6] 王 丹. P2P系统资源查询机制研究综述[J]. 计算机科学, 2004, 31(9): 57-59.
WANG Dan. The research overview of resource search mechanisms in P2P system[J], computer science, 2004, 31(9): 57-59

[7] 张 骞, 张 霞, 刘积仁, 等. 混合P2P环境下有效的查询扩展及其搜索算法[J]. 软件学报, 2006, 17(4): 782-793.
ZHANG Qian, ZHANG Xia, LIU Ji-ren, et al. Query expansion and its search algorithm in hybrid peer-to-peer networks[J]. Journal of Software, 2006, 17(4): 782-793.

(下转第121页)