文章编号:1001 - 9081(2009)10 - 2652 - 03

基于 Gossip 协议的流媒体播放机制的研究与改进

乔志伟,彭 俊,徐汀荣
(苏州大学 计算机科学与技术学院,江苏 苏州 215006)
(qiaozhiwei353@126.com)

摘 要:针对 Gossip 协议数据传播随机性与不确定性问题,提出了一种混合结构方式。该方式将非结构化 P2P 网络和结构化 P2P 网络相结合,通过增加数据片预取调度算法,弥补了 Gossip 协议数据传输的随意性。实验结果表明,此方式提高了节点成功播放率,降低了系统开销。

关键词:对等网络;流媒体;Gossip 协议;分布式散列表
中图分类号: TP393.02 **文献标志码**:A

Research and improvement of streaming media system based on Gossip protocol

QIAO Zhi-wei, PENG Jun, XU Ting-rong
(*School of Computer Science and Technology, Soochow University, Suzhou Jiangsu* 215006, *China*)

Abstract: The Gossip-based P2P streaming media system has become the mainstream of P2P streaming media systems. But the inherent drawbacks of Gossip protocol—the randomness and uncertainties of data dissemination influence the P2P streaming media system. Based on the Gossip, an approach combining unstructured and structured P2P network was proposed to cover the shortage of the data dissemination based on Gossip and improve the efficiency of streaming media system. Finally the approach has been proved to be reliable through simulation.

Key words: Peer-to-Peer (P2P); streaming media; Gossip protocol; Distributed Hash Table (DHT)

0 引言

随着流媒体技术的飞速发展,P2P 流媒体技术越来越受到人们的关注。相对于传统的集中式客户/服务器(C/S)模型,P2P 弱化了服务器的概念,系统中的各个节点不再区分服务器和客户端的角色关系,每个节点既可请求服务,也可提供服务,节点之间可以直接交换资源和服务,而不必通过服务器。用户可以根据他们的网络状态和设备能力,与一个或几个用户建立连接来分享数据。这种连接能减少服务器的负担和提高每个用户的视频质量,即使是大量的用户同时访问流媒体服务器,也不会造成服务器因负载过重而瘫痪。

近年来,基于 Gossip 协议[1]的 P2P 流媒体系统已成为 P2P 流媒体系统的主流。但 Gossip 协议本身的随机性与不确定性,造成了不能保证高播放连续度。因此,我们基于传统 Gossip 协议的流媒体传播机制,根据控制拓扑模型和数据拓扑模型,提出改进的 CGossip 流媒体传播方法,即采用非结构化 P2P 网络与结构化 P2P 网络相结合的方法,在非结构化 P2P 覆盖网之上进行数据调度,在结构化 P2P 覆盖网之上进行数据预取,弥补了 Gossip 协议数据传播随机性的缺陷,从而保证流媒体系统能保持高播放连续度。

1 P2P 流媒体

P2P 流媒体系统的划分有很多种,按照数据的接受和传输方式,将 P2P 流媒体应用分为 3 类:第一类为单发送端多接受端,比如 Narada[2];第二类为多发送端单接受端,比如微软研究院开发的 CoopNet[3]系统;第三类为多发送端多接受端,比如 BitTorrent。而按其工作方式可分为两类:一类是基于树状多播,其中典型的有 ZigZag[4]及 SplitStream[5]结构模型;另一类是基于 Gossip 协议,典型的是实时流媒体播放系统 CoolStreaming[6]。

基于树状多播的 P2P 流媒体系统将网络中所有节点组织成一棵多播树,树的根节点是媒体发布源,数据分片总是从多播树的父节点向其子节点传播直到叶节点。基于多播树的方法可以最小化系统中多余的数据传播,并能保证每个数据分片可以传播到系统中每个节点,但它的缺陷是:除叶节点以外任何一个系统节点的失效都将导致多播树分裂为两棵,而其中一棵在分裂后不能得到任何数据。因此,多播树极易分裂,且维护多播树的开销巨大,造成树状多播方法不适合于高动态性的因特网环境。另外,树状多播方法的带宽利用率一般较低,原因有两方面:一方面是多播树的叶节点只下载不上传,是纯粹的带宽消费者,对系统没有贡献;另一方面是多播树的父节点限制了其所在子树的最大输入带宽,因此多播树中带宽瓶颈节点到处存在。

从本质上看,基于树状多播的方法实际上是属于 P2P 流媒体系统沿袭传统流媒体技术的一种过渡。当前,基于 Gossip 协议的 P2P 流媒体机制是研究热点。

2 Gossip 协议

Gossip 协议是一种洪泛信息发布机制。它的工作原理是:系统中每个节点维护一定数量的邻居信息,被发布的信息体在源端被分成许多小的信息片断,随机把这些信息片段发布给组通信中的部分主机,此时各个主机可能拥有一部分信息体或者没有信息;然后主机之间相互交换彼此没有的信息,最终所有的主机都得到完整的信息体。

相比树状多播方法,基于 Gossip 协议的网状多播方法的优势主要表现在高容错性与高带宽利用率两方面。邻居维护

收稿日期:2009 - 04 - 09;**修回日期**:2009 - 06 - 04。

作者简介:乔志伟(1984 -),男,江苏无锡人,主要研究方向:对等网络、网络安全; 彭俊(1986 -),男,湖北仙桃人,硕士研究生,主要研究方向:对等网络; 徐汀荣(1958 -),男,江苏苏州人,教授,主要研究方向:网络、数据库、数据挖掘。

的灵活性与数据传播的随机性使得基于 Gossip 协议的 P2P 流媒体系统不会因节点失效而导致显著的性能下降,从而良好地适应了高动态性的因特网环境。基于 Gossip 的系统以其出色的播放效果和较低的延迟已经在实际运行中得到了一定的证实。

因此,基于 Gossip 协议的 P2P 流媒体系统更适用于大规模、高动态的 Internet 网络环境,凭借本身的技术优势成为 Internet 流媒体领域的主流。但是,传统的 Gossip 协议仍然存在缺点:随着发送消息数目的增加,消息扩散的负载也随之增加;当网格规模很大时,消息发送的失败率会明显增加。

3　CGossip 设计

针对 Gossip 协议的不足,我们提出了改进的 CGossip 流媒体播放机制,即采用基于分布式散列表(Distributed Hash Table,DHT)的数据预取方法来弥补 Gossip 协议传播数据的缺陷,从而改善了流媒体系统,使之能保持高播放连续度。基于 DHT 的定位功能,从媒体源发布出来的每一个数据分片都会被分布式地存储在 k 个网络节点中。系统中每个节点不断预测哪些数据分片可能被数据调度算法遗漏,如果有认为遗漏的数据分片,就启动数据预取算法通过 DHT 快速查找并获得它们,从而保证连续播放。

CGossip 的机制主要由 3 部分构成:覆盖网的构造、数据调度算法和数据预取算法。CGossip 系统的节点软件架构如图 1 所示。其中包括:1)网络接口,即节点与覆盖网之间的接口;2)缓冲区,其中的数据按照等时间间隔的数据块存放,方便调度;3)调度器,负责从建立连接的网络邻居中周期性地获取数据可用性信息,从而安排到哪里去取哪些数据分片;4)播放器,用来播放流媒体;5)备份器,负责当前节点数据分片的备份,只要当前节点正常工作,其他节点就能从当前节点的按需数据备份中获取它所负责存储的数据分片;6)组间关系器,负责对节点的加入、退出、失效等进行处理。

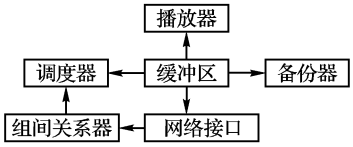

图 1　CGossip 系统的节点软件架构

3.1　覆盖网的构造

当前非结构化 P2P 网络具有网络拓扑简单、实现难度小、高容错性和良好的自适应性等特点,所以目前基于 Gossip 协议的 P2P 流媒体系统通常采用非结构化的 P2P 网络作为其底层结构。但是,由于运行在非结构化 P2P 网络上的 Gossip 协议固有的随机性,不能保证高播放连续度,本文设计了一个混合式 P2P 覆盖网,组合了非结构 P2P 覆盖网和结构化 P2P 覆盖网。

结构化 P2P 覆盖网是用来实现高效数据定位的 DHT。首先,基于 DHT 的分布式算法,通过分布式散列 Hash 函数,将节点和数据映射到一个常规地址空间(见图 2);然后通过路由算法查找该节点。采用 DHT 结构有明显的优点:一方面能够实现节点的动态加入和退出,保证节点之间的均匀性和自组织能力;另一方面实现对目标节点的快速路由发现和路由查找,减少状态维护开销和转发开销。比较典型的例子是 Chord[7]、Pastry[8]、Can[9],它们之间的区别在于具体的路由策略和发现方式不同。

在 CGossip 系统中,每个节点维护两张表:一张为邻居表,保存一定数量的节点信息;另一张为 DHT,其内容如下。

1)DHT 节点,包含 $\log N$ 个按层排列的 DHT 节点。N 是覆盖网最多能容纳的节点个数。

对于某个节点 n,n 是该节点的 ID,它的第 i 层 DHT 节点唯一需要满足的条件是必须落在区间中,所有节点的 ID 均是模 N 后的结果。可见节点 n 在选择它的 DHT 节点时拥有很高的灵活性,所以这里设计的 DHT 是松散组织的。

2)连接邻居,包含非结构 P2P 覆盖网上 M 个邻居。当前节点与这 M 个邻居建立 TCP 连接,周期性的数据交换仅发生在当前节点和其连接邻居之间。如果当前节点发现某个邻居已失效或者给自己提供的数据很少,就选取时延最小的节点来替换这个邻居。

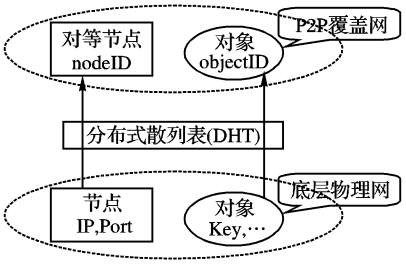

图 2　DHT 在 P2P 系统中的位置

3.2　数据调度算法

数据调度算法是在控制拓扑模型上进行调度。控制拓扑主要用来帮助成员反复交换状态信息,使得每个节点从其邻居节点中获得尽可能多的具有紧迫性和稀缺性的数据分片,从而提高拓扑网络的稳定性。

数据调度算法工作在非结构 P2P 覆盖网之上,它从建立连接的网络邻居中周期性地获取数据可用性信息,从而规划数据分片的获取过程。综合考虑数据分片的稀缺性与紧迫性,根据一个分片的稀缺性与紧迫性计算出获取它的优先权,优先权越高的数据分片越早获得,从而有效减少不能按期播放的数据分片数目。

每个节点和其连接邻居周期性地交换各自缓存中的数据信息,交换的周期称为调度周期。每个调度周期中,节点的数据调度器检查连接邻居的缓存中有哪些当前节点还没得到数据分片。算法具体如下:假设数据分片为 D_i,每个数据分片的提供者集合 $s_1,s_2,s_3,\cdots,s_m$,D_i 的优先值为 P_i,即此数据分片在所有拥有数据分片的节点缓冲中的位置综合值 $P_i = l_{i1} \times l_{i2} \times \cdots \times l_{in}$。数据片 D_i 的提供节点 $n_{i1},n_{i2},\cdots,n_{in}$,$t$ 为数据分片最早获取时间,T 是从 s_{ij} 处获取数据分片 D_i 的预期时间,$Trans(j)$ 为节点 j 的发送速率,$Q(j)$ 为节点 j 处的预期排队时间。$NSupply$ 为数据分片的提供节点。

算法 1　数据调度算法

计算在一个调度周期 m 内最多能接受的分片数 $\max(m)$;

```
for i = 1 to max( m )
  t = ∞ ;                                  //设置 D_i 的最早获取时间
  for j = 1 to k
    T = 1 / Trans(s_ij)
    if T + Q(s) < t and T + Q(s) < Q;
      t = T + Q(s);
      NSupply = s;
      if ( NSupply ! = null)
        Q(NSupply) = t;
```

3.3　数据预取算法

数据调度算法是在数据拓扑模型上进行调度,数据拓扑

模型主要用于解决数据的传输问题。

数据预取算法工作在结构化 P2P 覆盖网即 DHT 之上。依靠 DHT 所提供的高效定位功能,从媒体源发布出来的每一个数据分片都会被分布式地存储在 k 个网络节点中,k 也是模拟实验中的重要参数。系统中每个节点不断预测哪些数据分片很可能被数据调度算法遗漏,如果有认为遗漏的数据分片,就启动数据预取算法快速查找并获得它们,从而保证连续播放。所有尚未获得的 ID 小于分割线 L 对应的数据分片为 LD_i 值的数据分片,都被预测为遗漏数据,它们的总数目为 $NMiss$,如图 3 所示。

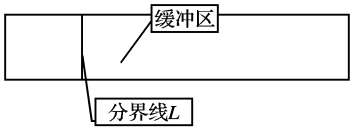

图 3 分界线基本原理

数据预取算法在一个调度周期内最多能获取 m 个数据分片,根据 $NMiss$ 与参数 m 的关系,数据预取算法采取不同行为:

情形 1 如果 $NMiss = 0$,不需要启动数据预取算法;

情形 2 如果 $NMiss \leqslant m$,启动数据预取算法并行获取所有遗漏的数据分片;

情形 3 如果 $NMiss > m$,不启动数据预取算法,因为遗漏分片太多,预取将带来过大开销。

具体算法如下。假设被预测为遗漏的数据分片为 D_i,对每个数据 D_i,首先需要定位到 k 个备份了 D_i 的节点,然后选择能以最大发送速率传送 D_i 的节点作为 D_i 的预取提供者。当节点 N 向一个应当备份 D_i 的节点 N_i 索取 D_i 时,有可能 N_i 也没有获得 D_i,此概率为 $\frac{1}{2}$,一般认为 N 和 N_i 有等价的概率获得 D_i,N 从 k 个应当备份 D_i 的节点处都不能获得 D_i 的概率是 $(\frac{1}{2})^k$,此概率值非常小,所以绝大多数情况下遗漏数据的预取操作都能成功。$DSupply$ 为数据分片的预取节点。

算法 2 数据预取算法

```
for i = 1 to m;
并发发送 k 条查询消息查询数据分片 D_i,得到 k 个节点 N_i;
R_i = 0;
//设置 D 的最大接受速率为 0;
for j = 1 to k;
  If( N_j has D_i ) and Rate(N_j) > R_i;
  // N_j 的发送速率大于 R_i;
  DSupply = N_j;
```

4 实验分析

我们通过模拟实验评估改进的 CGossip 系统。通过实验,对比原来基于 Gossip 协议的 P2P 流媒体系统和本文提出的 CGossip 系统,评估改进的 CGossip 系统。本文模拟实验环境是用 MyEclipse6.5 开发了 CGossip 系统的 Java 原型程序。模拟网络允许的最多节点数为 1000,每个节点邻居个数为 4,每个节点数据备份在 4 个节点,即 $k = 4$。数据调度周期为 1 s,节点每秒播放数据分片数为 10,缓冲区大小为 600 MB。

下面通过实验结果分析 CGossip 系统的性能。首先考察节点的成功播放率。从图 4 中可以观察到,当节点数为 1000 时,随着播放时间的增加,CGossip 系统的节点成功播放率明显比原来普通的基于 Gossip 协议的 P2P 流媒体系统提高大约 10%。

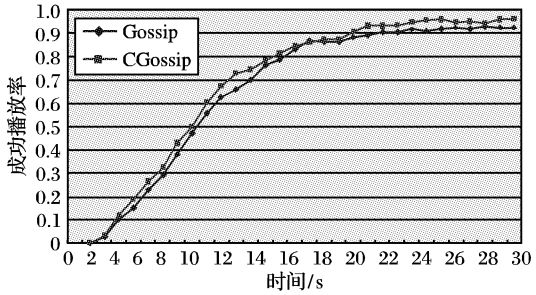

图 4 节点成功播放率

从图 5 可以看出,在节点数分别为 200、400、600、800、1000情况下的平均路由长度,采用 CGossip 的系统明显比 Gossip 的系统平均路由降低了路由耗费,提高效率约 4%。

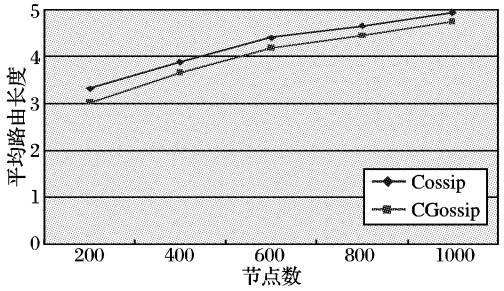

图 5 平均路由长度

考虑到流媒体播放过程中数据缓冲区的作用,实验模拟了数据缓冲区中信息交换带来的开销占总开销的比率,CGossip 的系统在前半段时间内开销比率比 Gossip 的系统效率提高约 25%,在后半段时间也比原来比率小,如图 6 所示。

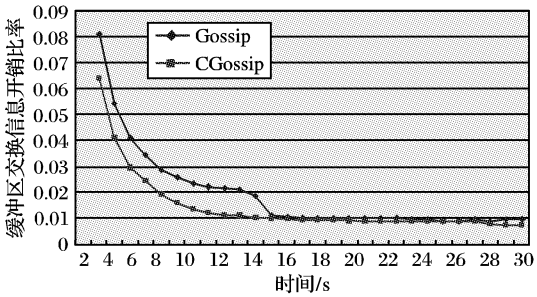

图 6 缓冲区交换信息开销比率

5 结语

本文在基于 Gossip 协议的 P2P 流媒体机制上,考虑到 Gossip 协议本身的不足,即数据传播的随机性与不确定性,提出改进的 CGossip 机制。在控制拓扑模型和数据拓扑模型的基础上,将非结构化 P2P 网络和结构化 P2P 网络相结合,执行数据调用算法和数据预取算法。最后通过模拟实验表明,CGossip 系统比基于 Gossip 协议的 P2P 流媒体系统有更好的播放效果。

参考文献:

[1] KERMARREC A, MASSOULIE L, GANESH A. Rliable probabilistic communication in large-scale information dissemination systems, MSR-TR-2000-15[R]. Cambridge: Microsoft Research, 2000.

[2] CHU YANG-HUA, RAO SAN-JAY, ZHANG HUI. A case for end system multicast[C]// Proceedings of the ACM Sigmetrics. New York: ACM Press, 2000: 1 - 12.

(下转第 2658 页)

2)解析和翻译结构良好的 XMPP 协议包。

设计跨协议的企业分布式即时通信系统客户端时尽量做到以下要求:

1)降低客户端的复杂度,提供友好的客户端界面,方便用户的使用;

2)客户端可定制性,使用户可以根据个人兴趣定制个性化的客户端界面。

本文所述的基于 XMPP 跨协议的企业分布式即时通信系统实现了即时通信的基本功能,包括即时信息的交互、文件的传送及语音通话、好友信息列表的维护以及个人信息的维护等。同时,企业即时通信系统实现与企业现有系统的无缝对接和集成,统一了企业资源的信息管理平台。

3 应用案例分析

基于 XMPP 的跨协议的企业分布式即时通信系统,针对教育行业的应用需求,开发了校园通即时通信系统。校园通即时通信系统整体架构如图 6 所示。

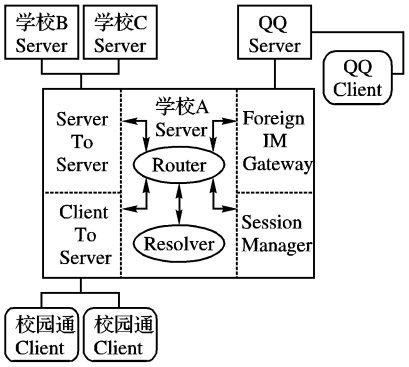

图6 校园通即时通信系统架构

如图 6 所示,每个学校部架设学校内部的即时通信服务器(实现分布式架构设计):

1)同一学校的用户通过学校内部的即时通信服务器进行即时信息的交互,采用客户端与服务器之间信息传递的交互模式;

2)不同学校间的用户通过路由服务(Router 和 Resolver)实现即时信息的交互,交互模式增加了一层服务器与服务器的信息交互;

3)校园通用户与 QQ(以 QQ 为例)用户的交互,通过网关协议层(Foreign Gateway)进行XMPP协议与QQ即时消息协议的转换,从而实现校园通用户与 QQ 用户的互联互通和即时消息的交互,交互模式中又增加了服务器与网关协议的信息交互。

根据在学校实际应用情况的信息反馈,校园通即时通信系统可以同时允许约 10000 人在线,可以进行即时的文本信息的交互、文件的传输以及语音通话等,实现了与 QQ、MSN 等平台的互联互通(在图 6 中仅说明了与 QQ 的互联互通)。在实现主要功能的前提下,系统也体现出了良好的稳定性及可靠性。同时,校园通系统在设计开发过程中,预留了接口,便于以后系统功能的扩展与维护和与已有教学信息管理系统的集成。

4 结语

本文设计并实现的基于 XMPP 跨协议的企业分布式即时通系统,实现了不同学校间文本信息的交互、文件的传送、语音通话等基本功能,体现出了分布式的特点;兼容并支持多种协议,实现与 MSN、QQ 等即时通信平台的互联互通,体现了该系统跨协议的特点。该系统成功应用于教育行业,满足即时信息交互的分布性及多样性;采用安全协议与加密相结合的技术,确保了即时通信系统的可靠性及稳定性。校园通即时通信系统只是在教育行业中的成功应用范例,根据是实际需求可以应用在政府及规模较大的企业当中,构建相应的即时通信平台。

XMPP 是即时通信的标准协议,但它不仅仅只用在即时通信领域,它可以作为一种标准的数据传输协议,可以广泛应用在其他方面。根据企业即时通信实际的应用需求,也可以对 XMPP 进行扩展。这将是下一步的研究内容。

参考文献:

[1] 张文茂, 章淼, 毕军, 等. 互联网即时消息(Instant Messaging, IM)的研究现状与展望[J]. 小型微型计算机系统, 2005, 20(5): 56 - 61.

[2] 刘影, 季波. 企业即时通信系统的应用研究[J]. 现代商贸工业, 2007, 19(6): 202 - 204.

[3] 杜松波. 企业即时通讯系统服务器的设计与实现[D]. 成都: 电子科技大学, 2004.

[4] RFC3920, Extensible messaging and presence protocol (XMPP): Core[S], 2004.

[5] Why your business should use enterprise instant messaging now[EB/OL]. [2009 - 01 - 01]. http://www.instantmessagingplanet.com/public/article.php/3491996.

(上接第 2654 页)

[3] PADMANABHAN V, WANG H J, CHOU P A, *et al*. Distrubuting streaming media content using cooperative network[C]// Proceedings of the 12th International Workshop on Network and Operating Systems Support for Digital Audio and Video. New York: ACM Press, 2002: 177 - 186.

[4] TRAN D, HUA K, DO T. Zigzag: An efficient peer-to-peer scheme for media streaming[C]// Proceedings of IEEE INFOCOM'03. San Francisco: IEEE Press, 2003:

[5] CASTRO M, DRUSCHEL P, KERMARREC A-M, *et al*. Splitstream: High-bandwidth content distribution in a cooperative environment[C]// Proceedings of IPTPS'03, LNCS 2735. Berlin: Springer, 2003: 292 - 303.

[6] ZHANG X, LIU J, LI B, *et al*. CoolStreaming/DoNet: A data-driven overlay network for peer-to-peer live media streaming[C]// Proceedings of the IEEE INFOCOM'05. Miami: IEEE Press, 2005: 2102 - 2111.

[7] STOICA I, MORRIS R, KARGER D, *et al*. Chord: A scalable peer-to-peer lookup service for internet applications[C]// Proceedings of the 2001 ACM SIGCOMM Conference. New York: ACM Press, 2001: 149 - 160.

[8] ROWSTRON A, DRUSCHEL P. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer system[C]// IFIP/ACM International Conference on Distributed System Platforms (Middleware). Heidelberg: Springer, 2001: 329 - 350.

[9] RATNASNMY S, FRANCIS P, HANDLEY M, *et al*. Application-level multicast using content-addressable networks[C]// SIGCOMM. New York: ACM Press, 2001: 161 - 172.