

基于邻近度的 P2P 路由算法^{*)}

王晓斌 卢显良 侯孟书 周 旭

(电子科技大学计算机科学与工程学院 成都 610054)

摘 要 邻近度路由技术对 P2P 网络的路由性能影响很大。本文提出一种新的分布式结构化 P2P 路由算法 FD-PNS。FD-PNS 路由算法采用邻近度路由技术,节点在转发一个查询请求后触发路由表的维护任务,有针对性地对本次转发使用的路由表项进行优化,以减小搜索和定位的路由延迟,提高系统的整体性能。

关键词 对等网络,邻近度,路由算法

Proximity-based P2P Routing Algorithm

WANG Xiao-bin LU Xian-liang HOU Meng-shu ZHOU Xu

(School of Computer Science & Engineering, UESTC, Chengdu 610054, China)

Abstract Recent work confirms that this property of Proximity Neighbor Selection(PNS) has a significant impact on peer-to-peer network routing performance. This paper presents a novel effective P2P route algorithm: FD-PNS. The PNS technology is adopted by the FD-PNS algorithm. In FD-PNS, the routing table maintenance task is invoked after a node has forwarded a request and the entry used at this time is optimized. So this decreases the latency of searching and locating file, and improves the performance of whole system.

Keywords Peer-to-peer network, Proximity, Routing algorithm

1 引言

在结构化 P2P 网络中,节点之间通过弱连接形成 overlay 网络(overlay network)。节点平均在 $O(\log N)$ 跳内就可完成对象的定位操作,因此具有良好的可扩展性,非常适合构建大规模的网络应用。

然而在 overlay 网络中相邻的节点,其在底层网络中可能并不相邻甚至相隔很远,这样导致 overlay 网络中两个节点间的路由延迟与其在底层网络中的路由延迟有很大的差异。在 overlay 网络中相邻节点间的一跳(one hop),可能在底层网络中需要跨越多跳才能到达目的地。因此,在 overlay 网络中,节点路由表的建立和维护十分关键:只有节点路由表项的内容正确地反映节点之间在底层网络中的拓扑关系,才能最终减少应用层的路由延迟,提高 P2P 存储系统的性能。

Pastry^[1]是经典的结构化 P2P 路由算法。Pastry 中原有的周期性随机选择路由表项进行维护的方法虽然也引入了邻近度邻居选择技术,但是这种路由表维护算法存在以下几个缺点^[2-5]:

盲目性:随机选取路由表中的路由表项进行优化的方式缺乏针对性,比较盲目。

优化周期长:由于优化周期长度固定不变,路由表无法以最快的速度将自己调整到优化的状态。

优化幅度较小:该方法从自己的路由表中随机地选取一项(设该项处在第 i 行),比较自己到该项路由表中第 i 行的各项的距离。如果距离更近,则用新的节点替换自己路由表中相应项的内容。这样每次的节点数都较少,优化幅度较小。

针对以上几个问题,设计了基于邻近度的路由表维护算法 FD-PNS(Forward-Driven PNS)。

2 FD-PNS 算法的设计思想

在 DHT 覆盖网络中,节点的路由表记录了在上层网络

中与自己直接相邻的节点的位置信息,系统消息依靠在节点间不断转发实现最终的定位。如果节点路由表中记录的逻辑相邻节点在底层网络中的距离也是邻近的话,同样次数的逻辑层跳转跨越的实际底层距离将大大缩短,带来路由性能的提高^[6,7]。基于这种思路设计了 FD-PNS 路由表维护算法。

FD-PNS 算法的基本设计思想是节点在转发一个查询请求后触发路由表的维护任务,有针对性地对本次转发使用的路由表项进行优化。如果某节点正好处于通向一个热门文件的路由路径上,这时路由表中的某项可能会被频繁地使用,这种由转发驱动的路由表维护算法使该路由表项能得到及时的优化,避免了随机选取路由表项进行优化的盲目性。此外,一个路由表的某项经过数次优化之后将达到或者最大程度地接近最优值,这时再启动路由表的优化算法也难以进一步优化该项,反而会增加节点和网络的开销。为了避免这种情况,FD-PNS 算法采取可变周期地运行路由表维护任务来不断地优化路由表项的 proximity 性质,这样也比定长周期的优化调整到优化状态的速度更快。同时为了判断节点是否活跃,并且减少失效节点在一定时间内被修复后再次加入网络所带来的额外开销,FD-PNS 算法定期发送检测消息以检查节点是否活跃,并根据具体情况维护路由表,使查询消息能正确地被路由到目的地。

3 FD-PNS 原理

FD-PNS 路由表维护算法分为三个部分:节点的加入、后期维护、节点失效的判断。

3.1 节点的加入

系统使用分布式散列函数为新加入的节点分配一个 nodeID。这时,节点的路由表(routing table),叶子集合(leaf set),邻居集合(neighborhood set)均为空。系统需要为该节点构造路由表、叶子集合和邻居集合,并通知系统中的其他节

^{*)} 电子信息产业发展基金资助项目,编号:[2006]117;电子科技大学青年基金资助。王晓斌 博士生,副教授,研究方向为计算机网络、数据挖掘、P2P 计算;卢显良 教授,博士生导师,研究方向为计算机网络、操作系统;侯孟书 博士,副教授,研究方向为分布式存储、P2P 计算。

点该新节点的到来。

设新节点为 X。X 首先向一个在底层网络中靠近自己的系统节点 A 发出加入系统的消息,这个加入消息的目的地址被设为 X。A 收到这个加入消息后,将这个消息按照 Pastry 路由算法向目的地 X 转发,最后加入消息将到达一个 nodeID 最接近 X 的系统节点 Z。在从 A 到 Z 途中遇到的每个节点都将自己的路由表发给 X, X 使用路径中遇到的第 i 个节点的第 i 行路由表信息填充自己的路由表的第 i 行。完成路由表构造后, X 向自己路由表中第 i 行的所有节点发送该行的内容,该操作旨在向这些节点通知 X 的加入,并在节点间交换彼此的路由信息。这些节点收到 X 发来的路由表项后,测量自己到表项中各节点的距离,若到其中某个节点的距离比自己路由表中对应项的距离近,则用该节点替换自己路由表里的相应节点。由于 A 是距离上最靠近新节点 X 的节点,因此使用 A 的邻居集合初始化自己的邻居集合。而 Z 是系统中 nodeID 最接近 X 的节点,所以使用 Z 的叶子集合来初始化自己的叶子集合。至此,节点 X 完成自己路由表的初始化并通知必要的系统节点,加入了系统。

3.2 后期维护

节点加入系统后,还需要不断地优化路由表,使其中各项始终指向距离自己最近的节点。

首先定义一些符号:

R_n 代表 nodeID 为 n 的节点的路由表, $R_n(i, j)$ 代表路由表中第 i 行第 j 列表项; $d(n, m)$ 代表两个节点 n 和 m 之间的距离。

设节点 n 收到对标识符为 k 的对象的查询请求, n 查询 R_n 并选择表中节点 m 转发消息。设 $shl(n, k) = r$, 则 $m = R_n(r, kr)$ 。

由路由表的构造方法可知:

对 $\forall i > r+1, j \in [0, 2^b-1]$, 有 $shl(R_n(i, j), n) = r, R_n(i, j)r+1 = R_n(r, kr)r+1$

即 R_n 中 $r+1$ 行以后表项的 nodeID 都符合 R_n 中 m 所在项的要求,可以用来替换 R_n 中的 m 。FD-PNS 正是利用了这个特点,从 R_n 中 $r+1$ 行以后的表项中寻找比 m 距离 n 更近的节点,来优化表项 $R_n(r, kr)$ 的内容。

3.3 节点失效的判断

为了判断节点是否存在未报告的失效或者异常退出, FD-PNS 算法中每个节点定期给其路由表中的所有节点发一个 Ping 消息,以判断这些节点是否依然活跃可用于路由。

4 实验及其相关分析

为了测试 FD-PNS 路由表维护算法对较大规模系统环境下路由性能的改变,我们通过编写仿真程序,模拟一个有 1000 个节点的广域网环境中的 FD-PNS 测试系统。模拟系统中的 $b=4, |L|=16$ 。

4.1 路由性能

在 DHT 覆盖网络中,文件查询消息是通过在覆盖网络的节点间转发实现文件定位,这使得消息沿 overlay 路由路径走过的距离要比源节点和目的节点在底层网络中的直接距离要长。用“latency stretch”表示这两个距离之间的比值,它的值永远大于等于 1。

对实验系统从初始状态开始,分别使用周期性随机选择路由表项优化的方法和 FD-PNS 算法进行节点路由表的优化操作。在系统运行过程中,节点随机地发起文件查询请求,记录系统的平均 latency stretch 值随系统运行时间的变化趋势。

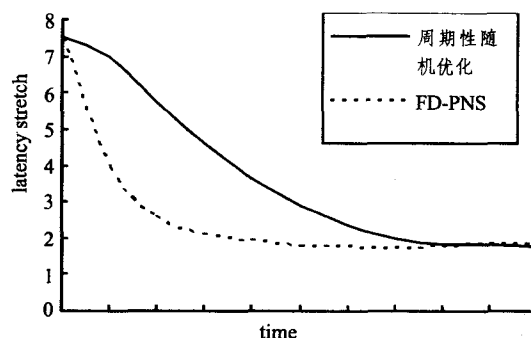


图1 latency stretch 随时间变化图

图1是系统的 latency stretch 值随运行时间的变化趋势。从图中可以发现两种算法的优化性能的差异。周期性随机优化算法由于采用周期性随机选择路由表项进行优化,优化效果逐步累积, latency stretch 变化曲线下降速率相对稳定,路由表达到最优状态的速度较慢。而对 FD-PNS 算法来说,在系统运行初期,节点每次转发都将触发对路由表的优化操作,针对性强,且每次都测量较多节点,优化幅度较大。节点路由表以较快的速度被调整到最优值,表现在 latency stretch 曲线上呈前陡后缓的变化趋势,比使用周期性随机调整的算法提前达到稳定状态。

4.2 优化开销测试

为了不断地优化路由表的 proximity 性质,节点需要不断地测量自己到某些节点的距离,以发现更接近自己的节点,替换路由表中相应项。我们将节点发起的对其他系统节点的距离测量操作次数作为衡量路由表优化算法的系统开销的指标。通过记录一个新加入系统的节点距离测量操作次数随系统运行时间的变化情况,我们对周期性优化算法和 FD-PNS 两种优化算法的系统开销进行了比较,见图2。

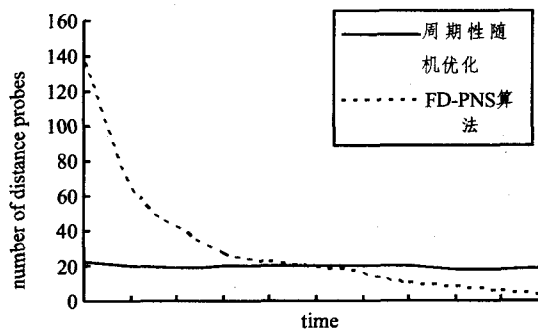


图2 优化开销随时间变化图

从图2中可以知道,周期性维护算法的优化开销在文件的整个生命周期里保持相对稳定,且维持在较低的水平。而 FD-PNS 算法在节点加入系统的初期为了达到快速优化节点路由表的目的,每一次路由转发都会触发优化算法,且每次需要测量的节点数都比较多,所以优化开销较大。但是,初期优化开销增加带来的好处是优化效果明显,节点路由表迅速被调整到最优状态。根据 FD-PNS 算法的设计,节点每次测量的相邻节点数目将随着优化次数增加而减少,故优化开销曲线下下降迅速,很快就降到与周期性优化算法同样的水平。在路由表达到最优状态后,FD-PNS 优化操作启动的频度将逐渐降低,优化开销曲线继续下降,低于周期性优化算法。到后期,FD-PNS 只需要极少的开销即可保持节点路由表最优。

结束语 邻近度路由技术是结构化 P2P 网络中路由表

构造和维护的最常用技术,本文在对 Pastry 路由表维护算法进行深入研究和分析的基础上,设计了一种新的基于邻近度的 P2P 路由表维护算法 FD-PNS。PNS-PGrid 对 P-Grid 的路由表构造进行了改进,并引入了节点就近加入机制、节点失效处理机制和路由表维护机制,从而使 P2P 网络逻辑拓扑尽可能反映底层网络的物理拓扑。仿真实验结果表明,PNS-PGrid 算法具有较小的延迟延伸率,从而有效地减少了路由延迟,提高了路由性能。

参考文献

- [1] Rowstron A, Druschel P. Storage management and caching in PAST, a large-scale, persistent peer-to-peer storage utility//

(上接第 31 页)

弈的 Nash 均衡理论来分析和优化网络资源分配的方法受到了众多学者的关注。因为市场机制中的博弈理论非常适合解决分布式网络资源分配问题,所以在研究 P2P 网络的对称点激励机制时博弈论也得到了大量的使用。但 P2P 环境区别于其他分布式系统的一个重要因素就是其网络环境受到网络参与者的心理活动的影响。如果能将对用户节点的行为和兴趣趋向进行分析,并合理利用经济学理论中对各种心理模型已有的大量成果,结合博弈论来建立适当的激励机制将是非常有意义的。另外博弈中迭代算法的高效性和效用函数的有效性是影响博弈应用的关键因素,应根据 P2P 流媒体系统的内在特征来研究这两个问题,目前的文献中相关研究还不够。

(3)同样地,在当前的 P2P 流媒体激励机制中,从对称点安全性方面出发的考虑较少。本文提到的激励机制几乎没有涉及到安全方面的问题,但实际上在所有 P2P 的协作体系中,安全保障都扮演着重要角色,例如采用信誉机制可能带来串通舞弊的问题。围绕如何评估共享资源的真实性和可靠性,研究如何提供数字证书、对称节点认证、安全存储等,具有重要意义。

(4)已有研究^[7,30]将激励机制和 P2P 流媒体覆盖网络的构建联系起来。P2P 流媒体覆盖网络的构建有很多需要考虑的问题,比如初始父节点的选择、节点失效时拓扑重建、底层物理网络拓扑匹配等。文献[30]中利用 P2P 流媒体的特性构造了一个树型的覆盖网络,但树型拓扑结构有着天然的缺陷,如对网络变化的适应性差、带宽递减、叶子节点不做贡献等。将 free-riding 问题的解决和网络的特性以及网络拓扑生成相结合,形成一个更优的覆盖网络构建机制还需要进一步的研究。

参考文献

- [1] Figueiredo D R, Shapiro J K, Towsley D. A public good model of availability in peer-to-peer systems. Technical Report 04-27. CSE Dept, Michigan State University, 2004
- [2] Veciana G d, Yang X. Fairness, incentives and performance in peer-to-peer networks. Allerton Conference on Communication, Control and Computing, 2003. www.cs.huji.ac.il/labs/danass/p2p/resources/fairness-incentives-performance.pdf
- [3] Bernardo EA, Huberman A. Free riding on Gnutella. Technical Report, SSL-00-63. Palo Alto: Xerox PARC, 2000
- [4] Asvanund A, Clay K, Krishnan R, et al. An empirical analysis of network externalities in peer-to-peer music-sharing networks. Information Systems Research, 2004, 15(2):155-174
- [5] Krishnan R, Smith M D, Telang R. The economics of peer-to-peer networks. Journal of Information Technology Theory and Application (JITTA), 2003, 5(3):31-44
- [6] Hardin G. The tragedy of the commons. Science, 1968, 162: 1243-1248
- [7] Habib A, Chuang J. Incentive mechanisms for peer-to-peer media streaming//Int'l Workshop on Quality of Service (IWQoS). Los Alamitos, CA: IEEE Computer Society Press, 2004:171-180
- [8] Golle P, Leyton-Brown K, Mironov I. Incentives for sharing in peer-to-peer networks. ACM Electronic Commerce (EC '01),

Proceedings of the 18th ACM Symposium on Operating Systems Principles. Banff, 2001:188-201

- [2] Aberer K. P-Grid: A self-organizing access structure for P2P information systems
- [3] Zhang H, Goel A, Govindan R. Incrementally improving look-up latency in distributed hash table systems In ACM Sigmetrics. California, USA, 2003
- [4] Plaxton C G, Rajaraman R, Richa A W. Accessing nearby copies of replicated objects in a distributed environment. Theory of Computing Systems, 1999, 32:241-280
- [5] Ratnasamy S, Handley M, Karp R, et al. Topologically-aware overlay construction and server selection//Proceedings of IEEE INFOCOM '02. 2002
- [6] 周旭. 面向 Internet 的大规模分布式存储技术研究. 博士论文. 电子科技大学, 2004
- [7] 侯孟书. 基于 P2P 的分布式存储研究. 博士论文. 电子科技大学, 2005

Tampa, Florida, Oct. 2001:264-267

- [9] Ma R T B, Lee S C M, Lui J C S, et al. A game theoretic approach to provide incentive service differentiation in P2P networks. ACM SIGMETRICS, New York, June 2004:189-198
- [10] Feldman M, Lai K, Stoica I, et al. Scalable and robust incentive techniques for P2P networks//ACM Conference on Electronic Commerce (EC '04). New York, May 2004:102-111
- [11] Ranganathan K, Ripeanu M, Sarin A, et al. To share or not to share: An analysis of incentives to contribute in collaborative file sharing environments//Workshop on Economics of Peer-to-Peer Systems, Berkeley, California, June 2003. <http://www2.sims.berkeley.edu/research/conferences/p2pecon/papers/sl-ranganathan.pdf>
- [12] Buragohain C, Agrawal D, Suri S. A game theoretic framework for incentives in P2P systems//Proc. of the 3rd Int'l Conf. on Peer-to-Peer Computing (P2P 2003). Los Alamitos: IEEE Press, 2003:48-56
- [13] Chu Y, Aganjam A, Ng T, et al. Early Experience with an Internet Broadcast System//USENIX Annual Technical Conference. 2004:155-170
- [14] Banerjee S, Bhattacharjee B, Kommareddy C. Scalable application layer multicast. In ACM Sigcomm, 2002:205-220
- [15] Kosti D, Rodriguez A, Albrecht J, et al. Bullet: high bandwidth data dissemination using an overlay mesh //Proceedings of the nineteenth ACM SOSP. 2003:282-297
- [16] hua Chu Y, Zhang H. Considering altruism in peer-to-peer internet streaming broadcast. In NOSSDAV, 2004:10-15
- [17] hua Chu Y, Chuang J, Zhang H. A case for taxation in peer-to-peer streaming broadcast//Workshop on Practice and theory of incentives in networked systems (PINS). 2004:205-212
- [18] Yuen S, Li B. Strategyproof mechanisms for dynamic multicast tree formation in overlay networks. In IEEE Infocom, Mar. 2005:2135-2146
- [19] Jannotti J, Gifford D K, Johnson K L, et al. Overcast: Reliable multicasting with an overlay network. In ACM OSDI, 2000: 197-212
- [20] Castro M, Druschel P, Kermarrec A, et al. SCRIBE: A large-scale and decentralized application-level multicast infrastructure. IEEE JSAC, 2002, 20:1489-1499
- [21] Padmanabhan V, Wang H, Chou P. Resilient Peer-to-peer Streaming. Microsoft Research, Tech. Rep. MSR-TR-2003-11. 2003
- [22] Castro M, Druschel P, Kermarrec A, et al. SplitStream: High-bandwidth Content Distribution in Cooperative Environments //Proceedings of SOSP. 2003:298-313
- [23] Ye Song, Makedon F. Collaboration-Aware Peer-to-Peer Media Streaming. ACM Multimedia 2004, 2004:412-415
- [24] Wilcox-O'Hearn B. Experiences deploying a large-scale emergent network// Druschel P, Kaashoek F, Rowstron A, eds. Proc. of the 1st Int'l Workshop on Peer-to-Peer Systems. Berlin: Springer-Verlag, 2002:104-110
- [25] Vishnumurthy V, Chandrakumar S, Sier E. Karma: A secure economic framework for p2p resource sharing // 1st Workshop on Economics of Peer-to-Peer Systems, June 2003. <http://www.cs.cornell.edu/People/egs/papers/karma.pdf>
- [26] Wang W, Li B. Market-driven Bandwidth Allocation in Selfish Overlay Networks//Proc. of INFOCOM 2005. 2005:2578-2589
- [27] Tan G, Jarvis SA. A Payment-based Incentive and Service Differentiation Mechanism for Peer-to-Peer Streaming Broadcast//14th IEEE International Workshop on Quality of Service (IWQoS-06). Yale University, New Haven, CT, USA, 2006:41-50
- [28] Rodriguez P, Tan S-M, Gkantsidis C. On the feasibility of commercial, legal p2p content distribution. In ACM/SIGCOMM CCR, 2006:75-78
- [29] Keidar I, Melamed R, Orda A. EquiCast: Scalable Multicast with Selfish Users. In PODS, 2006:63-71
- [30] Shrivastava V, Banerjee S. Natural Selection in Peer-to-Peer Streaming: From the Cathedral to the Bazaar//Proceedings of ACM NOSSDAV. Washington, USA, June 2005:93-98
- [31] Mannak R, de Ridder H, Keyson D V. The Human Side of Sharing in Peer-to-Peer Networks//2nd European Symposium on Ambient Intelligence. Eindhoven, The Netherlands, 2004:59-64