3TNet 视频点播系统中用户点播行为仿真及应用*)

徐 锦 朱 明 郑 烃

(网络传播系统与控制联合实验室 网络传播系统与控制安徽省重点实验室 合肥 230027)

摘 要 通过分析 3TNet 视频点播(VOD)系统中的媒体服务情况,集中研究了用户点播模式、媒体热度分布,根据实际系统中用户点播分布的研究结果构建了一个用户点播行为的仿真平台。该平台描述了用户点播行为的一般数学模型,刻画了用户进入系统的行为分布特点。实验表明,VOD 用户点播行为仿真平台能够客观真实地反映用户的点播行为,并通过构建一个仿真的视频点播系统来验证此平台的可用性,利用此仿真视频点播系统比较了两种部署策略的优劣。

关键词 视频点播,仿真平台,坐标轴变换,数学模型

Simulating User Behavior of Video-on-demand Systems of 3TNet and Application

XU Jin ZHU Ming ZHENG Quan

(Joint Lab of Network Communication System and Control, Key Lab of Anhui, Hefei 230027, China)

Abstract Analyzing media server of 3TNet video-of-demand (VOD) system, the study focuses on user access patterns, popularity distribution, with the result, a simulation platform for user behavior was proposed. A general mathematics model for user behavior in VOD system was built to present the distribution of user visiting action. The simulation indicates that the platform objectively reflects the user action in VOD system, and using this platform simulates a VOD system to prove its usability, comparing two deploy algorithms's performance.

Keywords Video-on-demand system, Simulation platform, Axis-transformation, Mathematical model

1 引言

计算机和通信技术的快速发展使得视频点播服务成为可能,且将很快成为高速网络中最重要的服务方式之一。由于媒体数据的海量特性和视频服务器服务能力的限制,典型 VOD 系统均须采用一定的流调度算法来管理系统和提高服务能力。研究用户点播行为,而后根据用户点播特性来合理建模,能够为上述算法的测试及验证提供一个合理的平台。所以,关于用户点播行为的仿真研究至关重要。本文着重研究 VOD 仿真平台的建模过程及其应用。

Yu 等人研究了国内一个大型 VOD 系统,指出新加入的影片对于媒体热度分布具有强烈影响[4]。Frank T. Johnsen 着重分析新闻媒体的用户点播行为,提出对媒体点播行为发生器 MediSyn[1]的一些改进措施,使其能够适用于对新闻媒体负载的建模[6]。以上研究分析了用户点播行为所具有的一般特征,并未设计出一种可用的用户点播行为仿真平台。本文通过分析 3TNet 视频点播(VOD)系统中用户的点播特性,着重研究用户点播行为的仿真平台设计及其应用。

2 系统架构描述

3TNet 是国家 863 项目"十五"重大专项高性能宽带信息网的简称,3TNet 代表 T 比特级光传输系统、T 比特级自动交换光网络(ASON)和 T 比特级双协议栈路由器。图 1 所示为 VOD 系统的整体架构。图中 CS 为内容服务器,存放所有的影片数据(完整的影片);CDP 为内容分发平台,存放 40%的影片数据(以数据块方式存放),起二级缓存的作用;MS 为媒体服务器,存放 20%的影片数据(完整的影片),为机顶盒提供影片数据流;BMG 为机顶盒,解码并播放影片数据流;ASON 为自动交换光网络。

本文研究数据全部来自上海长宁区实际运营中的 3TNet 视频点播系统,通过对近三个月内 20 多万不同用户对 5000 多部不同影片超过 400 万个点播请求的数据进行用户点播行为的研究。下面从用户点播模式、媒体热度分布方面分别介绍仿真平台的建模过程。

3 用户点播行为建模

3.1 用户点播模式

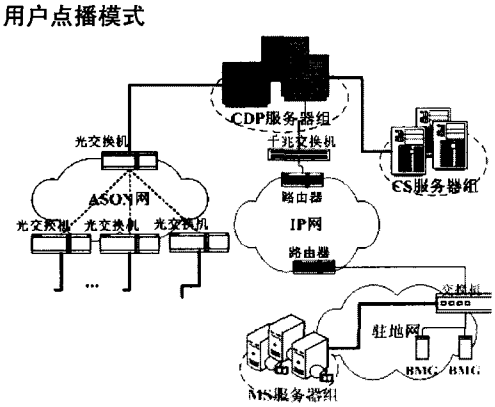

图 1 VOD 系统架构

*)国家 863 重大专项《高性能宽带信息网内容分发平台工程化》(2005AA103310)。徐 锦 硕士研究生,主要研究方向为多媒体网络中的性能测试;朱 明 博士生导师,教授,主要研究方向为宽带多媒体服务、Internet 多媒体服务等;郑 烃 硕士生导师,副教授,主要研究方向为网络多媒体。

用户点播模式即在一段时间内用户访问系统的分布，它反映了用户访问系统与时间的相互关系。了解用户的点播模式可以对系统的媒体服务情况有个基本认识。本文通过对一天内每小时、一周内每天用户访问分布，得到用户点播模式的基本特征。

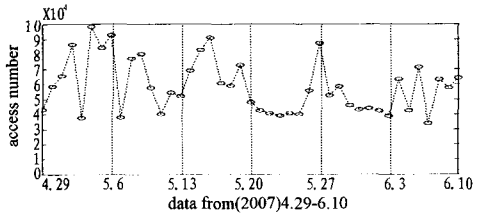

图 2　六周内用户每周的点播分布

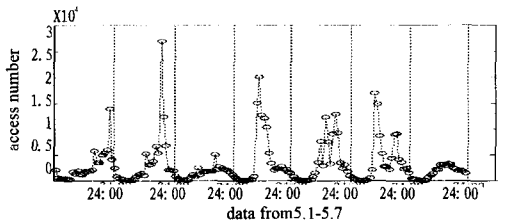

图 3　一周内用户每天的点播分布

首先分析六周之内用户点播模式的变化。其每日点播分布如图 2 所示，以每周周日作为分隔。从图中可看出，尽管用户每天的请求数目变化很大，但基本上前三个工作日的总点播量要大于每周的后四日。这与 Yu[4] 的研究结果恰恰相反，这是由 3TNet 视频点播业务所服务的用户每周的工作习惯所决定的，周末时外出游玩，所以点播率反而会下降。本文的仿真平台主要针对一天内用户点播行为，所以可以不考虑每天之间的点播行为变化。

下面着重研究一天内用户访问模式所具有的特征。从图 3 可知，用户每天的点播模式具有很清晰的特征，在用户一天内的点播分布中，总是围绕着几个峰值，而这些峰值一般都发生在正午(12PM-2PM)及晚上(8PM-11PM)。文献[1]使用 Possion 分布来描述一天内的用户点播模式，但观察图 3 可看出，Possion 分布不能反映出其所具有的多峰值特征，所以本文考虑用正态分布来描述用户每日的点播模式。但显然，一个正态分布也不能反映出其多峰值特征，由此用几个不同正态分布所构成的综合分布来描述用户每日的点播分布。只需确定每组正态分布的均值、方差及其之间的相对概率，即可得到一个综合的分布函数。表 1 所示为用来仿真一天内用户点播模式所需的数据，假设一天内的总点播量为 $M=10$ 万。

表 1　正态分布相应参数

组	1	2	3	4
μ	10	12	21	23
σ	5	3	4	1
比率	5%	24%	16%	55%

其形成的分布如图 4 所示。适当调整每组正态分布的均值及方差即可得到不同的每日点播模式，各个时刻的点播个数亦可计算得出，y 轴为每个时刻点播个数占总点播数的概率，由此可以很容易计算出每个时刻的点播个数。如 t 时刻的点播个数占总点播数的概率为 p_t，则 t 时刻的用户点播个数 N_t 为：

$$N_t = M \times p_t \tag{1}$$

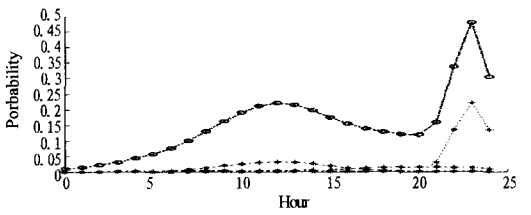

图 4　建模得到的每日点播分布

图 4 建模所得的用户每日点播模式，其分布反映了用户一天内点播行为所具有的特点：在正午时会达到每日点播的第一个峰值，第二个峰值在晚上出现，验证了本文建模方法的正确性。

3.2　热度分布

影片的热度表明影片吸引用户的程度。大多数系统普遍使用 Zipf 分布来描述影片热度的变化，Zipf 分布即第 i 部最热影片的点播次数与 $1/i^{\partial}$ 成正比，若用对数坐标来描述的话，Zipf 分布就为一条直线，其中斜率 ∂ 越大，表明越多的点播个数集中在热门的那些影片上。文献[1]中指出，当描述较长一段时间内影片的热度分布时，它并不满足 Zipf 分布，Zipf 分布只能近似描述影片热度的变化。研究时间持续越长，系统实际热度分布与 Zipf 分布的偏离就越大，最终将无法用 Zipf 分布来描述影片热度分布。

下面分析 3TNet 视频点播系统中影片热度分布情况。以用户对某部影片的点播次数作为这部影片的热度，假定在一天时间内，每部影片的热度保持不变，其热度分布如图 5 所示。这是一天(5.4)之内的影片热度分布，它很好地服从 $\partial=1.3775$ 的 Zipf 分布，只存在一些很小的波动。

图 6 所示为整个研究期间(三个月内)的影片热度分布，它并不完全符合 $\partial=0.8285$ 的 Zipf 分布，基本上是一个拱形。这是因为，当用 Zipf 分布来描述影片热度变化时，前提是在研究的期间内每部影片的热度保持不变，这在短期间内是可行的。因为短期内基本上认为每部影片的热度保持不变，但不能够很好地描述较长时间内的影片热度分布。本文提出一种坐标轴变换的 Zipf 分布，来描述较长时间内的影片热度分布。坐标轴变换的 Zipf 分布就是分别对 x 轴和 y 轴做一个坐标变换，这样在对数坐标轴下，这种分布仍然可以用一条直线来描述，如下所示：

$$x_f = \frac{x + f_x}{f_x + 1} \tag{2}$$

$$y_f = \frac{y + f_y}{f_y + 1} \tag{3}$$

其中 f_x, f_y 为坐标轴变化因子。$f_x = f_y = 0$ 就是没有经过坐标轴变换的 Zipf 分布。图 7 所示为经过坐标轴变换的影片热度分布，它基本上可用一条直线来表示。可看出，在短时间内(如一天)，$\partial > 1$；而在较长的时间内，$\partial < 1$，用变换后 Zipf 分布来描述热度分布将更加精确。

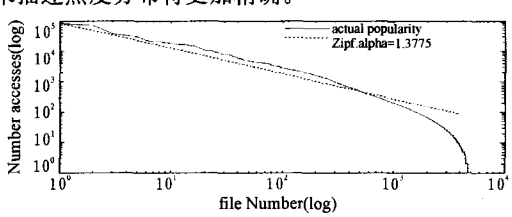

图 5　对数坐标下一天内(5.4)影片的热度分布与 Zipf 分布的契合

由于本文的用户点播行为仿真主要是针对一天的，因此影片的热度变化模型可以用 Zipf 分布来描述。

3.3 影片长度

在 3TNet 视频点播系统中，所有影片长度基本相同。本文对于影片长度的建模假设所有影片均为 2h，并且对所有影片的点播均从片头开始，点播长度在 0～7200s 之间随机产生，并随机分配到对每个文件的点播行为中。

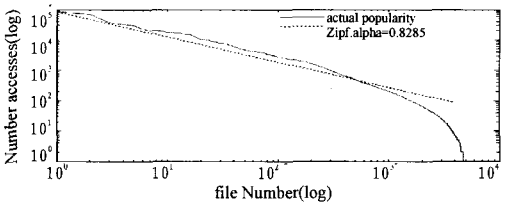

图 6 对数坐标下整个研究期间内影片的热度分布与 Zipf 分布的契合

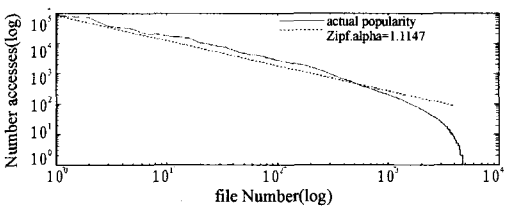

图 7 坐标轴变换后的影片热度分布与 zip 分布的契合

4 仿真平台的测试及应用

用户点播行为仿真平台是基于上述数学模型实现的。下面是仿真平台所生成的一天内用户点播行为的典型数据和使用其生成的点播数据来搭建一个视频点播系统的仿真应用。

4.1 用户点播行为仿真

对一天内的用户点播行为进行仿真。图 8 所示为仿真平台生成的用户点播行为强度分布，这与实际系统中用户的点播行为相似，证明了模型的可用性。图 9 所示为影片的热度分布，它严格地服从 Zipf 分布，符合一般的用户点播规律。

4.2 仿真平台的应用

下面利用仿真平台生成的点播数据来搭建一个视频点播系统，验证两种不同的部署算法的优劣。仿真时间为一天；5000 部影片，其中热门影片为 500 部。采取全存策略，对非热门影片的第一种存储策略为：MS 存储所有影片的片头，片尾按照指定的复制份数 N 存储到随机选择的 N 个 PN 节点中；第二种非热门影片的存储策略为：把影片分为 4 部分 P_1，P_2，P_3，P_4，按照用户的每日点播列表（用仿真平台来生成用户的点播列表），统计出 4 部分影片的点播次数。MS 选择存储点播次数最大的部分 $P_i(i=1,2,3,4)$，影片剩下部分按照指定的复制分数 N 存储到随机选择的 N 个 PN 节点中。仿真系统配置环境如表 2 所示。

下面通过对系统的本地数据击中率及用户请求的响应时间两方面来比较两种部署策略对系统性能的影响。图 10 所示为本地数据击中率的比较，其中 B 为第二种部署策略的本地数据击中率，可看出，第二种部署策略的每个 PN 节点的本地数据击中率基本维持在 98%以上，说明用户请求的数据中，98%的数据本地都有存储，远远高于第一种的 94%，体现了部署策略的优势。图 11 所示为请求响应时间的比较，其中 B 为第二种部署策略的请求响应时间。可看出，第二种部署策略有效提高了用户的请求响应时间。而且请求数目越大，两种部署策略的响应时间差别越大，第二种部署策略更能满足大量数据请求的处理，减少节点间数据的传送，可以缓解网络带宽的压力。

仿真平台的搭建对于评价各种算法的优劣提供了一个便利的平台，有利于对算法的性能进行分析。

表 2 系统配置环境

PN 节点数目	15
链路带宽	5000kbps
链路延迟	10ms
影片数目	5000
码率	6Mbps
请求数目	90000
仿真时间	1(day)
影片等级数目	4
各等级比例	1:1:4:4
各等级头部大小	{5529600000 1382400000 1382400000 1382400000} (Byte)
数据分块大小	9.28MB
数据块复制份数	4

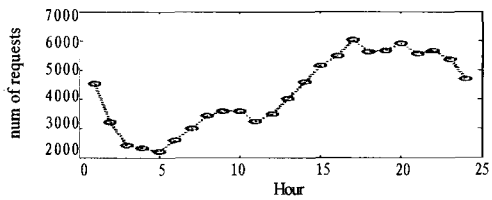

图 8 用户点播强度分布

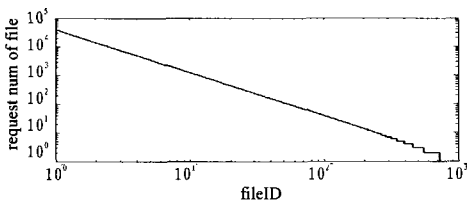

图 9 影片热度分布

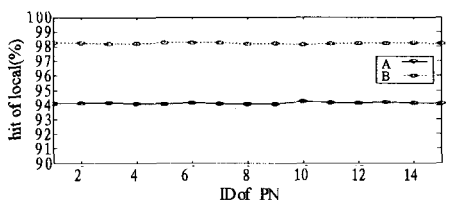

图 10 两种存储策略下本地击中率比较

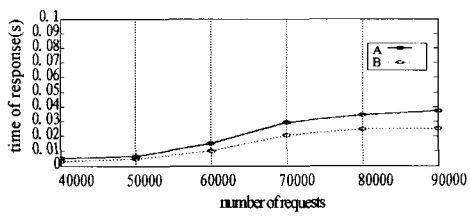

图 11 两种存储策略下请求响应时间比较

结束语 搭建仿真平台的目的在于方便测试和评价各种算法的优劣。本文首先分析了现有系统用户点播行为的服务情况，然后构建了一个用户点播行为的仿真平台，并应用其搭

（下转第 188 页）

表 2 转换后的格式

Webpage	Users
frontpage	a, e, g, k, …
news	b, c, …
tech	c, j, …
local	d, j, …
opinion	f, h, i, j, …
…	…

按照表 2 的格式来找最大频繁闭项目集。设滑动窗口的窗口大小为 4，最小支持度为 40%，使用改进的 Mafia 算法在得到事务内的最大频繁项目集的同时还得到了每一个集合中的每一项在原数据库中所在事务号的交集(对应算法 1)。再进行第二步，将找到的最大频繁项目集进行事务间的转换(对应算法 2)，得到的结果如表 3。

表 3 使用滑动窗口转换到的事务间最大频繁项目集

No.	Users(sub_window)	CUI
1	e, d, i, h, c, l, f, n, m[1]	2 3 4 5 7 14
2	a, l, f, m[1]	3 4 5 6 9 10
3	g, j, d, o, h, c, l, n, m[1]	1 2 3 4 5 7
4	k, d, c, l, n, m[1]	1 2 3 4 13 14
5	d, b, c, l, n, m[1]	1 2 4 5 7 13
6	a, l, f, m[2]	2 3 4 5 8 9
7	a, l, f, m[3]	1 2 3 4 7 8
8	p, l, f, n, m[4]	1 2 3 4 7 10

前面已经设最小支持度为 40%，此处设最小置信度为 60%。对于用户来说，存在一个容错的问题。即用户允许网站向他们推荐的网页中存在并不是他们兴趣所在的网页，但是要有一个限度。这个限度就是置信度。例如：第 6 行的交集有 67%是包含在第 1 行中的，那么就有 $e, d, i, h, c, l, f, n, m[1] \Rightarrow a, l, f, m[2]$ 置信度为 67%。

根据上述的寻找关联规则的方法，可以得到关联规则：

$e, d, i, h, c, l, f, n, m[1] \Rightarrow a, l, f, m[2]$，置信度为 67%。

$g, j, d, o, h, c, l, n, m[1] \Rightarrow a, l, f, m[3]$，置信度为 83%。

$k, d, c, l, n, m[1] \Rightarrow p, l, f, n, m[4]$，置信度为 67%。

需要注意的是生成前两个关联规则右边都是 a, l, f, m，其中第二个关联规则的置信度 83%，要大于第一个关联规则的置信度 67%。所以舍掉第一个。则最终生成的关联规则为：

$g, j, d, o, h, c, l, n, m[1] \Rightarrow a, l, f, m[3]$，置信度为 83%。

$k, d, c, l, n, m[1] \Rightarrow p, l, f, n, m[4]$，置信度为 67%。

从此可以看出，如果用户 l 看了网页 3，那么就可以推断网页 1、2、4、5、7、13、14 中有 l 想了解的信息，并推荐给用户 l。

结束语 本文提出了一种新的基于 Web 事务间关联规则的挖掘方法。以直接找用户之间关联规则的思想为基础，首先提出了用改进的 Mafia 算法找到最大频繁项目集以及对应的 CUI，然后以 CUI 为依据来对项目集进行由事务内到事务间的转换，相比从直接生成事务间项目集的找事务间关联规则的方法要简单且高效，能够推荐给用户更多包含他们感兴趣信息的网页，并随着网站数据的不断增加，仍然继承了 Mafia 算法对大数据集处理效率高的优点。

参 考 文 献

[1] Ting I H, Kimble C, Kudenko D. Applying Web Usage Mining Techniques to Discover Potential Browsing Problems of Users// Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference. 2007:929 - 930

[2] Baeza - Yates R, Hurtado C, Mendoza M. et al. Modeling user search behavior // Comput. Sci. Dept., Chile Univ., Chile, IEEE Web Congress, 2005. LA-Web 2005. Third Latin American

[3] Tung A K H, Lu Hongjun, Han Jiawei, et al. Efficient Mining of Intertransaction Association Rules. IEEE Transactions on Knowledge An Data Engineering, 2003, 15(1)

[4] Chen Jian, Yin Jian, Tung A K H, et al. Discovering Web Usage Patterns By Minig Cross-transaction Association Rules // Proceedings of the third international conference on machine learning and cybernetics. Shanghai, 2004:26-29

[5] Yang Wanzhong, Li Yuefeng, Xu Yue. Granule Based Inter - transaction Association Rule Mining. 2007 IEEE, DOI 10. 1109/ICTAI. 2007, 143

[6] Burdick D, Calimlim M, Gehrke J. MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases. 1063-6382/01 2001 IEEE

[7] Buzikashvili N. Sliding Window Technique for the Web Log Analysis. WWW 2007. May, Banff, Alberta, Canada. ACM 978-1 —59593-654-7/07/0005

[8] Tanasa D, Trousse B. Advanced Data Preprocessing for Intersites Web Usage Mining. Intelligent Systems, IEEE 2004, 19 (2):59-65

[9] 张波，巫莉莉，周敏. 基于 Web 使用挖掘的用户行为分析. 计算机科学，2006，33(8)

[10] Agrawal R, Srikant R. Fast algorithms for Minging Association Rules // Proceedings of the 20th VLDB Conference. Santiago, Chile, 1994

(上接第 94 页)

建了一个仿真的视频点播系统，验证了两种部署算法的优劣，但本文的仿真平台只针对一天内的用户点播行为，下一步可继续研究长时间内用户点播行为的仿真。

参 考 文 献

[1] Tangy W, Fuz Y, Cherkasovay L. Amin Vahdatz; Long-term Streaming Media Server Workload Analysis and Modeling[J]// Proceedings of the 13th International Workshop on Network and Operating Systems Support for Digit. June 2003

[2] Cherkasova L, Gupta M. Analysis of enterprise media server workloads: access patterns, locality, content evolution, and rates of change [J]. IEEE/ACM Transactions on Networking (TON), 2004, 12(5)

[3] Vilas M, Paneda X G, Garcia R, et al. User behaviour analysis of a video-ondemandservice with a wide variety of subjects and-lengths[J]//EUROMICRO. IEEE Computer Society, 2005

[4] Yu Hongliang, Zheng Dongdong, Zhao B Y, et al. Understanding user behavior in large-scale video-on-demand systems[J]//ACM SIGOPS Operating Systems Review Proceedings of the 2006 EuroSys Conference EuroSys '06. April 2006

[5] Johnsen F T, HafsØe T, Griwodz C. Analysis of Server Workload and Client Interactionin a News-on-Demand Streaming System[J]//IEEE ISM. San Diego, CA, USA, December 2006

[6] Frank J T, Trude H, Carsten G, et al. Workload Characterization for News-on-Demand Streaming Services [J] // Performance, Computing, and Communications Conference, IPCCC 2007. IEEE Internationa, April 2007