基于主题和物理位置相近原则的层次化对等语义覆盖网络结构

于 婧　张建辉　顾小卓　汪斌强
(国家数字交换系统工程技术研究中心　郑州　450002)

摘　要：对等语义覆盖网络构建主要采用索引和超级节点的方法，不可避免地存在瓶颈问题，同时忽略了覆盖网络与实际网络拓扑一致性问题对语义覆盖网络性能的重要影响。该文提出的对等语义覆盖网络结构 TPPH 充分结合了结构化 P2P 网络高效的定位和非结构化 P2P 网络的复杂检索功能，采用分布式哈希表机制将相同主题节点组织成主题区域，在同一主题区域内通过物理位置相近原则进行群的划分，从而在物理网络拓扑基础上建立语义 P2P 覆盖网络结构。性能分析和仿真实验表明，该结构显著提高了查全率并缩短了平均查询时延，是一种支持复杂查询、高性能的语义覆盖网络结构。

关键词：对等网络；语义覆盖网络；拓扑一致性；物理位置相近
中图分类号：TP393　　**文献标识码**：A　　**文章编号**：1009-5896(2008)08-1999-05

A Hierarchical P2P Semantic Overlay Network Architecture Based on Topic and Physical Proximity

Yu Jing　Zhang Jian-hui　Gu Xiao-zhuo　Wang Bin-qiang
(*National Digital Switching System Research Center*, *Zhengzhou* 450002, *China*)

Abstract: The main methods to construct P2P overlay network are index and super-node, which introduce bottleneck problem and ignore the significant influence of topology aware problem. A new P2P overlay network TPPH is established. It combines the virtue of efficient locating performance of structured P2P network and complex searching function of unstructured P2P network. Based on physical network topology, it partitioned nodes into topic domain by utilizing distributed hash table mechanism, and according to physical-proximity principle it clustered nodes within a domain. Analysis and simulation results indicate that TPPH can dramatically increase the recall and decrease the searching latency simultaneously, thus it is complex searching available and a high performance semantic overlay network architecture.

Key words: Peer-to-Peer (P2P) network; Semantic overlay network; Topology aware; Physical proximity

1　引言

语义对等网络(semantic P2P[1])是结合了 semantic web 和 P2P 技术的对等网络。语义对等网络中内容搜索是基于语义覆盖网络(Semantic Overlay Network，SON[2, 3])进行的。好的语义覆盖网络结构不但能提高路由效率，还需要满足可维护性和可扩展性要求。基于索引式 P2P 网络[4]构建的 Edutella[5]存在单点失效及性能瓶颈的问题。基于结构化 P2P 网络[6, 7]构建的 pnear[8]不能很好的支持复杂检索。沿用 Gnutella[9]机制的语义覆盖网络采用洪泛或随机游走[10]的搜索策略使查询消息以指数级别增长。另外，上述语义覆盖网络结构都忽略了覆盖网络与物理网络不匹配问题。

本文提出的基于主题和物理位置相近原则的层次化语义覆盖网络结构 TPPH(Topic and Physical Proximity based, Hierarchical)不存在索引及超级节点的概念，充分利用了结构化 P2P 高效的定位性能及非结构化 P2P 网络的自治性及复杂检索功能，同时考虑了物理网络与覆盖网络拓扑一致性的问题，是一种支持复杂查询的、高性能的语义覆盖网络结构。

本文在第 2 节介绍 TPPH 结构及构建；路由算法及性能分析在第 3 节给出；第 4 节进行与 Gnutella 系统的比较仿真；最后在第 5 节总结全文。

2　TPPH 结构

TPPH 结构设计的出发点是依据当前互联网的体系架构构建语义覆盖网络，依据主题和物理位置进行节点划分，将拥有相同主题的节点聚集在一起，将物理位置相近的节点组织在一起。

2.1 层次结构

TPPH 可以分为组群覆盖层和 DHT 主题覆盖层。其中，组群覆盖层实现基于网络物理拓扑的组群的划分，DHT 主题覆盖层实现基于 DHT 算法的主题路由和定位功能。图 1 为当前互联网典型框架结构，T_2^{A1} 表示网络 A 中主题为 2 编

2007-01-08 收到，2007-12-31 改回
国家 973 计划(2007CB307102)资助课题

号为 1 的节点，图 2 是对图 1 以 TPPH 机制组织形成的结构示意图。在组群覆盖层，相同主题且位置相近的节点构成组群(Cluster)，相同主题的组群构成区域(Domain)，位于不同主题区域但位置相近的组群之间通过超链接相连。

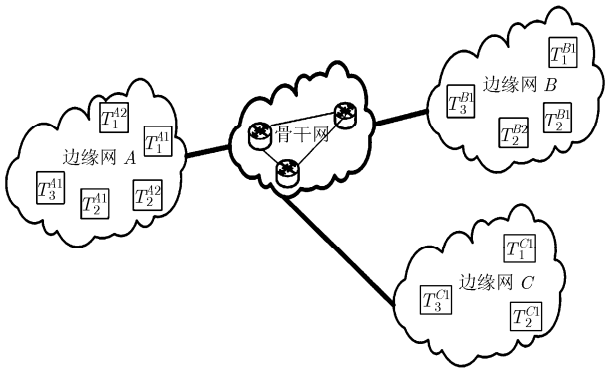

图 1 互联网典型框架结构示意图

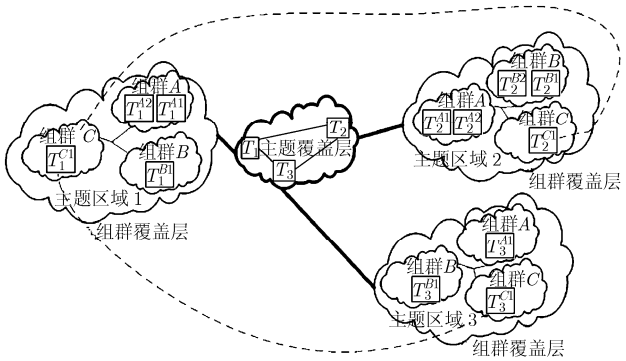

图 2 基于图 1 的 TPPH 结构示意图

2.2 TPPH 构建

2.2.1 节点命名 节点包含的资源具有特定的主题，假设可以由信息检索和数据挖掘领域中成熟的分类算法[11, 12]或者由用户对本地资源进行统计分类。一个节点可以包含多个主题的资源。$P_n=\{(T_i,\lambda_i),i=1,2,\cdots\}$ 将节点 n 描述为带权的主题的集合。其中，T_i 是节点所属主题类别，λ_i 是该类别的资源占节点总资源的比例。

定义 HASH 函数 $F(T_i)$，将主题类别 T_i 映射到标识符 b_i 上。主题覆盖层 DHT 节点定义为区域虚拟节点 DN(Dummy Node)，DN 的节点标识符就是对应主题的标识符，也是该区域的标识符(domain id)。组群标识符(cluster id)用区域标识符和组群标识来表示。组群内节点的节点标识符则应用组群标识符和节点标识表示。系统划分及节点命名如图 3 所示。

2.2.2 主题覆盖层 主题覆盖层可由已有 DHT 结构构成。当具有主题 T_i 的节点进入网络时，主题覆盖层没有与之对应的 DN，则该节点自动代理 DN 功能，依照主题覆盖层的结构进行新节点的加入过程。

2.2.3 组群覆盖层 组群覆盖层的组群划分依据物理距离相近原则进行。本文以节点间的时延为度量，设置节点间时延的门限值为 T_{thresh}，节点间时延低于该门限则将其划分到一个组群内。每个组群中有一个指定节点 DP(Designated Peer)，DP 是映射到区域虚拟节点 DN 上的。DP 要保存本区域其它 DP 信息以及与相邻区域物理位置相近组群的超链接信息。

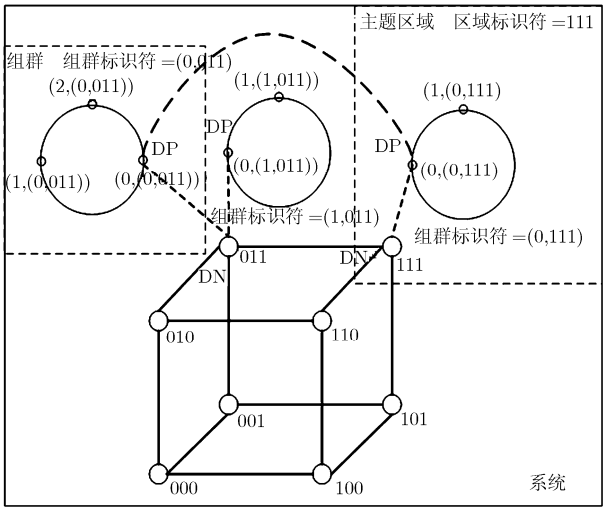

图 3 系统划分及节点命名图

3 TPPH 路由算法

TPPH 结构中主要的路由机制包括主题间路由、主题内路由以及组群内路由。主题间路由在主题覆盖层实现，采用具有较高的路由灵活性的超立方体结构 HyperCup[13]组织；主题内路由通过各组群 DP 间通信进行；组群内则采用灵活及可靠的环状结构，依据 Chord 路由算法进行路由。

3.1 节点路由信息

TPPH 结构中，节点分为 3 类：普通节点，DP 以及 DN。DN 是虚拟节点，所有相关的信息存储在与之对应的 DP 上。对于普通节点和 DP，则对应所属每一主题，都需要维护以下路由信息：

(1)普通节点位于组群内部，它主要维护在组群内部选路的路由信息及与 DP 的连接信息：

(a)Intra-cluster 路由信息：同 Chord 协议选路基本路由表(finger table)，见文献[6]；

(b)DP 信息：本组群 DP 信息。

(2)DP 负责组群内部节点与群外节点的通信，它需要维护的路由信息包括：

(a)Intra-cluster：同一 cluster 内的节点的链接信息，同普通节点 intra-cluster 信息；

(b)intra-domain：同一 domain 内的所有其它组群的 DP 信息；

(c)inter-domain：相邻区域物理位置相近组群的超链接信息，也就是该组群的 DP 信息。

DP 通过 inter-domain 信息与相邻主题区域内物理位置相近的组群相连。这可以使得在主题间路由时一直遵循物理相近原则，提高路由效率。表 1 为图 3 中节点(0,(0,011))也是该组群 DP 的路由信息表。

表 1 节点(0,(0,011))路由信息表

Intra-cluster 路由信息		
successor	Finger table	
	Int.	Succ.
(1,(0,011))	[1,2)	(1,(0,011))
	[2,1)	(2,(0,011))
Intra-domain 路由信息		
DP	Latency	
(0,(1,011))	100	
Inter-domain 路由信息		
(0,(0,111))		

3.2 路由算法

TPPH 路由算法的基本思想是：根据查询消息所属主题类别将消息送至对应的主题区域，然后在主题区域内进行搜索。查找时除依据主题类别限制搜索范围外，还充分利用物理位置相近原则展开。

TPPH 路由算法

假设 P 是当前节点；输入查询消息 Q=Query(message, Topic(Q), Src, TTL)，其中 message 是查询语义描述，对应主题类别标识符 Topic(Q)，Src 是查询消息发起节点，TTL 是该查询消息最大存活时间。

算法描述

(1)P 接收到查询消息 Q=Query(message, key(Q), Src, TTL)；

(2)若消息没有到期，并且查询主题与 P 所处主题相同，则在该主题区域内根据 Intra-cluster 信息进行组群内查找，然后通过 P 所在组群的 DP 根据 Intra-domain 信息将消息送至该主题区域内的其它组群进行查询；

(3)若消息没到期，查询主题与 P 所在主题不相符，则通过 P 所在本组群 DP 根据 inter-domain 信息通过主题间路由机制将查询 Q 发送到符合查询主题的区域；

(4)若消息到期，则丢弃。

3.3 TPPH 路由算法性能分析

路由算法性能主要由它的存储开销、时间开销来衡量。首先给出一些相关定义及假设条件：D 代表超立方体维数；N 代表系统内总主题数；假定查询主题是 d，发起查询的节点主题不同于查询主题；区域 d 节点数为 N_d，组群数为 C_d；区域 d 组群内平均节点数为 N_d^c。

3.3.1 时间开销　TPPH 完成一次查询所需的时间开销 T_{total} 由以下几部分组成：

$$T_{\text{total}} = T_{n,\text{DP}} + T_{\text{inter-domain}} + T_{\text{intra-domain}} \tag{1}$$

其中 $T_{n,\text{DP}}$ 是组群内节点 n 到 DP 的时延；$T_{\text{inter-domain}}$ 是 DP 依据 inter-domain 信息到达符合查询主题的邻居节点的时延，称之为主题间查询时延；$T_{\text{intra-domain}}$ 是主题内查询时延。

定义主题区域间超链接平均时延为 $T_{\text{hyperlink}}$，则根据超立方体结构路由，得

$$T_{\text{inter-domain}} \le \log_D^N \times T_{\text{hyperlink}} \tag{2}$$

区域 d 中组群内部搜索的时延 $T_{\text{intra-cluster}}$ 满足

$$T_{\text{intra-cluster}} \le 2 \times \log N_d^c \times T_{\text{thresh}} \tag{3}$$

记区域不同组群 DP 之间的时延为 $T_{\text{inter-cluster}}$。假设 $\forall i, j < C_d$，$T_{\text{thresh}} < T_{\text{DP}_i,\text{DP}_j} < C_d \times T_{\text{thresh}}$。则

$$T_{\text{inter-cluster}} < C_d \times T_{\text{thresh}} \tag{4}$$

则主题内查询时延

$$\begin{aligned} T_{\text{intra-domain}} &\le \max(T_{\text{inter-cluster}}) + \max(T_{\text{intra-cluster}}) \\ &< (2 \times \log N_d^c + C_d) \times T_{\text{thresh}} \end{aligned} \tag{5}$$

据式(1)，式(2)，式(5)得 TPPH 完成一次查询所需的时间开销：

$$T_{\text{total}} < (2 \times \log N_d^c + C_d) T_{\text{thresh}} + \log_D^N \times T_{\text{hyperlink}} \tag{6}$$

式(6)中第 1 项表示的是主题内路由的时间开销。主题内节点数 N_d 固定的情况下，T_{thresh} 越小，对应组群内节点数 N_d^c 越少，而主题内组群数 C_d 则增大，因此区域内路由的时间开销是由多个互相制约的参数共同决定的。第 2 项表示的是主题间路由的时间开销。主题区域间超链接时延 $T_{\text{hyperlink}}$ 越小，T_{total} 越小，因此 DP 选取与自己有最短时延的相邻主题 DP 做超链接可以有效缩短查询时延。

3.3.2 存储开销　以区域 d 组群 c 为例，组群内普通节点只需要维护 $O(\log N_d^c)$ 个表项的组群路由信息以及本组群 DP 信息。DP 除了需要维护组群内路由信息外，还需维护 intra-domain 和 inter-domain 信息，而它的 intra-domain 信息是该区域 C_d 个组群 DP 信息，inter-domain 信息是 D 个相邻区域物理位置相近组群的 DP 信息。若节点同时属于多个主题，它维护的信息是它位于每个主题需维护信息的总和，即它的存储开销 $S = \sum S_i$，S_i 代表节点在所属第 i 个主题内的存储开销。

3.4 节点动态加入和退出

TPPH 结构中，节点加入前并不是按照自己是普通节点，DP 或者 DN 加入的，而是在加入系统的过程中由系统当前情况决定节点的类别。

(1)节点 n 加入系统的过程描述如下：

(a)根据自己所属的主题类别 T_i 确定节点标识符，通过引导节点查找标识符为 b_i 的超立方体节点 DN，若该 DN 不存在，则创建主题标识符为 b_i 的 DN,，将节点 n 绑定到该 DN；

(b)返回该 DN 对应的所有组群 DP 地址信息到节点 n；节点 n 根据到 DP 的时延确定加入组群的位置，建立 intra-

cluster 信息列表，加入过程结束；否则，节点 n 做为 DP 构建新的组群；

(c)创建节点 n 到该区域内其它组群 DP 的链接，建立 intra-domain 信息列表；

(d)创建节点 n 到相邻区域内最近组群 DP 的链接，建立 inter-domain 信息列表。

(2)节点 n 退出系统的过程描述如下：

(a)在 cluster 内部按照 chord 结构退出原则进行退出操作，若节点是该组群的 DP，还需要进行 DP 工作交接；若该节点是组群内的最后一个节点，则撤销该组群；

(b)若该组群是区域上的唯一组群，则按照超立方体结构节点退出原则执行主题 DN 的退出操作。

3.5 TPPH 路由优化策略

3.5.1 BDP(Backup DP)机制 在系统运行的过程中，DP 根据自身情况，如带宽、处理速度等信息，主动地在本组群内建立备用 DP(简称 BDP)。建立 BDP 之后，DP 要将路由信息发送到各个 BDP 上，并通知组群内节点 BDP 的位置信息。组群内节点在进行主题间路由时，可以任意选择 DP 或者 BDP 作为它的第 1 跳节点。BDP 机制可以有效地进行数据分流，减轻了 DP 的负担。同时，在 DP 非正常退出时，可以迅速的进行角色切换。

3.5.2 small world[14]特性的应用 Small world 特性从网络层面来讲是指网络中存在大量的短链，使得任意两个节点可以通过这些短链连接起来，从而大大的降低两点间的距离。TPPH 结构中，具有多个主题的节点将自己所属主题按照 bloom filter[15]机制形成主题向量，在所属组群内发布。节点发起查询需要进行主题间路由时，首先在主题向量内查找具有目标主题的多主题节点，然后通过多主题节点直接到达目标主题区域，而不再需要通过超立方体的主题间路由，从而大大降低了路由时延。

4 仿真实验

在仿真实验中，使用 GT-ITM[16]拓扑发生器以 TS[17]模型生成网络节点规模为 1000 的网络随机拓扑图。采用查全率、搜索平均时延、查询产生的平均消息数这 3 种指标对语义查询性能进行评价。查全率是指搜索到的符合条件的文档占网络中所有相关文档的比例[18]。以搜索时延与查全率的比值来衡量搜索平均时延。以查询产生的消息数与查全率的比值衡量查询操作引起的网络负载的变化。

为测试 TPPH 拓扑的有效性，以 $T_{\text{thresh}}=40$，70 和 100 的 TPPH 与 Gnutella 网络进行对比。按照参数为 $\alpha=1.2$ 的 Zipf 分布[19]为每个节点分布主题和文档。假设每个节点最多有 4 个主题，每个主题下最多有 100 个文档。所有结果是在查询次数为 100000 次下的平均值。

图 4(a)所示为 TTL 从 1 到 10 对应的查全率。可以看出，在 TTL≤6 时，TPPH 较之 Gnutella 查全率有了显著的提高。TTL=3 时，最高可达 75%，而在 TTL>7 之后，Gnutella 的查全率高于 TPPH 将近 20%。

图 4(b)显示了在相同查全率的情况下，TPPH 和 Gnutella 的平均查询时延对比。可以看出，TTL<5 时，TPPH 比 Gnutella 在相同查全率的情况下查询时延平均降低到原来的一半，最好情况下达到 40%。而在 TTL 超过 7 之后，Gnutella 的查询范围基本遍历整个网络，从而大大的增加了查全率，从而对比 TPPH 降低了每查全率的平均查询时延。

图 4(c)显示了 TTL 从 1 到 10 在相同查全率的情况下查询产生的平均消息的数量。结合图 4(a)可以看出，Gnutella 的消息数量是以指数级别增长的，而 TPPH 则是线性增长，所以 TPPH 能显著的降低 P2P 网络的消息负载。

另外，由图 4 可以看出 T_{thresh} 为 40 时查全率较高，平均查询时延较低，但对应的消息数也是最高的。因此在划分主题内组群时，需要确定合理的参数，使得在该条件下系统具有较高的查全率、较低的查询时延以及较少的消息数量。

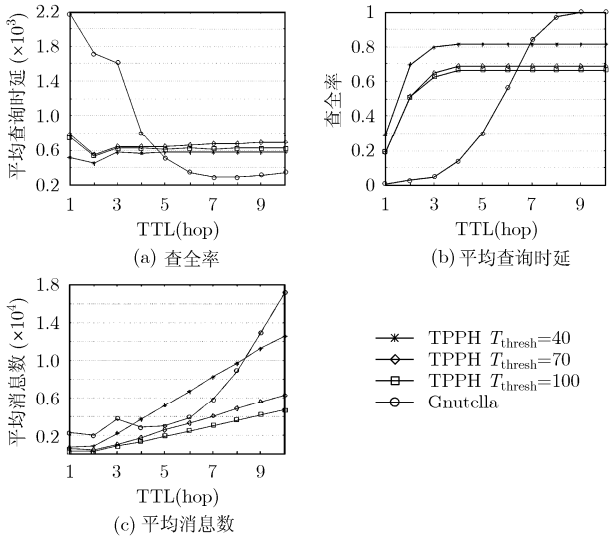

图 4 TPPH 与 Gnutella 的性能对比

5 结束语

语义覆盖网络结构是当前语义 P2P 研究的热点，TPPH 不存在任何索引及超级节点的概念，是利用了结构化 P2P 高效的定位性能同时结合物理位置相近性原则建立的语义覆盖网络结构。它通过基于 DHT 路由的主题区域划分及相对物理位置的组群划分机制，将结构化 P2P 高效的路由与语义 P2P 的复杂检索功能融合，为语义对等网络提供了一种基于复杂查询条件的、有效搜索数据对象的方法。仿真结果表明，在 TTL 较小情况下，TPPH 结构能够显著的提高系统查询性能。

参 考 文 献

[1] Haase P, Agarwal S, and Sure Y. Service-oriented semantic

peer-to-peer systems. In Web Information Systems Engineering--Workshop Proceedings, Australia, 2004, 11: 46–57.

[2] Crespo A and Garcia-Molina H. Semantic overlay networks for P2P systems. Technical report, Stanford University (2002).

[3] Noervaag K, Doulkeridis C, and Vazirgiannis M. Semantic overlays for P2P web searching. Technical report, AUEB, 2005.

[4] Napster Inc. The Napster Homepage. http://free.napster.com/, 2001.

[5] Wolpers M, Siberski W, and Schmitz C, *et al*.. Super-peer-based routing and clustering strategies for RDF-based peer-to-peer networks. WWW 03, 2003: 536–543.

[6] Stoica I, Morris R, and Karger D, *et al*.. Chord: A scalable peer-to-peer lookup service for Internet applications. In Proceedings of ACM SIGCOMM'2001, San Diego, CA, August 2001, 31(4): 149–160.

[7] Ratnasamy S, Francis P, and Handley M, *et al*.. A scalable content addressable network. In Proceedings of ACM SIGCOMM'2001, San Diego, CA, August 2001, 31(4): 161–172.

[8] Siebes R. Pnear: combining content clustering and distributed hash tables. In The Second International Workshop on to-Peer Knowledge Management (P2PKM05), San Diego, CA, 2005. http://www.p2pkm.org/ downloads/ Siebes2005.pdf.

[9] Gnutella Inc. The Gnutella Homepage. http://www.gnutella.com/, 2001.

[10] Gkantsidis C, Mihail M, and Saberi A. Random walks in peer-to-peer networks. In: Proc. of the IEEE INFOCOM 2004. New York: IEEE Press, 2004: 120–130.

[11] Ricardo B Y and Berthier R N. Modern Information Retrieval. New York: Addison Wesley, 1999: Chapter 2.

[12] Witten I and Frank1 E. Data mining: Practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann, 1999: Chapter 4.

[13] Schlosser M, Sintek M, and Decker S, *et al*.. HyperCuP—hypercubes, ontologies and efficient search on P2P networks. In International Workshop on Agents and Peer-to-Peer Computing, Bologna, Italy, July 2002: 112–124.

[14] Kleinberg J. The small-world phenomenon: an algorithmic perspective. Cornell Computer Science Technical Report, 99–1776, 2000.

[15] Andrei B and Michael M. Network applications of bloom filters: A survey. *Internet Mathematics*, 2003, 1(4): 485–509.

[16] Georgia Institute of Technology. Modeling Topology of Large Internetworks. http://www.cc.gatech.edu/ projects/gtitm/, 2000.

[17] Palmer C R and Steffan J G. Generating network topologies that obey power laws. In: Kero TEF, ed. Proceedings of the IEEE Global Telecommunications Conference, San Francisco, CA: IEEE Computer Society Press, 2000: 434–438.

[18] 陈汉华, 金海等. SemreX:一种基于语义相似度的 P2P 覆盖网络. 软件学报, 2006, 17(5): 1170–1181.

[19] Huberman B A and Adamic L A. Zipf's law and the Internet. Glottometrics 3, 2002, 3: 143–150.

于　婧: 女，1979 年生，博士生，研究方向为对等网络体系结构及路由.

张建辉: 男，1977 年生，博士生，研究方向为 IP 网络路由协议.

顾小卓: 女，1978 年生，博士生，研究方向为 IP 网络安全.

汪斌强: 男，1963 年生，教授，博士生导师，主要研究方向为宽带 IP 网络.