大规模层次化视频点播存储系统的设计与管理*

李 勇 吴 飞 陈福接

(长沙工学院计算机系 长沙 410073)

Email: yli@nudt.edu.cn

摘要 近来计算机和通信技术的发展使得视频点播(video-on-demand,简称VOD)在技术和经济上成为可能.连续媒体的特性使得VOD系统需要大规模的存储服务器.层次化存储体系是减少系统费用的合理方案.文章提出了一种层次化的存储模型和磁盘cache的概念.根据这个模型,提出了基于访问频率的替换算法,并对算法的有效性进行了模拟和分析.结果表明,这种算法解决了LFU(least frequently used)算法中的"cache污染"(cache pollution)问题,能较好地适用于连续媒体数据应用.

关键词 连续媒体,视频点播,层次化,存储服务器,替换算法.

中图法分类号 TP393

网络和存储技术的进展使得多媒体在线(on line)访问成为可能.其中,视频点播(VOD)是一个典型的应用范例,成为目前分布式多媒体研究的热点[1].对于VOD存储服务器的设计,许多人已经做了很多出色的研究工作[2~6].这些研究在媒体流的实时性分析、缓冲区管理、磁盘调度和磁盘阵列的分布和优化上取得了一定的成果,但是,它们大都局限于单节点上小规模的视频服务器的研究.Brubeck在文献[5]中提出了一种分布的VOD体系结构,它将影片分布到许多规模较小的视频服务器中.本文研究一种大规模的集中式VOD系统,它在一个中心服务器上存储了大量的影片,能支持很大规模的客户数目.服务器还可以在网络上交换数据,和异地的服务器协同工作.大规模VOD系统对信息的提供者和消费者都更具吸引力,而且也有利于资源的优化使用.

由于连续媒体具有实时连续特征,大规模VOD系统需要巨大的存储容量.如一部100分钟的影片,采用4Mbps的MPEG-II压缩标准需要3GB的存储容量,1 000部影片就要3TB的存储容量.

根据表1所示的参数,如果将影片存储在内存中,系统存储费用将达到600万美元,假设系统支持1 000个用户,每个用户存储费用为6 000美元,这在经济上是不可行的.如果使用磁盘存储影片,系统存储费用为150万美元,用户平均费用为1 500美元.如果使用光盘,系统存储费用为1.5万美元,用户平均费用仅为15美元,这在经济上最具可行性.但是,光盘的性能太低,不能满足带宽的需求.可见,单一的存储器件难以满足其性能和价格上的双重要求.为了设计既能满足带宽要求,费用又合理的存储服务器,层次式存储方案是一种合理的选择.

表1 存储器件参数表

设备	访问延迟(ms)	带宽(Mbps)	价格($/MB)
内存	约0.01	800~2 000	2
磁盘	10~20	200~400	0.5
光盘	30~50	10~20	0.005

层次化存储系统是将全部的数据存储在低速的后援存储器中,而把常用的数据存放在具有较高速度的存储器中,如果大部分的访问能在高速器件中满足,存储系统的总体性能接近于高速存储器件,而平均价格接近于后

* 本文研究得到并行与分布处理国家重点实验室基金资助.作者李勇,1970年生,博士生,主要研究领域为分布式多媒体,高性能视频服务器,连续媒体网络传输.吴飞,1968年生,博士后,主要研究领域为计算机组织与系统结构,分布式多媒体技术.陈福接,1935年生,教授,博士生导师,主要研究领域为高性能计算机体系结构,分布式多媒体技术.

本文通讯联系人:李勇,长沙 410073,长沙工学院计算机系研究生队

本文1998-01-22收到原稿,1998-05-04收到修改稿

援存储器件，性能价格比就能提高.能否达到这样的效果将取决于数据访问的局部性和替换算法的有效性.

本文提出一种层次化存储系统的模型和磁盘cache的概念.对VOD系统的数据访问特性和数据局部性进行了分析，根据VOD系统和传统的计算机系统的数据特性的不同，开发了新的替换算法.文章第1节提出了系统模型，并分析了VOD影片数据的访问特性和访问的局部性.第2节描述了层次化存储系统的替换算法，并对算法进行了模拟和分析.第3节是全文的总结.

1　VOD数据访问特性和系统模型

我们的VOD系统模型是处于分布环境下的层次式结构.如图1所示，各种类型的服务器分布在广域网上(如Internet).图中文档服务器(AS)提供后援存储服务，全部的影片都存储在其中，包括不常被访问的冷门电影.它可以是本地的，也可以是异地的，它们之间可以交换数据.视频服务器(VS)存储的是热门影片，它向客户提供点播的实时服务.客户端是机顶盒(set-top-box)或多媒体电脑，它通过高速网络与代理服务器(EFS)相连，客户向代理服务器发出请求，代理服务器决定是否接收请求以及和哪一个特定的视频服务器相连，然后由视频服务器为客户提供实时服务.视频服务器和本地文档服务器构成层次化结构，巨大的影片库存储在由光盘或磁带组成的后援存储器中，由磁盘构成的二级存储器存储的是访问概率较大的热门影片.我们称之为"磁盘cache".用户点播的数据由视频服务器从磁盘读到内存缓冲区中，再通过高速网络发送给客户端.

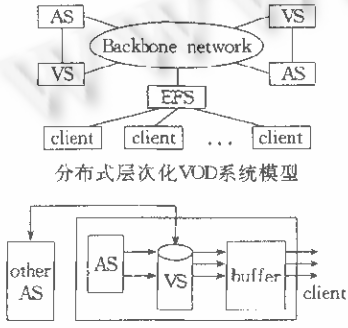

图1　系统模型和服务器结构

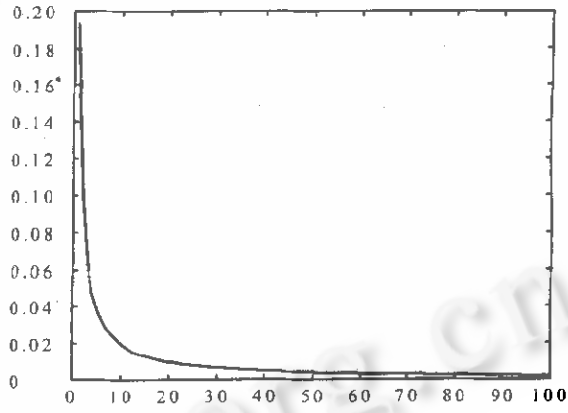

图2　Zipf法则下的访问

层次化存储系统的关键在于数据访问的局部性，由于目前还没有大规模投入实际运行的VOD系统，我们无法统计出确切的影片访问特征，但是，我们可以利用影像出租的统计资料，为VOD系统的设计服务.研究表明，如果按访问概率从大到小进行排序，影片的访问概率基本上符合Zipf法则，即N部电影排序后，第i部电影的访问概率$f(i)=C/i, i=1,2,\ldots,N$，其中$1/C=\sum_{i=1}^{N}\frac{1}{i}$.

如图2所示，客户的点播集中在前面的一小部分热门影片，体现了影片访问的局部性.当影片库容量增大时，这种局部性表现得更加明显.有研究表明，随着单位时间内访问人数的增加，热门影片的访问概率将变得更大，从而访问的局部性就更强.这样，大规模VOD系统采用层次化存储是合适的.考察一部影片在不同的时间内的访问概率，虽然每部影片的访问概率有不同的变化特点，但它们都有上升、平稳和下降3个阶段，我们分别称之为成长期、成熟期和衰老期.显然，成长期的影片存入磁盘cache最有利.

热门影片体现的访问局部性是层次化VOD存储系统设计和管理的依据，下一节将利用这种数据的局部性和影片访问概率的变化规律开发合适的替换算法.

2　层次存储系统的管理

层次化存储系统的管理主要解决访问不命中的问题.在层次化的VOD系统中，这个问题主要体现在磁盘数

据的选取上，即磁盘 cache 的替换算法上. 在传统的计算机系统中，cache 替换算法主要有 FIFO(first in first out)，LRU(least recently used)和 LFU(least frequently used). FIFO 算法最简单，它只是将最先进入 cache 的数据替换出来，而没有考虑数据的使用情况；LRU 算法利用了数据的时间局部性，将最近最少使用的数据替换出 cache；LFU 算法考虑了数据的使用频率，将使用频率最小的数据替换出 cache.

在 VOD 系统中，影片的数据尺寸很大，数据访问一般是按顺序的，时间局部性不强，但存在访问局部性，所以FIFO和 LRU 算法都不适用于磁盘 cache. LFU 算法考虑了数据的访问频率，但存在所谓的 cache 污染问题(cache pollution)，即一个曾经访问多次而又不再使用的数据不能及时替换出 cache，因而降低了 cache 的命中率[7].

我们根据 VOD 系统影片数据使用周期较长，访问频率变化有一定的规律这些特点开发了一种周期频率预测算法(periodic frequency prognosis，简称 PFP). 将时间轴按一定的长度划分成一个个的周期：$T_1, T_2, \ldots$. 只用前一个周期和本周期的访问频率来决定是否替换. 这样，较早的访问频率被屏蔽掉了，就可以避免 LFU 算法中的 cache 污染问题. 算法的主要思想是利用访问频率的变化趋势，将成长期的影片尽早换入磁盘，将衰老期的影片尽早换出磁盘. 我们在周期的中间时刻使用现行频率法，在周期的结束时刻使用预测法来提高替换算法的命中率.

用 T_i 表示第 i 个周期(i 为自然数). 设 P_k 为第 k 部电影. 总数为 N 的影片库可以表示为

$$S=\{P_k \mid k=1,2,\ldots,N\}.$$

设 f_{ki} 为 P_k 在 T_i 中的访问概率. 热门电影表示为 $S_{\text{hot}}=\{P_k \mid f_{ki}>\Delta\}$，冷门电影表示为 $S_{\text{cold}}=\{P_k \mid f_{ki}<\Delta\}$. 其中 Δ 为电影访问概率的阈值. 我们的目标就是将热门电影放到磁盘中去，但在算法中，并不固定一个热门电影集，而是统计影片的访问频率，用访问频率替代影片的访问概率.

设 N_{ki} 为 P_k 在 T_i 中的访问次数；F_{ki} 为 P_k 在 T_1 到 T_i 中总的访问次数，即 $F_{ki}=\sum_{j=1}^{i} N_{kj}$；$W_{ki}$ 为 P_k 在 T_i 比 T_{i-1} 访问次数的增量，即 $W_{ki}=N_{ki}-N_{ki-1}$. 每次请求到来时，都统计总访问次数. 我们为每部影片保留一个三元组 $\langle F_{ki}, F_{ki-1}, F_{ki-2}\rangle$，再根据它来计算 N_{ki} 和 W_{ki}：$N_{ki}=F_{ki}-F_{ki-1}$，$W_{ki}=N_{ki}-N_{ki-1}$.

(a) 在每个周期的中间，我们使用请调法.

设现在时刻为 t，第 k 部影片总的访问次数为 F_k，本周期到现在为止的访问次数为 N_k.

$N_k=F_k-F_{ki-1}$，我们将磁盘 cache 中的影片根据其访问频率的变化分为上升集和下降集.

$$\text{上升集 } S_u=\{P_k \mid W_{ki}>0\}, \qquad \text{下降集 } S_d=\{P_k \mid W_{ki}<0\}.$$

将后援存储器中处于上升期、访问频率大于阈值 Δ 的影片定义一个换入集，即

$$S_{\text{in}}=\{P_k \mid N_{ki}>\Delta, W_{ki}>0\}.$$

当对第 K 部影片的请求到来时，如果不命中，

若 $P_k \in S_{\text{in}}$，则满足换入条件，如果磁盘有足够的空间就换入，否则，按以下顺序换出影片.

(1) S_d 中当前时间周期内的访问频率加上上一周期的访问频率($N_{k-1}+N_i$)最小者为 $P_{k'}$，若 $N_{ki}>N_{k'i}$，则将 $P_{k'}$ 换出，退出，否则，转(2)；

(2) S_u 中当前时间周期内的访问频率加上上一周期的访问频率($N_{k-1}+N_i$)最小者为 $P_{k'}$，若 $N_{ki}-\Delta>N_{k'i}$，则将 P_k 换出.

(b) 在周期的结束时刻，我们使用预调法.

我们用一阶线性法则预测下一个时间周期的访问频率 $N_{ki+1}=N_{ki}+W_{ki}$，

并定义换入阈值 Δ_{in} 和换出阈值 Δ_{out}，

满足换入条件 $N_{ki+1}>\Delta_{\text{in}}, W_{ki}>0$ 的影片列入换入集 S_{in}，

满足换出条件 $N_{ki+1}<\Delta_{\text{out}}, W_{ki}<0$ 的影片列入换出集 S_{out}，

在特定的时候，将换出集中的影片换出，按 N_{ki+1} 的大小顺序将换入集中的影片依次换入，直到磁盘空间用完为止.

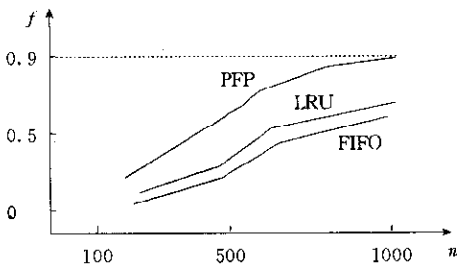

图3　替换算法性能模拟结果

我们用符合 Zipf 法则的随机访问序列简单地模拟了 3 种替换算法的效率，结果(如图 3 所示)显示，周期频率预测替换算法(PFP)命中率最高，FIFO 和 LRU 命中率较低，但两者差别不大. 在 FIFO 替换算法中，热门影片会随时间的推移被替换出磁盘 cache，它的命中率最低. LRU 将最近最少使用的影片替换出磁盘 cache. 热门影片因为经常使用，可以保留在磁盘 cache 中，但是冷门影片一定会替换入磁盘 cache，这必将替换出热门影片，从而增加了替换次数，降低了命中率. 基于访问频率的算法，可将热门影片保留在磁盘 cache 中，而且冷门影片因为访问频率低而不能替换入 cache，不会挤占热门影片的磁盘空间，所以命中率最高.

3　结束语

在大规模的 VOD 系统中，影片数据具有访问局部性. 利用这种访问局部性，我们提出了一种层次化的存储体系. 根据 VOD 影片数据访问频率变化特性，我们开发了周期频率预测替换算法(PFP). 这种替换算法解决了 LFU 算法中的 cache 污染问题. 模拟实验显示，与 LRU 算法和 FIFO 算法相比，该算法的效率最高. VOD 层次化存储系统还有许多内容需要研究，如，磁盘 cache 的大小设置、后援服务器服务策略等. 客户访问模型对系统性能影响很大，目前还没有精确的模型来描述，需要进一步深入地研究.

参考文献

1　Gemmell D J *et al*. Multimedia storage servers: a tutorial. IEEE Computer, 1992,28(5):40～49
2　Rangan P V, Vin H M, Ramanathan. Design an on-demand multimedia service. IEEE Communication Magazine, 1994, 30(7):56～64
3　Lougher P, Shepherd D. The design of a storage server for continuous media. The Computer Journal, 1993,36(1):32～42
4　Banu Ozden, Rajeev Rastogi, Avi Silberschatz. On the design of a low-cost video-on-demand storage system. Multimedia System, 1996,4(1):40～54
5　Brubeck D W, Rowe L A. Hierarchical storage management in a distributed VOD system. IEEE Multimedia, 1996,3(3): 37～47
6　吴飞，陈福接，王朴. 一个支持并行流的多媒体服务系统. 软件学报，1997,8(5):327～334
(Wu Fei, Chen Fu-jie, Wang Pu. A multimedia service system supporting parallel streams. Journal of Software, 1997,8 (5):328～334)
7　Ramakrisna Karedla, Love J S, Wherry B G. Cacheing strategies to improve disk system performance. IEEE Computer, 1995,27(3):38～46

Design and Management of Large Scale Hierarchical VOD Storage System

LI Yong　WU Fei　CHEN Fu-jie

(*Department of Computer Science　Changsha Institute of Technology　Changsha*　410073)

Abstract　Recent advances in computing and communication technologies have made VOD (video-on-demand) technically feasible and economically. The VOD system needs massive storage server because of the characteristic of the continuous media. Hierarchical storage is a good solution for the system under low cost. In this article, the authors first propose a hierarchical architecture model and the concept of disk cache, then develop a replacement algorithm based on access frequency by the model, finally analyse and simulate the algorithm. The results show that the algorithm avoids cache pollution problem in LFU (least frequently used) algorithm and well suits the applications of the continuous media data objects.

Key words　Continuous media, VOD (video-on-demand), hierarchical, storage server, replacement algorithm.