第 32 卷 第 20 期
2010 年 10 月

武 汉 理 工 大 学 学 报
JOURNAL OF WUHAN UNIVERSITY OF TECHNOLOGY

Vol. 32 No. 20
Oct. 2010

DOI: 10.3963/j.issn.1671-4431.2010.20.025

基于数据对象的访问模型的研究

刘 群[1], 冯 丹[2], 李 坚[2]
(1. 华中科技大学网络与计算中心, 武汉 430074; 2. 华中科技大学计算机学院, 武汉 430074)

摘 要: 随着计算机技术的发展和应用的普及, 基于对象存储(Object-based Storage, OBS)技术逢时崛起, 成为下一代网络存储的主流。构建基于 OBSS 的元胞自动机(OBSSCA)框架模型, 并在此基础上分析数据对象的访问频率对系统的影响, 结合数据对象访问的特征和主动性, 通过机械学习适当调整数据对象的访问行为频率, 使系统朝着稳定方向发展。

关键词: 元胞自动机; 基于对象存储系统; 数据对象; 访问频率
中图分类号: TP 333 文献标识码: A 文章编号: 1671-4431(2010)20-0118-05

Research on Based on Data Object Access Model

LIU Qun[1], *FENG Dan*[2], *LI Jian*[2]
(1. Network and Computing Center, Huazhong University of Science and Technology, Wuhan 430074, China;
2. School of Computer, Huazhong University of Science and Technolog, Wuhan 430074, China)

Abstract: With the development of computer technology and the popularization of its application, Object Based Storage (OBS) is an emerging technology and will become the next wave of network storage technology. We construct a general model frame of OBSSCA which provides the foundation for the analysis of a case. The data object access behavior analyzes the effect of the access frequency of data object on OBSS, and then suitably adjusts the access frequency of data objects in accordance with their characteristics through mechanical study and activity so as to improve the system stableness.

Key words: cellular automata; object-based storage system; data object; access frequency

网络化已是目前存储系统发展新趋势, 大量的 NAS(Network Attached Storage)和 SAN(Storage Area Network)等网络存储出现, 使存储系统从传统的集成式计算机系统中独立出来, 与计算脱离。但随着源源不断增加的数据和不断增加的物理设备, 对这些网络存储体系架构提出了巨大挑战, 期待着相对愚钝的存储设备能转变为智能型和自主管理, 这样基于对象存储(Object-based Storage, OBS)[1,2]产生了, OBS 吸收 NAS 和 SAN 优点, 具有 SAN 的高性能和 NAS 的数据共享和安全性, 采用对象接口, 具有属性[3], 用于描述对象的特征与信息。将保存对象的存储设备称为对象存储设备(Object-based Storage Device, OSD), OSD 是一个智能设备, 维护所有与数据空间分配、空闲空间等有关的元数据管理, 并有能力承担更多的智能处理, 由这些设备构成了基于对象存储系统(Object-based Storage System, OBSS)[4]。

收稿日期: 2010-06-20.
基金项目: 国家自然科学基金(60873028), 湖北省科学研究项目(2009053)和华中科技大学实验技术研究项目(2009041、2010027).
作者简介: 刘 群(1969-), 女, 博士, 工程师. E-mail: liliuqun@mail.hust.edu.cn

众所周知, 数据存储规模越来越大, 网络存储结构也日益复杂, 其演变是一个繁杂的过程, 不同的存储系统有着不同的运动过程, 即使相同的系统中在不同的条件和规则下, 也会产生不同的结果。OBSS 是一个时空耦合系统, 一切存储过程、存储现象和存储表现既包括了时间、空间上的性质, 又包含状态的性质, 只有同时将时间、空间和状态纳入统一的基础之中, 才能真正认识 OBSS 的本质规律。

元胞自动机(Cellular Automata, CA)[5,6] 不同于其他的动力学模型, 由元胞(Cell)、网格(Lattice)、邻居(Neighbor)和规则(Rule)组成, 表现为一个时空离散的动力系统, 其散布于规则网格中的每一元胞均取有限的离散状态, 各元胞遵循相同的演化规则进行同步演化, 且仅和它相邻的元胞发生相互作用, 大量元胞通过简单的相互作用而构成动态系统的演化过程。它不是由严格定义的物理方程或函数确定, 而是用一系列模型构造的规则构成, 其特点是时间、空间、状态都离散, 每个变量只取有限多个状态, 且其状态改变的规则在时间和空间上也是局部的。因此, 该文引入元胞自动机理论, 研究 OBSS 的演化规律。

1　元胞自动机框架模型

OBSS 是一种具有可扩展、主动智能存储系统, 它能够在初始状态下, 事先不知道对象的平均到达率和平均数据请求长度, 将对象的历史访问信息记录在对象的属性中, OSD 可以通过统计得到负载的特征, 通过策略的自主学习促使不断演变, 规则也在系统的需求和统计中完善, 根据外界应用变化调整数据对象的布局、负载均衡, 随着系统的扩展并控制 OBSS 不稳定向平衡发展, 使系统总体性能达到最优。

由于元胞自动机是一种时空离散的动态模型, 它可以模拟信息在不同 OSD 中的传递、存储等现象。鉴于元胞自动机在模拟 OBSS 复杂现象应用中特点, 针对具有丰富语义的对象接口和存储过程, 提出一个基于 OBSS 的元胞自动机(OBSSCA)框架模型, 该模型不仅对 OBSS 进行静态描述, 还可以动态地模拟, 形象地模拟 OBSS 的演变和运动, 从而在深层次上揭示海量存储系统的复杂过程中规律。

为了便于元胞自动机能够模拟和分析 OBSS, OBSSCA 须将元胞自动机构成基本要素进行相应的扩展和存储专业化。如图 1 所示。

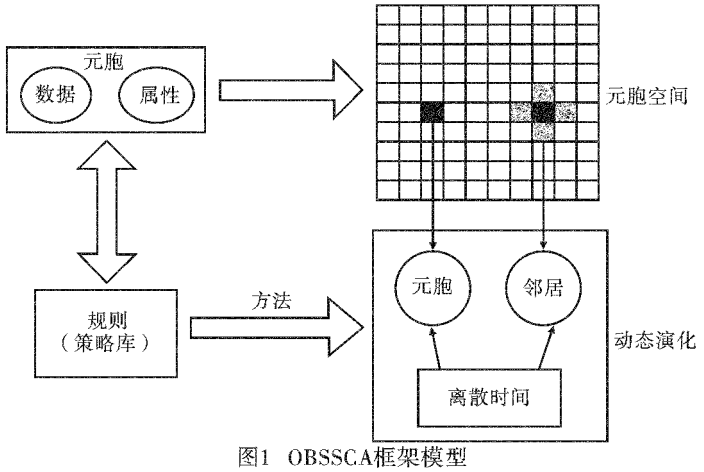


图1　OBSSCA框架模型

1) 元胞与状态

在 OBSSCA 中, 元胞可以被赋予为对象, 由不同的对象构成不同的模型。这样, 在元胞含义的定义基础上, 相应元胞的状态也确定下来, 如数据有被访问和未被访问 2 种状态, 存储设备则有储存和未储存 2 种状态等等, 但由于系统中每一个对象拥有着不同的状态, 并且相互关联、相互作用, 联动变化, 构成一个多元变量的元胞自动机模型, 同时, 在 OBSS 中, 元胞的状态还可作为对象的一个属性, 则元胞的状态变化还会进一步扩展, 如迁移、复制等。

2) 元胞空间

在 OBSSCA 中, 元胞空间的概念可以转变为笛卡儿坐标 OBSS, 对于常用的二维元胞自动机可以用网格来表示。在这个意义的转变过程, 元胞空间被赋予空间长度的概念。不同的空间尺度的概念, 会影响整个模型的其他方面, 如规则。动力学边界问题一直是比较难以解决的问题, 在 OBSSCA 中, 常用的周期型、反射型均不适合, 因而, 在 OBSS 采用定值型, 即将研究对象作为一个独立的个体, 利用某种概率控制系统的随机性, 同时采用较大空间使研究区域位于空间的中部, 尽量避免边界情况的出现。

3) 邻居定义

元胞自动机模型中的邻居概念, 在 OBSS 中体现为同类对象之间的近邻关系, 并在一定程度上有着相互作用的关系。在标准元胞空间中, 元胞的邻居常常被定为平衡对称的构型, 而在 OBSS 中同类对象实体间相互作用的复杂性, 邻居的构型可能是非平衡、非对称的多种形式, 如 OSD 的存储按照某一策略分配数据对象, 而这种策略要考虑当前的 OSD 数量、负载情况等; 另外, 在标准元胞自动机模型中, 邻居定义具有齐性,

所有的元胞邻居定义一致,然而在 OBSSCA 模型中,其中 OSD 的 CPU 处理能力和存储容量均有可能不同,或者根据不同的存储实体定义不同的邻居构型,形成不同的邻居半径。

4) 规则

规则是元胞自动机模型的核心,它决定元胞自动机的动态演绎过程,OBSSCA 的规则为一个策略库,集中体现了空间存储实体的相互作用,这种相互作用根据不同的条件、不同的应用被赋予不同的应用含义,也就是说,OBSSCA 的规则是对象海量存储系统特征和规律在局部和微观上的体现。但它与宏观上的规律并不相同,如当系统某一 OSD 负载过重时实现数据对象的迁移,根据分配策略,在逻辑上无法确切地判断那些 OSD 将会造成负载过重、何时过重等宏观特征。因此这些宏观特征只有在模型中运行才能动态地表现出来。

2 基于数据对象的访问模型

存储系统中访问数据对象的行为具有明显的复杂特性,即简单的局部规则能产生整体的复杂行为。在 OBSS 中,储存的对象是一个个数据,它们局部行为相对比较简单,只有读、写或不访问,可将其行为概率同样作为对象的一个属性。它们的行为除了本身外,还与其他数据对象的访问也有着关系,虽然影响数据对象访问的行为作用机制和演化规律比较简单,但经过一段时间的演化后,整个存储系统会产生截然不同的整体行为。通过简单的元胞与简单规则产生复杂的现象,从而分析这个复杂系统。

1) 元胞 数据对象。

2) 元胞空间 系统的存储空间,由于网格越大,计算机模拟的速度越慢,但太小又无法模拟结果,因此选取 50× 50 的网格。

3) 邻居 为了简化问题,采用典型的 Moore 型,即相邻的 8 个元胞为邻居。

4) 元胞状态 元胞状态是考察元胞某方面特征时的取值,根据数据对象的访问行为和访问频率建立二维元胞状态空间(S),第一维是数据对象的访问行为,其中包括读(s_{11})、无访问(s_{12})、写(s_{13}),即 $S_1=\{s_{11}, s_{12}, s_{13}\}$;第二维是访问频率,其中包括低($s_{21}$)、中($s_{22}$)、高($s_{23}$),即 $S_2=\{s_{21}, s_{22}, s_{23}\}$。因而这个元胞空间为 $S_1=\{s_i, s_j\}$,其中 $s_i \in S_1$, $s_j \in S_2$。考虑元胞 2 个方面的状态,共有 3× 3= 9 种组合,即元胞的状态总数共有 9 种。

5) 规则 规则是一个从中心元胞的邻居状态到中心元胞下一时刻状态的映射。一般情况下,元胞下一时刻的状态受其自身状态、邻居状态和控制变量的影响,可表示为

$$\begin{cases} s_m^{t+1} = F(s_m^t, s_{mn}^t, C) \\ s_{mn}^t = (s_{mn(1)}^t, \cdots, s_{mn(k)}^t) \\ s \in S_1 \times S_2 \end{cases} \tag{1}$$

式中,s_m^{t+1} 表示元胞空间中位置为 m 的元胞在 $t+1$ 时刻的状态;F 为元胞演变规则;s_m^t 表示元胞空间中位置为 m 的元胞在时刻 t 的状态;s_{mn}^t 表示位置为 m 的元胞在其邻居 n 在 t 时刻的状态;C 为控制变量;k 为邻居元胞的个数;S_1 和 S_2 表示元胞的 2 个方面的状态。

在 OBSS,由于数据对象存储到介质上是受磁盘的磁头寻道和旋转物理运动所决定,若相邻请求之间存在较大的寻道和旋转延时,会导致磁头臂的无规则移动,影响着整个存储系统的性能。因此,采用 I/O 合并方法以提高系统性能,它是将分解到磁盘上的子命令按照操作类型进行合并,即当前为读操作时,接下来的邻居也为读操作,这样合并 I/O 可以减少 I/O 服务时间,能提高系统性能;若当前为读操作时,接着的邻居却为写操作,就会产生较大寻道和旋转延时,降低系统性能。设数据对象的访问行为受到访问频率的一定概率影响,当大概率的高访问频率时,则此访问行为的频率一定高,而小概率的高访问频率时,访问行为的频率则不高;同样,当大概率的低访问频率时,则此访问行为的频率一定低,而小概率的低访问频率时,访问行为的频率则不低。同时,对系统而言,其读、写请求受应用程序的影响,但通常数据对象的读行为的概率较大,这是因为在存储系统中,读行为占所有访问行为的 60% ~ 70%[7]。因此,该模型所设计的规则是以提高系统性能为原则,当前元胞能够根据自身特性和机械学习,以设定的概率系数进行读、写或无访问。具体规则如表 1 所示,表 1 中 ρ 为调整概率,它是受邻居元胞中上一时刻占多数访问行为的频率影响的概率。μ 为适

应因子, 适应因子是通过机械学习而得某种策略, 用来增加或减少访问行为的概率。当无热点时, 设 μ 大于 0, 而有热点时, 则 μ 小于 0。表中所有概率值均在[0, 1]之间, 若小于 0 的值按 0 处理, 而大于 1 按 1 处理, 并且每一个元胞转化的 3 个行为的概率总和为 1。

表 **1**　数据对象访问行为的元胞自动机演化规则

邻居元胞 t 时刻多数访问行为		元胞 $t+1$ 时刻的状态
无热点	读	以概率 $(\rho+\mu)$ 读, 以概率 $\frac{(1-\rho-\mu)}{2}$ 无访问, 以概率 $\frac{(1-\rho-\mu)}{2}$ 写
	无访问	以概率 $\frac{(1-\rho)(1+\mu)}{2}$ 读, 以概率 ρ 无访问, 以概率 $\frac{(1-\rho)(1-\mu)}{2}$ 写
	写	以概率 $\frac{(1-\rho)(1-\mu)}{2}$ 读, 以概率 $\frac{(1-\rho)(1+\mu)}{2}$ 无访问, 以概率 ρ 写
热点	读	以概率 ρ 读, 以概率 $\frac{(1-\rho)(1+\mu)}{2}$ 无访问, 以概率 $\frac{(1-\rho)(1-\mu)}{2}$ 写
	无访问	以概率 $\frac{(1-\rho)(1+\mu)}{2}$ 读, 以概率 ρ 无访问, 以概率 $\frac{(1-\rho)(1-\mu)}{2}$ 写
	写	以概率 $\frac{(1-\rho+\mu)}{2}$ 读, 以概率 $\frac{(1-\rho+\mu)}{2}$ 无访问, 以概率$(\rho-\mu)$写

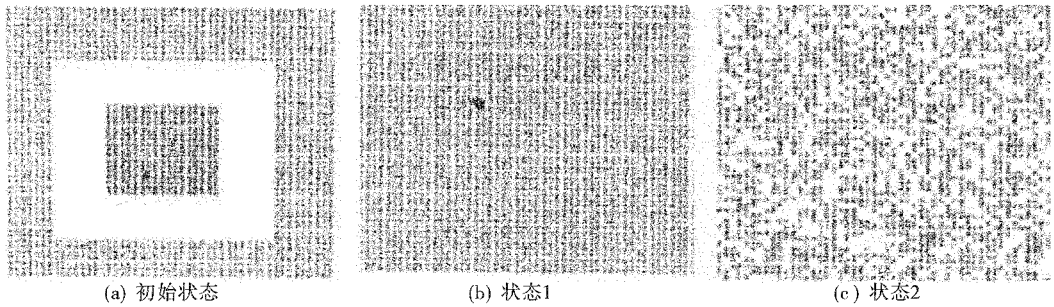
当系统无访问热点时, 依据上一时刻的多数元胞的访问行为适当调整访问行为的概率。假若上一时刻 8 个元胞中多数行为为读, 根据存储系统中读行为的概率较大之特性, 则此刻通过适应因子适当增加读行为的概率, 以$(\rho+\mu)$概率进行读, 以$[(1-\rho-\mu)/2]$的概率进行写或无访问; . 当上一时刻 8 个元胞中多数行为是无访问时, 此刻仍用 ρ 概率进行无访问, 以大于写行为概率进行读行为; 当上一时刻 8 个元胞中多数行为是写行为时, 采用 ρ 概率进行写行为, 以大于读行为概率进行无访问, 这样防止磁头臂的无规则移动, 以提高系统性能。

然而, 当某一个数据对象的读行为超过阈值时成为热点, 则需要将它复制迁移到相邻的存储对象中, 以适应因子适当增加写行为概率进行调节。因此, 当有热点时, 若上一时刻 8 个元胞中多数行为为读, 此刻仍用 ρ 概率进行读, 以$[(1-\rho)(1-\mu)/2]$的概率进行写行为, 此概率大于无访问的概率; 对于上一时刻 8 个元胞中多数行为是无访问, 此刻仍用 ρ 概率进行无访问, 以大于读行为概率进行写行为的概率; 当上一时刻 8 个元胞中多数行为是写行为时, 适当增加写行为概率, 采用$(\rho-\mu)$概率进行, 这样不仅消除热点, 而且还可提高系统性能。

系统初始状态分为随机状态或者对称状态2 种[8]。假设初始状态为对称状态, 因为对称状态中读、写和无访问 3 种访问行为是对称分布, 能够非常直观地显示各自访问频率。设输入参数为五元向量{ 初始状态, μ, $s21$, $s22$, $s23$}。不同的演化规则是由不同的数据对象访问行为来决定, 不同的演化规则对 OBSS 有着不同的影响。当 μ= 0, 即无适应系数, 初始状态选为对称初始状态(如图 2(b) 所示) , 中间深颜色区域表示读行为, 空白区域表示无访问、四周淡颜色区域表示写行为, 此时读行为频率低, 无访问频率中, 写行为频率高。根据数据对象不同的访问频率分 2 种情形进行模拟, 图 2(b) 状态 1 表示{ 对称, 0, 1, 1, 1} , 它是以 1 的概率在低、中、高频率中访问, 也就是说原来访问行为低频率依然低, 而高频率访问行为会一直延续下去, 这样系统演化趋向是数据对象都是写行为, 这是因为确定性规则和对称规则中不均匀分布所决定的。图 2(c) 的状态2 表示{ 对称, 0, 0. 2, 0. 5, 0. 8} , 低访问频率只有概率 0. 2, 中访问频率有概率 0. 5, 而高访问频率也只有 0. 8 概率, 这样, 低访问频率的读行为逐渐加大, 数据对象以一定的调整概率调整各自的访问行为, 系统需要一定的步数演化, 使 3 种数据访问行为的数量分布较为均匀。

以下考察适应因子对数据对象的影响程度。假定访问频率$\{s21, s22, s23\} = \{0.2, 0.5, 0.8\}$, 根据 μ 4 种情况进行模拟实验, 输入参数分别为{ 对称, 0. 1, 0. 2, 0. 5, 0. 8}、{ 对称, 0. 9, 0. 2, 0. 5, 0. 8}、{ 对称, – 0. 1, 0. 2, 0. 5, 0. 8}、{ 对称, – 0. 9, 0. 2, 0. 5, 0. 8}, 图 3~ 图 6 分别相应的模拟实验中访问行为数量的演变过程, 其中这个数量的变化与编写代码有一定关系。

图 3 和图 4 分别表示 μ= 0. 1 和 μ= 0. 9 时 3 种访问行为的数量变化过程, 当 μ= 0. 1 时, 在模拟 50 步过程中, 也就是系统在较长时间演变过程中, 读行为由初时的低频率逐步转为高频率, 而初时为高频率的写



(a) 初始状态 (b) 状态1 (c) 状态2

图2 初始状态及2种情况的演变

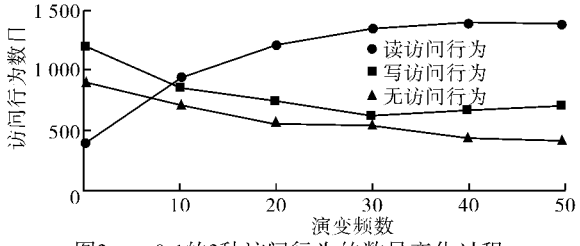


图3 μ=0.1的3种访问行为的数量变化过程

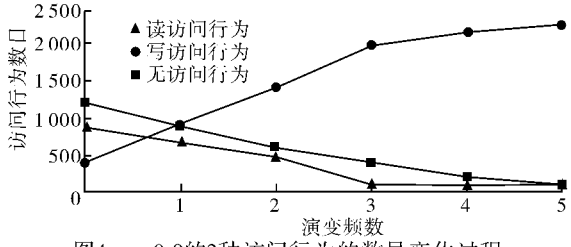


图4 μ=0.9的3种访问行为的数量变化过程

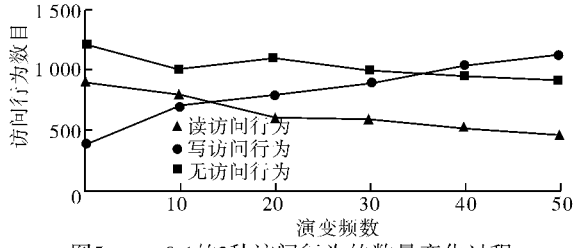


图5 μ=-0.1的3种访问行为的数量变化过程

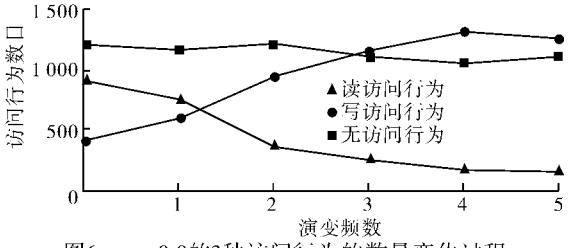


图6 μ=-0.9的3种访问行为的数量变化过程

行为则逐步转为中频率; 当 $\mu = 0.9$ 时, 则只需模拟 5 步过程, 系统在短时间内演变为低频率的读行为转为高频率, 易出现热点。

图 5 和图 6 分别表示 $\mu = -0.1$ 和 $\mu = -0.9$ 时 3 种访问行为的数量变化过程, 由于局部有热点, 需要适当将数据迁移, 增加写行为, 增加当 $\mu = -0.1$ 时, 在较长时间内, 系统读行为由初时的低频率逐步缓慢转为高频率, 而初时为高频率的写行为则逐步缓慢低转为中频率; 而当 $\mu = -0.9$ 时, 则在短时间内系统低频率的读行为转为了高频率, 出现新的热点。因此, 适应因子(以 μ 的绝对值衡量)过大仍会造成系统的不稳定, 它需要根据当时实际访问行为和访问频率, 并统计数据对象属性中原有的访问行为, 通过机械学习选择一个适当的数。由此可以看出, 数据对象访问行为按照某种已确定的策略规则进行演变, 而且演变结果与初始状态紧密相关, 不同的初始状态会有不同的结果, 导致系统无法预测, 这也体现了 OBSS 的复杂特性。

3 结 语

在OBSS 中, 数据对象的访问频率是影响系统的关键因素, 当某个访问行为的频率过高, 特别当访问行为的频率集中于某个或某些数据对象中, 易形成热点访问, 导致系统不稳定, 反之越稳定。同时消除热点的适应因子也不能过大, 过大仍然会造成系统的不稳定。这需要策略通过机械学习, 不断地调整访问行为的概率, 并依据周围数据对象的访问行为动态自适应地调整自身的访问行为。正是因为数据对象主动服务和策略的自我学习, 逐步形成了一个自管理系统。

参考文献

[1] Mesnier M, Ganger G R, Riedel E. Object-based Storage[J]. Communications Magazine, IEEE, 2003(41): 84-90.

[2] Liu Qun, Feng Dan. An Approximate Analytic Performance Model of Object-based Storage[C]//Proceedings of the International Conference on Computational Science and Its Applications, Glasgow, UK, 2006: 8-11.

[3] Mesnier M, Ganger G R, Riedel E. Object-based Storage: Pushing More Functionality into Storage[J]. Potentials, IEEE, 2005, 24(2): 31-34.

(下转第 127 页)

表 1 发送强度和成功率测试图

	科特邮件群发器	MassiveMails (λ= 14.00 封/s)	MassiveMails (λ= 1.50 封/s)	MassiveMails (λ= 1.26 封/s)
用时/s	36.0	35.7	333.3	398.0
发送成功数/封	294	330	496	500
发信成功率/%	60	66	99.2	100

5 结 语

邮件群发功能是各类国际会议组织工作中非常重要的支撑技术。针对学术会议组织特殊需求, 提出并实现了一种新的面向学术会议组织工作的多功能邮件群发器, 支持与多种典型会议投稿系统及其数据库的连接与互动。实验测试表明系统能够有效地完成会议组织中与邮件发送相关的各项组织工作, 并保证发送成功率, 从而提高会议组织效率。

参考文献

[1] Mike A, Robert B. The State of the Email Address[J]. ACM SIGCOMM Computer Communication Review, 2004, 35(1): 29-36.

[2] 李 理. 群发邮件系统的设计及其在实现数字化校园社区中的应用[D]. 沈阳: 东北大学, 2004.

[3] Hämäläinen H, Tarkkonen J, Heikkinen K, et al. Use of Peer-review System for Enhancing Learning of Programming[C]// 2009 9th IEEE International Conference on Advanced Learning Technologies (ICALT 2009), 2009: 658-660.

[4] 李立功, 赵 扬. MySQL 程序设计与数据库管理[M]. 北京: 科学出版社, 2001.

[5] Masuda N, KimJ S, Kahng B. Priority Queues with Bursty Arrivals of Incoming Tasks[J]. Physical Review E-statistical, Nonlinear, and Soft Matter Physics, 2009, 79(3): 036106.

(上接第 122 页)

[4] Liu Qun, Feng Dan, Qin Ling jun, et al. A Framework for Accessing General Object Storage. Proceedings of the 2006 International Workshop on Networking, Architecture, and Storages (IWNAS2006). Dalian, China: IEEE. 2006: 145-148.

[5] Wolfram S. Theory and Applications of Cellular Automata[M]. Singapore: World Scientific, 1986.

[6] 李才伟. 元胞自动机及复杂系统的时空演化模拟[D]. 武汉: 华中科技大学图书馆, 1997.

[7] Gibson G A, Nagle D F, Amiri K, et al. A Case for Network-attached Secure Disks. CMU SCS Technical Report CMU-CS-96-142, 1996.

[8] 应尚军, 魏一鸣, 范 英, 等. 基于元胞自动机的股票市场投资行为模拟[J]. 系统工程学报, 2001, 16(5): 382-388.