一种基于多重覆盖的结构化 P2P 搜索*)

冯国富[1] 姜玉泉[1] 张金城[1] 顾 庆[2] 陈道蓄[2]

(南京审计学院信息科学学院 南京 210029)[1] (南京大学计算机科学与技术系 南京 210093)[2]

摘 要 本文提出了一种基于多重覆盖的结构化 P2P 搜索策略,mDOS(multi-Dimensional Overlays based Search)。mDOS 模拟社会网络中的小世界模型,根据结点的多重属性将结点组织成为多维树型结构。每一维上的任意两个结点根据其在树型结构中的位置形成语义距离。按照覆盖网络中实际距离与语义距离之间的一定函数关系构造多重覆盖网络。邻居结点和目标结点之间在网络中的实际距离可以通过他们的语义距离估计得到,形成基于结点属性的路由方向感。两个结点的有效距离取值多维中的最短语义距离。多重覆盖中的一个结点可以桥接在不同维上的邻居结点,缩小结点之间的平均有效距离,进而缩短搜索路径长度。mDOS 还引入索引内容"懒惰维护"机制和路由表"被动修改"机制以克服结点动态性带来的覆盖网络维护代价。最后的实验表明 mDOS 可以降低搜索路径长度,降低维护代价,提高可用性。

关键词 结构化 P2P,小世界,多重覆盖网络,语义距离,路由方向感

A Multi-dimensional Overlays based Structured Peer-to-Peer Search

FENG Guo-Fu[1] JIANG Yu-Quan[1] ZHANG Jin-Cheng[1] GU Qing[2] CHEN Dao-Xu[2]

(School of Information Science, Nanjing Audit University, Nanjing 210029)[1]

(Department of Computer Science and Technology, Nanjing University, Nanjing 210093)[2]

Abstract In this paper we propose a structured Peer-to-Peer Search, mDOS (multi-Dimensional Overlays based Search). mDOS simulates the Small World model in the social network, arranging the peers in multi-demensional tree structures according to peers' multi-attributes. The semantic distance between any two peers in any dimension is formed according to their positions in the tree. Multi-dimensional overlays can be constructed in therms of the relation between the virtual distance in the overlay and the semantic distance in tree structure. Then the virtual distance in overlay between the neighbors and the destination can be estimated according to the semantic distance, forming the attribute based sense of routing orientation. Because the effective distance between any two peers is the closest semantic distance among the distances in all dimensions, a peer can bridge the neighbors in different dimensions, shortening the average effective distance, and shortening the average search path length further. And the mechanisms of "lazily-maintained" and "passively-modified" are adopted in mDOS to deal with the overhead of overlay maintenance. The final simulation experiment shows that mDOS can improve the usability and facilitate the topology maintenance with the low search latancy.

Keywords Structured Peer-to-Peer, Small world, Multi-demensional overlay, Semantic distance, Sense of routing orientation

1 引言

Peer-to-Peer(P2P)系统根据体系结构可以分为集中式、层次式和全分布式三种。集中式 P2P 和层次式 P2P 采用一些中心服务器来维护结点信息和数据信息,因此容易带来单点容错和性能瓶颈问题。全分布体系结构 P2P 没有集中结点,通过消息传递实现信息维护和内容查询。当前全分布的 P2P 系统主要分为基于 DHT(Distributed Hashable Table)的结构化 P2P(DHTs)和基于消息泛洪(flooding)的无结构 P2P。结构化 P2P 系统通常首先将结点组织成为具有规则形状的覆盖网络,将内容对象或者关键字-地址索引存储到可预知的结点,然后基于确定的拓扑实现查询消息在结点之间的路由搜索。结构化 P2P 搜索响应快,消息复杂度小,伸缩性好。但由于 P2P 系统中的结点动态性强,结点频繁加入、退出,要耗费巨大的网络开销来维护既定拓扑和确定的索引存放位置。无结构化 P2P 没有存放位置和网络组织的限制规则,主要通过在结点之间转发大量的查询消息实现内容的匹配检索。全分布的无结构 P2P 网络维护代价小,但会带来巨大的消息开销,造成网络流量负担。

本文提出了一种可以进行灵活自组织的结构化 P2P 覆盖网络构造方法和资源发现策略,mDOS(multi-Dimensional Overlay Search)。mDOS 以启发式小世界模型为依据构造基

*)本文得到江苏省高校自然科学基金(06KJD520090)、国家自然科学基金(60573106,60573131)、国家"八六三"高技术研究发展计划基金项目(2006AA01Z199)和国家重点基础研究发展规划(973)(No. 220CB312002)资助。冯国富 博士,讲师,主要研究领域为分布式计算、协议测试和计算机审计;姜玉泉 教授,主要研究领域为计算机审计;张金城 教授,主要研究领域为计算机审计和 CIMS;顾 庆 副教授,主要研究方向为分布式计算和软件工程;陈道蓄 教授,博士生导师,主要研究领域为分布式计算与软件工程。

于多维结点属性的多维覆盖网络，能够降低搜索延迟。以“懒惰维护”机制管理关键字-地址索引（或内容对象），以“被动修改”机制维护覆盖网络，能够克服传统 DHTs 中结点动态性带来关键字维护和覆盖网络维护负担。

本文内容组织结构如下：第 2 部分回顾了已有工作；第 3 部分描述了如何构造小世界覆盖网络，并给出了路由策略；第 4 部分给出了实验结果并和相关工作作了比较；最后是结论。

2 相关工作

2.1 小世界

近年来，针对小世界网络的研究受到了图论、统计物理学、计算机网络、生态学、社会学以及经济学等各个不同领域的广泛关注。小世界网络由密集的短程连接和稀疏的远程连接构成，任意两结点间的最短路径很短，这给 P2P 搜索带来有益的启发。

1967 年著名的“六度分离”实验很好的说明了小世界现象[1]。实验表明，在美国任意两个人之间，通过“熟人”关系，平均需要 6.5 步就可以关联起来。Watt 等人[2]提出了一种小世界模型，结点顺次连接成环，每个结点连接最近的 k 各邻居结点，并与其它每一个结点以概率 P 决定是否相连，该模型较好的模拟了社会网络中的小世界特征。Kleinger[3]首先构造了一个二维网格，其中任意结点维护四条到达最近邻居的短程连接和一条远程连接，可以证明，使用贪婪算法，在 O(log$N$)跳内可以将消息路由到任意结点。

文[4]中给出了一种构造小世界网络的启发模型，指出构造小世界网络需要符合的几个要素特征，包括结点的多维标识，基于标识聚集成簇，层次式的簇组织方式等。本文 mDOS 也是受该模型启发，使之发展成为切实可用的 P2P 协议方案。

2.2 Distributed Hashable Table

典型的 P2P 系统有 Chord[5]，CAN[6]，Pastry[7]和 Tapestry[8]等。它们根据哈希物理地址生成的标识符号，分别将结点组织成为基于环，基于网格，基于超立方体的拓扑结构。哈希内容关键字得到和结点相同空间上的标识符，并定义后继结点（successor）是结点标识符大于等于关键字 K 标识符的第一个结点，每个关键字都保存在它的后继结点中。然后根据基于拓扑的路由算法实现高效定位。这些系统在查找方面都具有良好性能，搜索路径长度基本上都达到 O(logN)复杂度。但以上基于 DHT 的结构化 P2P 系统也存在一些问题。

首先，以上结构化 P2P 系统用以实时方式处理结点的加入和退出，维护覆盖网络要耗费巨大的网络开销。这一方面在于，在基于 DHT 的 P2P 系统中，标识符哈希空间要远远大于实际的地址空间，内容必须存放于其后继结点，内容的存放因而对结点的顺次连结顺序具有依赖性。结点退出必须将其所负责内容转移到其后继结点，结点加入必须从其前驱结点获取所负责的内容；否则，非实时的操作将会导致命中失败。这种结点地址的非连续性和内容存放规则的确定性带来结点顺次连接的需求，进而带来覆盖网络实时性维护的要求。另一方面，P2P 系统中的结点具有高度动态性，频繁的加入、退出，实现结点的搜索、数据的转移，必然带来巨大的网络开销。

其次，以上基于 DHT 的 P2P 系统缺乏必要的多副本冗余容错机制。在不明显增加存储和维护开销的前提下，适当增加副本冗余能够提高搜索性能，同时能够避免当结点加入退出时对内容索引的实时维护。

最后，正如文[9]所描述，结构化 P2P 采用了一维线性化名字空间，破坏了许多现实中有用的层次性语义，比如行政地理信息，文件夹层次包含信息等，使得一些系统失去可用性。比如失去了内容上的临近依赖关系使内容预取变得不可行，失去地理上的邻近关系，近距离结点之间的相互代理缓存变得困难。

考虑到结构化 P2P 以上固有问题，结合小世界启发模型，我们引入语义树的概念。根据结点属性，P2P 系统中的每个结点与语义树的一个叶子结点相对应。每个结点属性都与语义树上的一条路径相对应，利于克服传统一维线性化结点命名方式对结点层次性语义属性的损害。结点之间形成基于语义树的相对距离，指导查询消息路由，实现高效定位。位于同一叶子的结点集相互协作，能够消除内容存放对结点顺序的严格依赖；在叶子结点集内灵活部署冗余副本，能够避免对结点加入退出实时处理带来的内容维护和网络维护负担，能够提高系统可靠性，利于负载均衡。

2.3 基于小世界的资源发现策略

目前基于小世界的 P2P 搜索算法主要分为基于小世界的启发性算法和基于小世界理论模型的搜索算法。

Freenet [10]通过大量的文件复制和 LRU 存储空间管理算法使得具有相似 id 的文件聚集，以启发性的模拟小世界现象。但是，Freenet 大量的沿途复制文件会浪费大量的网络带宽，热点文件的大量复制往往会替换冷门对象出系统，不能保证冷门对象的命中结果。

文[11,12]等受小世界网络启发，将具有相同兴趣的结点聚集成簇，作为路由选路的首选范围。这种方式能够改善盲目搜索系统的性能。但是它并没有给出完善的小世界拓扑构造方法，不能适应用户兴趣的变化，不能对兴趣以外内容的搜索发挥效用。

Symphony[13]借助 W-S 理论模型方面的工作来构造 P2P 系统覆盖网络并实现内容查找。但总的说来，理论模型只能提供构造符合小世界特征拓扑结构的方法，能够缩短泛洪方式查找的平均路径长度，但对于查询消息路由并不能提供方法策略上的帮助，并不能提供选择优化路径的方法。

本文根据[4]从社会网络得出的小世界模型特征，为每个结点分配多维独立的标识符，结点根据标识符形成独立的簇集。每一维上的簇集使用层次方式进行组织，形成语义层次的距离。根据语义层次的距离，结点之间添加紧密或者稀疏的连接，形成路由方向感。此外，一个结点拥有多维标识，具有多个独立的语义层次距离，路由总是按照最短的距离进行查询消息转发，有效地实现资源查找。

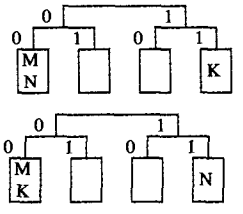

图 1 语义树示例

3 mDOS (multi-Dimensional Overlay Search)

这一部分着重描述小世界覆盖网络的构造方法和查询消息的路由策略。

3.1 多维标识生成

定义系统中成员结点为一个二元组 Peer(IP, attribute)。其中 IP 是结点的物理 IP 地址，attribute 是一个 D 元组(x_1, x_2, x_3, …, x_D)，其中 x_i 是结点在第 i 维的属性。

P2P 系统中的一条连接〈Peer1, Peer2〉，是存储于 Peer1，关于 Peer2 的属性信息。Peer2 也即可由 Peer1 直接到达的邻居结点。一个结点上连接的集合构成路由表。

查询是用户在一维或几维上给出的目标资源属性描述的关键字集合。定义查询为一个 D 元组 Query(x_1, x_2, x_3, …, x_D)，其中 x_i 是结点第 i 维上的属性，至少在一维上不可以为缺省值。

查询结果是结点 IP 地址的集合，这些结点的属性和查询相匹配。

文[4]研究表明，小世界网络中的结点，并非仅仅由一维名字空间所标识，而是具有多维属性特征，这些属性包括自身所具有的个体特征以及所属组织的团体属性。

这里用一致性哈希函数，比如 SHA-1，通过哈希 IP 地址为每个结点分配第一维上的标识符。由于一个结点的各维属性之间是独立的，一个结点的标识符在各维之间也是独立的；而且，在每一维上哈稀生成的所有标识符之间也是独立的，因此，通过再次哈稀标识符可以得到下一维的标识符。标识符的长度应该足够大，以使得两个结点在同一维映射到同一标识符的概率足够小。

3.2 层次组织与语义树

根据每一维上的不同特征可以将系统中结点逐步进行层次化的分组归类。最顶层包含整个系统，接下来的每一层都是对上一层的细化。细化后的集合，成员之间的特征更加相似，在概念特征上更加具体，相应地，细化之后的集合具有更好的可认知性和可控性。这样，得到一种在概念特征语义上逐渐具体，用于对结点进行渐近分组的层次性树形结构，我们称之为语义树。语义树是一种基于节点属性的概念上的虚拟结构，和混合体系结构的 P2P 不同，语义树只是通过属性体现，不需要维护相应的连接。可以对系统中的结点进行逐步划分，直到叶子簇中的结点能够对其邻居成员很容易的就能熟悉掌握为止，尽管原则上根据每个结点个体特征的唯一性，最终都可以将系统细化到一个结点一个簇的程度。文[4]中也给出了一个社会网络中比较合理的最小簇规模 g 的上限 $g<100$。在本文，我们定义最小簇规模 g 的范围为 $15<g<32$，如果超出这个范围，就要进行相应地拆分或合并操作。

每个结点根据其标识属性与语义树上一个确定的叶子簇相对应，一条从根结点到叶子的路径标志了该叶子簇内成员的属性。两个叶子簇内成员的属性，或者说两个结点所属叶子簇在语义树上的相对位置确定了两个结点间联系的紧密程度。两个叶子簇所在的最小共同簇越小，两者之间的联系就越紧密；相反，如果两者所属的最小共同簇越大，两者之间的关联就越稀疏。我们称这种决定两个结点联系紧密程度的相对位置为语义距离。定义有：

定义 1 最近共同祖先(Least Common Ancestor) 两个结点的最近共同祖先是指语义树上这两个结点所在叶子所拥有的最小深度的祖先结点。

定义 2 语义距离(Semantic Distance) 在语义树中，处于语义树同一叶子的两个结点之间的语义距离为 1；之外的两个结点，最近共同祖先的层次每升高一层，两者之间的语义距离就增加 1。语义距离表示了语义树上两个叶子簇成员之间特征的相似程度，以及带来的两者之间联系紧密程度。如图 1 左图所示，结点 M、N 之间的语义距离为 1，M、K 之间的语义距离为 4。

本文定义语义树为一棵二叉树。如果分别用 0 和 1 代表二叉树的左子树、右子树，那么一个结点标识符的前 H 位唯一的确定了一条从语义树根结点到叶子簇的路径。具有相同路径的结点组成一个叶子簇，彼此之间具有更多的相似属性。假设每个结点获取的标识符在哈希空间上足够随机，每个标识符在每一位上取值 0 或 1 的概率也足够均匀，那么，系统中的所有成员结点将会被均匀的分布到语义树的叶子簇中去。由于语义树随簇成员的加入退出做动态调整，造成叶子簇的路径长度有所差异，因此，每个结点在每一维上记录 H，代表所在叶子簇的路径长度，标识符的前 H 位也就表达了该结点在语义树上的位置。

3.3 语义距离与连接密度

如果两个结点的语义距离越小，说明两个结点所在的最小共同组就越小，两个结点相互连接的概率就越大，相互熟知的概率相应的就越大。根据文[4]，在小世界社会网络中成员之间的连接概率 p 是相对层次距离 h 的函数，$p_{\#}=e^{-\alpha h}$，其中 $\alpha=0.92$。在文[14]中也统计发现，两个结点之间存在连接的概率 p 与所在的共同最小组的成员数量 g 呈接近反比，$p_{\#}\sim g^{-\beta}$，$0.75<\beta<1$。这都说明，如果小世界网络中的两个结点拥有更小的语义距离，两个结点所在的最小共同组就越小，两个结点的联系就越紧密。

因此，借鉴社会小世界网络方面的工作，定义结点 i 和 j 之间的连接概率 P_{ij} 为语义距离 d_{ij} 的函数：

$$P_{ij}=\begin{cases}P_i, & d_{ij}=1\\ P_o/2^{d_{ij}-2}, & d_{ij}>1\end{cases} \quad (1)$$

其中，P_i 和 P_o 值可调，P_i 值一般接近 1，P_o 一般取值 0.25。

3.4 路由表与路由算法

路由表：每个 mDOS 结点维护一张路由表，即邻居信息的集合。路由表的行数与所采用的语义维数 D 相同，每一行上的结点标识符采用同一维标准。每一行由结点在该维上的标识符和邻居表项组成。每个邻居表项由四部分组成，包括该邻居结点所在语义树叶子簇的路径长度 H，邻居标识符到 IP 地址的映射，上次选中该邻居时的在线状态，以及发现该邻居结点失效的时间戳。时间戳用于决定何时淘汰失效邻居表项，将在后面作具体描述。

路由：如果当一个结点不能满足查询，就要从路由表里面选择一个与目标结点语义距离最近的邻居结点转发查询。这是因为，通过网络构造过程可以知道，两个结点之间的连接密度与二者之间的语义距离成反比，这种不同的语义距离拥有不同连接密度的特点形成路由方向感，即与目标结点语义距离越近的结点，与目标结点拥有更为紧密的关联，会以更大的概率命中目标或者缩短到目标结点的距离。

mDOS 并不以实时的方式来处理结点的退出。当转发查询失败，便认为该结点暂时离开系统，标记该结点状态，再次选取到目标结点语义距离最近的在线邻居结点继续转发查询。

推论 1 在一个拥有 N 个结点的 mDOS 系统中，平均搜索路径长度不大于 $O(\log N)$ 跳。

证明：首先考虑一维语义树的情况。根据公式(1)，与一个结点语义距离为 d 的邻居数目为 CP_o。因此，只要 $P_o>1/C$，我们总能找到这样一个邻居结点，它与目标结点在相同的

语义子树。即每一次路由转发，至少能缩小一半的搜索范围，与目标结点的语义距离至少减少 1。而且，距离越近，连接的密度越大，搜索能够收敛。所以，基于一维语义树的搜索路径长度与语义树高度相关联，所以搜索路径长度复杂度为 O(log$N$)。

另一方面，mDOS 采用多维语义树能够加强搜索效率，因为路由所采用的语义距离是各维语义距离中的最小值。如图 1 所示，M、N 之间使用的语义距离是左图中的 1，而非右图中的 4。而且，多维语义树能够桥接远程结点。如图 1 所示，结点 N、K 在任一维上的语义距离为 4，而综合二维语义距离 $d_{KN}=d_{MN}+d_{MK}=2$，即两边之和小于第三边。所以，采用多维语义树能够缩短查询结点和目标结点间的语义距离，能够提高每一步路由转发过程中接近目标的效率，从而提高 mDOS 搜索性能。

因此，mDOS 中的搜索路径长度不大于 O(log$N$)跳。

3.5 自组织和自适应性

结点加入：当一个结点第一次加入 mDOS 系统，需要初始化路由表。这里假设新加入的结点，通过“半径递增”广播或者从一些知名的管理结点获取一个已知的 mDOS 在线结点。

假设第一次加入系统的结点 A，而且已经知道系统中的一个结点 X。使用 SHA-1 算法哈希 A 结点 IP，可以得到一个 128 位的标识符，标识符前面 H 位就是所在语义树的表述。然后从 X 结点路由表里面匹配和 A 结点语义距离最小的 Y 结点，并将加入消息转发给 Y 结点。Y 结点继续转发该加入消息，直到找到这样一个结点 Z：Z 结点标识符的前 H 位和 A 结点标识符的前 H 位完全匹配，其中 H 是从语义树根结点到 Z 结点所在叶子簇的路径长度。A 结点从 Z 结点路由表中拷贝所有关于簇成员的路由表项，并将自己通告给所有簇成员，标志结点 A 找到了所处的叶子结点，加入了一个协作组。然后，从邻居结点的路由表开始可以逐次获取关于系统所有成员的路由表项。根据 A 结点和其它系统结点的语义距离按一定的概率决定是否保留该结点的路由表项，完成簇之间互连的路由表项。

结点退出：如果给路由表中一个邻居结点转发消息而不能得到正确响应，便认为该邻居结点暂时失效，记录结点失效的时间。考察两次失效的时间间隔，如果大于某一阈值，便认为该结点永久退出系统。搜索与失效结点同一叶子簇的成员，来填补该失效路由表项。如果修改路由表项的结点和失效结点处于同一叶子簇，则还需通知其它所有同簇将该结点从簇中剔出。这种被动非实时方式的路由表项淘汰机制能够削减大量网络维护方面的开销。

在结点加入系统或者修改路由表项的时候都需要发现某个特定的叶子簇，以加入该协作组或者替换簇中的失效结点。这种情况下，查询消息只要给出叶子簇在语义树中的路径，也即簇内成员标识符的前面 H 个二进制位，便可以返回关于该簇所有成员的路由表。

对于短暂离线后再上线的结点不做具体处理。同样，对于结点暂时的离线，系统也不做任何特殊的操作。这样可以在一定程度上避免结点频繁离线上线给实时维护带来的开销。

簇拆分：随着一些结点的加入和永久退出，叶子簇成员也在做动态变化。过多的簇成员会消耗太多的空间用于簇内连接，过少的簇成员又不足以完成簇内成员之间的协作功能。叶子簇应该做动态的合并与拆分调整，使成员个数在一个可控和可用的范围。

当一个结点加入系统，要统计其协作组，即所在叶子簇的成员数量。如果成员数量超过某一范围，便触发拆分操作。拆分步骤比较简单，首先将路由表中所有标志叶子簇路径长度的 H 替换为 $H+1$；其次，根据簇内连接概率 P_i 和簇间连接概率 P_o 调整路由表中的涉及到的表项。这样，原先叶子簇中的成员分别根据第 $H+1$ 位的值分为 1 和 0 的两个分支子树。系统其它结点路由表中关于这些成员的表项，在后面再次访问的时候修改 H 的值。由于远程连接概率与语义距离呈反比，所以其它远程连接不需要调整。

簇合并：当结点发现其它簇成员永久失效，要通知其它成员更新路由表。同时要统计本叶子簇成员数以及邻居叶子簇成员数。如果两个叶子簇的成员总数低于某一下限就需要进行合并。合并过程和拆分过程相逆，就是修改叶子簇属性路径长度，用 $H-1$ 代替 H。这样，标识符前 $H-1$ 位相同，第 H 位不同的两个叶子簇合并成为一个新的叶子簇。相应地，修改之后的簇成员之间用稠密的簇内连接取代原先的簇间连接。

前面所有的前提假设就是随机分配的标识符在任意位取值 0 或者 1 的概率完全相同，而且永久失效结点的标识符在整个标识符空间上也绝对分布均匀。但或许有时候事实并非完全如此。为了克服非绝对均匀的缺陷，拆分阈值可以设置一定弹性，以避免语义距离为 2 的两个叶子簇成员数量出现差别迥异的情况。

通过以上可以看出，mDOS 中路由表的修改和覆盖网络的维护并不和传统的 DHTs 一样，并非结点一离开就要作实时调整。这种索引内容“懒惰维护”的存储维护策略和路由表项“被动修改”的覆盖网络策略能够把 DHTs 从沉重的覆盖网络维护中解放出来，削减维护开销。

3.6 层次性语义与语义树

大多现实应用中的数据都是以层次形式组织的，例如文件组织，资源组织，地理位置信息等等。DHTs 一维名字空间将所有的元素线性离散化，破坏了元素之间的隐含语义关系，使得许多传统的提高可使用性的操作不能实施，比如没有了内容之间的依赖关系和分类属性关系无法实现内容的预取，没有了结点之间的物理邻近关系便无法实现可用的缓存机制。

mDOS 用层次语义树来组织结点，有利于保持原先系统中元素间所隐含的描述信息。当上面所采用的二叉语义树不足以表达实际的层次语义分类时，可用标识符中几个连续的二进制位来表达一个语义树中的结点，使得语义数具有任意数量的分支出度。也可以根据实际的应用需求事先定义语义树，然后将标识符映射到语义树。这与前面根据标识符形成的二叉语义树在搜索方面没有什么大的不同，除了需要在路由表中附带少量额外的语义树信息。关于语义树的预先定义可以参考本体分类方面的相关资料。

基于多维语义树路由机制的 mDOS，可以采用任意的一维或者多维进行路由，也可以根据其中的不同优先级路由，以实现不同的实际需求。比如位置维路由优先的策略可以减小实际的底层网络开销，语义维优先可以返回更多相近相关的内容。基于层次语义的路由方式能够明显提高结构化 P2P 系统的灵活性和可用性。

4 实验

在这一部分，我们用 C 语言编程做了模拟实验，并对 mDOS 的性能与 Pastry 做了比较。由于结构化 P2P 系统中所查询的数据存在就一定能够搜索成功，因此考察性能指标主要为搜索路径长度。

图 2 展示了 mDOS 一维语义信息下的搜索路径长度。可以看出，路径长度随系统规模完美的呈现 $O(\log N)$ 复杂度。图 3 显示了语义信息的维数对搜索路径的影响。不限制路由表项数目，增加维数明显缩短搜索路径长度。在路由表项数目固定的情况下，搜索性能与簇内连接概率 P_i 存在关联。这是因为，多维语义本身能够加强搜索性能，所以在稍微增加维数的情况下，搜索路径长度缩短。但是，随着维数增多，用于簇内连接的路由表项数目增加，而用于簇间连接的表项数目将明显减少，反而会增加簇间查找代价。正如图 3 所显示，在 $P_i=1$ 时，路径先下降后上升；当 $P_i=0.5$，起点路径长度较大，但是路径长度随维数增多持续下降。为了便于和 *Pastry* 比较，下面在不特别指明的情况下，语义信息维数都取 2，簇内连接概率 $P_i=1.0$，簇间连接概率 $P_o=0.25$，以使得路由表项在相同系统规模下保持与 Pastry 相当的数目。

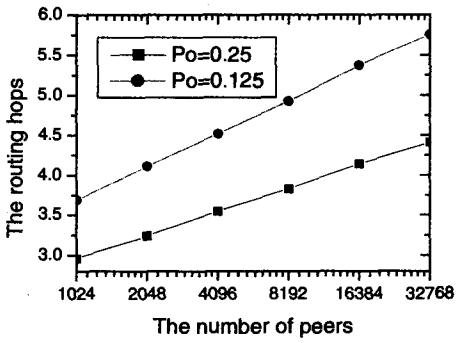

图 2 搜索路径长度与系统规模之间的关系
($P_i=1.0$, D=1)

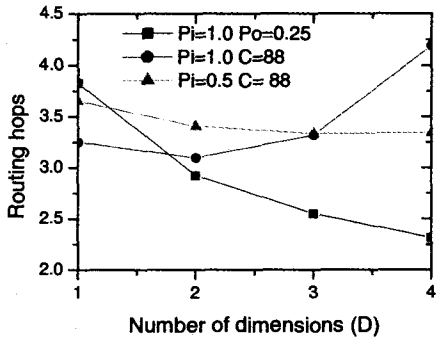

图 3 搜索路径长度与语义维数之间关系(C 为每个结点的路由表项数目)

图 4 是 mDOS 的搜索路径长度随系统规模的曲线，以及和 Pastry 的比较。曲线显示，mDOS 搜索路径长度随系统规模的增长速度和预期结果一致。又由于多维信息语义树的引入加强了 mDOS 的搜索效率，性能效果比 Pastry 高 10%左右。图 5 是路径长度的分布图。和预期结果一致，系统最大搜索路径长度不大于语义树的层数 10。mDOS 大部分搜索路径长度位于 3 跳，而 Pastry 的大多数路径长度为 4，这也是之所以图 4 显示 mDOS 比 Pastry 搜索效率高的原因。

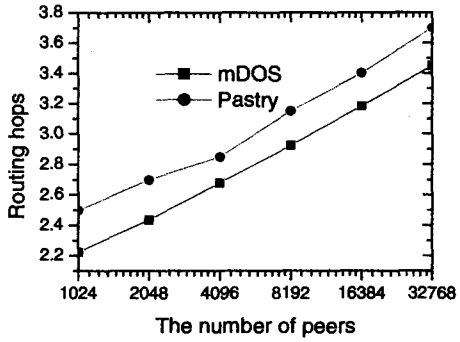

图 4 mDOS 与 Pastry 平均搜索路径长度比较

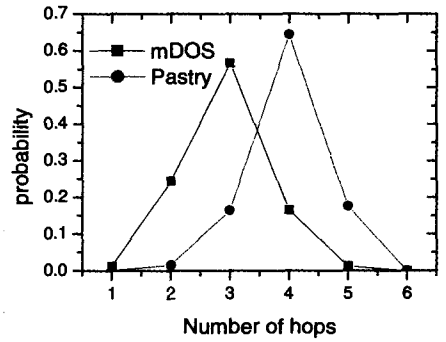

图 5 搜索路径长度分布(系统规模为 2^{13} 个结点)

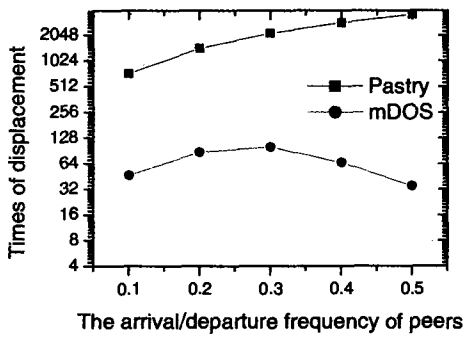

图 6 数据位置移动次数与结点到达/离开频率 R 之间的关系比较
(查询按照泊松分布产生，每秒产生一次查询；路由表项永久失效的阈值设定为 2 小时；结点具有相同的加入退出频率，平均为 R 次每秒；网络规模 2^{13})

结构化 P2P 中，一个结点维护其后继关键字。一个结点退出就要将其负责的内容转移给其后继结点，一个结点加入就要从其前驱结点获取其该负责的关键字。结点的每次加入和退出都引起关键字及其内容的转移，而且每次转移都对应着网络拓扑的重构。这给物理网络带来巨大负担。mDOS 采用“懒惰维护”法，结点暂时离开并不立即做实时处理，其功能由其协同成员完成。只有当结点第一次加入或者永久离开系统时才做相应的关键字转移操作。此外，mDOS 采用“被动修改”的方式维护网络连接，能够明显节省网络维护费用。图 6 是 mDOS 和 Pastry 在不同加入/退出速率下关键字转移次数的比较。可以看出，mDOS 能够明显改善传统结构化 P2P 普遍存在的结点动态性强、网络维护代价高的缺点。mDOS 转移关键字的次数在一定范围内独立于结点的加入/退出频率。而 Pastry 以及传统结构化 P2P 的网络维护代价与结点加入/退出频率呈正比关系。图 6 还显示，随着结点加入/退出频率

(下转第 51 页)

Sensor Networks: a Voronoi Approach. In: Proc. 28th IEEE Conference on Local Computer Networks (LCN2003), Bonn/Konigswinter, Germany, 2003
10 Huan P S, Ramamritham K. Scheduling messages with deadlines in multi-hop real-time sensor networks. In: Proc. The 11th IEEE Real Time and Embedded Technology and Applications Symposium, San Francisco, CA, Mar 2005. 415~425
11 Florens C, McEliece R. Packet distribution algorithms for sensor networks. In: IEEE INFOCOM, 2003
12 Rhee I, Warrier A, Xu L. Randomized dining philosophers to TDMA scheduling in wireless sensor networks: [Technical report]. Computer Science Department, North Carolina State University, Raleigh, NC, 2004
13 Hohlt B, Doherty L, Brewer E. Flexible Power Scheduling for Sensor Networks. In: Proc ISPN 2004
14 Wu Hejun, Luo Qiong, Xue Wenwei. Distributed Cross-Layer Scheduling for In-Network Sensor Query Processing. In: Proc. the Fourth Annual IEEE International Conference on Pervasive Computing and Communications (PERCOM'06) Pisa - Italy, March 2006. 13~17
15 Mihail L, Sichitiu. Cross-Layer Scheduling for Power Efficiency in Wireless Sensor Networks. In: Proc. INFOCOM'04
16 Gandham S, Zhang Ying, Huang Qingfeng. Distributed Minimal Time Convergecast Scheduling in Wireless Sensor Networks. In: Proc. The 26th International Conference on Distributed Computing Systems, Lisboa, Portugal, July, 2006
17 Choi W, Das S K. A Novel Framework for Energy-Conserving Data Gathering in Wireless Sensor Networks. In: Proceedings of IEEE INFOCOM, Miami, Florida, Mar 2005
18 Iyer R, Kleinrock L. QoS Control for Sensor Networks. In: IEEE International Communications Conference (ICC 2003), Anchorage, AK, 2003
19 Kay J, Frolik J. Quality of Service Analysis and Control for Wireless Sensor Networks. In: 1st IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS 2004), Ft Lauderdale, FL, Oct 2004
20 Liang Biyu, Frolik Jeff, Sean Wang X Sean. A Predictive QoS Control Strategy for Wireless Sensor Networks. In: The 1st Workshop on Resource Provisioning and Management in Sensor Networks (RPMSN '05) in conjunction with the 2nd IEEE MASS, Washington, DC, Nov 2005
21 Ray S, Carruthers J B, Starobinski D. RTS/CTS-induced Congestion in Ad Hoc Wireless LANs. In: Proc. Wireless Communications and Networking Conference (WCNC), March 2003

(上接第 36 页)

的增长，mDOS 进行关键字转移的次数呈现先增长后降低趋势。这是因为，当加入/退出频率增快时，失效结点在被淘汰之前再次进入系统，相应的信息不需要做移动。

下面的实验利用语义树来解决 P2P 系统的物理临近性问题，演示了语义树对实际应用在可用性方面的提高。取一块平面，一分为二，分别代表左子树和右子树；依次继续，分为一系列的小方格。小方格代表了结点位置，并与语义树上的一个叶子结点相对应，形成一棵能够表达相对位置的语义树。实验分别采用三种路由策略实现搜索来验证搜索策略对物理临近性的影响。这三种策略分别是：基于前面设计的一维的标识符语义距离、基于一维的几何距离、基于前两者距离的乘积。表 1 是其中的实验结果。从应用层跳数来看，前两者相当，第三者最短，这是因为第三者采用了两维语义树的缘故。从相对距离，即路由过程经过的几何距离之和与查询结点和目标结点之间几何距离的比值来看，后两者效果较好，因为根据相对位置形成的语义树在路由过程起辅导标准作用，而基于标识符语义树的路由策略忽视了结点之间的物理距离因素，使得逻辑上的一跳对应物理上的多跳，更远的物理距离。在实际应用中还可以根据需要设计语义树，以在结构化 P2P 系统中保留更多应用描述性信息。可以采用其中任意的一维或者多维作为路由的度量标准，也可在各维之间规定不同的优先级，以实现不同的应用需求。

表 1 不同路由策略下搜索路径长度与相对距离比较

系统规模	路由依据					
	语义距离		几何距离		语义距离 * 几何距离	
	Hops	相对路径	Hops	相对路径	Hops	相对路径
2^{10}	2.91	2.54	2.67	1.64	2.13	1.21
2^{12}	3.53	3.69	3.19	2.1	2.58	1.69
2^{14}	4.14	4.72	3.72	2.57	3.06	2.2

此外，mDOS 中的热点内容不仅仅由一个结点负担，而是由协作的几个结点共同负担，因此，mDOS 对热点访问具有适应性，比之传统的 DHTs 更利于负载均衡。

结论 本文给出了一种 Internet 环境下完全分布的、容错的、可扩展的 P2P 搜索系统，mDOS，介绍了基于小世界构造自组织覆盖网络的方法以及查询消息自适应路由的策略。在查找路径、容错、维护代价等方面与 Pastry 的实验比较说明，mDOS 具有如下优点：mDOS 具有很高的搜索效率，具有 $\log(N)$ 搜索路径长度；mDOS 以"懒惰维护"机制管理关键字，以"被动修改"维护覆盖网络，能够克服传统 DHTs 中结点动态性带来关键字维护和覆盖网络维护负担；mDOS 在叶子内结点之间实现协同备份，对结点离线具有很好的容错性，利于负载均衡；mDOS 采用层次性语义树组织结点，利于保持许多有价值的应用描述性信息。接下来我们将着手 mDOS 的实现，并加强可用性方面的工作。

参 考 文 献

1 Milgram S. The small world problem. Psychology Today, 1967, 67(1)
2 Wattz D J, Strogatz S H. Collective dynamics of small world networks. Nature, 1998, 393
3 Kleinberg J. The small-world phenomenon: an algorithmic perspective. In: Proc. 32nd ACM Symposium on Theory of Computing (STOC 2000), 2000
4 Watts D J, Dodds P S, Newman M EJ. Identity and search in social networks. Science, 2002, 296: 1302~1305
5 Stoica I, Morris R, Karger D, et al. Chord: a scalable peer-to-peer lookup service for internet applications. In: Proc. ACMSIGCOMM, 2001
6 Ratnasamy S, Francis P Mhandley, Karp R M. A scalable content-addressable network. In: Proc. ACM SIGCOMM 2001, 2001
7 Rowstron A I T, Druschel P. Pastry: Scalble, decentralized object locateon, and routing for large-scale peer-to-peer systems. In: IFIP/ACM International Conference on Distributed Systems Platforms, 2001
8 Hildrum K, Kubiatowicz J D, Rao S, et al. Distributed object locateon in a synamic network. In: Proc. 14th ACM Symposium on Parallel Algorithms and Architecures (SPPAA), 2002
9 Keleher P J, Bhattacharjee B, Silaghi B D. Are Virtualized Overlay Networks Too Much of a Good Thing? IPTPS, 2002. 225~231
10 FreeNet. http://freenet.sourceforge.net
11 Sripanidkulchai K, Maggs B, Zhang Hui. Efficient Content Location Using Interest-based Locality in Peer-to-Peer Systems. Infocom, 2003
12 Iamnitchi A, Ripeanu M, Foster I. Locating Data in (Small-World) Peer-to-Peer Scientific Collaborations. In: 1st International Workshop on Peer-to-Peer Systems (IPTPS'02), Cambridge, MA, March 2002
13 Manku G S, Bawa M, Raghavan P. Symphony: Distributed hashing in a small world. In: 4th USENIX Symposium on Internet Technologies and Systems, USITS, 2003
14 Kleinberg J. Small-world phenomena and the dynamics of information. Advances in Neural Information Processing Systems (NIPS) 14, British Columbia, Canada, 2001