高性能存储系统的读请求的建模和预测*)

魏文国
(广东技术师范学院电子信息工程系　广州 510665)

摘　要　为了改善高性能计算系统的I/O性能，本文设计和实现了一个读请求的空间和时间特性的综合预测模型—马尔可夫空间模型与自回归集成的移动平均时间模型，该模型在应用程序执行过程中被自动识别；每个新的请求到达时模型参数被实时递归地估计，并根据应用程序本身和运行环境自适应改变。经过实验检验我们的建模和预测比较有效，测试例子表明时间模型的预测值与观测值的相对误差的平均值为0.16%，并且预测的稳定性较好；空间模型的预测准确性也较好。

关键词　并行文件系统，读请求建模，预测，马尔可夫空间模型，读请求的时间模型

Automatic Modeling, Forcasting of Read Request about High Performance Storage System

WEI Wen-Guo
(Department of Electronic Information Engineering, Guangdong Polytechnic Normal University, Guangzhou 510665)

Abstract　In order to improve I/O performance of high performance computing systems, this paper designs and implements an integration modeler to model and predict for read request interarrival times and spatial characteristic; models are automatically identified during application execution, model parameters are recursively estimated in real-time for every new I/O arrival, adapting to changes that are intrinsic to the running application, the implementation shows that our modeler has good veracity and validity, we compare and calculate 2 examples and know the average relative forecast error of interarrival times modeler is 0.16%, forecast is steady; and Markov model spatial predictions also has good veracity.

Keywords　Parallel file system, Modeling of read request, Markov spatial model, Read request interarrival times model

1　背景和相关工作

读预取和延迟写是隐藏或者减少I/O延迟的两个标准技术，因为处理器负载的波动和操作系统队列延迟；加上应用程序周期性的检查点和嵌套的循环结构，使得I/O请求到达具有突发性；并行程序的文件访问的时间与空间特性比串行程序更复杂，因为多级别的I/O库的原因，不仅I/O模式具有突发性，时间模型也不规则。为了有效地预取，需要考虑这些变化的条件。我们的研究从Nancy Tran博士的工作[1]中取得灵感，参考和借鉴了他的部分研究方法，组合时间序列和马尔可夫空间预测模型来获得及时的预取，该方法更加周密、有效。

2　在线读请求的空间和时间建模

2.1　读请求的空间特征建模——马尔可夫模型

马尔可夫模型通过状态模型来描述复杂的I/O行为，该模型中状态代表固定大小的文件块，边的权值表示相关文件块被相继访问的概率。

我们使用马尔可夫模型来描述应用程序的I/O访问的空间流，具有 m 个状态的马尔可夫模型可用状态转换矩阵 $P=(p_{ij})_{m*m}$ 表示，其中每个元素 p_{ij} 表示从状态 i 转换到状态 j 的概率，并且转换的概率仅与当前的状态 i 有关。

读请求的马尔可夫空间模型预测主要包括：贪婪预测和路径预测两种方法，我们在此只介绍贪婪预测方法。

贪婪预测策略：对每个当前状态，预测总是取最大可能的下一个状态，预测 N 步就是重复 N 次前面的步骤。其算法如算法1所示：

算法1：读请求的马尔可夫空间模型贪婪预测算法

输入：状态转换矩阵 $P=(P_{ij})_{M*M}$，预测的步数 N 及初始状态 s_{i_0}
输出：预测的 N 块序列 $\{s_{i_1},s_{i_2},\cdots,s_{i_N}\}$
Begin
　(1)求矩阵 P 的第 i_0 行的最大值 $p_{i_0i_1}=\max\{p_{i_01},p_{i_02},\cdots,p_{i_0M}\}$，第 i_1 行对应的块 s_{i_1} 就是预测的第一块。
　(2)求矩阵 P 的第 i_1 行的最大值，$p_{i_1i_2}=\max\{p_{i_11},p_{i_12},\cdots,p_{i_1M}$，第 i_2 行对应的块 s_{i_2} 就是预测的第二块。
　……
　(n)求矩阵 P 的第 i_{N-1} 行的最大值 $p_{i_{N-1}i_N}=\max\{p_{i_{N-1}1},p_{i_{N-1}2},p_{i_{N-1}M}\}$，第 i_N 行对应的块 s_{i_N} 就是预测的第 N 块。
　$(n+1)$前面 n 步得到的块 $s_{i_1},s_{i_2},\cdots,s_{i_N}$ 就是预测的 N 块序列 $\{s_{i_1},s_{i_2},\cdots,s_{i_N}\}$。
End

2.2　读请求的时间序列模型

时间序列预测模型使用统计方法分析按时间排序的读请求的观测值序列的依赖关系。基于"将来的I/O是过去的延伸"的假设，通过分析观测值序列内在的依赖关系来预测将来的I/O行为。通过观测值之间的相关度来识别I/O序列的结构。

自回归、集成的移动平均时间序列预测模型为I/O提供复杂的统计建模方法，包括平稳的、变化的和具有周期(季节)的三类时间序列。直观上说，平稳的时间序列是指其观测值的时间间隔序列围绕一个恒定的平均值变化；变化/非平稳的时间序列则不围绕平均值变化，而是呈献出增加或者减少的时间间隔趋势；若其观测值序列有规律地以特定的模式重复就称为周期/季节性时间序列。

为了刻画I/O请求的平稳时间序列 $\{f(t)\}_{t=1,2,\cdots}$，我们使用自回归，集成的移动平均 $ARMA(p,q)$ 模型，即第 t 步读操作发生的时刻 $f(t)$ 可以根据前 p 个观测值(自回归部分 AR，即等式1的第一行)和过去的 q 个扰动值(移动平均部分

*)基金项目：中国教育、研究网格(ChinaGRID)项目(CG2003-CG005)。魏文国　副教授，研究方向为集群，计算机网络和高性能计算。

MA，即等式1的第二行）等的线性组合来预测，如等式1[1]所示：

$$f(t)=a_0+a_1f(t-1)+a_2f(t-2)+\cdots+a_pf(t-p)+e(t)+b_1e(t-1)+b_2e(t-2)+\cdots+b_qe(t-q) \quad (1)$$

非平稳序列的 $ARMA(p,d,q)$ 模型是：对时间序列经过 d 阶正常的差分（时间序列的相邻观测值依次相减）之后，最后的时间序列有 p 项自回归部分和 q 项移动平均部分。季节性序列的 $ARMA(p,d,q)\times(P,D,Q)_s$ 模型是：在 $ARMA(p,d,q)$ 模型的基础上添加了季节性部分：$(P,D,Q)_s$，该部分经过 D 阶跨距为 S 的差分后，再经过 d 阶正常的差分，观测值的相关度结构是 p 项自回归部分和 q 项移动平均部分。季节性序列模型也称为一般的自回归。集成的移动平均模型，因为它包括所有的三类时间序列（平稳、非平稳和季节性）及其组合。

3 自动模型识别

自回归、集成的移动平均模型使用自相关度函数（autocorrelation function，简称为 ACF）和偏自相关度函数（partial autocorrelation function，简称为 PACF）来识别模型（有关概念请参考文[1]）；并使用 Haar 小波变换自动检测 I/O 行为的突然变化[2]。

3.1 自动区分 ACF 和 PACF 的衰减类型

当衰减具有单调性时，主要分为缓慢衰减，指数衰减和突然衰减三类。识别信号的衰减类型可以采用衰减计数测试法：差分 ACF 或者 PACF 序列并找到高频信号的位置，若高频信号单调递减，并且高频信号的次数小于等于 2 则为“突然衰减”；否则若高频信号的次数大于等于 10 则为“缓慢衰减”或者“指数衰减”。当偶然出现非单调时则采用其他方法。

3.2 通过 ACF 和 PACF 来识别时间序列的模型结构

经验表明大多数 ACF 和 PACF 主要有两类行为：呈指数函数衰减或者显著地突然下降。根据衰减类型决定模式结构的算法如图 1 所示。

```
If (ACF或者PACF中存在高频信号)
{
   if (ACF中高频信号的衰减是突然衰减) then
      模式＝MA(q)
   else if (PACF中高频信号的衰减是突然衰减) then
      模式＝AR(p)
   else if (ACF和PACF中高频信号的衰减都是指数衰减) then
      模式＝ARMA(1, 1)
   end if
}
else
   p = q = 0
end if
```

图 1 读请求的时间序列模式识别算法

4 实时估计模型参数

实时参数估计要能适应程序执行过程中 I/O 的时间序列的变化，包括数据依赖和资源依赖两种情况。参数值的估计可以采用最小平方法：最小化估计值和实际观测值之差的平方和。对非平稳和周期性的时间序列，需要先转换成平稳序列，然后使用最小平方法的参数估计。实际应用时我们使用递归的、扩展的最小平方算法，递归减少计算开销，最小平方来最小化估计误差。最小平方基于两个假设：模型是线性的并且估计误差服从正态分布（具有 0 均值和常量方差）。

分析观测值和预测值的差，若差值表现为可区分的，非随机的模式则意味着模型结构没有捕获所有的明显相关性。当该情况发生时，检查差值的相关度函数，模型被迭代地再识别直到差值没有明显的相关度为止。一般对差值使用 χ 方差来验证是否符合正态分布。

5 N-步预测器

5.1 平稳的时间序列的 N-步预测器

按照最小平方预测的原则，从第 t 步开始，1-步预测是第 $t+1$ 步的新观测值的条件期望，因此假设时间序列模型如等式 1 所示。1-步预测标记为：$\hat{f}(t+1)=E[f(t+1)\mid f(t),f(t-1),\cdots]$。

等式 1 除了 $e(t)$ 之外，其他的项在第 $t-1$ 步都是已知的。$e(t)$ 服从高斯分布 $N(0,\sigma^2)$，我们用第 $t-1$ 步的期望值 0 代替 $e(t)$。因此，1-步预测值是过去的时间序列和噪声的组合，权重参数分别为 a_i 和 b_i（a_i,b_i 已经在建模，参数估计过程中求得）。即有：

$$\hat{f}(t)=a_0+a_1f(t-1)+a_2f(t-2)+\cdots+a_pf(t-p)+b_1e(t-1)+b_2e(t-2)+\cdots+b_qe(t-q)$$

与 1-步预测器类似，平稳的时间序列的 N-步预测可表示为[1]：

$$\hat{f}(t+1)=a_0+a_1f(t)+a_2f(t-1)+\cdots+a_pf(t-p+1)+0+b_1e(t)+b_2e(t-1)+\cdots+b_qe(t-q+1)$$

$$\hat{f}(t+2)=a_0+a_1f(t+1)+a_2f(t)+\cdots+a_p^*f(t-p+2)+0+b_10+b_2e(t)+\cdots+b_qe(t-q+2)$$

$$\cdots$$

$$\hat{f}(t+n)=a_0+a_1f(t+n-1)+a_2f(t+n-2)+\cdots+a_pf(t-p+n)+0+b_10+b_20+\cdots+b_qe(t-q+n)$$

为了计算预测值 $\hat{f}(t+j)$ 的自回归部分，以前的预测值 $\hat{f}(t+j-1),\hat{f}(t+j-2)$ 等被作为将来的到达时刻的期望值使用。为了计算移动平均部分，将来的高斯噪声项 $e(t+j),e(t+j+1)$，被它们的期望值 0 代替。

5.2 非平稳的和季节性时间序列的 N-步预测器

非平稳的和季节性时间序列的 N-步预测包括两个阶段：

(1)正如前面小节介绍的，对最终变换形成的平稳时间序列作 N-步预测：$\hat{f}(t+1),\hat{f}(t+2),\cdots,\hat{f}(t+n)$。

(2)对前一步的预测结果，运用非平稳的和季节性时间序列变换成平稳的时间序列的逆序的逆操作，则可求得非平稳的和季节性时间序列的预测值。

6 实验与测试结果

对读请求的时间预测模型的验证如下：为了检验上述模型的有效性，我们用 C++ 开发了应用层的“读请求的 ARMA 预测”软件包。借鉴伊利诺斯州大学从各种真实的科学计算实例中提取的 I/O trace[3]，我们下载“模拟原子核和电子之间交互作用的并行程序 HTF”的 I/O trace 文件，将这个 I/O trace 简称为 HTF。

HTF 记录了 444 个 I/O 访问的观测值的时间序列（时间单位：毫秒），部分数据如下，可以看出这些数据有显著的差异。

663859, 12235591, 40805921, 20083, 4976, 20043, 41033415, 259, 267……

我们的系统识别出该 I/O 序列满足的模型为：ARMA $(0,0,0)X(1,1,0)_{73}$，周期的长度为 73，与实际情况完全吻合。图 2 画出了一个周期内 HTF 的观测值与预测值的比较数据。经过计算得到：444 个预测值与观测值的相对误差的平均值为 0.16%，相对误差的平均值较小说明预测较准确，其中相对误差的最大值为 13.1%，最小值为 0.033%；444 个

预测值与观测值的相对误差的方差为 18.67，相对误差的方差小说明预测较稳定！

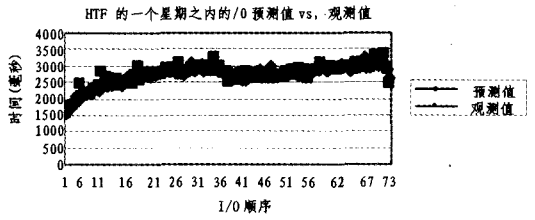

图 2　HTF 的一个周期之内的 I/O 序列的观测值与预测值

对读请求的马尔可夫空间预测模型的验证如下：不失一般性，我们取磁盘仿真工具 diskSim[4] 使用的两个真实的 I/O 跟踪文件 atlas10k. trace 和 atlas_III. trace 作为测试用例，这两个 trace 是运行 OLTP 的数据库系统的基准测试程序 TPC-C[5] 捕获的 trace，但是它们分布在不同的物理磁盘 Quantum Atlas10K 和 Atlas III 上，即该跟踪文件在不同磁盘上的空间访问特征不一样。我们从这两个 trace 文件中分别提取所有的读操作的起始位置（即磁盘块标号）形成两个序列，即 Atlas10k 的读操作的起始位置序列是：3962268，3962296，2924420，2924276，2000388，8692968，……；Atlas III 的读操作的起始位置序列是：7563004，11589236，11589108，11589040，14486396，……。将上述两个序列分别作为输入参数被我们用 C++ 开发的应用层“读请求的 Markov 空间预测”软件包调用，使用贪婪预测算法作 3-步预测。我们分别重复 1 次、2 次、……、9 次上述序列得到的实验结果如图 3 所示，可以看出随着 I/O 序列的重复次数越来越多，该模型预测的准确性越来越高，重复 1 次预测的准确性约为 48%，重复 9 次预测的准确性约为 87%，并且在两个仿真磁盘上的表现都很一致。

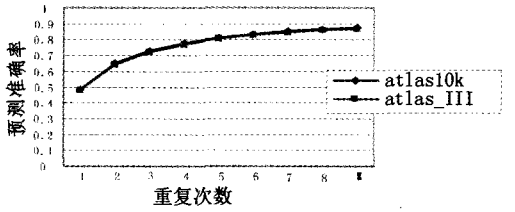

图 3　基于 3-步贪婪预测算法的马尔可夫空间预测模型的准确度

小结　本文设计和实现了一个软件框架——读请求的自动建模器：通过对读请求的空间和时间特性进行综合建模；模型参数被在线、递归地估计；根据新的参数预测将来的读请求。经过实验检验我们的建模和预测比较有效；并且预测的准确性和稳定性都较好。将来的工作是将读请求的空间和时间模型更好地集成，取得令人满意的效果。

参 考 文 献

1　Tran N, Reed D A. Automatic ARIMA Time Series Modeling for Adaptive I/O Prefetching [J]. IEEE Transactions on Parallel and Distributed Systems, 2004, 15(4): 362～377

2　Strang G, Nguyen T. Wavelets and Filter Banks [M]. Wellesley-Cambridge Press, 2002. 102～120

3　I/O traces [EB/OL]. http://www.renci.unc.edu/Project/IO/traces/experimentEntry.asp. 2005～04

4　The DiskSim Simulation Environment. Available at : http://www.pdl.cmu.edu/DiskSim/. 2005, 9

5　Transaction Processing Performance Council. Available at : http://www.tpc.org/. 2005, 9

（上接第 259 页）

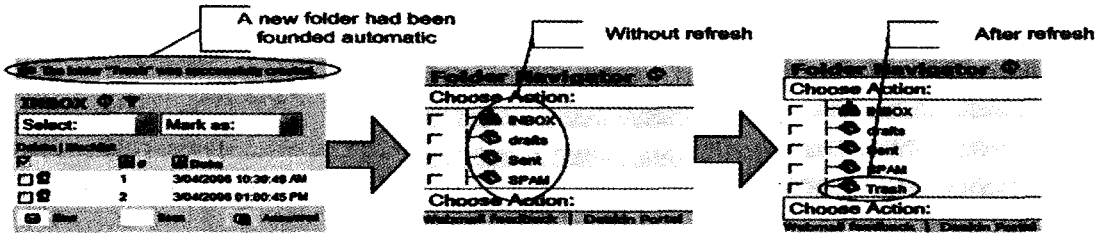

图 9　问题 3 示意图

参 考 文 献

1　Kirakowski J. The Software Usability Measurement Inventory: Background and Usage. In: Usability Evaluation in Industry, Taylor and Francis

2　Bos R, van Veenendaal E P W M. For quality of Multimedia systems: The MultiSpace approach. In: Information Management, May 1998

3　Bass L J, John B E, Kates J. Achieving usability through software architecture: [Technical Report CMU/SEI-2001-TR-005]. Carnegie Mellon University/Software Engineering Institute 2001

4　ISO/IEC FCD 9126-1. Information technology - Software product quality - Part 1 : Quality model. International Organization of Standardization, 2001

5　ISO 9421-10. Ergonomic requirements for office work with visual display terminals (VDT's) -Part 10 : Dialogue principles, International Organization of Standardization, 1994

6　ISO 9241-11. Ergonomic requirements for office work with visual display terminals (VDT's) -Part 11 : Guidance on usability, International Organization of Standardization, 1995

7　Ritter F E, Baxter G D, Jones G, et al. Supporting Cognitive Models as Users 2000 ACM, 1073-0516/00/0600-0141

8　Maloner-Krichmar D, Preece J. A multilevel analysis of sociability, usability, and community dynamics in an online health community. ACM Transactions on Computer-Human Interaction, 2005, 12(2): 201～232

9　Vanderdonckt J, Chieu Chow Kwok, Bouillon L, et al. Interaction Model-based design, generation and evaluation of virtual user interfaces. 2004 ACM 1-58113-845-8/04/0004

10　Vandervelpen C, Coninx K. Towards model-based design support for distributed user interfaces. In: NordiCHI '04, Tampere, Finland, October, 2004

11　Koski T. Hidden Markov Models for Bioinformatics. Kluwer Academic Publishers

12　Lai Xiangwei, Yang juan, Qiu Yuhui, et al. A HMM based Adaptation Model Used in Software Usability Monitoring, ICYCS, 2006

13　Baxter G D. Perspection on CHI. Addison-Wesley, 1999

14　McGregor J D, Sykes D A. A Practical Guide to Testing Object-oriented Software. Addison-Wesley, 2001

15　Bevan N. Quality and usability: a new framework. In: van Veenendaal E, McMullan J, eds. Achieving Software Product Quality, Tutein Nolthenius, 's Hertogenbosch, The Netherlands, 1997