

多媒体文件欢迎度的分析和预测*

李金龙, 潘 嵘

(中山大学软件研究所, 广东 广州 510275)

摘 要: 多媒体文件的欢迎度是随时间而变化的, 分析预测欢迎度并将其应用到节目存储管理中可以降低 VoD 系统的性价比。对用户点播数据的分析表明, 大部分多媒体节目欢迎度变化过程包括增长、平稳和下降三个阶段。文章探讨了完整变化过程的阶段划分算法, 以及节目所处阶段的判断算法, 并用回归分析的方法探讨了各个阶段欢迎度的分析和预测。

关键词: VoD; 趋势预测; 阶段划分

中图分类号: TN949 **文献标识码:** A **文章编号:** 0529-6579 (2003) S1-0261-03

VoD 就是根据用户的要求播放节目的服务。由于用户在地理上的分散性, 采用分布式视频服务器 (DVS) 相对于集中式视频服务器能有效降低系统成本。文献 [1, 2] 论述了层次图结构和网状结构的分布式服务器中节目的存储策略, 文献 [3] 论述了冗余层次图结构 DVS 的节目存储和轮换策略。但节目欢迎度都采用齐夫定律模拟, 没有体现出节目欢迎度随时间而变化这个特征。

文献 [4] 指出, 通过分析 VoD 的点播记录, 可以得到节目的点播次数的序列, 这个点播次数序列可以看作时间序列, 可以用时间序列趋势分析的方法分析和预测请求频率。每个节目的点播次数序列的变化趋势和同类节目的历史趋势具有相似性, 并提出按是否考虑历史数据、是否固定预测模型提出了四种预测方式: Fixed Regression Selecting (FRS), Continuous Regression Updating (CRU), Historical Updating (HU), Continuous Regression with Historical Updating (CRHU)。但是文献 [4] 认为多媒体文件的欢迎度总体上呈下降趋势的, 仅讨论了点播次数下降阶段的趋势分析和预测。

经对中山大学软件所 VoD 系统的用户点播数据进行细致分析, 发现大部分节目欢迎度的变化过程均由增长、平稳和下降三个阶段组成, 我们把这三个阶段称为完整变化过程。本文针对这个规律, 讨论了欢迎度完整变化过程的分析和预测。

1 欢迎度的度量方法

为了对 VoD 节目点播的趋势进行有效的分析及预测, 必须选择适当的量来代表节目的欢迎度。

文献 [4] 直接分析预测节目的点播次数, 但在用户个数变化较大的 VoD 系统中, 点播次数不能准确地度量欢迎度。

由于总点播次数综合了所有用户的点播量, 所以 [节目点播次数/总点播次数] 能克服用户点播能力及兴趣的差异对点播次数的影响, 较为准确地度量节目的实际欢迎度。由图 1 可以看出, [节目点播量/点播总量] 的变化曲线相对比较光滑, 更适合于分析和预测。

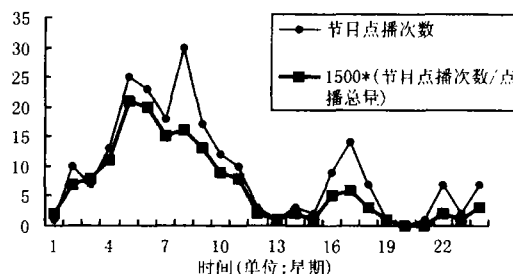


图1 欢迎度度量选取对比图

Fig.1 Metric selection of program's popularity

2 欢迎度变化过程的阶段划分

2.1 阶段简介

通过观察和分析软件所 VoD 系统中节目的欢迎度变化趋势, 发现大部分多媒体文件要经过被用户知晓、大量点播、逐渐过时三个阶段, 欢迎度要经过增长、平稳振动、下降三个阶段。节目加入 VoD 系统后, 逐渐在用户群体中流传, 其欢迎度逐渐增长。这个增长过程与新产品的推销类似, 是有饱和值的增长时间序列^[5]。欢迎度达到饱和值后, 节目在用户群体中的欢迎度趋于稳定。经过一段时

* 收稿日期: 2003-03-28

作者简介: 李金龙 (1975 年生), 男, 硕士生; E-mail: lucida163@163.com

期的大量点播后,节目逐渐过时,欢迎度随着下降。我们把欢迎度要经历的三个阶段说明如下:

- (1) 节目加入 VoD 系统后,随着节目在用户中的流传,欢迎度迅速增长的时期称为增长期;
- (2) 增长期结束后,欢迎度要经历一个基本平稳的时期,称为平稳期;
- (3) 平稳期结束后,欢迎度在较长时期内缓慢下降,称为下降期。

图 2 欢迎度变化过阶段划分的示意图。

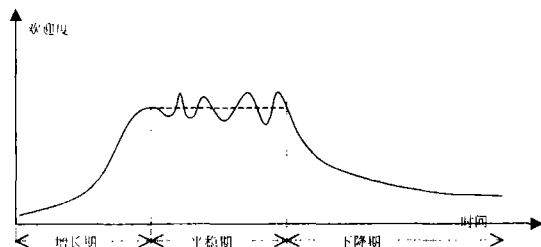


图 2 VOD 节目欢迎度阶段划分的示意图

Fig.2 Sketch map of program's stage dividing

2.2 完整欢迎度变化过程的阶段划分算法

我们把欢迎度经历了增长、振动、下降三个阶段的变化过程称为完整变化过程。设在时间 $(t_1, t_2, \dots, t_n)$ 上的欢迎度序列 $(y_1, y_2, \dots, y_n)$ 是一完整变化过程,本文引用了有序样本聚类的最优分割法^[6]对其进行阶段分割,采取了以下步骤:

- (1) 确定节目欢迎度大于最小欢迎度的第一个时间点 t_a ,及节目欢迎度大于最小欢迎度的最后一个时间点 t_e ;

- (2) 使用有序样本聚类的最优分割法对 $[t_a, t_e]$ 内的样本按时间顺序分割为 3 部分,记为 $G_1 = [t_a, t_b]$, $G_2 = (t_b, t_c)$, $G_3 = [t_c, t_e]$ 。

- (3) 分别求出 G_1, G_2, G_3 内的欢迎度均值 $\bar{y}_1 = \frac{1}{b-a} \sum_{i=a}^b y_i$, $\bar{y}_2 = \frac{1}{c-b} \sum_{i=b+1}^{c-1} y_i$, $\bar{y}_3 = \frac{1}{e-c} \sum_{i=c}^e y_i$ 。

- ① 若 $\bar{y}_1 \geq \bar{y}_2$ 或 $\bar{y}_3 \geq \bar{y}_2$,则不能划分为上述三段;

- ② 否则,此节目欢迎度变化过程可以划分为增长、平稳、下降三个阶段,各阶段起始点的确定方法如下:

增长期开始点: t_a ;

第一个欢迎度大于等于 $\bar{y}_2$ 的时间点为增长期结束(平稳期开始)的时间点,记做 $t_b = \min\{t_i \mid y(i) \geq \bar{y}_2\}$;

最后的欢迎度大于等于 $\bar{y}_2$ 的时间点为平稳期结束(下降期开始)的时间点,记做 $t_c = \max\{t_i \mid y(i) \geq \bar{y}_2\}$;

由于上升期和下降期的欢迎度均值要小于平稳期均值,所以上述算法能准确划分阶段^[6]。

2.3 节目所处阶段的判断

当前欢迎度较大的节目,其所处阶段的判断过程如下。

通常情况下,节目加入 VoD 系统后,欢迎度首先要增长一段时期,即进入增长期。判断的方法是,如果节目加入系统的时间较短,且当前有增长趋势,则说明节目处于增长期。增长期结束平稳期开始的判断较为简单,因为增长期趋于结束时,曲线的斜率会急速地降低,如果用增长曲线建模,增长期结束时增长曲线将逼近饱和值。平稳期结束、下降期开始的判断较难,因为平稳期趋势线本身是有振动的。经分析实际数据发现,平稳期振幅一般不超过平稳期平均值的 20% ~ 30%。所以本文的策略是,当欢迎度小于平稳期平均值的 70% 时,认为已进入下降期,把最后一个欢迎度等于平稳期均值的时刻点认为是下降期的开始点。

3 各阶段模型的选择

我们用统计回归的方法来分析各阶段的特征。SSE (残差平方和) 是衡量拟合效果的重要指标^[4], SSE 的计算方法引自文 [7]。为了便于分析比较,我们用线性模型的 SSE 除以所有模型的 SSE,得到了标准化 SSE。本文使用中山大学软件所 VoD 系统的历史数据,分别就三种节目(流行歌曲、经典歌曲、电影)的点播历史进行了分析。

由表 1 可以看出,流行歌曲最适合的模型是 Gompertz,而对于经典歌曲和电影,线性模型是最适合的。可能的原因是,流行歌曲在传播的过程中,其对用户的吸引力也在增加;而经典歌曲和电影加入 VoD 系统前,用户已经通过各种渠道对其有相当的了解,对用户的吸引力是基本不变的,而且同一用户重复点播经典歌曲和电影的可能性较小,所以其欢迎度的增长率变化不大。

表 1 增长期的标准化 SSE

Tab.1 Normalized SSE of grow phase

SSE	流行歌曲	经典歌曲	电影
线性模型	1	1	1
Gompertz	0.72	1.38	1.64
Logistu	0.96	1.47	3.66

GM (1, 1) 对平稳期的拟合效果比线性模型要好,多项式模型虽然拟合数据的效果较好,但不适合于长期预测。

表2 下降期的标准化 SSE
Tab.2 Normalized SSE of decline phase

SSE	流行歌曲	经典歌曲	电影
线性模型	1	1	1
指数模型	0.93	7.32	1.44
幂模型	5.32	7.43	9.64
指-指 1,5	0.93	1.70	0.89
指-幂 1,5	0.47	1.62	0.91

关于指-指、指-幂模型的定义及拟和方法见文献[4]。由表2可以看出,对流行歌曲的下降期,指-幂模型相对于其他模型有较大优势;对电影和经典歌曲的下降期,指-指、指-幂模型差别不大,线性模型表现出了很好的效果。可能的原因是,流行歌曲在欢迎度下降的过程中,其对用户的吸引力逐步减小;而经典歌曲和电影节目对用户的吸引力是基本不变的,所以其欢迎度的变化也比较稳定。

4 测试结果

本文采用综合预测性能最好的 CRHU 方式^[4],对软件所 VoD 系统历史数据做了模拟测试,和 LFU 策略进行了比较。

我们以星期为时间单位,用前 k 个星期的数据预测每个节目下个星期的欢迎度。用前 k 个星期的点播频率作为 LFU 策略的估计值。为了便于与预测结果比较,将 LFU 的估计值标准化,使所有节目的估计值之和为 1。标准化的方法是对每个节目的估计值除以所有节目的估计值之和。

图3中,横轴表示用于预测的时间长度 k ,纵轴表示所有节目的误差平方和。可以看出,预测结果比 LFU 的估计值更准确地反映了节目的欢迎度。

5 讨论

本文通过对 VoD 系统历史数据的分析,提出

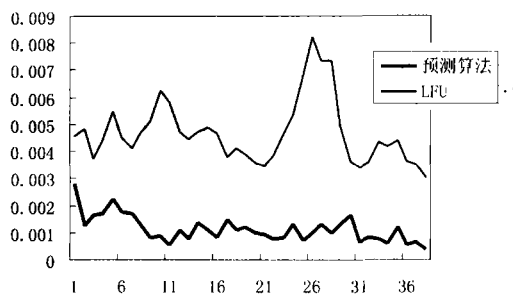


图3 预测误差与 LFU 误差比较

Fig.3 Sum error of predict algorithm and LFU

大部分多媒体文件的欢迎度变化过程可分为增长、平稳和下降三个阶段,设计了阶段的划分算法;并探讨了各个阶段适合的预测模型。实际系统的测试结果表明,用趋势分析预测欢迎度的方法,比 LFU 有较大的优势。

参考文献:

- [1] BISDIKIAN C C, PATEL B B. Issues on movie allocation in distributed video-on-demand systems[J]. IEEE International Communications Conference, 1995, 250 - 255.
- [2] BISDIKIAN C C, PATEL B B. Cost-based program allocation for distributed multimedia on demand systems[J]. IEEE Multi Media, 1996, 62 - 72.
- [3] SHAHABI C, ALSHAYEJI M H, WANG S M. A redundant hierarchical structure for a distributed continuous media server [C]. Proceedings of the IDMS'97, 1997, 1 - 25.
- [4] DANNY M P Ng, ERIC W M, WONG K T, et al. Tend analysis and prediction in multimedia-on-demand systems[J]. Proceedings of International Conference on Communications, Helsinki, 2001: 1292 - 1298.
- [5] 封建强. 商业和经济预测中的时间序列模型[M]. 北京: 中国人民大学出版社, 2002.
- [6] 张尧庭, 方开泰. 多元统计分析引论[M]. 北京: 科学出版社, 1982.
- [7] 张小蒂. 应用回归分析[M]. 杭州: 浙江大学出版社, 1991.

Analysis and Prediction of Multimedia Popularity in VoD Systems

LI Jin-long, PAN Rong

(Software Institute, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: The popularity of multimedia contents changes along with time. Most of the Multimedia contents's popularity changing procedure can be divided in to three phases: grow phase, stable phase and decline phase. In this article, partition algorithm of the whole procedure and Judge algorithm of current files are proposed, Analysis and prediction of multimedia popularity is studied in regression method also.

Key words: VoD; phase partition; trend analysis and prediction