ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.11, November 2007, pp.2851−2862
DOI: 10.1360/jos182851

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种支持多维资源描述的高效P2P路由算法*

宋 伟+, 李瑞轩, 卢正鼎, 於光灿

(华中科技大学 计算机科学与技术学院,湖北 武汉 430074)

An Efficient P2P Routing Algorithm Supporting Multi-Dimensional Resource Description

SONG Wei+, LI Rui-Xuan, LU Zheng-Ding, YU Guang-Can

(College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

+ Corresponding author: Phn: +86-27-67111986, Fax: +86-27-87544285, E-mail: rxli@hust.edu.cn, http://idc.hust.edu.cn

Song W, Li RX, Lu ZD, Yu GC. An efficient P2P routing algorithm supporting multi-dimensional resource description. *Journal of Software*, 2007,18(11):2851−2862. http://www.jos.org.cn/1000-9825/18/2851.htm

Abstract: Analyzing the existing P2P (peer to peer) routing algorithms, Flabellate Addressable Network (FAN) routing algorithm, an efficient second-moment-based resource routing algorithm supporting multi-dimensional resource description is proposed. Peers are mapped into a multi-dimensional Cartesian space with FAN routing algorithm that manages the subspaces and searches resources based on the peers' second-moment. The routing efficiency of FAN algorithm is up to $O(\log(N/k))$. When a peer joins and leaves the FAN network, the cost for updating routing messages is $O(k\log(N/k))$. The experimental results show that FAN routing algorithm has advantages of high efficiency of routing and low cost of network maintenance, and is an efficient structured P2P resource routing algorithm supporting multi-dimensional resource description. Some improved routing algorithms based on CAN (content-addressable network) can also be implemented in FAN network, and they can obtain better routing efficiency and lower maintenance cost.

Key words: P2P (peer to peer); FAN (flabellate addressable network) routing algorithm; second-moment locating; resource search; multi-dimensional resource description

摘 要: 在分析现有 P2P(peer to peer)路由算法的基础上,提出了一种基于二阶矩定位、支持多维资源数据描述的高效资源路由算法——FAN(flabellate addressable network)路由算法.FAN 算法将节点映射到统一的多维笛卡尔空间,并以节点相对空间原点的二阶矩作为子空间管理和资源搜索的依据.FAN 路由算法具有 $O(\log(N/k))$的高路由效率,在节点加入和退出 FAN 网络时,更新路由信息的代价为 $O(k\log(N/k))$.实验结果表明,FAN 路由算法具有路由效率高、维护代价小的优点,是一种 P2P 环境中支持多维资源数据描述的高效结构化资源路由算法.而且,目前部分基于 CAN(content-addressable network)网络的改进算法也可以在 FAN 网络中适用,并获得更好的路由效率和更低的维护代价.

关键词: P2P(peer to peer);FAN(flabellate addressable network)路由算法;二阶矩定位;资源搜索;多维资源描述

* Supported by the National Natural Science Foundation of China under Grant Nos.60403027, 60773191 (国家自然科学基金); the Natural Science Foundation of Hubei Province of China under Grant No.2005ABA258 (湖北省自然科学基金); the Open Foundation of State Key Laboratory of Software Engineering of China under Grant No.SKLSE05-07 (软件工程国家重点实验室开放基金)

Received 2006-02-13; Accepted 2006-11-03

中图法分类号: TP393 文献标识码: A

Peer-to-Peer(简称 P2P)系统中每个节点既是客户机,又是服务器,资源没有存储在集中服务器上,而是存储在各自分散的节点上.因此,在 P2P 环境中如何高效地进行资源搜索,是 P2P 研究中的一个关键问题.随着 P2P 应用的大量普及,基于资源单维数据(关键字)描述搜索资源已经不能满足应用的需要.越来越多的网络应用要求 P2P 系统采用多维数据描述资源并基于资源的多维描述进行资源搜索.例如,在科技文献共享应用中,文献资源采用标题、作者、摘要、关键词、参考文献等多维数据描述,在这种 P2P 应用中需要提供一种高效的、可扩展的、易维护的、支持多维数据描述的 P2P 资源路由算法.

设计一种高效、易维护的P2P环境中支持资源多维数据描述的路由算法是我们希望解决的问题.与单一关键字描述共享资源相比,采用多维数据对资源进行描述,可以使资源访问者按照自己的要求,基于多维数据发出查询请求,实现更加准确和高效的资源搜索.在已有的P2P路由算法中,CAN(content-addressable network)[1]是最成熟并被研究最多的一种支持多维数据描述的结构化P2P路由算法.CAN针对资源采用多维数据描述的特点,提出了一种多维笛卡尔空间的划分策略,实现了P2P环境中的支持多维数据描述的资源路由.但是,CAN存在网络结构维护代价大、路由效率相对较低的缺陷.

本文提出一种P2P环境中支持多维数据描述的资源路由算法——FAN(flabellate addressable network)算法,将资源映射到d维笛卡尔坐标映射空间中,以空间节点坐标相对原点的二阶矩进行子空间的划分和资源路由.与同样支持多维数据描述的CAN[1]相比,FAN算法具有路由效率高、维护代价小的优点.同时,FAN算法还具有很好的可扩展性和可移植性,目前,结构化P2P环境中的资源路由算法都可以方便地修改成支持FAN的资源路由算法.

1 相关工作

如何高效、可靠地搜索共享资源,一直是 P2P 领域中的研究重点.目前,国内外在这方面的研究成果也很多.从拓扑结构来看,P2P 网络可以分成结构化和非结构化两大类,不同网络结构在路由算法方面存在很大差异;从路由算法方面来看,又可以根据资源描述方式的不同分为支持单维描述和多维描述的路由算法,各种算法在维护代价和路由效率上也有很大区别.目前,国内外主要的 P2P 路由算法包括以下几种类型:

第 1 类是非结构化P2P网络中的路由算法,其中最具代表性的就是Napster和Gnutella.Napster[2]是最早的P2P应用程序之一,Napster利用一个类似于服务器的节点,集中提供节点标识和资源索引信息.Napster网络并不提供节点的逻辑命名机制,仍然以IP地址和端口作为节点标识,完全由服务器来完成资源注册和搜索功能,因此,Napster对中心服务器有很大的依赖性.Napster的资源搜索仍然属于一种集中式的资源搜索策略.Gnutella[3]与Napster不同,它没有集中式服务器,是一种分布式的P2P网络.Gnutella利用洪泛策略进行路由选择[4].在Gnutella系统中,对等点只能确定与之直接通信的对等点所在位置,它所进行的路由选择具有很大的随机性,路由效率不稳定.

第 2 类是结构化 P2P 网络中的路由算法.在结构化 P2P 网络中,各个节点有其自身的逻辑标识,映射到特定的逻辑空间中来进行对等网络的管理以及资源的搜索.根据资源描述的维度(映射逻辑空间的维度),可以将目前结构化 P2P 网络中采用的路由算法分为支持资源单维数据描述和多维数据描述的路由算法两类.

目前,支持单维数据描述的结构化P2P路由算法有很多,Chord[5]和Tapestry[6]是其中典型的两种路由算法.Chord是MIT提出的一种基于DHT(distributed hash table)的资源路由算法.在Chord网络中,通过一致Hash函数[7]散列节点的IP地址,为每个节点分配m位的标识符.Chord按照节点标识的顺序形成一个逻辑环形拓扑结构.Chord资源路由算法具有$O(\log n)$的路由效率.Tapestry网络的路由算法类似于无类域间路由(classless inter-domain routing,简称CIDR)的最大前缀匹配,在Tapestry中,按照层次组织每张邻近映射表,节点通过邻近映射表保存节点之间的邻近关系,利用本地路由映射表,节点把消息按照目的地址逐位向前传递.Tapestry路由算法同样具有$O(\log n)$的路由效率.

CAN[1]是最具代表性的支持多维数据描述的结构化P2P路由算法.CAN中每个节点资源映射成d维笛卡尔空间中的一个坐标点,并管理包括自身坐标的一个虚拟坐标区域.每个节点维护一个坐标路由表,路由表保存相邻节点的坐标区域信息.通过邻近区域路由表以及贪婪转发机制,CAN网络实现消息转发直至到达目标节点.CAN路由算法具有$O(dn^{1/d})$的路由效率.CAN网络在节点频繁加入、退出,邻近节点同时失效等情况下的维护代价较大,而且CAN网络存在一个节点管理多个临时区域的可能.

目前,越来越多的P2P应用需要提供支持多维数据描述的资源搜索机制.国内外在P2P环境中基于资源多维描述的路由方面进行了一些研究及相关改进,并取得了一些研究成果.例如:文献[8]提出了基于直方图的分层Top-k查询算法;文献[9]改进了R*-tree[10],提出一种P2P环境中支持多维距离查询的路由算法;文献[11]提出了支持多维属性距离查询的Mercury;文献[12]基于skip graphs[13]设计了一种支持多维数据查询的ZNet算法;文献[14]基于FISSIONE[15]设计了一种支持多维数据查询的Armada路由算法;还有其他一些P2P环境中支持多维数据查询的路由算法[16−18].目前,支持多维查询的P2P路由算法还存在着一定的局限和不足.例如:Top-k查询算法[8]适用于非结构化的纯P2P网络;改进自R*-tree的路由算法[9]适合进行有关距离等空间信息的查询;pSearch对搜索资源需要全局统计信息的支持;Mercury对每种属性设置Hub进行处理,不能支持资源描述具有太多的属性;Armada利用树结构实现基于多维数据描述的资源查询,但是树结构本身也需要较大的维护开销;ZNet算法利用Z曲线对底层DHT空间进行划分,在节点失效情况下,维护代价相对较大;类似于文献[16,17]中采用的语义聚类方式进行资源路由,需要对资源进行语义提取和聚类处理.目前,国内外许多关于P2P环境中的多维查询研究[9,19,20]都是基于CAN展开的,但是,这些改进算法并不能解决CAN网络本身路由效率不高和维护代价相对较大的问题,因此,有必要提供一种支持资源多维数据描述的高效通用路由算法.本文提出了一种新颖的支持资源多维数据描述的路由算法,在这种路由算法中,节点映射到统一的d维笛卡尔空间,以节点相对坐标原点的二阶矩作为子空间划分和路由的依据.由于二维情况下子空间的形状是一个个的扇形,因此称这种路由算法为扇形路由算法,简称FAN路由算法,称支持FAN路由算法的P2P网络为FAN网络.部分支持多维查询的CAN改进算法(例如文献[9,20])在FAN网络中也是适用的,而且由于FAN本身的优势,这些改进的路由算法在FAN网络中也会有更好的表现.

2 支持多维数据描述的资源路由算法

2.1 FAN网络中的概念

为了叙述方便,首先给出 FAN 网络中涉及的基本概念.为了进行基于多维数据描述的资源路由算法研究,把节点映射到一个多维笛卡尔空间,并在多维映射空间内进行资源路由,首先给出映射空间的定义.

定义 1. 若 A 是一个满足 FAN 网络的 m 维笛卡尔坐标映射空间,则 A 须满足如下条件:

(1) $\exists l>0$,使得$\forall P\in A, P=(x_1,x_2,\ldots,x_m)$,有 $0\le x_i\le l$;

(2) $\forall P=\{(x_1,x_2,x_3,\ldots,x_m)|0\le x_i\le l\}$,有$P\in A$.

FAN网络采用满足定义 1 的m维笛卡尔空间作为映射空间.FAN网络中任一节点采用m维数据描述,利用统一Hash函数映射到映射空间中,得到唯一的坐标点$P(x_1,x_2,x_3,\ldots,x_m)$,采用节点相对原点的二阶矩作为资源路由的基础.下面给出节点P的二阶矩定义.

定义 2. m维映射空间中节点$P(x_1,x_2,x_3,\ldots,x_m)$的二阶矩为

$$M_P=x_1^2+x_2^2+\ldots+x_m^2=\sum_{i=1}^{m}x_i^2.$$

在 FAN 路由算法中,需要将映射空间划分为坐标子空间来进行管理和路由.下面给出坐标子空间的定义.

定义 3. 若映射空间A中的一个子集$B(r_1,r_2)$满足:

(1) $\forall P\in B(r_1,r_2)$,有$r_1<M_P\le r_2$

(2) $\forall P$若满足$r_1<M_P\le r_2$,有$P\in B(r_1,r_2)$

则称$B(r_1,r_2)$为映射空间A的一个子空间.设$B_1(r_1,r_2)$和$B_2(r_3,r_4)$是映射空间A的两个子空间,若满足条件$r_1=r_4$或者

$r_2=r_3$,则称子空间B_1和B_2是邻接子空间.整个映射空间A可以被划分为若干个邻接子空间.

定义 4. 设节点P的二阶矩为M_P,$B(r_1,r_2)$是FAN网络映射空间中的一个子空间,则P到子空间B的距离$d=\min(|x-M_P|, r_1<x\le r_2)$.由点到子空间距离的定义可知,子空间内的点到子空间的距离为 0.

2.2 FAN路由算法描述

下面详细介绍 FAN 路由协议及 FAN 网络中节点加入、退出的处理.为了讨论方便,假设 FAN 路由算法是基于 m 维的多维数据描述,也即映射后的映射空间是 m 维的.

2.2.1 FAN 算法中的资源路由

在FAN网络中,任意节点$P(x_1,x_2,x_3,\ldots,x_m)$被唯一地映射到某个子空间$B(r_1,r_2)$(FAN网络中的节点加入、退出策略保证了子空间不会重叠,且覆盖整个m维空间),每个子空间最多包含k个节点.FAN网络中每个节点维护一张路由表,路由表信息包括节点坐标、节点IP地址、节点所属子空间以及路由信息的获得时间.子空间用二阶矩上下界描述,节点定期探测路由表中节点的可达性进行路由表的更新,节点存储所有相同子空间内节点信息.因此,在FAN算法中,查询某一个节点等价于查找目标节点所属的子空间.节点还会存储邻接子空间的节点信息.FAN网络中任一节点P查询某一目标节点Q的路由算法见算法 1.

算法 1. FAN 网络中节点 P 查询目标节点 Q 的路由算法.

输入:发起路由请求的源节点 P,路由目标节点 Q.

输出:若目标节点存在,则返回 Q;否则,返回包含 Q 的子空间内的某个节点(为节点加入算法提供服务).

routePeer(P,Q)

1. $Q(x_1,x_2,x_3,\ldots,x_m)=Hash(Q)$; $M_P=x_1^2+x_2^2+\ldots+x_m^2=\sum_{i=1}^{m}x_i^2$ //计算得到Q的二阶矩M_q
2. min*Distance*=*get*Min*Distance*(P,M_q) //取P所属子空间与目标节点Q的最小距离min*Distance*
3. If (min*Distance*==0) then
4. return *getPeer*(Q) //P,Q 在相同的子空间中,在 P 的路由表中返回节点 Q
5. else
6. P'=*get*Min*DistancePeer*(P,M_q) //取P路由表中与Q距离最小的子空间对应的节点P'
7. *routePeer*(P',Q) //递归调用 *routePeer*(P',Q)

算法 1 保证每次查询都在二阶矩方向上向目标节点逼近,最终到达覆盖目标节点的子空间.假设 FAN 网络包含 N 个节点,每个子空间内最多包含 k 个节点,则子空间的数目为 $O(N/k)$(第 3.3 节的实验将给出相关验证).设 FAN 网络中共有 M 个子空间,源节点位于从坐标原点出发第 l 层子空间,目标节点出现在 M 个子空间中的概率相等,则 FAN 网络的平均路由跳数如式(1)所示.而 $M\sim O(N/k)$,因此,FAN 网络的路由效率也为 $O(N/k)$.

$$Hops=\frac{1}{M}\left(\sum_{i=1}^{l}(l-i)+\sum_{i=l+1}^{M}(i-l)\right)=\frac{1}{M}\left(\frac{l(l-1)}{2}+\frac{(1+M-l)(M-l)}{2}\right)$$
$$=\frac{M^2-2Ml+2l^2+M-2l}{2M}=\frac{M}{2}-l+\frac{l^2}{M}+\frac{1}{2}-\frac{l}{M} \quad (1)$$

如果仅通过两个邻接子空间来传递路由消息,随着FAN网络中节点的增加,子空间数目将有所增加,路由效率将会大为降低.因此,从路由效率考虑,需要实现二阶矩方向上的跳跃查询,提高查询效率.下面提出扩展邻接子空间的概念.图 1 给出了二维情况下FAN网络中的子空间及路由示意图.任意子空间中不仅存储邻接子空间信息,而且会存储在二阶矩两个方向上相邻 2^j层的子空间信息,称这些子空间为扩展邻接子空间.例如,子空间B_1,B_2, B_4,B_5都是B_3的扩展邻接子空间.

FAN 网络中,节点路由表存储所有扩展邻接子空间的节点信息.路由表信息具有以下两个性质:

性质 1. 对称性.若子空间B_1的路由表中包括子空间B_2的信息,则在子空间B_2的路由表中必然包括子空间B_1的信息.

证明略.

性质 2. 传递性.若在子空间B_1的路由表中有子空间B_2的信息,B_2为B_1的 $2^j(j>1)$层扩展邻接子空间,则在子空间B_1的路由表中必然存在子空间B,B为B_1的 $2^{j'}(j'<j)$层扩展邻接子空间,B_2也是B的扩展邻接子空间,称这种性质为路由表的传递性.

证明:FAN网络中子空间路由表传递性的证明等价于对于任意的$j>1$,必然存在分解 $2^j=2^{j'}+2^{j''}$.j为奇数时,存在j'使得$2^j=2^{2j'+1}=2^{2j'}+2^{2j'}$;$j$为偶数时,存在$j'$使得 $2^j=2^{2j'}=2^{2j'-1}+2^{2j'-1}$.综上,性质 2 得证. □

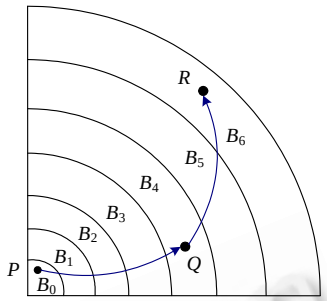


Fig.1 Subspaces and routing in a 2-dimensional FAN network

图 1 二维情况下 FAN 网络子空间及路由

由于FAN网络子空间路由表信息具有上述的对称性和传递性,因此可以保证子空间在只知道两侧邻接子空间的情况下,通过反复调用路由探测算法获得整个网络的扩展邻接子空间信息,分析引入扩展邻接子空间后的改进FAN路由算法的路由效率和维护代价.同样,设定FAN网络中节点总数为N,每个子空间最多容纳k个节点,子空间总数为M,$M\sim O(N/k)$.采用扩展邻接子空间后,FAN网络中的节点路由表中最多包含 $2\log_2(M/2)$个扩展邻接子空间信息,最少包含$\log_2(M)$个扩展邻接子空间信息.因此,FAN网络子空间的扩展邻接子空间的数目为$O(\log(N/k))$.分析FAN网络的路由效率,假设每个子空间在二阶矩方向上是等大小的.FAN网络中每次路由消息都是在扩展邻接子空间中进行传递,假设源节点和目标节点相隔 T 个子空间,则整个路由过程可以表示为 $T=2^{j_0}\pm 2^{j_1}\pm ...\pm 2^{j_m}$ 的代数分解,其中分解的第 s 项表示路由的第 s 次消息传递是向相隔 2^{j_s} 扩展邻接子空间进行的,正号表示向二阶矩增大方向进行路由,负号表示向二阶矩减小方向进行路由.下面证明 FAN 路由效率为 $O(\log(N/k))$.

首先证明上述路由的代数分解中各项绝对值是严格单调递减的.考察其中任意两次相邻路由 2^{j_s} 和 $2^{j_{s+1}}$,每次路由选择都选择路由表中的最近扩展邻接子空间,所以当第 s 次路由选择 2^{j_s} 时:

1) 若第 s 次路由的起点子空间包含与 2^{j_s} 同向的扩展邻接子空间 $2^{j_{s+1}}$ (没有超出映射空间范围),那么目标节点与第 s 次路由的起点子空间相隔子空间的绝对值小于 $2^{j_s}+2^{j_{s-1}}$,否则,第 s 次路由将优先选择 $2^{j_{s+1}}$ 进行路由,因此第 s+1 次路由时,目标节点和路由起点相隔子空间绝对值小于 $2^{j_{s-1}}$,因此,第 s+1 次路由选择时相隔子空间数不会大于 2^{j_s};

2) 若第 s 次路由的起点子空间不包含与 2^{j_s} 同向的扩展邻接子空间 $2^{j_{s+1}}$ (超出映射空间范围),则第 s+1 次路由时,目标节点与路由起点相隔子空间的绝对值小于 $2^{j_{s+1}}-2^{j_s}=2^{j_s}$,因此第 s+1 次路由选择时相隔子空间数也小于 2^{j_s},否则超出 FAN 网络映射空间范围.

综上可知,路由的代数分解 $T=2^{j_0}\pm 2^{j_1}\pm ...\pm 2^{j_m}$ 中,各指数项严格单调递减.因此代数分解中的每一项大于等于其后各项之和,每次路由传递都使得路由源节点和目标节点之间的距离至少减半,任意两个子空间间隔 T 的节点之间路由,路由效率为 $O(\log(T))$.FAN 网络的子空间总数为 M,$M\sim O(N/k)$,因此,FAN 网络中路由效率为 $O(\log(N/k))$.采用扩展邻接子空间后,FAN 网络路由效率由 $O(N/k)$下降到 $O(\log(N/k))$,路由效率得到极大的提高.查询过程中也可以采用类似于 CAN 中的贪婪转发机制发出查询请求,以进一步提高查询效率.

2.2.2 FAN 算法中节点的加入

FAN网络初始化时,整个映射空间只有一个子空间$B(0,ml^2)$.当节点P加入FAN网络时,P首先和FAN网络中的某个节点Q相连,然后节点P通过节点Q路由到M_P所对应的子空间B,加入B中.设FAN网络中每个子空间最多容纳k个节点.下面给出节点P加入FAN网络的算法,并分析节点加入操作的代价,同时说明节点的加入操作不会破坏FAN网络子空间的独立性(m维空间中任意两个子空间区域不重叠)和覆盖性(m维空间中不存在某个坐标点不属于任一子空间,即有限个子空间覆盖整个m维映射空间).

算法 2. 节点 P 加入 FAN 网络的算法.

输入:待加入的节点 P,与 P 相连的 FAN 网络中的节点 Q.

输出:节点 P 加入后的 FAN 网络.

peerJoin(*P*,*Q*)

1. *Q*′=*routePeer*(*Q*,*P*)
2. if (*getAreaPeerCount*(*Q*′)<*k*) then
3. *joinArea*(*P*,*Q*′) //子空间节点数目小于 *k*,*P* 直接加入 *Q*′所在子空间
4. else
5. min*area*=*get*Min*PeerAdjacentArea*(*Q*′) //得到两个邻接子空间节点数目较小的邻接子空间
6. end if
7. if (min*area*'s *peer amount*<*k*) then
8. *joinArea*(*P*,*Q*′)
9. *adjustAreaPeer*(*P*,min*area*) //*P* 加入 *Q*′所在子空间,调整子空间,使得节点数目平均
10. else
11. *joinArea*(*P*,*Q*′)
12. *spaceSplit*(*P*) //*P* 加入 *Q*′所在子空间,分裂子空间使得新生成子空间节点数目平均
13. end if

在算法 2 中,待加入节点 *P* 首先路由到覆盖自身二阶矩的子空间 *B*,然后根据子空间 *B* 的节点数分两种情况进行处理:

(1) 如果 *B* 中节点数目小于 *k*,则节点 *P* 加入到 *B* 中,将节点信息传递给 *B* 中其他节点,同时获得其他节点信息,并从其他节点的路由表中获取扩展邻接子空间信息,*P* 根据路由表信息调用探测子空间算法确定正确的子空间路由信息,并且将自身信息通知路由表中相应扩展邻接子空间中的节点.

(2) 若 *B* 中节点数目等于 *k*,则在 FAN 算法中,路由效率与子空间的层数直接相关.因此,并不直接调用子空间分裂(spaceSplit)服务,而是根据相邻子空间中节点数目分两种情况进行处理:

① 如果 *B* 的两个相邻子空间中某一个子空间的节点数目小于 *k*,那么优先调用 adjustAreaPeer 服务,调整节点分配.设 *B* 的相邻子空间 *A* 的节点数目较少,包含 *j* 个节点,此时子空间 *B* 包括新节点 *P* 有 *k*+1 个节点.节点分配原则是两个相邻子空间的节点数目尽量相等,将子空间 *B* 中(*j*+*k*+1)/2 个与子空间 *A* 距离较近的节点移到 *A* 中,这些节点查询 *A* 内节点的路由表调整自己的路由表信息,并通知原路由表中节点自身子空间的变化,同时修改子空间 *A*,*B* 的相邻边界.在调整子空间的过程中,并不会改变子空间与其他子空间的独立性和覆盖性,调整后的子空间仍然相邻,而且总的覆盖区域没有发生变化,因此不会改变调整后两个相邻子空间之间的独立性和覆盖性关系.

② 如果两个相邻子空间的节点数都等于 *k*,这时就不能再向相邻子空间中增加节点了,子空间 *B* 调用 spaceSplit 服务,分裂成两个节点数目相等的邻接子空间,同时修改分裂后各个节点的路由表信息.分裂操作并不修改原子空间的上、下边界.例如,子空间 $A(r_1,r_2)$ 分裂成两个相邻子空间 $A_1(r_1,r)$ 和 $A_2(r,r_2)$,其中,$r_1<r<r_2$.由于 A_1,A_2 未改变子空间 *A* 的总覆盖区域,所以子空间分裂操作不会破坏子空间之间的独立性和覆盖性关系.

分析节点加入 FAN 网络空间的代价可以发现,在不需要修改子空间的情况下,只需为新节点建立路由表信息,同时更新子空间内其他节点的路由信息.在更新扩展邻接子空间节点路由表信息时,首先只在相隔较近的几个子空间(例如相隔 1,2,4)内更新节点信息,而不是立刻在全部的扩展邻接子空间中更新路由,这样做的目的是为了防止节点的反复加入、退出引起全网路由表的抖动,新加入节点的路由信息会在此后各个节点定期发起的更新路由表操作中扩展到全网.因此,这种情况下需要修改路由表数据量为扩展邻接子空间内的所有节点,为 $O(k\log(N/k))$.在需要修改子空间的情况下,调整后的两个子空间内的节点的路由表都要作相应的变化,这时的数据量为第 1 种情况的两倍,也是 $O(k\log(N/k))$.由此可见,在 FAN 网络中加入节点时,消息量为 $O(k\log(N/k))$,其中,*N* 为 FAN 网络节点总数,*k* 为子空间容纳的最大节点数.

2.2.3 FAN 算法中节点的退出

P2P环境中节点的退出是非常普遍的,同样,支持多维数据描述的CAN网络在节点退出时,有可能出现 1 个

节点管理多个临时区域的情况,以及在邻近区域节点同时离开时需要重构网络等问题[1].与CAN网络相比,FAN算法在节点退出的处理上相对简单,可以很好地解决上述问题.FAN网络中节点的退出分以下两种情况讨论:

(1) 退出的节点不是子空间的唯一节点.当节点发出退出请求,或者其他节点周期性地检查发现节点离线时,都会发起删除节点(RemoveNode)请求.这时,只需在子空间内的各个节点和子空间的扩展邻接子空间的节点处删除离开节点的路由信息即可.

(2) 退出节点是子空间 A 内的唯一节点.这时,A 的两个邻接子空间中节点较少的子空间 B 修改上、下界,接管 A.B 向所有自己路由表中的节点发起修改区间请求.此外,还需要将原来所有指向 A 的路由重新定位.A 和 B 是邻接子空间,那么,A 的所有扩展邻接子空间也必然是 B 的扩展邻接子空间的邻接子空间.因此,在 B 通知所有扩展邻接子空间修改空间范围的同时,B 的所有邻接子空间向自己的一个邻接子空间发出修改路由请求.由于新生成的子空间覆盖区域与节点退出前的两个子空间的覆盖区域完全相等,因此并不改变新生成的子空间与其他子空间的独立性和覆盖性关系.因此,子空间合并操作不会改变子空间的独立性和覆盖性关系.

第1种情况下需要修改离开节点所有扩展邻接子空间中节点的路由消息,第2种情况下需要修改路由信息的节点数是第1种情况的2倍(需要修改被删除子空间和将要覆盖子空间的扩展邻接子空间的路由信息).两种情况下节点退出的消息数量都为 $O(k\log(N/k))$.与CAN算法相比,FAN算法在节点离开的处理上代价较低,而且不会出现节点管理多个临时区域的情况.

3 仿真实验及结果分析

下面通过仿真实验的方法对FAN网络的路由效率和构造复杂度等方面进行验证.在现有的资源路由协议中,CAN是典型的支持多维数据描述的路由算法,所以在路由效率的实验中,主要对FAN和CAN算法进行对比.

3.1 实验设置

所有实验在一台PC机上完成,PC机的配置为CPU P4 2.0GHz,内存512MB,操作系统是Windows 2003.模拟程序采用Java编写.模拟程序模拟了FAN网络中所有的节点加入、退出、路由、子空间的合并、调整、分裂等操作.实验初期随机地进行节点加入、退出操作构建子空间分布,在达到预定节点数目的情况下,采集相应的实验数据.若实验的节点数为N,则实验初期节点加入操作为 $10N$次,节点的退出操作数为 $9N$次,每个节点发生加入、退出操作的次数服从均匀分布.表1是实验的基本参数设置.在实验环节中,FAN网络的实验结果由模拟实验得到,CAN网络的相关实验结果由CAN原文[1]得到.

Table 1 Parameter and settings in the simulations

表1 模拟实验参数设置

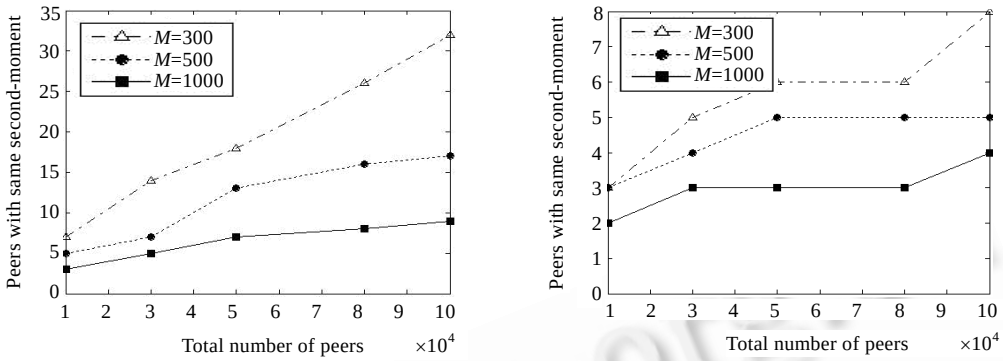
	Parameter	Value
N	Network size of FAN	5 000, 10K, 30K, 50K, 80K, 100K
D	Dimension of FAN mapping space	2,3,4,5
K	Maximal peers of subspace can contain	1,4,7,10

3.2 节点重复二阶矩实验

在FAN网络中,资源路由基于节点二阶矩进行,每个子空间内限制了节点的最大数目.但是,节点的 m 维坐标和节点的二阶矩并不是一一对应的.因此在FAN算法中,二阶矩相等的节点数目必须小于每个子空间容纳的最大节点数,否则,FAN算法将不能对子空间进行正确的管理.

设计仿真实验研究二阶矩的重复情况,以确定Hash函数映射样本空间、维度、节点总数以及最大二阶矩重复节点数目之间的关系.仿真实验中采用随机Hash函数分配节点的每个维度上的分量为[0,M]范围内的整数,实验节点数目为 10 000~100 000,针对不同的维度考察发生重复二阶矩的节点最大数目,得到实验结果如图2所示.从实验数据中可以发现,在低维(d=2)情况下,随机样本空间 M 越小,二阶矩重复节点随节点总数的增加变化越明显;而二阶矩重复节点数目越多,需要每个子空间容纳的最大节点数目(k)也越多,每个节点处的路由表中的表项也会越多.因此在低维情况下,FAN网络在每个维度上Hash函数样本空间的选择要适当.在高维

(*d*=5)情况下,基于各种 Hash 函数样本空间数目实验情况均显示出,随着节点总数的增加,二阶矩重复数目增长速度明显下降.



(a) *d*=2 (b) *d*=5

Fig.2 Max total number of peers with the same second-moment experiments

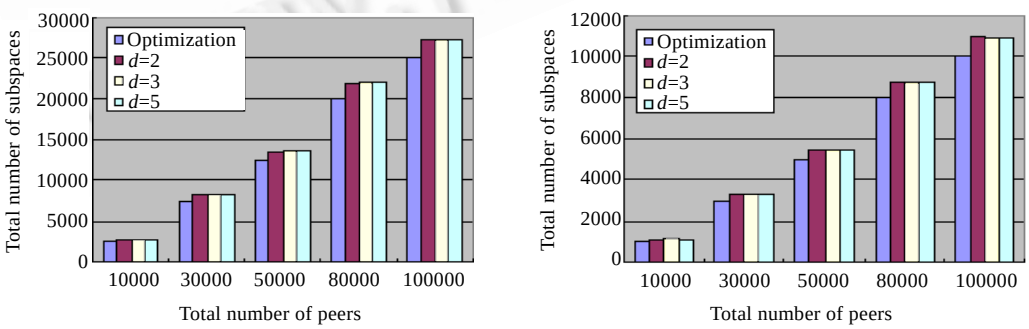
图 2 二阶矩重复最大节点总数实验

从实验中可以得出结论,FAN 算法在 Hash 函数样本空间要求和节点总数的限制上对资源高维数据描述有着更好的适应性.在低维数据描述(*d*=2)情况下,必须选择较大的 Hash 函数样本空间(不小于 500)以及较小的总节点数(不大于 50 000),以保证最大重复二阶矩节点数小于子空间容纳节点数(在实际系统中,*k* 取值一般小于 20);而在高维数据描述情况下,在最大二阶矩重复节点方面对于 Hash 函数样本空间的要求较小,并且可以适应大量节点的加入.

3.3 FAN网络子空间构造实验

前面已分析过,FAN 网络的节点路由效率为 $O(\log(N/k))$,节点加入、退出 FAN 空间代价为 $O(k\log(N/k))$,因此,FAN 网络的路由效率与子空间数目 N/k 相关,FAN 算法中节点加入和退出策略是否可以尽量减小网络中子空间数目,并保证子空间数目与 N/k 大致相等,对于 FAN 网络是非常关键的.设计仿真实验,测试子空间构造算法的工作效率.每个子空间最多容纳节点数目为 *k*,实验节点数目为 10 000~100 000,考察不同 *k* 值以及不同维度 *d* 情况下经过加入、退出稳定后 FAN 网络中子空间总数.

实验结果如图 3 所示.从实验数据中可发现,在不同 *k* 值和不同维度情况下,子空间数目基本上与最优值 N/k 相近.这说明 FAN 算法对节点加入、退出的高效管理,保证 FAN 网络可以按照 $O(\log(N/k))$的路由效率高效搜索资源;同时说明 FAN 的路由效率只与 FAN 网络的节点总数 *N* 和子空间容纳节点数 *k* 相关,而与资源描述维度无关,是一种稳定的路由算法.



(a) *k*=4 (b) *k*=10

Fig.3 Subspace constructing experiment

图 3 子空间构造实验

3.4 查询路由效率实验

路由效率是路由算法优劣的一项重要评判指标,CAN 是成熟的支持多维数据描述的 P2P 路由算法.设计仿真实验,对比 FAN 算法和 CAN 算法在相同网络条件下的平均路由效率.文献[1]中设计的 CAN 网络每个区域容纳的节点数目是 1,在文献[1]的改进算法中提到了在每个区间内容纳多个节点改进路由效率.但是,CAN 采用贪婪算法进行资源路由,若 CAN 网络的空间维度是 d 维,每个空间容纳的节点数目为 k,则每次路由会向 $O(d)$个相邻区域发送路由消息,而消息量为 $O(kd)$,因此在 CAN 中,k 不能设置得太大(文献[1]中给出的 MAXPEER 为 3 或 4),否则,网络中的消息量会大量增加.而在 FAN 网络中,每次路由是向特定的一个区域发送路由消息,消息量为 $O(k)$,所以,在 FAN 网络中可以设置比 CAN 网络更大的 k 值以获得更高的路由效率.

为了在同等条件下进行路由效率比较,设置 FAN 网络和 CAN 网络在每个子空间容纳的最大节点数相同的情况下进行实验.设置实验节点数目为 256~64K.FAN 和 CAN 网络的子空间最大节点数 k 为 1~4.由于 FAN 网络的路由效率为 $O(\log(N/k))$,与空间维度 d 没有关系,因此,FAN 网络采用 d=3 进行实验,在 FAN 网络中随机选择节点平均进行 $10N$ 次路由,统计平均路由跳数.CAN 网络实验数据从文献[1]中得到,实验结果如图 4 所示.

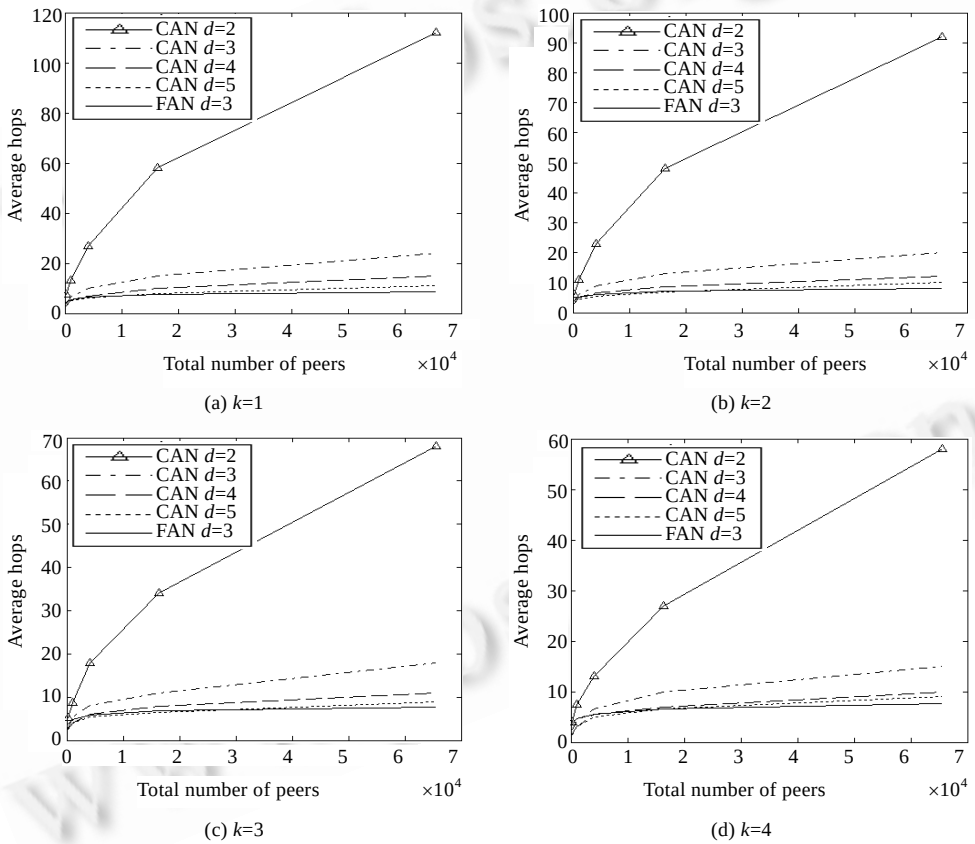


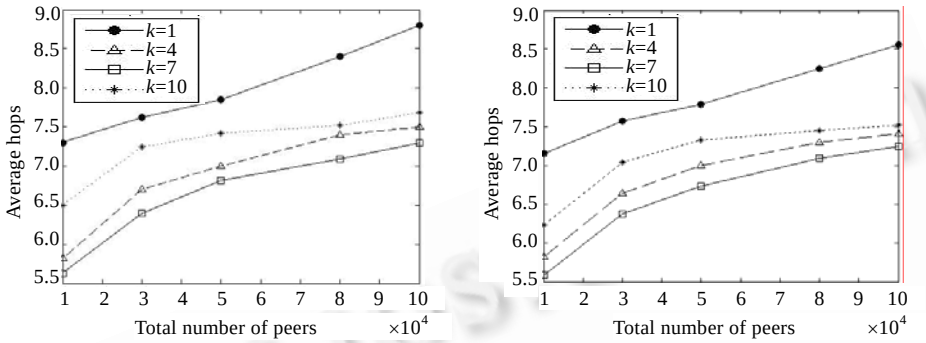
Fig.4 Experiments of FAN routing efficiency

图 4 FAN 路由效率实验

从实验数据中可以得出结论,在 k 值相同的情况下,FAN 网络可以获得更高的路由效率,且在低维情况下(d=2,3)优势更加明显.在加大节点数目的情况下,FAN 网络也可以获得稳定的路由效率.考虑到 FAN 网络中可以将 k 值设置得更大以获得更高的路由效率,因此可以得出结论,相对于同样支持多维数据描述的 CAN 路由算法,FAN 路由算法具有更高的路由效率.

3.5 k值对FAN路由效率的影响

设计仿真实验,考察子空间最大容纳节点数目对FAN网络路由效率的影响.仿真实验条件如下:实验节点数目为10 000~100 000,FAN网络映射空间维数为3,5.分别针对不同k值在FAN网络中随机选择节点进行路由查询,平均每个节点随机发起10次查询请求,统计平均路由跳数,得到实验结果如图5所示.从实验结果可以得出结论,FAN网络中路由效率随着k值的增大而加强,尤其是在节点数目较小的情况下,其改善更为明显.



(a) d=3 (b) d=5

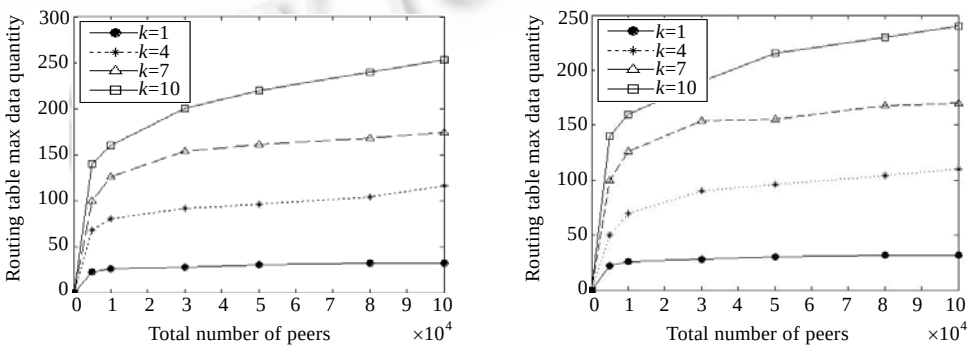
Fig.5 Experiments of different k value's effects on the routing efficiency of FAN network

图5 k值对FAN网络路由效率影响实验

第3.4节的实验证明相同k值情况下,FAN网络比CAN网络拥有更好的路由效率.由于随着k值的增加,CAN网络中消息量会大幅度地增加,因此,CAN网络中推荐的k值较小.但在FAN网络中,k值增大不会引起路由消息量的急剧增加,因此可以将k值设置得大一些,以便获得更好的路由效率,仿真实验中设置k=10.从实验数据中可以发现,FAN网络随着k值的增大获得更高的路由效率.综合第3.4节和第3.5节的实验可以得出结论,FAN是比CAN拥有更高路由效率的一种支持多维数据描述的P2P路由算法.

3.6 节点路由表数据量实验

从上述实验中可以发现,FAN具有比CAN更高的路由效率.理论上,FAN的路由效率是$O(\log(N/k))$,CAN的路由效率是$O(dn^{1/d})$.FAN能够获得更好的路由效率,除了网络构造上的原因以外,也是以每个节点的路由表信息量比CAN更大而获得的.在CAN网络中,每个节点存储$O(2d)$个邻居节点信息,而在FAN网络中,每个节点存储$O(k\log(N/k))$个扩展邻接子空间的节点信息.因此在FAN网络中,每个节点的路由表信息量比CAN更大.因此,需要考虑每个节点的路由表数据量是否在可以接受的范围内.在仿真实验中,节点数目为5 000~100 000,在d= 3,5时,考察不同k值情况下FAN网络中节点路由表最大数据量,实验结果如图6所示.



(a) d=3 (b) d=5

Fig.6 Experiments of data quantity in peer's routing table

图6 节点路由表数据量实验

结合第 3.5 节的实验结果可以得出结论,随着 k 值的增大,路由效率将会得到改善,同时,节点路由表的数据量也将会增大.为了获得较高的路由效率同时尽量减小节点路由表数据量,可以采用一些改进策略.例如,FAN网络中的节点随机选择存储邻接子空间中的部分节点信息,这样,采用贪婪转发的策略也可以保证覆盖所有扩展邻接子空间中的节点,在不影响路由效率的情况下,大幅度地降低路由表数据量.

4 结 论

本文提出了一种全新的支持多维数据描述的资源路由算法——FAN 算法.在这种路由算法中,节点映射到统一的 m 维笛卡尔空间,以节点相对原点的二阶矩作为子空间划分和路由的依据.本文详细介绍了 FAN 算法的资源路由协议,从理论上讨论了该路由协议的路由效率和节点的加入、退出策略,并通过仿真实验与 CAN 算法进行效率比较,给出了 FAN 算法中各个参数对路由效率和节点路由表数据量的影响.实验结果表明,FAN 算法是一种 P2P 环境中支持多维数据描述的高效、可行的路由算法.

FAN路由算法可以很好地为P2P环境中支持多维数据描述的资源搜索应用提供服务.例如,基于内容的资源共享应用、基于特征量的资源查找应用等.目前,部分基于CAN的改进多维查询路由算法[9,20]也可以在FAN网络中得到应用.在后续工作中,将针对具体P2P应用,利用FAN网络提供高效的路由服务,从而提高P2P应用的服务质量,目前正在进行相关方面的研究工作.FAN路由算法在维度变化的情况下难以适应,在后续的研究工作中将对此做进一步的研究.此外,在目前的实验过程中发现,FAN网络中相同子空间节点路由表以及扩展邻接子空间节点路由表中冗余数据量较大,这在一定程度上加大了各个节点的处理负担.在后续工作中希望可以改进路由表存储策略,以期在保持FAN网络路由效率基本不变的条件下,降低FAN网络中节点路由表的冗余数 据量.

References:

[1] Sylvia R, Paul F, Mark H, Richard K, Scott S. A scalable content-addressable network. In: Cruz R, Varghese G, eds. Proc. of the ACM SIGCOMM 2001. New York: ACM Press, 2001. 161−172.

[2] Napster. http://www.napster.com

[3] Gnutella. http://www.gnutella.com

[4] Yang B, Hector GM. Improving search in peer-to-peer networks. In: Rodrigues LET, Raynal M, Chen WSE, eds. Proc. of the 22nd IEEE Int'l Conf. on Distributed Computing Systems (ICDCS 2002). Washington: IEEE Computer Society, 2002. 5−14.

[5] Stoica I, Morris R, Karger D, Kaashoek MF, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for Internet applications. In: Govindan, ed. Proc. of the ACM SIGCOMM 2001. New York: ACM Press, 2001. 149−160.

[6] Zhao BY, Huang L, Jeremy S, Sean CR, Anthony DJ. Tapestry: A resilient global-scale overlay for service deployment. IEEE Journal on Selected Areas in Communications, 2004,22(1):41−52.

[7] Karger D, Lehman E, Leighton T, Levine M, Lewin D, Panigrahy R. Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the World Wide Web. In: Proc. of the 29th Annual ACM Symp. on Theory of Computing. New York: ACM Press, 1997. 654−663.

[8] He YJ, Wang S, Du XY. Efficient top-k query processing in pure peer-to-peer network. Journal of Software, 2005,16(4):540−552 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/540.htm

[9] Liu B, Lee WC, Lee DL. Supporting complex multi-dimensional queries in P2P systems. In: Proc. of the 15th IEEE Int'l Conf. on Distributed Computing Systems (ICDCS). New York: IEEE Computer Society, 2005. 155−164.

[10] Norbert B, Hans-Peter K, Ralf S, Bernhard S. The R*-tree: A efficient and robust access method for points and rectangles. In: Hector GM, Jagadish HV, eds. Proc. of the 1990 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1990. 322−331.

[11] Ashwin RB, Mukesh A, Srinivasan S. Mercury: Supporting scalable multi-attribute range queries. Computer Communication Review, 2004,34(4):353−366.

[12] Shu YF, Beng CO, Kian-Lee T, Zhou AY. Supporting multi-dimensional range queries in peer-to-peer systems. In: Germano C, Nathalie W, Marcel W, Nahid S, eds. Proc. of the 5th IEEE Int'l Conf. on Peer-to-Peer Computing (P2P). New York: IEEE Computer Society, 2005. 173−180.

[13] James A, Gauri S. Skip graphs. In: Proc. of the 14th Annual ACM-SIAM Symp. on Discrete Algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 2003. 384−393. http://www.ic.unicamp.br/~celio/peer2peer/skip-net-graph/skip-graph-aspnes.pdf

[14] Li DS, Lu XC, Wang BS, Su JS, Cao JN, Keith CCC, Hong VL. Delay-Bounded range queries in DHT-based peer-to-peer systems. In: Proc. of the 26th IEEE Int'l Conf. on Distributed Computing Systems (ICDCS). New York: IEEE Computer Society, 2006.

[15] Li DS, Lu XC, Wu J. FISSIONE: A scalable constant degree and low congestion DHT scheme based on Kautz graphs. In: Proc. of the 24th Annual Joint Conf. of the IEEE Computer and Communications Societies (INFOCOM). New York: IEEE Press, 2005. 1677−1688.

[16] Li M, Lee WC, Anand S. Semantic small world: An overlay network for peer-to-peer search. In: Proc. of the 12th IEEE Int'l Conf. on Network Protocols (ICNP 2004). New York: IEEE Computer Society, 2004. 228−238.

[17] Lakshmish R, Bugra G, Liu L. A distributed approach to node clustering in decentralized peer-to-peer networks. IEEE Trans. on Parallel and Distributed Systems, 2005,16(9):814−829.

[18] Tang CQ, Xu ZC, Mallik M. pSearch: Information retrieval in structured overlays, ACM SIGCOMM Computer Communications Review, 2003,33(1):89−94.

[19] Artur A, Xu ZC. Scalable, efficient range queries for grid information services. In: Ross LG, Nahid S, eds. Proc. of the 2nd Int'l Conf. on peer-to-peer Computing (P2P 2002). New York: IEEE Computer Society, 2002. 33−40.

[20] Murat D, Hakan F. Peer-to-Peer spatial queries in sensor networks. In: Nahid S, Ross LG, Germano C, eds. Proc. of the 3rd IEEE Int'l Conf. on Peer-to-Peer Computing (P2P 2003). New York: IEEE Computer Society, 2003. 32−39.

附中文参考文献:

[8] 何盈捷,王珊,杜小勇.纯 Peer-to-Peer 环境下有效的 Top-*k* 查询.软件学报,2005,16(4):540−552. http://www.jos.org.cn/1000-9825/16/540.htm

宋伟(1978－),男,湖北武汉人,博士生,主要研究领域为对等计算及安全.

卢正鼎(1944－),男,教授,博士生导师,CCF 高级会员,主要研究领域为分布式计算,软件集成环境,数据库系统,信息安全.

李瑞轩(1974－),男,博士,副教授,CCF 高级会员,主要研究领域为分布式计算,分布式系统安全.

於光灿(1974－),男,博士生,主要研究领域为分布式异构系统,协同设计.