P2P网络下一种基于DHT的视频点播方案

江 庆[1],钟尚平[1],吕建明[2]
(1.福州大学 数学与计算科学学院,福建 福州 350001;2.中科院 计算技术研究所,北京 100080)

摘 要:在基于P2P的视频点播系统中,节点邻居选择策略对服务质量有很大的影响。提出一种基于DHT(Distributed Hash Table)的P2P覆盖网络下视频点播(Video-on-demand)的解决方案。通过网络坐标系统的拓扑发现能力,充分结合DHT网络高速搜索和VoD视频点播的特性、构造一种具有高效邻居选择能力、高用户自由度、高可靠性、扩展性的体系架构。针对架构设计中资源发布/分发、资源的搜索、视频点播的实现等关键问题提出了解决方案,分析了系统的特性。

关键词:点对点网络;视频点播;分布式哈希表
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2008)07-0229-04

A P2P Video-on-Demand Scheme Based on DHT

JIANG Qing[1], ZHONG Shang-ping[1], LÜ Jian-ming[2]
(1. College of Mathematics and computer Science, Fuzhou University, Fuzhou 350001, China;
2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: The neighbor-selection strategy in P2P VoD system is an important factor for its QoS. In this paper, a VoD (Video-on-demand) scheme based on DHT(Distributed Hash Table) under P2P overlay network is proposed. Via exploiting the topology discovery ability of distributed network coordinate system and integrating the fast-searching property of DHT with the trait of VoD system, a user-free framework with high robustness and scalability has been designed. Aims at the key issues of the user-free framework, has given the solutions for users to publish and demand video resources.

Key words: P2P; video on demand; DHT

0 引 言

近年来,P2P流媒体服务作为P2P网络下的一种典型应用,越来越得到业界的关注。许多P2P流媒体服务已经走上商业运营的轨道。

当前的P2P流媒体服务类型,分为直播与点播(VoD)两类,现有的系统大量地集中在视音频的直播上。直播的特点是用户的请求具有高度的同步性,即在同一时刻,用户请求数据的内容大致相同,因此,这些数据可以被广泛地分布在边缘节点的缓存中,在满足节点自身需要的同时,可供其它节点使用。但是,VoD服务与直播服务有很大的不同。

它的特点表现在3个方面[1]:

1)请求的异步性(Request asynchrony)。用户的请求在不同的时刻到达,所以不同的用户请求的是不同的数据分片,因此,采用简单ALM(应用层组播)[2,3]的流媒体组播模式不能满足所有的用户需求。

2)节点的动态性(Peer dynamics)。节点任意地加入/离开系统,并随时可能失效,所以需要一种高效、灵活的机制适应这种动态环境[4]。

3)不可预期的交互(Unpredictable interactivity)。节点随时都可能进行VCR操作(快进、快退、暂停,恢复播放),这就意味着要频繁地改变为节点提供流数据的"源"。

目前,P2P VoD的解决方案有:对直播采用的ALM进行改进的方案[2];基于多层P2P覆盖网络构建的分布式CDN[5];基于DHT分布式存储[1]。

DHT分布式哈希表,是结构化P2P覆盖网络的基本技术。覆盖网络的拓扑被严密地控制,文件或内容(或指向它们的指针)被放置在具体指定的节点上。这类系统通常通过分布式路由表,提供内容到节点(例如,节点的地址)的映射方法。因此,在此系统下,能够快速、高效地定位内容对应的节点(一般在$O(\log N)$跳内),这为VoD服务的交互性需求提供了高效实现的方法。典型的DHT系统有:Chord[6],CAN[7],Pas-

收稿日期:2007-10-20

作者简介:江 庆(1981-),男,硕士研究生,研究方向为P2P网络与网络安全;钟尚平,博士,教授,硕士生导师,研究方向为网络安全与数字图像处理等。

try[8]等。

文中提出方案的框架分为 3 个层次,底层为物理网络层,提供最基本的网络主机互联;其上为 P2P 覆盖网络层,是基于 DHT 的 P2P 覆盖网络,它负责维护分布式路由表,提供节点登陆、搜索功能;最上层为服务层,提供流媒体发布/分发服务、流媒体节目搜索服务及视频点播服务。系统结构如图 1 所示。

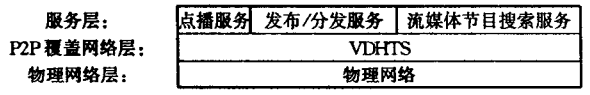


图 1 系统结构

1 基于 DHT 的 VoD 分布式存储方案

文中的分布式存储方案(VDHTS)基于 DHT 覆盖网络构建。覆盖网络中的每个节点都有一个特定的 ID,只要指定一个搜索关键字 S 就能通过 DHT 的搜索机制找到 ID 与 S 最接近的一个(或多个)节点(其搜索机制参考文献[6~8])。媒体文件在系统中按照一定的时间长度(比如说 5 分钟)划分为一个个数据分片,数据分片被分散存储于 VDHTS 的节点中,每个节点为系统划分出一定大小的外存共享空间,以存储数据分片。

VDHTS 的设计要满足以下两种需求:

(1)由于不同网络范围对不同流媒体资源需求往往具有很大的差异,盲目地发布与分发流媒体资源只会造成不必要的资源浪费,系统应该能够按需求将资源发布与分发到相应的网络范围,因此 VDHTS 应能提供搜索特定网络范围节点的功能。

(2)当某个节点请求点播服务时,总是希望能最快地找到拥有相应资源的节点,并能从这些节点高速地下载数据[9]。

对于需求(1),要将资源发布或分发到相应的网络范围就必须有一种机制能支持快速地搜索相应网络范围的节点,即坐标优先搜索 CFS(Coordinate First Search)。而对于需求(2),要快速搜索拥有相应资源的节点,而资源在系统中是以资源的哈希值为标识的 Resource ID。因此,这种搜索是资源 ID 优先搜索 RKFS(Resource ID First Search)。要在同一系统中实现这两种不同的搜索,可在 VDHTS 空间中设置两类不同的节点。

VDHTS 中的节点通过如 GNP[10],Vivaldi[11]的网络坐标系统确定自身在网络中的坐标,再通过 Hilbert 的空间填充曲线(SFC)坐标转化将多维的坐标值转化为一维,跟据这种坐标信息,系统能自适应地优化数据分片在不同 SFC 范围内副本数量。

在 VDHTS 中节点分为两种:实节点与虚节点。实节点代表的是一台接入 VDHTS 的计算机,表示物理节点在 DHT 空间中的存在。实节点 ID 的设置使它适合进行 CFS 搜索。而虚节点表示的是 VDHTS 中资源的存在,虚节点 ID 的设置适应 RKFS 搜索以方便于点播节点快速定位到所需资源。实节点与虚节点之间是 1:n 的关系。这个关系实际上反映的是物理节点与它存储的资源的关系。

所有实节点组成的集合为实节点簇 RPC(real peer cluster),虚节点组成的集合为虚节点簇 VPC(virtual peer cluster)。如图 2 所示的 Chord Ring 的地址空间,环的左半部分是 VPC,右半部分为 RPC。对于一台拥有数据分片 1,2,3,4 的计算机 A,它在 RPC 中映射了它的实节点地址 RP(A),在 VPC 中映射了它拥有的 4 个数据分片的地址 VP(1)~ VP(4)。

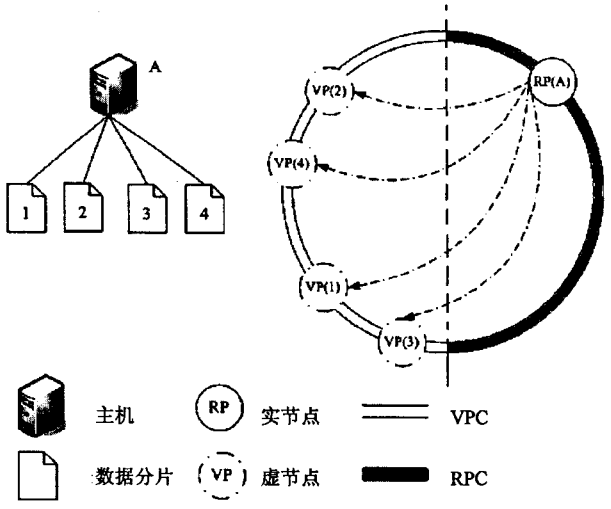
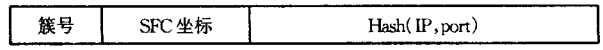


图 2 Chord Ring 的地址空间

那么实节点与虚节点是如何区分的呢?这体现在节点 ID 的结构上。如图 3 所示,簇号标识节点的类型,实节点的簇号为 0,虚节点的簇号为 1 。节点 SFC(space filling curve)坐标标识节点的网络位置,它是通过特定系统(如 GNP[10]等)得到的节点多维网络位置,通过 Hilbert 空间填充曲线映射到一维的 ID 中,这种转化后的坐标能够保持原来多维空间中节点的相对位置(即多维空间中接近的节点在转化后的 SFC 坐标系中依然是接近的)。Hash(IP,port)为节点 IP 和端口的哈希值,此段的设置是为了统一 ID 长度,并平衡节点的负载。Hashed seg info/Hashed Media info 为该虚节

实节点的 ID 结构:



虚节点的 ID 结构:

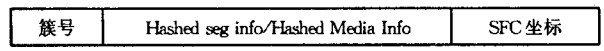


图 3 节点 ID 的结构

点存储的媒体分片的哈希值或媒体元信息哈希值。

2　资源的发布与分发

2.1　流媒体资源的发布

把内容发布分为两块来看:流媒体资源索引发布与流媒体分片发布。

流媒体资源索引不包含资源的实在内容,只是流媒体资源的一些元信息(如资源的码率、长度、分片总数、各分片的哈希值等),相当于P2P文件共享系统Bittorrent[12]中的种子文件内容。

这里假定一个物理节点P,它具有一个未发布的媒体文件M,文件被划分为k个分片(seg[1]～seg[k])。发布M的流程如下:

1) P在DHT网络中根据P的SFC坐标及IP,Port生成ID,以实节点加入RPC。

2) 发布M的流媒体分片。

a. 对M的k个分片seg[i](i=1…k),计算其哈希值hash(seg[i])。

b. 对M的k个分片seg[i],以其哈希值hash(seg[i])作为虚节点ID中的Hash seg info,在DHT上以虚节点加入VPC。

3) 发布M的流媒体资源索引。

a. 设定M的检索关键字Mmkey[i](这里的检索关键字可由发布者设定多个,也可以直接从媒体文件名中提取,第i个关键字用Mmkey[i]表示),并计算其哈希值hash(Mmkey[i])。

b. 对M的内容进行哈希运算,得到hash(M)。

c. 对每个M的检索关键字Mmkey[i],将hash(Mmkey[i])作为虚节点ID中Hashed Media Info,在DHT上以虚节点加入VPC;将hash(M)作为虚节点ID中Hashed Media Info,在DHT上以虚节点加入VPC。

2.2　流媒体资源的分发

以上的发布过程,只是简单地将发布者拥有的资源注册到DHT网络中,在发布之后,P拥有的资源M就可以被用户搜索到,并为用户提供点播服务,但系统中M的副本数仍为1。在VDHTS中,通过复制(replica)与缓存(cache)来增加副本。这里设计了3种不同的方式实现这种策略:

* 基于实节点路由表的数据复制RPFR (Real Peer Finger table based Replica)。RPFR基于如下假设:与特定媒体资源M的发布者网络距离越近的节点,对M的需求越高。即对特定媒体的需求具有一定的地域性。因此,在进行首次资源分发时,应更多地考虑网络距离与发布者近的节点。

DHT覆盖网络大多是基于Kleinberg的小世界原理构建,每个节点了解的近节点数量大于远节点,这种关系反映在节点的路由表里(如Chord的FingerTable)。然而,在一般的DHT协议中节点在DHT ID空间上的距离不能反映节点的物理距离,但在VDHTS的RPC中实节点ID的次高位是由SFC坐标构建的,SFC能够在一定程度上反映节点的物理距离。因此,某个实节点的DHT路由表项可以看作是以与实节点物理距离由近到远排序的。

RPFR算法如下:

```
Algorithm RPFR(M)
  For(i = 0 to n)
    If( RP(A).finger[i].ID.CluserID <> 0 ) //若表项对应的是虚节点,跳过
      Continue;
    Destination = RP(A).finger[i].address
    For(j = 0 to k)
      Send(destination, M.Seg[j])
    Send(destination, media info of M)
```

表1　RPFR算法中符号的含义

A	发布资源的节点A
RP(A)	A在VDHTS中的实节点
.finger	实节点的路由表(finger[i]为路由表的第i项,共有n项)
.finger[i].ID	路由表第i项对应节点的ID(finger[i].ID.ClusterID为该节点的簇号)
.address	节点的地址,如(IP, port)
M.Seg[i]	媒体文件的第i个分片,共k个
Media info of M	M的检索关键字等索引信息

RPFR在节点发布M之后运行。由于只有实节点ID能够反映节点的物理距离信息,RPFR必须从路由表中选出所有的实节点,这可以通过表项中节点ID的簇号(ClusterID)来分辨。然后将分片数据,索引信息等复制到这些实节点。在RPFR运行结束后,目标节点再通过上文中的方法在VPC中注册相应的虚节点,使副本数据在系统中可被搜索。

* 点播节点缓存数据DCS(Demander Cache Scheme)。与RPFR的主动复制不同,DCS是一种被动的缓存策略,当存储有数据分片seg[i]的主机A收到点播节点D1的请求后,开始向D1发送Seg[i],Seg[i]的数据在提供D1播放的同时,存储进D1节点的外存中。另外,A将D1记录在本地的一个DTable中,DTable的表项为<Hash(Seg[i]),D1>。当另有节点D2也向A请求Seg[i]数据,A将把DTable发送给D2,同时把D2加入到DTable中。这时D2就可以同时向A和D1请求数据了。当Seg[i]的数据发送完毕时,D1可以跟据自己共享外存空间的剩余容量,来决定再发布还是删除Seg[i]。

* 按点播节点的反馈信息Cache数据DFC(Demander Feedback Cache)。一个点播节点D对某一数据分片Seg[i]发出请求后,往往有多个源节点为其服务。若所有可用源节点提供的有效传输率低于流媒体

的码率时，点播节点将对 SFC 坐标最近的源节点 A 发出 Improve－QoS 请求，收到请求后 A 将在 RPC 空间中搜索 SFC 坐标与它最近的实节点，将 Seg[i]缓存到该节点上。

3 P2P 下基于 DHT 的视频点播算法

在用户注册进 DHT 覆盖网络后，即可通过视频点播的方式使用覆盖网络中的所有已发布的流媒体资源。进行点播的用户 D(Demander)首先在系统中搜索感兴趣的文件名或关键字，搜索将返回一个流媒体索引信息的列表。Demander 在列表中选取喜好的资源，即可开始缓冲播放。这两个过程归纳为两个算法 Search_Index(返回媒体索引列表)与 VoDRequest(请求媒体数据)。

Search_Index 算法由点播用户 D 调用，在 VPC 空间(ClusterID＝1)中搜索关键字 SearchKey，返回一个媒体索引列表(Index_List)。先在 DHT 空间中搜索节点 ID 离 SearchKey 最近的 k 个节点，返回节点的列表(PeerList)。("大多数 DHT 协议的查找算法经过修改都能满足返回多节点的需求"[1]。故不对此算法详述)接下来，在返回的节点列表中，筛选出 HashMediaInfo 段与 Hash(SearchWord)相同的虚节点，这些虚节点必定拥有相应 SearchKey 相关的媒体索引信息 Index_Info。对其发出 GetIndexInfo 请求，将返回的 Index_Info 加入 Index_List 中。之后，用户从 Index_List 选出想要点播的资源索引 Index_List[i]，作为参数，调用 VoDRequest 算法。

VoDRequest 算法输入参数为 Index_List[i](用户选中的流媒体资源的索引)和 PlayTime(当前播放时间，默认为 0)。索引列表中包含该流媒体的所有分片哈希值 Index_List[i].HashSeg。VoDRequest 算法跟据当前播放时间映射到相应的数据分片号，再跟据分片号在 Index_List[i]中找到相应的哈希值，用它生成 SearchKey，在 VPC 中查找包含离 SearchKey 最近的虚节点列表(共 t 项)。接着，筛选出 HashSeg 段与 Index_List[i].HashSeg[Seg_Entry]相同的虚节点，按离 D 节点的 SFC 坐标距离升序排序。最后，对 PeerList 中的节点请求数据。

在数据请求时，并不对 PeerList 中的所有节点发出请求，而是只对前 m 个节点发出请求，$m = \min(\{k \mid BW_{content} \geqslant \sum_{i=1}^{k} BW(P_i), k \leqslant R\})$，$R$ 为 PeerList 大小，$BW(P_1) \sim BW(P_R)$ 为降序排列的节点带宽，流媒体的码率为 $BW_{content}$。这里可能出现的最坏情况，$m = R$ 时，总体传输速率仍然低于 $BW_{content}$。这时 D 便会对列表中的第一个节点 P_1 发出 Improve－QoS 请求，以提高服务质量。

4 性能分析

本方案在基本 DHT 覆盖网络的基础上，跟据 VoD 服务的特点，将 DHT 空间分为两个簇，并加入了节点的网络坐标信息。这不但能让节点在请求流媒体数据时更加高效，而且在资源的分发上更具选择性，使媒体资源的分布更利于请求的本地化。接下来分析系统的性能特点：

* 健壮性：数据通过 RPFR，DCS，DFC 三个策略进行分发，使同一媒体分片在系统中具有多个副本。当其中的部分节点下线时，能自适应地产生新副本，保证系统的可用性。

* 高效性：利用高效的 DHT 路由算法，实现高速的流媒体索引、分片的搜索。另外，因为分片数据是长期存储于节点中，使得用户在播放特定的分片时，与该分片相关的有效源节点可以一直提供数据。所以在分片播放完毕或进行 VCR 操作之前，只需对下一分片的源节点进行搜索。这大大地减少了搜索次数。另外，通过 SFC 坐标进行内容的分发，使数据分片在系统中的分布更加合理，一定程度上提高了数据的本地化，使系统更加高效。

* 高可靠性：分片数据的传输采用"多对一"的方式，尽可能地保证了流传输速率大于流媒体码率，当服务质量不能满足节点的点播需求时，系统也能自适应地进行调整，保证点播的 QoS。

* 可扩展性：首先 DHT 路由算法本身具有很高的搜索效率，只需要 $O(\log N)$Hop(N 为覆盖网络的规模)即可找到目标节点。其次，消费者节点根据空闲带宽的多少，从待选节点中选出实际为其提供数据的源节点，这样就将带宽和负载分散到多个节点上。以上两点使系统规模扩大时，仍然保持良好的性能。

* 用户的自由性：加入系统的任一用户不但可以作为消费者使用 VoD 服务，还能作为服务的提供者自由地发布自己的媒体内容。

5 结束语

将一个流媒体点播服务构建基于 DHT 的分布式存储的覆盖网络上，通过一定的副本冗余和数据本地化策略提高系统的可用性，并保证点播服务的 QoS；利用 DHT 高效搜索的特点提高系统的性能。

此方案针对视频点播的特点而设计，提供了一个可行的系统原型。但还有许多地方需要完善。例如，

(下转第 236 页)

词组合结构后唯一的一个名词做下一个动词组合的主语。

例如:小明消费的地点是东大街。

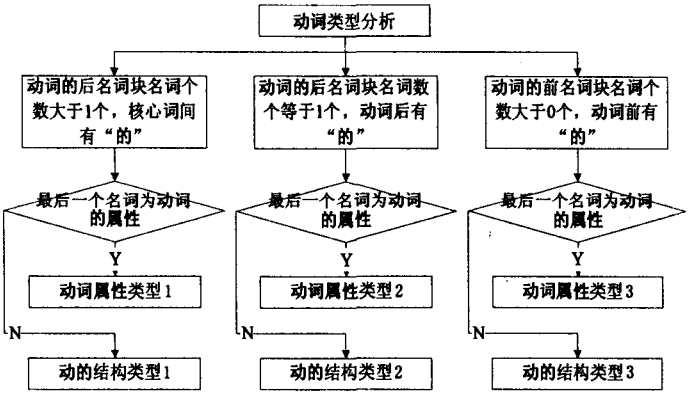
(3) 动词属性类型 3。

结构特征如"名词 + 的 + 动 1 + 名词 + 动 2"。这种结构前名词块的名词的个数大于 0,后名词块的名词的个数大于 0。动词组合结构的动词有主语,动词组合结构后唯一的一个名词做下一个动词组合的主语。

例如:小明的消费地点是东大街。

4.2 具体处理办法

实际上解决动词属性与"动的"结构耦合现象的关键在于分辨核心词是否是动词的属性。若是则进入到对应的动词属性处理模块,不是再进入到相应的"动的"结构处理模块。处理的总体流程图如图 2 所示。



"动的"结构、动词属性的分类流程

图 2 总体流程

5 结 语

语言学的研究发现了动词意义中丰富的语义、句法属性。结合这方面的研究成果,提出了动词属性的概念,并对动词属性的分类和语义特征进行讨论,将其运用到自然语言处理模型的具体实践当中,针对动词属性结构与"动的"结构存在耦合的现象提出了一种解决办法。当然,此解决方案还不能涵盖所有的耦合情况,中间的某些情况可能仍需要细分,并在以后调试过程中逐渐优化。

参考文献:

[1] 马庆株. 汉语动词和动词性结构(一编)[M]. 北京:北京大学出版社,2005.

[2] 吕叔湘. 汉语语法分析问题[C]//汉语语法论文集. 北京:商务印书馆,1979.

[3] 冯丽萍. 动词的语义特征及其在句子加工中的心理现实性[J]. 语言文字应用,2006,12:172-176.

[4] 陈昌来. 现代汉语语义平面问题研究[M]. 上海:学林出版社,2003.

[5] 石纯一,黄昌宁,王家钦. 人工智能原理[M]. 北京:清华大学出版社,1993.

[6] 裘荣棠. "动 + 的"短语的表意功能[J]. 修辞学习,1999(1):43-44.

(上接第 232 页)

现有方案中资源只能被添加入系统而未提供删除方案;方案不支持模糊关键字搜索等。

参考文献:

[1] Yiu W P K, Jin Xing, Gary H S C. Chan. Distributed Storage to Support User Interactivity in Peer - to - Peer Video Streaming[C]//ICC'2006. Istanbel:[s.n.],2006:55-60.

[2] 阳天保, 张修如, 贾丽会. 基于 P2P 视频点播系统模型及算法研究[J]. 计算机技术与发展,2006,16(10):45-48.

[3] Castro M, Druschel P, Kermarrec A, et al. Splitstream: High - bandwidth multicast in a cooperative environment[C]// ACM SOSP. New York, US: ACM,2003.

[4] Tran D, Hua K, Do T. ZIGZAG: An Efficient Peer - to - Peer Scheme for Media Streaming[C]//IEEE Infocom'2003. San Francisco, California, USA: IEEE,2003.

[5] Wan Kan Hung, Loeser C. An overlay network for replica placement within a P2P VoD network[J]. Journal of High Performance Computing and Networking,2004(7):1-10.

[6] Stoica I, Morris R, Karger D, et al. Chord: A Scalable Peer to peer Lookup Service for Internet Applications[C]//SIGCOMM' 2001. San Diego, Carifornia, US: ACM,2001.

[7] Ratnasamy S, Francis P, Handley M, et al. A Content Addressable Network[C]//SIGCOMM'2001. San Diego, Carifornia, US: ACM,2001.

[8] Rowstron A, Druschel P. Pastry: Scalable, distributed object location and routing for large - scale peer - to - peer systems [C]//Proceedings of IFIP/ACM Middleware 2001. Heidelberg, Germany: ACM,2001.

[9] Rowstron A, Druschel P. Storage Management and Caching in PAST, a Large - Scale, Persistent Peer - to - Peer Storage Utility[C]//Proc. of ACM Symposium on Operating Systems Principles (SOSP). Banff, Alberta, Canada: ACM,2001.

[10] Ng E. GNP sofware[EB/OL]. 2003. http://www - 2. cs. cmu. edu/eugeneng/research/gnp/software. html.

[11] Dabek F, Cox R, Kaashoek F, et al. Vivaldi: A Decentralized Network Coordinate System[C]// SIGCOM'04. Portland, Oregon, US:[s.n.], 2004.

[12] Cohen B. Incentives build robustness in Bit Torrent[J]. Economics of Peer - 2 - Peer Systems,2003(6):55-61.