

南京邮电大学学报(自然科学版)

2009 年 12 月 第 29 卷 第 6 期(总第 125 期)

目 次

学术论文

- 物联网的体系结构与相关技术研究 沈苏彬,范曲立,宗 平,毛燕琴,黄 维(1)
无线传感器网络中基于空间相关性的分布式压缩感知 胡海峰,杨 震(12)
基于 YC_bC_r 颜色空间的背景建模及运动目标检测 卢官明,郎苏娟(17)
基于 QoS 和 SLA 的网络计费系统设计 张登银,吴 超,程春玲(23)
基于 BFGS 方法的拥塞速率控制算法 魏 涛,张顺颐(28)
基于无理数的 DES 加密算法 王 静,蒋国平(31)
三维 MIMO 信道物理模型的统计特征 海 淩,张业荣(38)
一种新型的网络带宽最优分配机制 冯慧斌,张顺颐,刘 超,王 攀(43)
基于 CSP 的进程行为取证方法研究 孙国梓,俞 超,陈丹伟(48)
基于 RSSI 的无线传感器网络环境参数分析与修正方案 凡高娟,王汝传,孙力娟(54)
基于鲁棒性的移动自组织网络路由选择算法 徐占洋,张顺颐(58)
不同掺杂浓度的 $CdS:Mn/SiO_2$ 核壳纳米结构的光致发光 薛洪涛(64)
基于 K 近邻分类间隔的特征选择方法研究 李 云,张腾飞,杨文杰(68)
保单驱动索赔离散风险模型的精算量分布 徐小阳,唐加山(75)

研究报告

- 应用层组播的效率优化技术研究 饶 翔,张顺颐,许建真,陈 涛,周 篓(79)
随机工艺变化下互连线 ABCD 参数建模与仿真 张 瑛,王志功,方承志,杨恒新(85)
嵌入式通信服务器 E1 网络驱动程序设计与实现 高建国,戴海鸿(91)
一种支持大型多人在线游戏的覆盖网组播生成树算法 林巧民,王汝传,许棣华,林 萍(96)

附:2009 年(第 29 卷)总目次

第 29 卷卷终

本期责任编辑:胡长贵 本期英文审订:卢官明

期刊基本参数:CN 32 - 1772/TN * 1960 * b * A4 * 100 * zh * P * ￥8.00 * 1000 * 18 * 2009-12

JOURNAL OF NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS(NATURAL SCIENCE)

Dec. 2009 Vol. 29 No. 6 (Sum. No. 125)

CONTENTS

Papers

Study on the Architecture and Associated Technologies for Internet of Things	
.....	SHEN Su-bin,FAN Qu-li,ZONG Ping,MAO Yan-qin,HUANG Wei (1)
Spatial Correlation-based Distributed Compressed Sensing in Wireless Sensor Networks	
.....	HU Hai-feng,YANG Zhen (12)
Background Modeling and Moving Object Detection Based on YC_bC_r Color Space LU Guan-ming,LANG Su-juan (17)
Design on Network Billing System Based on QoS and SLA ZHANG Deng-yin,WU Chao,CHENG Chun-ling (23)
Congestion Rate Control Algorithm Based on BFGS Method WEI Tao,ZHANG Shun-yi (28)
Irrational Numbers Based on DES Encryption Algorithm WANG Jing,JIANG Guo-ping (31)
Statistical Characteristics of 3-D Physical MIMO Channel Model HAI Lin,ZHANG Ye-rong (38)
A Novel Network Bandwidth Optimal Allocation Mechanism	
.....	FENG Hui-bin,ZHANG Shun-yi,LIU Chao,WANG Pan (43)
Research on Forensic Methods of the Process Behavior Based on CSP SUN Guo-zi,YU Chao,CHEN Dan-wei (48)
Distance-assistant Node Coverage Identification Model for Wireless Sensor Networks	
.....	FAN Gao-juan,WANG Ru-chuan,SUN Li-juan (54)
Route Selection Based on Robustness for Mobile Ad hoc Network XU Zhan-yang,ZHANG Shun-yi (58)
Photoluminescence of CdS/SiO₂ Core-Shell Nanostructures with Different Manganese Concentration	
.....	XUE Hong-tao (64)
Feature Selection Based on Margin of K-Nearest Neighbors LI Yun,ZHANG Teng-fei,YANG Wen-jie (68)
Distributions of Actuarial Variables of Discrete Risk Model with Policies Driving Claims	
.....	XU Xiao-yang,TANG Jia-shan (75)

Technical Reports

Study of the Efficiency Optimization Technology of ALM	
.....	RAO Xiang,ZHANG Shun-yi,XU Jian-zhen,CHEN Tao,ZHOU Jun (79)
Modeling and Simulating the ABCD Parameters of Interconnects in the Presence of Random Process Variations	
.....	ZHANG Ying,WANG Zhi-gong,FANG Cheng-zhi,YANG Heng-xin (85)
Design and Implementation of E1 Network Driver for Embedded Communication Server	
.....	GAO Jian-guo,DAI Hai-hong (91)
The Design and Implementation of an Intelligent Game Engine Based on OGRE	
.....	LIN Qiao-min,WANG Ru-chuan,XU Di-hua,LIN Ping (96)

物联网的体系结构与相关技术研究

沈苏彬¹,范曲立²,宗平¹,毛燕琴¹,黄维²

(1. 南京邮电大学 软件学院,江苏南京 210003
2. 南京邮电大学 信息材料与纳米技术研究院,江苏南京 210046)

摘要:物联网技术已经引起国内学术界、工业界和新闻媒体的高度重视,当前物联网的定义、内在原理、体系结构和系统模型等方面还存在许多值得探讨的问题,通过对现有物联网技术文献和应用实例的分析,探讨了物联网与下一代网、网络化物理系统和无线传感器网络的关系;提出了物联网的服务类型和结点分类,设计了基于无源、有源和互联网结点的物联网的体系结构和系统模型;在总结物联网特征基础上,对物联网的研究提出了建议。

关键词:物联网;网络化物理系统;下一代网;产品电子标签;网络体系结构

中图分类号:TP393 文献标识码:A 文章编号:1673-5439(2009)06-0001-11

Study on the Architecture and Associated Technologies for Internet of Things

SHEN Su-bin¹, FAN Qu-li², ZONG Ping¹, MAO Yan-qin¹,
HUANG Wei²

(1. College of Software, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Institute of Advanced Materials, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: The technology of the Internet of Things (IOT) has attracted highly attention of academia, industry, and news media. There are still many open issues in the definition, internal principles, architectures and system models of IOT. Through the analysis of current technical materials and application cases of IOT, this paper discusses the relations between Next Generation Network, Cyber-Physical Systems, Wireless Sensor Network and the IOT. It proposes service types and node classification of IOT and designs the architecture and system model of IOT based on passive, active and internet nodes structure of IOT. After summary of the features of IOT, it proposes the suggestions on the researches and development of IOT.

Key words: internet of things; cyber-physical systems; network generation network; electronic product code; network architecture

1 物联网研究背景

随着政府对于物联网研究和开发的高度重视,物联网已经引起国内学术界、工业界和新闻媒体的高度重视,物联网研究和技术的报道已经十分普及。2009年10月,通过谷歌查询“物联网”关键词,获得了约500多万条结果,如表1所示。可以看出,对于物联网的关注程度相当于对“商品房价格”和“传感

器网络”的关注程度,远小于对“软件”(2亿多条)、“互联网”(1亿多条)、“奥运会”(约3千万条)的关注程度;大于对“物价指数”(约300万条)的关注程度,远大于对“第三代移动通信”(约71万条)的关注程度。

通过与物联网相关技术的搜索可以看出,虽然物联网的被关注程度远小于“网络安全”、“RFID (Radio Frequency IDentification)”、“下一代网络”和

“嵌入式系统”(超过或者接近1千万条)技术的关注程度,但是,与“传感器网络”的关注程度十分接近,而传感器网络发展历史远长于物联网。从以上搜索结果可以看出,物联网已经迅速成为当前具有影响力的技术。

表1 在谷歌网站上2009年10月的搜索结果

序号	关键词	搜索结果/条
1	软件	241 000 000
2	互联网	127 000 000
3	奥运会	29 100 000
4	网络安全	16 000 000
5	RFID	15 700 000
6	下一代网络	11 200 000
7	嵌入式系统	9 390 000
8	商品房价格	5 640 000
9	传感器网络	5 480 000
10	物联网	5 220 000
11	物价指数	2 730 000
12	第三代移动通信	705 000

社会各界在较短时间内对于物联网产生了极大的关注,说明许多人相信物联网可能对人类社会、人们日常生活产生巨大的影响。无论国内还是国外,物联网的研究和开发才处于起步阶段,有关物联网的定位和特征还存在一些混乱的概念,物联网的系统模型和结构尚没有形成标准,物联网的研究和开发在国内还存在一定程度的盲目性。

从科学的角度看,物联网的研究和开发存在一些值得思考的问题。例如,物联网是否就是传感器网络?什么是物联网研究和开发的核心技术?什么是物联网的创新技术?物联网与互联网存在哪些本质的区别?如何开展对我国经济和社会发展有价值的物联网研究和开发?

本文在分析物联网相关的技术和应用的基础上,试图回答以上有关物联网的问题;在分析和研究已有物联网技术方案的基础上,尝试提出一种物联网互连体系结构,用于指导物联网的理论研究;在分析和研究物联网应用实例的基础上,试图提出一种物联网系统模型,用于指导物联网技术标准的研究和应用系统的开发。在以上研究基础上,试图得出物联网不同于互联网的特征,从中推导出科学地开展物联网研究和开发的基本原则,为我国的物联网研究和开发提供有科学依据的参考。

2 物联网的基本概念

2.1 物联网的基本定义

按照国际电信联盟(ITU)的定义^[1],物联网主要解决物品到物品(Thing to Thing,T2T),人到物品(Human to Thing,H2T),人到人(Human to Human,H2H)之间的互连。

这里与传统互联网不同的是,H2T是指人利用通用装置与物品之间的连接,H2H是指人之间不依赖于个人电脑而进行的互连。需要利用物联网才能解决的是传统意义上的互联网没有考虑的、对于任何物品连接的问题。

物联网是连接物品的网络,有些学者在讨论物联网中,常常提到M2M的概念,可以解释为人到人(Man to Man)、人到机器(Man to Machine)、机器到机器(Machine to Machine)。实际上M2M所有的解释在现有的互联网都可以实现,人到人之间的交互可以通过互联网进行,最多可以通过其他装置间接地实现,例如第三代移动电话,可以实现十分完美的人到人的交互;人到机器的交互一直是人体工程学和人机界面领域研究的主要课题;而机器与机器之间的交互已经由互联网提供了最为成功的方案。本质上,在人与机器、机器与机器的交互,大部分是为了实现人与人之间的信息交互,万维网(World Wide Web)技术成功的动因在于:通过搜索和链接,提供了人与人之间异步进行信息交互的快捷方式。

我们认为,在物联网研究中不应该采用M2M概念,这是容易造成思路混乱的概念,应该采用ITU定义的T2T、H2T和H2H的概念。

2.2 物联网与下一代网络

按照ITU物联网研究组的研究结论^[1],物联网的核心技术主要是普适网络、下一代网络和普适计算。这3项核心技术的简单定义如下,普适网络,无处不在的、普遍存在的网络;下一代网络,可以在任何时间、任何地点,互连任何物品,提供多种形式信息访问和信息管理的网络;普适计算,无处不在的、普遍存在的计算。其中下一代网中“互连任何物品”的定义是ITU物联网研究组对下一代网定义的扩展,我们认为,这是对下一代网发展趋势的高度概括。从现在已经成为现实的多种装置的互连网络,例如手机互连、移动装置互连、汽车互连、传感器互连等等,都揭示了下一代网在“互连任何物品”方面的发展趋势。从以上的定义可以看出,下一代网络

在某种角度看,就是可以连接任何物品的物联网。

按照传统的定义,下一代网络是在任何时间、任何地点,以任何方式提供信息访问和管理的服务。传统意义上的下一代网侧重于为人提供方便的信息服务,所以,从网络服务角度看,下一代网络可以称为信息网络;而从互连角度看,这种传统的下一代网定义还是局限在传统互联网的范畴,仅仅强调人与人之间的信息交互。

我们认为,应该按照 ITU 的定义,把物联网研究和开发纳入下一代网的范畴,而不是把下一代网络仅仅作为引入 IP 核心网、移动性和个性化服务的网络,这样,下一代网可以真正推动人类社会发展。

人与人之间的信息交互是具有百年发展历史的电信网主要业务范畴,引入了物联网理念的下一代网,从根本上扩展了电信网的业务范畴,可以真正推动电信业务和电信网络的全面变革,可以为电信网(包括固定电信网和移动电信网)创造新的发展机遇。

2.3 物联网与 CPS

随着处理器、存储器、网络带宽等成本的下降,嵌入式系统广泛应用于许多领域,特别是广泛应用于各类物理设备中,例如飞机、汽车、家电、工业装置、医疗器械、监控装置和日用物品。国际上把利用计算技术监测和控制物理设备行为的嵌入式系统称为网络化物理系统^[2-3](CPS, cyber-physical systems)或者深度嵌入式系统(deeply embedded systems)^[4-5], CPS 也可以翻译为“物理设备联网系统”。

美国总统的科学技术咨询委员会(PCAST)在 2007 年 8 月发布的题为“挑战下的领导地位:在世界竞争中的信息技术研发”的咨询报告^[6]中,明确建议把 CPS 作为美国联邦政府研究投入最高优先级的课题,由此启动了美国高校和研究机构的 CPS 研发热潮。

PCAST 咨询报告认为^[6],CPS 的设计、构造、测试和维护难度较大、成本较高,通常涉及到无数联网软件和硬件部件,在多个子系统环境下的精细化集成。在监测和控制复杂的、快速动作的物理系统(例如医疗设备、武器系统、制造过程、配电设施)运行时,CPS 在严格的计算能力、内存、功耗、速度、重量和成本的约束下,必须可靠和实时地操作。绝大部分 CPS 系统都是安全关键的系统,必须在外部攻击和打击下能够继续正常工作。这种融合信息世界和物理世界的技术具备以下自身的特征:

(1) CPS 是未来经济和社会发展的革命性技

术。CPS 是信息领域的网络化技术、信息化技术,与物理系统中控制技术、自动化技术的融合。CPS 可以连接原来完全分割的虚拟世界和现实世界的关联,使得现实的物理世界与虚拟的网络世界连接,通过虚拟世界的信息交互,优化物理世界的物体传递、操作和控制,构成一个高效、智能、环保的物理世界。从这个角度看,CPS 技术是可以改变未来经济和社会发展的革命性技术。

(2) 信息材料本身就是一种 CPS 技术。材料技术与信息技术融合构成的信息材料技术本身就是一种 CPS 技术,它是最为基础的网络化世界与物理世界连接的技术。例如小型化、低成本、环保节能的新型材料传感器、显示器等技术,都是 CPS 发展中的关键技术。

(3) CPS 要求计算技术与控制技术的融合。为了把网络世界与物理连接,CPS 必须把已有的、处理离散事件的、不关心时间和空间参数的计算技术,与现有的、处理连续过程的、注重时间和空间参数的控制技术融合起来,使得网络世界可以采集物理世界与时间和空间相关的信息,进行物理装置的操作和控制。

(4) CPS 要求开放的嵌入式系统。CPS 系统中的计算技术主要是嵌入式系统,CPS 中的嵌入式计算系统不是传统的封闭性系统,而是需要通过网络,与其他信息系统进行互联和互操作的系统。CPS 要求的嵌入式系统是一种开放的嵌入式系统,需要提供标准的网络访问接口和交互协议、标准的计算平台和服务调用接口、标准的计算环境和管理界面。

(5) CPS 要求可靠和确定的嵌入式系统。CPS 把计算技术带入了与国家基础设施、人们日常生活密切相关的领域,CPS 大部分应用领域是与食品卫生一样的安全敏感的领域,CPS 的技术和产品需要经过政府严格的安全监督和认证。原来信息技术领域习以为常的“免责”条款将不再适用,CPS 技术和产品必须成为高可靠的、行为确定的产品,CPS 技术要求可靠和确定的嵌入式系统。

对照国际电信联盟有关物联网的定义以及 PCAST 咨询报告有关 CPS 定义,我们认为 CPS 是物联网的专业称呼,侧重于物联网内部的技术内涵;而物联网是 CPS 的通俗称呼,侧重于 CPS 在日常生活中的应用。从专业角度看,CPS 提供了物联网研究和开发所需要的理论和技术内涵;从应用角度看,物联网提供了 CPS 未来应用的一个直观画面,更加适合于普及 CPS 方面的科学知识。物联网的研究和

开发应该从 CPS 入手和深入,而 CPS 技术和产品的普及和应用可以从物联网角度介绍和举例。

2.4 物联网与无线传感器网络

由于目前对于物联网研究尚未深入,对于物联网的技术内涵也缺乏专业的研究,有些专业的或非专业的报道通常会把无线传感器网络作为物联网。实际上,只要略微查询一下专业学术刊物,研究无线传感器网络的定义,对比物联网定义,就可以得出较为科学的结论。

按照国内权威学术期刊的定义^[7],无线传感器网络是一种“随机分布的集成有传感器、数据处理单元和通信模块的微小节点通过自组织的方式构成网络”,它可以“借助于节点中内置的形式多样的传感器测量所在周边环境中的热、红外、声纳、雷达和地震波信号”,并且“传感器网络有着与传统网络明显不同的技术要求,前者以数据为中心,后者以传输数据为目的”。所以,无线传感器网络并没有赋予 T2T 的连接能力,更不具备与物理系统连接并且控制物理系统的能力。

我们认为,无线传感器网络仅仅是采集和传递数据,并没有涉及到物联网中的核心控制技术,也不具备 CPS 要求的高可靠性。所以,无线传感器网络并不是物联网,更不是网络化物理系统,无线传感器网络的相关技术在一定程度上可能支撑物联网的开发。

3 物联网体系结构

要深入研究物联网的体系结构,必须首先研究物联网已经构建的应用系统和应用实例。物联网已经在仓储物流^[8],假冒产品的防范^[9],智能楼宇、路灯管理、智能电表、城市自来水网等基础设施^[10-11],医疗护理^[12]等领域得到了应用。

人类社会在相当长时间内将面临两大难题:其一是能源短缺和环境污染;其二是人口老龄化和慢性病增加,物联网首要的应用在于能耗控制和医疗护理。人类社会目前遇到的问题是:恐怖活动和信任危机,物联网目前急需的应用在于安防监控、物品身份鉴别。另外,物联网在智能交通、仓储物流、工业控制等方面都有较大的应用价值。已经公开的物联网的应用实例基本上围绕这些应用领域。

3.1 物联网已有的体系结构

在公开发表物联网应用系统的同时,很多研究

人员也发表了若干个物联网的体系结构,例如物品万维网的(Web of Things, WoT)体系结构^[13],它定义了一种面向应用的物联网,把万维网服务嵌入到系统中,可以采用简单的万维网服务形式使用物联网。这是一个以用户为中心的物联网体系结构,试图把互联网中成功的、面向信息获取的万维网应用结构移植到物联网上,用于简化物联网的信息发布和获取。

物联网的自主体系结构^[14]是为了适应于异构的物联网无线通信环境而设计的体系结构。该自主体系结构采用自主通信技术。自主通信是以自主件(selfware)为核心的通信,自主件在端到端层次以及中间结点,执行网络控制面已知的或者新出现的任务,自主件可以确保通信系统的可进化特性。

物联网的自主体系结构如图 1 所示,包括了数据面、控制面、知识面和管理面,数据面主要用于数据分组的传递;控制面通过向数据面发送配置报文,优化数据面的吞吐量以及可靠性;知识面提供整个网络信息的完整视图,并且提炼成为网络系统的知识,用于指导控制面的适应性控制;管理面协调和管理数据面、控制面和知识面的交互,提供物联网的自主能力。

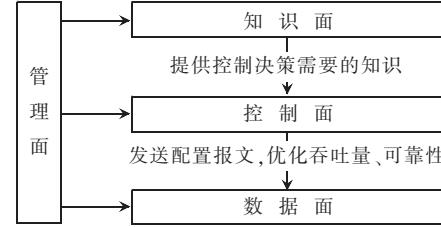


图 1 一种联网的自主体系结构

这里自主特征主要由 STP/SP 协议栈和智能层取代传统的 TCP/IP 协议栈,如图 2 所示,这里的 STP 和 SP 分别表示智能传送协议(Smart Transport Protocol)和智能协议(Smart Protocol),物联网结点的智能层主要用于协商交互结点之间 STP/SP 的选择,用于优化无线链路之上的通信和数据传送,满足异构物联网设备之间的联网的需求。

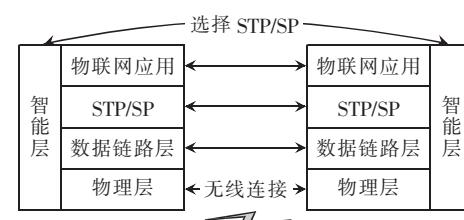


图 2 物联网自主体系结构的协议栈

这种面向物联网的自主体系结构涉及的协议栈较为复杂,只能适用于计算资源较为富裕的物联网

结点。

目前物流仓储的物联网应用都依赖于产品电子代码(EPC)网络^[8,15],该网络如图 3 所示,主要组成部件包括:产品电子代码(EPC),这是一种全球范围内标准定义的产品数字标识;电子标签和阅读器,电子标签通常采用射频标识(RFID)技术存储 EPC,阅读器是一种阅读电子标签内存储的 EPC 并且传递给物流仓储管理信息系统的装置;EPC 中间件,这

是一组具有特殊属性的程序模块或服务,用户可以根据某种应用需求定制和集成 EPC 中间件中的不同功能部件,其中最重要的部件是应用层事件(ALE),用于处理应用层相关的事件; EPC 信息服务(EPC-IS),该服务包括两个功能,一是存储 EPC 中间件处理的信息,二是查询相关的信息;对象名字服务(ONS),类似于域名服务器,其中的信息可用于指向某个存放 EPC 中间件信息的 EPC-IS 服务器。

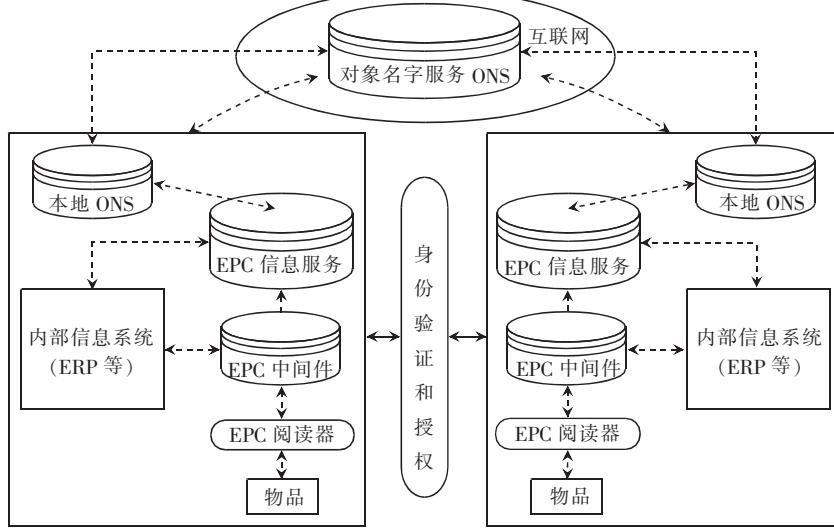


图 3 EPC 网络体系结构

EPC 网络包括 3 个层次:

(1) 实体和内部层次,该层由 EPC、RFID 标签、RFID 阅读器、EPC 中间件组成。这里的 EPC 中间件实际上屏蔽了各类不同的 RFID 之间的信息传递技术,把物品的信息访问和存储转化成为一个开放的平台。

(2) 商业伙伴之间的数据传输层,这层最重要的部分是 EPC-IS,企业成员利用 EPC-IS 服务器处理被 ALE 过滤之后的信息,这类信息可以用于内部或者外部商业伙伴之间的信息交互。

(3) 其他应用服务层,这层最重要的部分是 ONS,ONS 用于发现所需的 EPC-IS 的地址。EPC-global^[15](全球 EPC 管理机构,详见第 4 节) 委托全球著名的域名服务机构 VeriSign(威瑞信)公司提供 ONS 全球服务,全球至少有 10 个数据中心提供 ONS 服务。

以上都是分别从某个具体应用的角度给出了物联网的系统结构,这类结构还无法构成一个通用的物联网系统结构。下面给出一个通用物联网体系结构。

3.2 建议的物联网体系结构

我们认为,物联网体系结构设计应该遵循以下

5 条原则:(1) 多样性原则,物联网体系结构必须根据物联网结点类型的不同,分成多种类型的体系结构;(2) 时空性原则,物联网体系结构必须能够满足物联网的时间、空间和能源方面的需求;(3) 互联性原则,物联网体系结构必须能够平滑地与互联网连接;(4) 安全性原则,物联网体系结构必须能够防御大范围内的网络攻击;(5) 坚固性原则,物联网体系结构必须具备坚固性和可靠性。

以下从物联网的服务类型,结点分类和互连结构 3 个方面讨论物联网的体系结构。

3.2.1 物联网的服务类型

根据物联网自身的特征,物联网应该提供以下几类服务:

- (1) 联网类服务:物品标识、通信和定位;
- (2) 信息类服务:信息采集、存储和查询;
- (3) 操作类服务:远程配置、监测、远程操作和控制;
- (4) 安全类服务:用户管理、访问控制、事件报警、入侵检测、攻击防御;
- (5) 管理类服务:故障诊断、性能优化、系统升级、计费管理服务。

以上罗列的是通用物联网的服务类型集合,根

据不同领域的物联网应用需求,以上服务类型可以进行相应的扩展或裁剪。物联网的服务类型是设计和验证物联网体系结构和物联网系统的主要依据。

3.2.2 物联网的结点分类

为了构建物联网的体系结构,首先需要划分物联网中网络结点的类型。物联网结点可以分成无源 CPS 结点、有源 CPS 结点、互联网 CPS 结点,其特征从以下方面进行描述:电源、移动性、感知性、存储能力、计算能力、联网能力、连接能力,具体如表 2 所示。

表 2 物联网结点类型与特征

结点类型	无源 CPS	有源 CPS	互联网 CPS
电源	无	有	不间断
移动性	有	可有	无
感知性	被感知	感知	感知
存储能力	无	有	强
计算能力	无	有	强
联网能力	无	有	强
连接能力	T2T	T2T, H2T, H2H	H2T, H2H

无源 CPS 结点,就是具有电子标签的物品,这是物联网中数量最多的结点,例如携带电子标签的人可以成为一个无源 CPS 结点。无源 CPS 结点一般不带电源,可以具有移动性,具有被感知能力和少量的数据存储能力,不具备计算和联网能力,提供被动的 T2T 连接。

有源 CPS 结点,具备感知、联网和控制能力的嵌入式系统,这是物联网的核心结点,例如装备了可以传感人体信息的穿戴式电脑的人可以成为一个有源 CPS 结点。有源 CPS 带有电源,可以具有移动性、感知、存储、计算和联网能力,提供 T2T、H2T、H2H 连接。

互联网 CPS 结点,具备联网和控制能力的计算系统,这是物联网的信息中心和控制中心,例如具有物联网安全性、可靠性要求的,能够提供时间和空间约束服务的互联网结点就是一个互联网 CPS 结点。互联网 CPS 结点不是一般的互联网的结点,它是属于物联网系统中的结点,采用了互联网的联网技术相互连接,但具有物联网系统中特有的时间和空间的控制能力,配备了物联网专用的安全性和可靠的控制体系。互联网 CPS 结点具有不间断电源,不具备移动性,可以具有感知能力,具有较强的存储、计算和联网能力,可以提供 H2T、H2H 连接。

3.2.3 物联网互连体系结构

根据以上物联网结点的分类,可以进一步研究可能存在的连接类型,例如物联网结点之间存在无

源结点与有源 CPS 结点,有源 CPS 与有源 CPS 结点,以及有源 CPS 结点与互联网 CPS 结点之间的连接,这些类型的连接结构构成了物联网互连的体系结构。

由于物联网的异构性,我们建议的通用物联网体系结构由 3 部分构成:无源 CPS 结点与有源 CPS 结点互连结构,有源 CPS 结点与有源 CPS 结点互连结构,有源 CPS 结点与互联网结点互连结构。

无源 CPS 结点与有源 CPS 结点互连结构如图 4 所示,无源 CPS 结点通过物理层协议与有源 CPS 结点连接,例如通过 RFID 协议,有源 CPS 可以获取无源 CPS 结点上电子标签的信息。

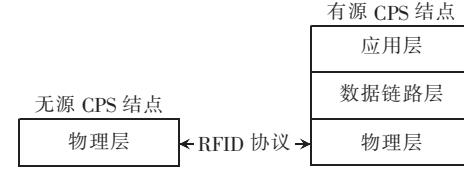


图 4 无源 CPS 结点与有源 CPS 结点互连结构

有源 CPS 结点与有源 CPS 结点互连结构如图 5 所示,有源 CPS 结点之间通过物理层、数据链路层和应用层的协议交互,实现有源 CPS 结点之间的信息采集、传递和查询。考虑到大部分有源 CPS 结点资源限制十分严格,有源 CPS 结点不适合配置已有的 IP 协议;配置的数据链路协议也应该是面向物联网的数据链路层协议,可以保证可靠、高效、节能地采集、传递和查询信息,满足物联网结点交互的应用需求。

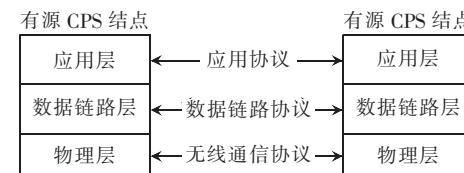


图 5 有源 CPS 结点与有源 CPS 结点互连结构

有源 CPS 结点之间的信息转发和汇聚可以通过应用协议实现,这样,可以按照应用需要,设计灵活的信息采集和转发的协议,不需要采用通用的、低效的互联网中的 IP 协议。

有源 CPS 结点与互联网 CPS 结点互连结构如图 6 所示,有源 CPS 结点需要通过 CPS 网关,才能连接互联网结点。CPS 网关实际上是一个有源 CPS 结点与互联网 CPS 结点的组合,其中实现了完整的互联网协议栈。这样,通过 CPS 网关,可以在应用层与互联网连接,实现物联网与互联网之间信息传递,以及物联网应用与互联网应用之间的互通、互连和互操作。这种互连结构可以允许不同类型的物联

网采用满足自身需要的联网结构,简化不必要的联网功能,降低网络系统的复杂性。不同的物联网联网技术,例如汽车电子联网技术、环境监测联网技术

等,可以采用适用于各自应用领域的有源 CPS 结点之间连接的协议结构,只需要通过 CPS 网关,就可以与互联网连接。



图 6 有源 CPS 结点与互联网 CPS 结点互连结构

在以上定义的物联网体系结构中,物联网物理层协议,提供在物理信道上采集和传递信息的功能,具有一定的安全性和可靠性控制能力;物联网数据链路层协议,提供对物理信道访问控制、复用,在链路层安全、可靠、高效传递数据的功能,具有较为完整的可靠性、安全性控制能力,可以提供服务质量的保证;应用层协议,提供信息采集、传递、查询功能,具有较为完整的用户管理、联网配置、安全管理、可靠性控制能力。

贴上电子标签或配置传感装置,改造成为 CPS 结点。例如牛角上贴上电子标签,奶牛也成为一个 CPS 结点,可以智能化管理奶牛的喂养和挤奶等操作;盲人穿着具有电子标签的鞋子,也可以成为一个 CPS 结点,与盲道上的电子标签阅读器协同操作,就可以指导盲人的行走。

构建物联网系统的第一步是标识物品,也就是表示世界上所有的物品,这里需要利用电子标签和传感器技术。而电子标签,特别是用于自然物品的电子标签,需要具备防水、耐磨、耐高温等特性,并且具备一定的电磁特征,这方面需要采用信息材料技术,这是属于物联网的最为基础的技术,是一个应用十分广泛的技术。在信息材料技术方面的任何突破,都会带来物联网产业的大幅度发展。因为电子标签和低端传感器是面广量大的产品,信息材料技术在降低成本、提供质量方面的任何改进,都会扩展物联网的应用面,降低物联网部署成本,提高物联网产业的收益。所以,信息材料技术的原始创新和自主创新,必定会带动我国经济和社会的发展。

标识物品的另外一项技术就是世界统一的物品编码技术。目前还没有针对物联网的全球物品编码技术,产品电子代码(EPC)是关于全球产品类电子代码编制的一个规范,EPC 由 EPCglobal(<http://www.epcglobalinc.org/>)负责标准化和应用^[15],它是国际物品编码协会 EAN 和美国统一代码委员会(UCC)的一个合资公司。它是一个受业界委托而成立的非盈利组织,负责 EPC 网络的全球化标准。EPC 网络由自动标识(Auto-ID)中心开发,其研究总部设在麻省理工学院,并且还有全球 7 所大学,美国麻省理工学院(MIT),英国剑桥大学(Cambridge),澳大利亚阿德莱德大学(Adelaide),日本庆应大学(Keio),中国复旦大学(Fudan),韩国信息与通信大学(ICU)和瑞士圣加仑大学(St. Gallen)的实验室参与。

4 物联网的系统模型

从抽象的物联网结点的互连结构可以提取出隐藏物联网背后的关键理论和技术,但这并不能完整反映出物联网系统实现中的关键技术,我们需要设计一个通用的物联网系统模型,进一步提取出物联网实现系统的关键技术和方法。在目前已发表的论文中还没有看到一个通用的物联网的系统模型,这样,难以指导物联网的研究和开发。

在前面提出的通用物联网体系结构的基础上,我们提出一个通用物联网的系统模型,试图通过物联网的系统模型,分析和梳理在实现物联网系统过程中涉及到的关键技术和方法。构建物联网通常需要分成标识物品、建立物品联网系统和建立物联网应用系统,以下将从这 3 个方面讨论物联网系统的设计和实现技术。

4.1 标识物品

世界上所有的物品可以简单分成人造物品和自然物品,人造物品包括食品、纺织品、其他日用品、货物、道路、桥梁、楼房、汽车、飞机、轮船、生产线等,通常在人造物品上贴上电子标签或者传感装置,就可以把人造物品改造成 CPS 结点。自然物品包括动物、植物、山峰、河流、湖泊等,这些自然物品也可以

4.2 建立物品联网系统

在完成物品标识之后,就可以建立物品联网系统,需要建立可以识别、验证和采集被标识物品的物联网结点,这个结点就是有源 CPS 结点。

为了实现有源 CPS 结点,首先需要设计和实现有源 CPS 结点与无源 CPS 结点、有源 CPS 结点与有源 CPS 结点之间的无线通信机制,以及基于信息编解码技术的物品识别机制。其次必须设计和实现通信信道复用机制,使得在一条信道上可以同时完成多个无源或者有源 CPS 结点的通信,例如 RFID 技术同时可以识别 100 多个具有 RFID 标签的物品。然后需要设计和实现通信信道上的可靠传输机制、实时传输机制,满足物联网对可靠性和实时性的要求。前面两个部分可以构成物联网系统中的联网系统(见图 7)。



图 7 有源 CPS 结点实现系统

4.3 建立物联网应用系统

由于物联网的特殊性,物联网应用系统需要分成两阶段建立:建立物联网应用平台,建立物联网应用系统。

在建立有源 CPS 结点的联网系统之后,就需要设计和实现有源 CPS 结点的网络配置、用户管理、结点控制、信息采集、信息传输和信息查询的功能,建立一个基本的物联网的应用平台,这也就是面向某个具体应用领域的物联网中间件(见图 7)。因为不同的应用领域对于结点控制的可靠性、实时性、安全性有不同的要求,所以,需要针对不同应用领域,设计和实现不同控制力度的应用中间件。设计和实现物联网应用的中间件,可以隔离物联网特定联网系统,满足快速应用开发的需求。

在设计和实现物联网应用中间件过程中,需要参照物联网相关领域的应用平台服务接口标准,如果是一个全新的物联网应用领域,可以在设计和实现物联网应用中间件过程中,提取与实现无关的部分,形成该领域的物联网应用平台服务接口技术规范。

在建立物联网应用中间件之后,就可以进一步

设计和实现物联网应用系统,包括基本应用系统和特定应用系统,如图 8 所示。基本应用系统可以包括物品命名管理系统、物品身份真伪验证系统、物联网系统管理等,特定应用系统可以包括仓储管理系统、楼宇监控系统、环境监测系统等。

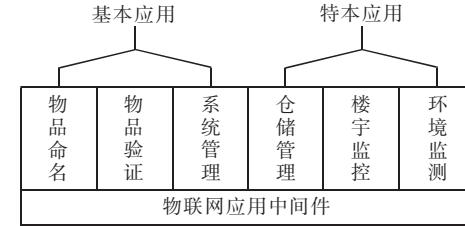


图 8 物联网应用系统逻辑结构

大部分物联网结点计算和存储资源都较为有限,物联网应用系统的部署是一个关键的技术。物联网应用系统需要区分应用系统的物联网端和互联网端,如图 9 所示,应用系统物联网端部署在有源 CPS 结点上,可以作为应用系统的客户端(采用客户机/服务器),也可以作为应用系统的对等端(采用 P2P 应用模式),但是,必须要求功能简洁可靠;应用系统互联网端部署在互联网 CPS 结点上,可以作为应用系统的服务器端(采用客户机/服务器),也可以作为应用系统的对等端(采用 P2P 应用模式),但都需要提供较为强大的存储和后端处理能力,满足物联网应用的需求。



图 9 物联网应用系统部署结构

5 物联网研究与开发面临的挑战

物联网研究和开发既是机遇,更是挑战^[16-17]。如果能够面对挑战,从深层次解决物联网中的关键理论问题和技术难点,并且能够将物联网研究和开发的成果应用于实际,则我们就可以在物联网研究和开发中获得发展的机遇。否则,物联网研究和开发只会浪费时间和资源,又一次错过了在科学和技术领域发展的机遇。

物联网研究和开发面临 3 个方面的挑战:基础研究方面的挑战,技术开发方面的挑战,以及示范系统构建和部署方面的调整。

5.1 基础研究方面的挑战

美国加州大学伯克利分校 Edward A. Lee 教授在分析了当今计算和联网方式与物理处理过程^[16],

提出了两者的差异:物理系统中的部件在安全性和可靠性方面的需求与通用计算部件存在质的差异;物理部件与面向对象的软件部件也存在质的差异,计算和联网技术采用的基于方法调用和线程的标准抽象体系在物理系统中无法工作。由此,Lee教授提出这样的疑问:今天的计算和联网技术是否能够为开发CPS系统提供足够的基础?其研究结论是:必须再造计算和网络的抽象体系,以便统一物理系统的动态性和计算的离散性。

如何再造计算和网络的抽象体系,这是物联网基础研究的核心内容,包括如何在编程语言中增加时序,如何重新定义操作系统和编程语言的接口,如何重新思考硬件与软件的划分,如何在互联网中增加时序,如何计算系统的可预测性和可靠性等。

5.2 技术开发方面的挑战

物联网技术开发中,面临诸多的技术开发方面的挑战。物联网是嵌入式系统、联网和控制系统的集成,它由计算系统、包含传感器和执行器的嵌入式系统等异构系统组成,首先需要解决物理系统与计算系统协同处理。在物联网环境下,事件检测和动作决策操作涉及到时间和空间^[18],这些操作必须准确和实时,以保证物联网操作中时间和空间的正确性。需要分析事件的时间和空间特性,设计面向物联网的、具有时间和空间条件限制的分层物联网事件模型。

物联网技术开发中,还需要建立物联网的可依赖性模型^[11],这也是进行物联网开发的一个挑战。采用传统的方法,分别评价、建模和仿真组成物联网的物理装置和网络部件,这样无法构造整个物联网系统的可依赖模型。必须建立物理装置和网络系统的相互依赖模型,其中包括构建定性的物联网交互依赖模型,构建量化的物联网交互依赖模型,按照物联网中的物理装置和网络部件属性描述物联网的可依赖性,验证这种可依赖性模型的正确性。

物联网技术开发中,需要面临如何构建面向物联网中间件的技术难题。中间件可以减少50%的软件开发时间和成本,由于CPS资源的限制、服务质量要求、可靠性要求等,通用的中间件无法满足CPS应用开发的需求^[19]。但是,重新开发一个面向CPS的中间件似乎难度较大,现代软件技术的一个基本原则是软件重用。所以,可以考虑采用面向应用领域的定制方法改造中间件。但是,改造一种结构复杂的、功能繁琐的通用中间件的成本是否一定小于构建一个结构简单的、功能简洁的专用中间件,

这是需要研究的问题。

物联网技术开发中面临许多挑战,例如提供安全、实时的数据服务技术^[20],物联网系统的正确性验证技术^[21]、嵌入式万维网服务开发技术、隐私保护技术^[22]以及安全控制技术^[23]等,这些技术是决定物联网技术能否得到广泛应用的关键技术。

5.3 示范系统建设的挑战

建设和部署物联网示范系统,在社会层面和技术层面都面临较大的挑战^[10]。物联网系统的典型示范系统,例如楼宇内部的照明、电表、街道路灯系统等,都会涉及到较为复杂的基本建设工程和公共设施工程。其次,消耗最多能源的、具有最大节能潜力的物品通常都是巨大的、昂贵的装置,改造这些装置面临很大的困难。另外,建设和部署物联网面临的较为直接的挑战是,如何让人们愿意使用并且可以维护物联网?这里不仅存在技术本身的问题,还存在如何进行培训、教育和普及物联网知识和技术的问题。

构建和部署物联网示范系统的技术层面的挑战包括通信基础设施、隐私保护和互操作性问题。物联网需要普适联网,对于公共设施的物联网需要在城市范围建立全覆盖的无线联网基础设施,而这种设施是无法在短时间建立的。如何经济有效地构建满足物联网需要的联网基础设施?这在技术上也是一个挑战。

无论是公共设施的物联网,还是企业专用的物联网,都需要提供严格的数据保护机制,否则,无论是公众,还是企业都不会接受物联网,不会使用物联网的相关应用的。从用户角度看,物联网应该是以用户为核心的网络,完全可以按照用户的意愿进行控制和操作。如何让用户信任物联网?这在技术上还是一个很大的挑战。

物联网提供的普适服务依赖于互操作性,它不仅依赖于网络运营商提供的标准服务质量,还依赖于跨域的命名、安全性、移动性、多播、定位、路由和管理,也包括对于提供公共设施的公平补偿。如何形成完整的物联网技术标准并且实现这些标准?这是一项十分具有挑战的工作。

6 结 论

物联网具有以下区别于互联网的特征:

(1)不同应用领域的专用性,不同应用领域的物联网,例如汽车电子领域物联网不同于医疗卫生

领域的物联网,医疗卫生领域的物联网不同于环境监测领域的物联网,环境监测领域的物联网不同于仓储物流领域的物联网,仓储物流领域的物联网不同于楼宇监控领域的物联网。由于不同应用领域具有完全不同的网络应用需求和服务质量要求,物联网结点大部分都是资源受限的结点,只有通过专用物联网技术才能满足物联网的应用需求。而互联网是通过TCP/IP技术互连全球所有的数据传输网络,虽然可以在较短时间实现了全球信息互连、互通,但是,也带来了互联网上难以克服的安全性、移动性和服务质量等一系列问题。物联网的应用特殊性以及其他特征,使得它无法再复制互联网成功的技术模式。

(2)高度的稳定性和可靠性,物联网是与许多关键领域物理设备相关的网络,必须至少保证该网络是稳定的,例如在仓储物流应用领域,物联网必须是稳定的,不能像现在的互联网一样,时常网络不通,时常电子邮件丢失等,仓储的物联网必须稳定地检测进库和出库的物品,不能有任何差错。有些物联网需要高可靠性的,例如医疗卫生的物联网,必须要求具有很高的可靠性,保证不会因为由于物联网的误操作而威胁病人的生命。

(3)严密的安全性和可控性,物联网的绝大多数应用都涉及到个人隐私或机构内部秘密,物联网必须提供严密的安全性和可控性。即物联网系统具有保护个人隐私、防御网络攻击的能力,物联网的个人用户或机构用户可以严密控制物联网中信息采集、传递和查询操作,不会由于个人隐私或机构秘密的泄露而造成对个人或机构的伤害。

根据以上物联网的特征,我们提出了对于物联网研究、开发的3点建议:

建议1:针对物联网不同应用领域的专用性,就要求研究和开发机构根据各自对于应用领域的理解能力和市场需求,客观地设定物联网应用领域,科学地设定物联网研究和开发的目标和内容,合理地部署物联网研究和开发的资源,切不可将学术界泛泛讨论的传感器网络作为即将带来人类社会巨大变化的物联网,误导政府决策和有限科研资源的分配。

建议2:针对物联网高度的稳定性和可靠性特征,要求重新研究、设计和开发已有的联网软件和应用软件,需要在现有的网络软件中引入自动控制领域的控制环路思想和方法,引入控制的机制和模型,需要把网络技术与控制技术有机融合,研究、设计和开发真正面向物联网的高稳定和高可靠的网络连接

和应用的软件系统。不可简单地将现有互联网上的联网软件和应用软件,直接应用于物联网环境,对社会造成较大的危害,从而阻碍物联网技术的应用和普及。

建议3:针对物联网严密的安全性和可控性特征,要求制定严格的、面向不同应用领域的物联网安全控制技术规范,并且设立由政府管理的物联网安全认证机构,对于实际应用的物联网系统进行严格的安全验证,防范由于物联网研究、开发或生产机构急功近利而对物联网用户造成危害。

本文在分析物联网相关文献、应用实例的基础上,初步研究了物联网定义、内涵、体系结构、实现系统以及面临的挑战,提出了通用的物联网体系结构,物联网系统模型,以及如何开展物联网研究和开发的建议。

物联网体系和技术博大精深,涉及到嵌入式系统、网络系统、控制系统、软件系统、安全系统等多种技术体系,需要我们长期研究和探索其中的理论和技术问题,本文仅仅对物联网进行较为初步的研究,难免存在一些不足,仅供物联网研究和开发人员参考。

物联网是一项庞大的系统,是未来十年或更长时间进行设计、部署、实现和完善的系统,需要分阶段有计划地开展相关的研究和开发工作。只要能够科学决策,长期、深入、持续地研究和开发物联网理论和技术,持之以恒,必能获得成功!

参考文献:

- [1] UIT. ITU Internet Reports 2005: The Internet of Things [R]. 2005.
- [2] SHA Lui, GOPALAKRISHNAN S, LIU Xue, et al. Cyber-Physical Systems: A New Frontier [C]//2008 IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing (sutc 2008). June 2008:1-9.
- [3] WOLF W. Cyber-physical Systems [J]. Computer, 2009, 42(3):88-89.
- [4] EASWARAN A, LEE Insup. Compositional schedulability analysis for cyber-physical systems [J]. SIGBED Review, 2008, 5(1):11-12.
- [5] TAN Ying, GODDARD S, PÉREZ L C. A prototype architecture for cyber-physical systems [J]. SIGBED Review, 2008, 5(1):51-52.
- [6] President's Council of Advisors on Science and Technology. Leadership Under Challenge: Information Technology R&D in a Competitive World, An Assessment of the Federal Networking and Information Technology R&D Program [EB/OL]. [2007-08-30]. http://ostp.gov/pdf/nitrd_review.pdf.
- [7] 任丰原,黄海宁,林闯. 无线传感器网络 [J]. 软件学报, 2003, 14

- (7):1282 – 1291.
- REN Fengyuan, HUANG Haining, LIN Chuang. Wireless sensor networks[J]. Journal of Software, 2003, 14(7):1282 – 1291.
- [8] YAN Bo, HUANG Guangwen. Supply chain information transmission based on RFID and internet of things[C]// ISECS International Colloquium on Computing, Communication, Control and Management. 2009, 4:166 – 169.
- [9] YAN Bo, HUANG Guangwen. Application of RFID and Internet of Things in Monitoring and Anti-counterfeiting for Products[C]// Proc of International Seminar on Business and Information Management. 2008, 1:392 – 395.
- [10] DOLIN R A. Deploying the “Internet of things”[C]// International Symposium on Applications and the Internet. 2006:216 – 219.
- [11] LIN Jing, SEDIGH Sahra, MILLER Ann. A general framework for quantitative modeling of dependability in Cyber-Physical Systems:a proposal for doctoral research[C]// Proc of 33rd Annual IEEE International Computer Software and Applications Conference. 2009: 668 – 671.
- [12] FREDERIX I. Internet of Things and radio frequency identification in care taking, facts and privacy challenges[C]// Proc of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology. 2009:319 – 323.
- [13] DUQUENNOY S, GRIMAUD J J G. Vandewalle. Smews:Smart and Mobile Embedded Web Server[C]// International Conference on Complex, Intelligent and Software Intensive Systems. 2009: 571 – 576.
- [14] PUJOLLE G. An autonomic-oriented architecture for the Internet of Things[C]// IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing. 2006:163 – 168.
- [15] ARMENIO F, BARTHEL H, DIETRICH P, et al. The EPCglobal Architecture Framework[EB/OL]. [2009-03-19]. <http://www.epcglobalinc.org/>
- [16] LEE E A. Cyber Physical Systems:design challenges[C]// 11th IEEE Symposium on Object Oriented Real-Time Distributed Computing (ISORC). 2008:363 – 369.
- [17] ABDELZAHER T. Research Challenges in Distributed Cyber-Physical Systems[C]// Proc of IEEE/IFIP International Conference on Embedded and Ubiquitous Computing. December 2008:5.
- [18] TAN Ying, VURAN M C, GODDARD S. Spatio-temporal event model for Cyber-Physical Systems[C]// Proc of 29th IEEE International Conference on Distributed Computing Systems Workshops. 2009:44 – 50.
- [19] DABHOLKAR A, GOKHALE A. An approach to middleware specialization for Cyber Physical Systems[C]// Proc of 29th IEEE International Conference on Distributed Computing Systems Workshops. 2009:73 – 79.
- [20] KANG Kyoungdon, SON S H. Real-time data services for Cyber Physical Systems[C]// Proc of the 28th International Conference on Distributed Computing Systems Workshops. 2008:483 – 488.
- [21] AKELLA R, McMILLIN B M. Model-Checking BNDC Properties in Cyber-Physical Systems[C]// Proc of 33rd Annual IEEE International Computer Software and Applications Conference. 2009: 660 – 663.
- [22] OLESHCHUK V. Internet of things and privacy preserving technologies[C]// Proc of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology. 2009:336 – 340.
- [23] de LEUSSE P, PERIORELLIS P, DIMITRAKOS T, et al. Self Managed Security Cell,a Security Model for the Internet of Things and Services[C]// Proc of First International Conference on Advances in Future Internet. 2009:47 – 52.

作者简介:



沈苏彬(1963 –),男,江苏南京人。南京邮电大学软件学院研究员,博士生导师。1984 年在南京工学院(现东南大学)计算机应用专业获工学学士学位, 1989 年在东南大学计算机应用专业获工学硕士学位, 2000 年在东南大学获博士学位。目前主要从事计算网络,下一代电信网以及网络安全的研究。

范曲立(1974 –),男,江苏无锡人。南京邮电大学信息材料与纳米技术研究院副院长、江苏省有机电子与信息显示重点实验室副主任,教授,博士生导师。2003 年在新加坡国立大学获博士学位。主要研究方向为有机电子材料,有机/无机杂化材料。

宗平(1956 –),男,江苏南京人。南京邮电大学软件学院教授。主要研究领域是智能数据处理、计算机网络、软件工程。

毛燕琴(1981 –),女,江西南昌人。南京邮电大学软件学院讲师。主要研究领域是计算机网络,网络安全。

黄维(1963 –),男,河北唐山人。南京邮电大学副校长,教授,博士生导师。国家重点基础研究发展计划(973 计划)首席科学家。(见本刊 2008 年第 4 期第 96 页)

无线传感器网络中基于空间相关性的分布式压缩感知

胡海峰¹, 杨震²

(1. 南京邮电大学 江苏省无线通信重点实验室, 江苏南京 210003
2. 南京邮电大学 信号处理与传输研究院, 江苏南京 210003)

摘要:提出了无线传感器网络中基于空间相关性的分布式压缩感知模型和分布式压缩感知算法,利用无线传感器网络节点间感知数据的空间相关性和联合稀疏模型,结合分布式压缩感知编解码算法,以能量有效的方式对无线传感器网络的感知数据进行压缩、重构。最后,通过仿真分析了分布式压缩感知重构误差和压缩比之间的关系,以及分布式压缩感知在能量有效性方面的性能,仿真结果表明分布式压缩感知以能量有效的方式满足了无线传感器网络中事件估计的精确度要求。

关键词:无线传感器网络; 分布式压缩感知; 空间相关性

中图分类号:TP393 文献标识码:A 文章编号:1673-5439(2009)06-0012-05

Spatial Correlation-based Distributed Compressed Sensing in Wireless Sensor Networks

HU Hai-feng¹, YANG Zhen²

(1. Jiangsu Key Laboratory of Wireless Communications, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Institute of Signal Processing and Transmission, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In this paper, spatial correlation-based Distributed Compressed Sensing (DCS) model and algorithm was presented in Wireless Sensor Networks (WSN), where spatial correlation and joint sparse models between the sensor nodes can be exploited in order to compress and reconstruct sensor observations in an energy efficient manner based on coding/decoding algorithm of DCS. Finally, the analysis of relationship between reconstruction error and compression ratio in DCS and Energy efficient performance of DCS are carried out in simulation. Simulation results show that DCS can achieve acceptable estimation accuracy in an energy efficient way.

Key words: wireless sensor networks; distributed compressed sensing; spatial correlation

0 引言

无线传感器网络(Wireless Sensor Networks, WSN)具有节点数目众多、资源受限,组成节点同构性等特点,如何利用WSN节点感知数据的时、空相关性,以能量有效的方式满足基于感知应用的WSN的QoS(服务质量)要求,是WSN应用亟待解决的问题。

针对上述问题,WSN在收集节点感知数据过程中应用数据融合技术是一种行之有效的手段。近年

来,一种新的数据融合理论——压缩感知CS(Compressed Sensing)^[1-2]逐渐发展起来。根据CS理论,只要信号在某些基上可以稀疏表示,就可以通过少量随机线性观测值来重构信号^[3-4]。压缩感知以优异的压缩性能、非自适应编码和编解码相互独立等特性在通信和信号处理领域得到广泛关注,已经成为国内外研究的热点^[5-7]。

相对于其他的数据融合技术,CS特别适合应用于WSN中,主要因为:(1)编码的低计算复杂度,信

号只需要在随机观测矩阵上进行线性投影,便可计算出压缩后的观测向量。(2) 优异的压缩性能,对于 k -稀疏的 N 维信号只需要 $m = ck$ ($c \geq 4$)维的观测向量($M \ll N$)便可重构信号。(3) 编解码相互独立性,对于相同的编码方案,可以采样不同的解码技术。以上优点使得CS特别适合应用在资源受限的WSN中,只需要明确节点感知数据是可压缩的,即在某些正交基上可以稀疏表达,节点侧便可运行低计算开销的编码算法,Sink节点通过收集节点感知数据的观测向量,运行较为复杂的CS解码算法,实现了以非协作方式,即节点间不需要进行数据交换,进行数据压缩和重构,显著减少了网络开销。

但是,CS理论一般研究如何利用单节点感知数据的内部相关结构(intra-signal)进行压缩编解码,考虑到WSN节点密集分布,以及节点有一定存储能力的特点,有必要进一步利用WSN中节点间感知数据的空间相关性(inter-signal)研究分布式压缩感知DCS(Distributed Compressed Sensing)算法^[8]。文献[9]基于Slepian-Wolf信息理论,建立DCS的观测率理论,证明了观测率的上界和下界,确定了分布式压缩感知DCS(Distributed Compressed Sensing)的信号压缩率;文献[10]针对环境监测、MIMO通信和语音信号排列等应用情景构造了不同的联合稀疏模型,并设计了相应的联合解码算法;文献[11]在文献[10]的基础上,改进了联合解码算法,提高了数据压缩率;文献[12]将DCS应用于WSN,阐述了DCS的安全性、容错性和信道容量自适应性等优点。但基于DCS的WSN数据融合技术研究尚处于起始阶段。因此有必要通过定义各节点基于空间相关性感知数据的联合稀疏性,研究DCS如何以非协作方式,通过对各节点的观测向量进行联合重构。

本文提出WSN中基于空间相关性的DCS模型和分布式压缩感知算法,利用WSN节点间感知数据的空间相关性和联合稀疏模型,以能量有效的方式对WSN感知数据进行压缩、重构,并通过仿真验证了模型的有效性和算法性能。

1 空间相关性的分布式压缩感知模型

WSN中sink节点搜集相关传感器节点(以下简称节点)的感知数据,对监控区域内发生的事件源 S (如目标跟踪、特定区域的物理量测量)进行估计,并使得估计结果的失真度(distortion)满足应用要求。

由图1可知,WSN监控区域内的事件源 S 会触发分布在事件区域EA(Event Area)的节点 n_i ($i = 1, 2, \dots, N$)获得信息数据 S_i ($i = 1, 2, \dots, N$),由于WSN节点密集分布在监控区域, S_i 之间以及 S_i 和 S 之间存在不同程度的空间相关性,在满足失真度要求的前提下,确定事件区域EA的大小,并利用WSN中节点间信息数据的空间相关性进行数据压缩和重构,在资源受限的WSN中有非常重要的意义。

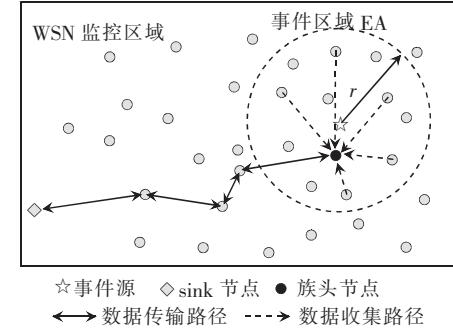


图1 WSN中基于空间相关性的分布式压缩感知场景图

假设事件源 S 所在位置为 $(0,0)$,事件区域EA中节点 $n_{(x,y)}, n_{(x_r,y_r)}$ 分别位于坐标 (x,y) 、 (x_r,y_r) ,其信息数据是 $S_{(x,y)}, S_{(x_r,y_r)}$,则变差函数定义为:

$$\gamma(r) = \frac{1}{2} E[(S_{(x,y)} - S_{(x_r,y_r)})^2] \quad (1)$$

其中, $(x - x_r)^2 + (y - y_r)^2 = r^2$,变差函数越小,数据之间的相关性越强。根据文献[13]的数学模型,在极坐标中定义节点 $n_{(r,\theta)}$ 的信息数据为 $S_{(r,\theta)}$,WSN监控区域事件 S 源触发的事件区域EA内,节点 $n_{(0,0)}$ 的信息数据 $S_{(0,0)}$ 和周围节点信息数据有如下相关性:

$$S_{(0,0)} = I_{(U=T)} Y + \\ I_{(U=H)} \int_0^\pi \int_r^\infty (S_{(r,\theta)} + Z) \delta(R=r) \delta(\Theta=\theta | R=r) r dr d\theta \quad (2)$$

其中, $U=H$ 表示 $S_{(0,0)}$ 由相邻节点 $n_{(r,\theta)}$ 的信息数据 $S_{(r,\theta)}$ 值获得,其概率为 $1-\beta$; $U=T$ 表示 $S_{(0,0)}$ 由随机变量 Y 获得,其概率为 β 。随机变量 Y 和 Z 的分布密度函数分别为 $f_Y(y)$ 和 $f_Z(z)$,且它们都与 $S_{(r,\theta)}$ 相互独立。变量 Y 反应了相邻节点数据存在无关性特征的情况;变量 Z 反应了在空间相关性情况下,相邻节点数据的差异,其中 $Z \sim N(0, \sigma_z^2)$,即:

$$f_Z(z) = \frac{1}{\sigma_z \sqrt{2\pi}} e^{-z^2/2\sigma_z^2}, \text{ 均方差 } \sigma_z \text{ 可由节点历史数} \\ \text{据统计求得。}$$

利用变差函数对WSN监控区域数据场空间相关特性进行分析,并根据应用所要求的误差门限值,

计算出事件区域 EA 分布范围,EA 范围内的节点 n_i ($i = 1, 2, \dots, N$) 形成簇,并选出节点 n_h ($h \in \{1, 2, \dots, N\}$) 作为簇头,由簇头收集 EA 范围内节点的感知数据 X_i :

$$X_i = S_i + N_i \quad (i = 1, 2, \dots, N) \quad (3)$$

其中, S_i 为节点 n_i 获得的信息数据, 观测噪声 N_i 为独立同分布的高斯随机变量。注意: X_i 是同一时刻各节点的感知数据。

设矢量 $X_N = (X_1, X_2, \dots, X_N)^T \in R^N$ 代表 EA 中 N 个节点的感知数据, 由于空间相关性, 感知数据在小波基 Ψ 上呈现 k -稀疏性^[13], 即:

$$X_N = \Psi \theta \quad (4)$$

其中, 变换系数向量(即稀疏系数向量) $\theta = (\lambda_0, \hat{\gamma}_0, \dots, \hat{\gamma}_{J-1})^T$, $\|\lambda_0\|_0 + \sum_{i=0}^{J-1} \|\hat{\gamma}_i\|_0 = k$ 。簇头获得了 X_N 后, 运行 DCS 编码算法, 将 k -稀疏信号 X_N 随机投影到观测矩阵 Φ 上产生了 M ($M = ck < N, 2 < c < 4$) 个观测数据 $Y_M = (Y_1, Y_2, \dots, Y_M)^T \in R^M$, 即:

$$Y_M = \Phi X_N = \Phi \Psi \theta = \Xi \theta \quad (5)$$

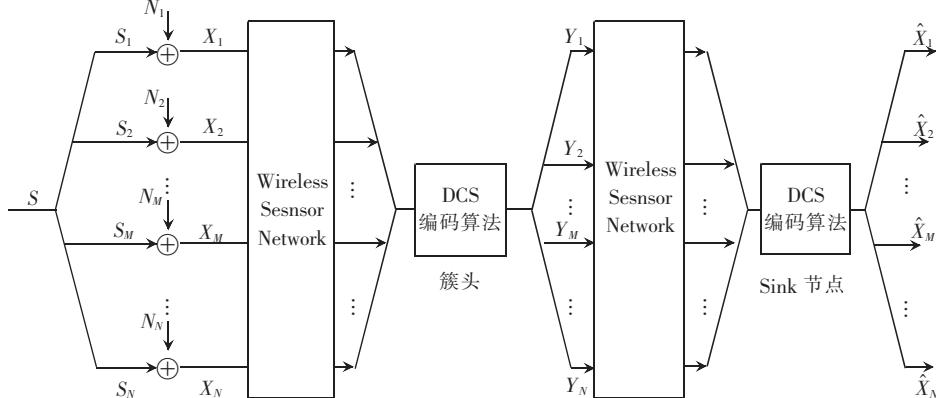


图 2 WSN 中基于空间相关性的分布式压缩感知模型

2 空间相关性的分布式压缩感知算法

首先, 计算事件区域 EA 的分布范围, 假设事件源 S 所在位置 $(0, 0)$ 有一虚拟节点 $n_{(0,0)}$, 其信息数据为 $S_{(0,0)}$, 事件源 S 触发的事件区域边界节点 $n_{(r,\theta)}$ 的信息数据 $S_{(r,\theta)}$ 满足 $|S_{(r,\theta)} - S_{(0,0)}| \leq \mu$, 其中 μ 是误差门限, 反应了不同位置的信息数据和事件源之间的差异, r 是事件区域 EA 的半径。于是由式(2)可得:

$$\begin{aligned} \gamma(r) &= \frac{1}{2} E[(S_{(r,\theta)} - S_{(0,0)})^2] = \frac{1}{2} E[Z^2] \\ &= \frac{1}{2} \int_{-\mu}^{\mu} z^2 \frac{1}{\sigma_z \sqrt{2\pi}} e^{-\frac{z^2}{2\sigma_z^2}} dz \end{aligned}$$

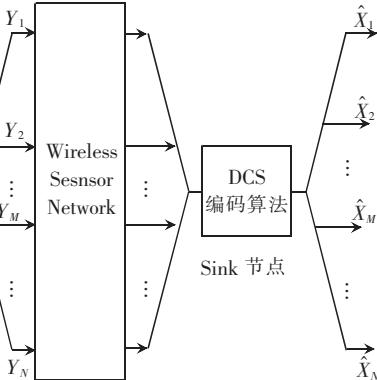
其中, $\Phi = (\varphi_1, \varphi_2, \dots, \varphi_N) \quad (\varphi_i \in R^M, i = 1, 2, \dots, N)$ 与 Ψ 互相不连贯(incoherence), 且 $\Xi = \Phi \Psi$ 为 DCS 矩阵。

簇头计算出 Y_M 后, 通过无线多跳方式把 Y_M 传送给 Sink 节点, Sink 节点根据各节点感知数据在小波基的 k -稀疏性, 以及随机观测矩阵与基矩阵的不连贯性, Sink 节点运行 DCS 解码算法:

$$\min \| \theta \|_1 \quad \text{subject to} \quad Y_M = \Phi \Psi \theta \quad (6)$$

通过式(6)的最优稀疏解 θ^* 可重构各节点感知数据为 $\hat{X}_N = (\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N)^T = \Psi \theta^*$ 。

从图 2 可知, 事件区域 EA 中的簇头把感知数据向量 $X_N \in R^N$ 经过 DCS 编码后, 生成观测数据向量 $Y_M \in R^M$ 并传递给 Sink 节点, X_N 因为空间相关性而呈现 k -稀疏性, $M = ck$ ($2 < c < 4$) 一般小于 N , 而且相对于事件区域的分布范围, 簇头距离 Sink 节点一般比较远, 从而节省了大量的传输能量。因此, WSN 中分布式压缩感知模型利用各节点感知数据的空间相关性, 以非协作方式, 通过对各节点的感知数据向量进行重构, 以能量有效的方式满足事件估计的精确度要求。



$$\begin{aligned} &= \frac{1}{2} \sigma_z^2 \operatorname{erf}\left(\frac{\mu}{\sqrt{2}\sigma_z}\right) - \frac{1}{\sqrt{2\pi}} \mu \sigma_z e^{-\frac{\mu^2}{2\sigma_z^2}} \\ &= \Psi(\sigma_z, \mu) \end{aligned} \quad (7)$$

其中, $\Psi(\sigma_z, \mu)$ 是 σ_z, μ 的函数, $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2} dt$ 。

通常 $\gamma(r)$ 有多种估计模型, 根据文献[14]选取 $\gamma(r) = c(1 - e^{-\lambda r^2})$, 其中参数 c 影响数据相关性的强弱, λ 反映了数据相关性随距离变化的快慢。参数 c 和 λ 取决于监控区域数据场空间相关特性, 代入式(7), 得

$$c(1 - e^{-\lambda r^2}) = \Psi(\sigma_z, \mu) \Rightarrow r = \left[\frac{1}{\lambda} \ln \left(\frac{c}{c - \Psi(\sigma_z, \mu)} \right) \right]^{\frac{1}{2}} \quad (8)$$

在各向同性统计过程中,根据节点历史数据统计规律,式(8)可以计算出不同误差门限 μ 条件下的事件区域EA的半径 r ,确定了事件区域EA的分布范围,从而Sink节点只需搜集EA中节点的感知数据,减少了处理数据的能量开销和执行时间。

其次,计算基于提升小波变换^[13]的小波基 Ψ 。设 X_N 对应某一时刻EA中节点集合 $E_J: \{n_1, n_2, \dots, n_N\}$ 的感知数据,设 $\lambda_J = X_N (\lambda_J \in R^N)$,根据文献[15]分布式小波变换算法,由节点地理位置不同,把 $E_J: \{n_1, n_2, \dots, n_N\}$ 节点集合分裂成 E_{J-1} 和 O_{J-1} ,对应的感知数据集合分别为 $\lambda_{J,E}$ 和 $\lambda_{J,O}$, T 表示提升小波变换,即:

$$\begin{cases} T(\lambda_J) = (\lambda_{J-1}, \gamma_{J-1}) \\ \gamma_{J-1} = \lambda_{J,O} - P(\lambda_{J,E}) \\ \lambda_{J-1} = \lambda_{J,E} + U(\gamma_{J-1}) \end{cases} \quad (9)$$

其中, γ_{J-1} 为小波系数集合, λ_{J-1} 为尺度系数集合, P 为线性预测算子, U 为线性更新算子,利用感知数据的空间相关性, $\lambda_{J,E}$ 可以很准确的预测 $\lambda_{J,O}$, γ_{J-1} 只包含很少的信息量。

对 $\lambda_i (i \in \{1, 2, \dots, J\})$ 经过 J 次递归提升小波变换,可得:

$$T^J(\lambda_J) = \{\lambda_0, \gamma_0, \gamma_1, \dots, \gamma_{J-1}\} \quad (10)$$

其中,小波系数集合 $\gamma_i (i \in \{0, 1, \dots, J-1\})$ 中包含很多非常小的元素,把 γ_i 中低于门限值的小波系数清零得到 $\hat{\gamma}_i$,从而使用具有稀疏结构的 $\hat{\gamma}_i$ 对原信号 λ_J 进行精确重构,设 $T^{-J}(\cdot)$ 为提升小波逆变换。因为预测和更新算子都为线性运算, $T^J(\cdot)$ 和 $T^{-J}(\cdot)$ 都为线性变换,可得:

$$T^{-J}(\lambda_0, \hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_{J-1}) = \hat{\lambda}_J \approx \lambda_J \Leftrightarrow \Psi\theta \approx \lambda_J = X_N \quad (11)$$

其中, $\theta = (\lambda_0, \hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_{J-1})^T$, $\|\lambda_0\|_0 + \sum_{i=0}^{J-1} \|\hat{\gamma}_i\|_0 = k < N$,因此,获得EA中节点的拓扑结构,可以求得 X_N 小波基 Ψ ,并且 X_N 在小波基 Ψ 上呈现 k -稀疏性。

综上所述,基于小波变换的分布式压缩感知算法表示如下:

Step1:根据式(8),Sink节点计算出事件区域EA的分布半径,并通过组播路由方式激活EA中 $n_i (i = 1, 2, \dots, N)$ 节点组成簇,并选出簇头 $n_h (h \in \{1, 2, \dots, N\})$,Sink传递给簇头随机种子 $s_M: \{s_1, s_2, \dots, s_M\}$ 。

Step2:簇头生成观测矩阵 $\Phi = R(s_M, \tau_N)$,其中 $R(\cdot)$ 为伪随机数发生函数, $\tau_N: \{\tau_1, \tau_2, \dots, \tau_N\}, \tau_i$ 为 n_i 的编号。

Step3:簇头与簇内节点通信,获得某一时刻EA内节点的感知数据 $X_N = (X_1, X_2, \dots, X_N)^T$,运行分布式压缩感知编码算法 $\mathbf{Y}_M = \Phi X_N$,产生了 $M (M = ck < N, 2 \leq c \leq 4)$ 个压缩的观测数据 $\mathbf{Y}_M = (Y_1, Y_2, \dots, Y_M)^T$ 。

Step4:簇头将 \mathbf{Y}_M 传输到Sink节点。

Step5:Sink节点生成相同的观测矩阵 $\Phi = R(s_M, \tau_N)$,并通过式(11),获得小波基 Ψ 变换矩阵,运行分布式压缩感知解码算法:

$$\min_{\theta} \|\theta\|_1 \quad \text{subject to} \quad \mathbf{Y}_M = \Phi \Psi \theta \quad (12)$$

通过求解 l_1 最优化问题,得到稀疏小波系数全局最优解 $\theta^* = (\lambda_0, \hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_{J-1})^T$,精确重构感知数据集合 $\Psi \theta^* = \hat{X}_N$ 。

3 仿真实验

本文仿真主要考虑基于空间相关性的分布式压缩感知算法中重构误差和观测数据个数的关系,以及相对于直接传输感知数据的分簇路由算法,分布式压缩感知DCS在能量有效性方面的性能。

仿真以matlab为工具,用二维高斯分布来模拟WSN中空间相关性数据源,节点随机均匀分布在 $50 \text{ m} \times 50 \text{ m}$ 事件区域内,事件区域内分别随机分布60、90、120个节点,分别比较传输的观测数据个数和重构误差之间的关系。在指定时刻每节点产生1bit数据,簇头获得所有节点的感知数据后,运行DCS编码算法,产生观测数据,并传输到Sink节点,由Sink运行DCS解码算法,以重构事件区域中节点的感知数据,并使用信号最大值对均方根RMS归一化来计算重构误差。

空间相关性数据的稀疏性是有损的,稀疏性越大,分布式压缩感知的观测数据个数越小,忽略的小波系数越多,重构误差越大。从图3可知,不同节点数情况下,Sink解码算法的归一化重构误差随观测数据个数增加呈下降趋势,当观测数据和感知数据个数接近时,重构误差趋向零。定义观测数据个数和感知数据个数之比为压缩比。从图4可知,DCS以增加归一化重构误差为代价减少传输的观测数据量,而且,当节点数越多,数据间的空间相关性越大,稀疏性越好,相同压缩比情况下,归一化重构误差随着节点数增加而减少,当节点数为120,压缩比为0.2对应的归一化重构误差小于0.05,因而,基于空间相关性DCS在满足应用误差要求的同时,对感知数据进行了很大的压缩。

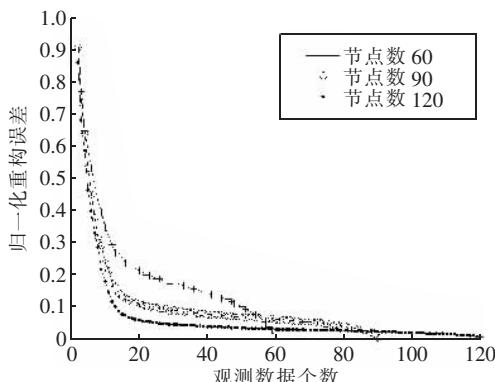


图3 归一化重构误差和观测数据个数的关系(不同节点数)

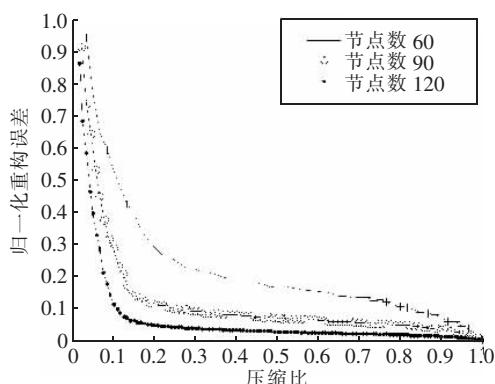


图4 归一化重构误差和压缩比的关系(不同节点数)

为了比较不同节点数情况下能量开销和重构误差的关系,仿真中簇头距离Sink节点的距离为100 m,单位数据量的传输能量消耗为: $\beta d^\gamma + \varepsilon$,其中 $\gamma = 2$, $\beta = 100 \text{ pJ}/(\text{bit} \cdot \text{m}^{-2})$, $\varepsilon = 100 \text{ nJ}/\text{bit}$, d 为节点间传输距离。

在计算能量开销时,因为节点侧的DCS编码算法计算开销很低,故只考虑簇头传输观测数据到Sink的传输能量,从图5可知,能量开销随着重构误差的增加呈下降趋势,当重构误差为零时,即观测数据和感知数据个数相等时,对应的能量开销为直接传输感知数据的分簇路由算法所消耗的能量,因此DCS可以根据应用所能容忍重构误差,最大程度地降低能量开销,同时又能满足事件估计的精确度要求。

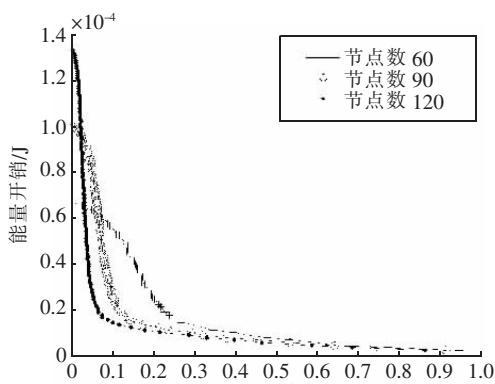


图5 能量开销和归一化重构误差的关系(不同节点数)

4 结束语

本文在WSN中基于空间相关性的分布式压缩感知DCS模型的基础上提出了DCS算法,利用WSN节点间感知数据的空间相关性和联合稀疏模型,以能量有效的方式对WSN感知数据进行压缩、重构。并通过仿真分析了基于空间相关性的DCS算法中重构误差和观测数据个数的关系,以及DCS在能量有效性方面的性能。结果表明基于空间相关性的DCS以能量有效的方式使得事件估计的失真度满足应用要求。下一步研究方向将着重考虑节点感知数据的时间相关性,建立基于时空相关性的DCS模型,进一步提高DCS算法性能。

参考文献:

- [1] DONOHO D. Compressed sensing[J]. IEEE Trans on Information Theory, 2006, 52(4):1289 – 1306.
- [2] CANDÈS E, ROMBERG J, TAO T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information[J]. IEEE Trans on Information Theory, 2006, 52(2):489 – 509.
- [3] TROPP J, GILBERT A. Signal recovery from random measurements via orthogonal matching pursuit[J]. IEEE Trans on Information Theory, 2007, 53(12):4655 – 4666.
- [4] CANDÈS E, ROMBERG J, TAO T. Stable signal recovery from incomplete and inaccurate measurements[J]. Comm Pure and Appl Math, 2006, 59:1207 – 1223.
- [5] CHARTRAND R. Nonconvex compressed sensing and error correction[C]//Acoustics, Speech and Signal Processing (ICASSP). Hawaii, USA, 2007:889 – 892.
- [6] ELAD M. Optimized Projections for Compressed Sensing[J]. IEEE Trans on Signal Processing, 2007, 55(12):5695 – 5702.
- [7] LUSTIG M, SANTOS J M. Application of compressed sensing for rapid MR imaging[J]. Magn Reson Med, 2007, 58:1182 – 1195.
- [8] BARON D, WAKIN M B, SARVOTHAM S. Distributed Compressed Sensing[R]. Rice University, 2006.
- [9] BARON D, DUARTE M F. An information-theoretic approach to distributed compressed sensing[C]//Proc 43th Allerton Conf Communication, Control and Computing. Monticello, IL, 2005.
- [10] BARON D, DUARTE M F, SARVOTHAM S, et al. Distributed compressed sensing of jointly sparse signals[C]//Proc 39th Asilomar Conf Signals, Systems and Computers. Pacific Grove, CA, 2005:1537 – 1541.
- [11] WAKIN M B, SARVOTHAM S, DUARTE M F, et al. Recovery of jointly sparse signals from few random projections[C]// Proc Workshop on Neural Info Proc Sys(NIPS). Vancouver, 2005.

(下转第22页)

基于 YC_bC_r 颜色空间的背景建模及运动目标检测

卢官明,郎苏娟

(南京邮电大学 通信与信息工程学院,江苏南京 210003)

摘要:高斯混合模型广泛应用于基于背景建模的运动目标检测中。首先在 YC_bC_r 颜色空间采用自适应高斯混合模型对背景的每个像素建模;然后,对输入的当前帧图像的每一像素值与该像素点对应的高斯混合背景模型的各个高斯模型进行比较,将前景运动区域(包括运动目标、投射阴影)从场景中提取出来;最后,采用局部二元图(Local Binary Pattern,LBP)来提取纹理特征,利用背景在阴影覆盖前后的纹理相似性去除投射阴影,同时结合阴影的空间几何特性优化运动目标检测结果。实验结果表明,该算法能有效地检测出投射阴影和运动目标,具有较高的实际应用价值。

关键词:运动目标检测; YC_bC_r 颜色空间;高斯混合模型;阴影检测

中图分类号:TN919.8

文献标识码:A

文章编号:1673-5439(2009)06-0017-06

Background Modeling and Moving Object Detection Based on YC_bC_r Color Space

LU Guan-ming, LANG Su-juan

(College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Gaussians mixture model (GMM) has been widely used for moving object detection based on background modeling. In this paper, the background is firstly modeled using adaptive Gaussian mixture models in YC_bC_r color space, and the foreground regions including moving objects and cast shadow are extracted from current frame by comparing the each pixel with Gaussian model. Then, the texture of little patches is represented by local binary patterns and the cast shadow is detected and eliminated based on the texture similarity between shadow region and corresponding region in the background. Finally, the geometric features of cast shadow are imposed to further improve the performance of moving object detection. Experimental results demonstrate the proposed algorithm can effectively detect cast shadow and moving object, and has higher practicability.

Key words: moving object detection; YC_bC_r color space; Gaussian mixture model; shadow detection

0 引言

在智能化的视频监控系统中,监视场景中的运动目标的检测与识别是十分关键的技术。但从目前的研究成果来看,快速精确的运动目标检测仍然是一个值得研究的课题。现有的运动目标检测方法主要有光流法^[1-2]、帧间差分法^[3]和背景相减法^[4-6]。对于摄像头固定的情形,背景相减法是目前最常用的运动目标检测方法。其基本思想是将当前帧图像

与事先存储或者实时得到的背景图像相减,若对应像素的差值大于某一阈值,则判此像素属于运动目标上的一个像素,阈值操作后得到的结果直接给出了运动目标的位置、大小、形状等信息。背景相减法的关键问题是如何建立背景模型和实时更新模型参数以适应背景变化,这些背景变化包括场景的光照变化和场景构成的改变。此外,检测运动目标时,运动目标投射的运动阴影也易被误检测为运动目标的一部分,这将影响检测结果的精确性,使检测到的目

标与真实目标形状相差很大。这虽然对目标检测来说影响不大,至少检测到了运动区域,但对后续的处理如目标识别、行为判断等应用会产生很大的影响。因此,如何去除运动阴影也至关重要。

在基于背景相减法的运动目标检测算法中,为了减少动态场景变化对于运动目标检测的影响,大部分研究人员曾致力于开发不同的背景模型,如 Haritaoglu 等^[7]用一段场景背景图像序列中每个像素的最小、最大亮度值和亮度最大时间差分值为场景中的每个像素进行统计建模,并且进行周期性更新;Stauffer 与 Grimson^[8]提出一种高斯混合模型(Gaussian Mixture Model, GMM),将像素的亮度分布用几个高斯函数的加权和来近似,以同时处理多种背景变化。其中,高斯混合模型可以利用在线估计来更新模型参数,以克服光照变化、背景扰动等影响,自适应性较强,被广泛用于背景建模^[9~11]。但高斯混合背景模型无法消除阴影的干扰,因此找到消除阴影的方法变得十分重要。

文献[12]对多种阴影检测抑制算法进行了比较全面的分类比较。运动阴影检测方法可分为两类:基于模型的方法^[13]和基于特征的方法^[14~16]。基于模型的方法是利用场景、运动目标、光照条件的先验知识建立阴影模型,通常只适合在一些特定的场景条件下使用。另外,该方法计算复杂度高,不适合于实时的应用。基于特征的方法是利用阴影的几何特性、亮度、色度等信息来检测阴影,对不同场景及光照条件有较强的鲁棒性。现有的大部分基于特征的方法都采用了 RGB 颜色空间或 HSV 颜色空间^[14~16]。本文提出了一种基于 YC_bC_r 颜色空间的高斯混合背景建模、阴影抑制及运动目标检测方法:首先在 YC_bC_r 颜色空间采用高斯混合模型对背景的每个像素建模;然后,对输入的当前帧图像的每一个像素值与该像素点对应的高斯混合背景模型的各个高斯模型进行比较,将前景运动区域(包括运动目标、阴影)从场景中提取出来;最后,采用局部二元图(Local Binary Patterns, LBP)来提取纹理特征,利用背景在阴影覆盖前后的纹理相似性去除阴影,同时结合空间几何特性优化运动目标检测结果。

1 基于 YC_bC_r 颜色空间的背景建模

1.1 颜色空间的选择

本文选择 YC_bC_r 颜色空间的理由是:(1) RGB 颜色空间是不均匀的颜色空间,两个颜色之间的知

觉差异与颜色空间中两点间的欧氏距离不成线性比例,而且 R、G、B 值之间的相关性很高,对同一颜色属性,在不同条件(光源种类、强度和物体反射特性)下 R、G、B 值很分散,对于识别某种特定颜色,很难确定其阈值和其在颜色空间中的分布范围;(2) HSV 颜色空间接近人眼感知色彩的方式,但是从 RGB 颜色空间到 HSV 颜色空间的转换较为复杂,H 和 S 均为 R、G、B 的非线性变换,存在奇异点,在奇异点附近即使 R、G、B 的值有很小变化也会引起变换值有很大的跳动,并且在亮度值和饱和度较低的情况下,采用 HSV 颜色空间计算出来的 H 分量是不可靠的;(3) 在 YC_bC_r 颜色空间中,色度分量和亮度分量是相互独立的,而且 YC_bC_r 颜色空间与 RGB 颜色空间存在一种线性变换关系,转换较为简单,常用于视频编码压缩领域;(4) 在数字视频监控系统,除了进行运动目标检测、识别等操作以实现自动报警等功能以外,还要实现对监控视频进行压缩存储与传输的功能,采用 YC_bC_r 颜色空间就无需再作颜色空间的转换。

1.2 模型描述

在一段时间内,监控场景中每个像素位置出现背景的概率比出现前景的概率要大,基于此,本文采用对背景进行统计建模的方法。对于摄像机固定的应用场合,如果背景完全静止,背景图像的每个像素点可以用一个高斯分布来描述。但背景场景往往不是绝对静止的,例如,由于树枝的摇动,背景图像上的某一像素点在某一时刻可能是树叶,可能是树枝,也可能是天空,每一状态像素点的颜色值都是不同的,背景常呈现多模态特性。所以,用一个高斯模型来描述是不能反映实际背景的。高斯混合背景建模的主要思想是利用 K 个高斯分布对像素点的颜色历史信息进行统计,不同的高斯模型表征像素点的不同颜色分布,每个高斯模型根据它表征当前颜色信息的频度而被赋予权值 w_k。对视频帧中的某个像素(x, y), t 时刻的观察值 X_t = [I_t^y, I_t^{c_b}, I_t^{c_r}]^T 是一个 3 维矢量,其中 I_t^y, I_t^{c_b} 和 I_t^{c_r} 分别是 t 时刻图像在像素(x, y)点处的 Y、C_b、C_r 3 个分量值。如果将该像素的所有历史值用 K 个高斯模型来近似,那么观察到的当前像素值的概率为:

$$P(X_t) = \sum_{k=1}^K w_{k,t} p(X_t | \mu_{k,t}, \Gamma_{k,t}) \quad (1)$$

式中,K 是高斯模型的个数,K 的取值取决于系统的计算能力,通常取值在 3~5 之间,本文取 K = 3。

$w_{k,t}$ 是 t 时刻第 k 个高斯模型的权值, 满足 $\sum_{k=1}^K w_{k,t} = 1$, 它的大小体现了当前用该高斯模型表示像素值时的可靠程度。 $\mu_{k,t}$ 和 $\Gamma_{k,t}$ 分别是 t 时刻第 k 个高斯分布的均值矢量和协方差矩阵。由于是对彩色图像进行建模, 且认为 Y, C_b, C_r 3 个通道是相互独立的, $\mu_{k,t}$ 和 $\Gamma_{k,t}$ 可写为如下形式:

$$\mu_{k,t} = [\mu_{k,t}^Y, \mu_{k,t}^{C_b}, \mu_{k,t}^{C_r}]^\top \quad (2)$$

$$\Gamma_{k,t} = \begin{bmatrix} (\sigma_{k,t}^Y)^2 & 0 & 0 \\ 0 & (\sigma_{k,t}^{C_b})^2 & 0 \\ 0 & 0 & (\sigma_{k,t}^{C_r})^2 \end{bmatrix} \quad (3)$$

式中, $p(X_t | \mu_{k,t}, \Gamma_{k,t})$ 为 t 时刻第 k 个高斯模型分布概率密度函数, 定义为:

$$p(X_t | \mu_{k,t}, \Gamma_{k,t}) = \frac{1}{(2\pi)^{\frac{3}{2}} |\Gamma_{k,t}|^{\frac{1}{2}}} e^{-\frac{1}{2} [X_t - \mu_{k,t}]^\top \Gamma_{k,t}^{-1} [X_t - \mu_{k,t}]} \quad (4)$$

1.3 模型参数初始化及调整

建立高斯混合模型的关键是确定参数 $w_{k,t}, \mu_{k,t}$ 和 $\Gamma_{k,t}$ 。当输入新的序列图像时, 图像中每一点的高斯混合模型的参数应该及时更新。理想的情况是在每一时刻 t , 用包含最新观察数据的一个时间段内的数据, 采用 EM(Expectation Maximization, 期望最大化)算法^[17]对每一点的高斯混合模型参数进行估计, 但这个过程非常耗时。因此, 本文采用在线 K -均值算法^[8]来近似地估计出模型参数, 其过程描述如下:

(1) 初始化模型: 将采集到的第一帧图像的每个像素点的像素值作为均值, 再赋以较大的方差和较小的权重, 本文取 $w_{k,t} = 0.05, \sigma_{k,t}^j = 30$, 其中, $j = \{Y, C_b, C_r\}$ 。

(2) 模型学习(确定参数): 将当前帧的对应像素的像素值与已有的 L 个 ($L \leq K$) 高斯模型作比较, 若同时满足 $|I_t^j - \mu_{k,t}^j| < 2.5\sigma_{k,t}^j$, 其中, $j = \{Y, C_b, C_r\}$, 则调整第 k 个高斯模型参数和权重。否则, 若 $L < K$, 则增加一个新的高斯模型; 若 $L = K$, 则用新的高斯模型代替优先级最低的高斯模型。新的高斯模型取每个像素点的像素值作为均值, 再赋以较大的方差和较小的权重。这样, 通过不断的训练学习, 最终得到混合高斯分布模型。

各高斯模型的优先级 p_k 定义如下:

$$p_k = \frac{w_{k,t}}{(I_t^Y - \mu_{k,t}^Y)^2 + (I_t^{C_b} - \mu_{k,t}^{C_b})^2 + (I_t^{C_r} - \mu_{k,t}^{C_r})^2} \quad (5)$$

各高斯模型参数和权重的调整算法描述如下。

对每一帧输入图像, 将图像的每点像素值与该像素点对应的高斯混合模型的各个高斯模型进行比较, 若满足 $|I_t^j - \mu_{k,t}^j| < 2.5\sigma_{k,t}^j$, 其中, $j = \{Y, C_b, C_r\}$, 则该点像素值与第 k 个高斯分布匹配, 令 $M_{k,t} = 1$; 否则, 令 $M_{k,t} = 0$ 。对该高斯模型的权重作如下的调整:

$$w_{k,t} = (1 - \alpha)w_{k,t-1} + \alpha(M_{k,t}) \quad (6)$$

式中, α 是权值更新率, 或称模型学习率, 表示背景模型更新快慢的常数。如果 α 取值比较小, 则模型适应环境变化的能力就低, 能适应缓慢的环境变化, 或者说, 只要给足够的时间, 该模型终究能适应环境的变化; 如果 α 取值比较大, 则模型适应环境变化能力强, 但容易受噪声影响, 不够稳定。本文取 $\alpha = 0.3$ 。

对于未匹配的高斯分布, 其均值和方差不作更新; 对于匹配的第 k 个高斯分布, 其均值和方差作如下的更新:

$$\mu_{k,t} = (1 - \rho)\mu_{k,t-1} + \rho X_t \quad (7)$$

$$(\sigma_{k,t}^Y)^2 = (1 - \rho)(\sigma_{k,t-1}^Y)^2 + \rho \cdot [(I_t^Y - \mu_{k,t}^Y)^2 + (I_t^{C_b} - \mu_{k,t}^{C_b})^2 + (I_t^{C_r} - \mu_{k,t}^{C_r})^2] \quad (8)$$

$$(\sigma_{k,t}^{C_b})^2 = (1 - \rho)(\sigma_{k,t-1}^{C_b})^2 + \rho \cdot [(I_t^Y - \mu_{k,t}^Y)^2 + (I_t^{C_b} - \mu_{k,t}^{C_b})^2 + (I_t^{C_r} - \mu_{k,t}^{C_r})^2] \quad (9)$$

$$(\sigma_{k,t}^{C_r})^2 = (1 - \rho)(\sigma_{k,t-1}^{C_r})^2 + \rho \cdot [(I_t^Y - \mu_{k,t}^Y)^2 + (I_t^{C_b} - \mu_{k,t}^{C_b})^2 + (I_t^{C_r} - \mu_{k,t}^{C_r})^2] \quad (10)$$

式中, $\rho = \alpha \cdot p(X_t | \mu_{k,t}, \Gamma_{k,t})$ 。

1.4 高斯混合背景模型的建立

高斯混合模型模拟的是前景分布和背景分布, 当输入图像的某个像素值与该像素点对应的高斯混合模型的第 k 个高斯分布匹配时, 需判断第 k 个高斯分布代表了前景分布还是背景分布。当一个高斯分布的优先级 p_k 较高时, 它代表背景的可能性就越大。故将 K 个高斯分布按优先级 p_k 从高到低的次序排序, 选择前 B 个高斯分布作为背景分布, B 按下式来确定。

$$B = \arg \min_b \left(\sum_{k=1}^b w_{k,t} > T_1 \right) \quad (11)$$

式中, T_1 为阈值, 它表示能真正作为背景的像素点占总像素点的最小比例。如果 T_1 选取过小, 则背景模型为单一高斯分布, 这种情况下, 采用最大的概率分布可节省计算时间; 如果选取较大, 则背景模型就是多高斯分布的混合模型, 能反映重复性的移动背景(如树叶的摇晃、旗帜的飘动等)。本文取 $T_1 = 0.8$ 。

2 前景运动区域提取

背景模型建立以后,就可以对输入的含有运动目标的视频进行前景区域提取。如果 X_i 与表示背景模型的某个高斯分布匹配,即满足 $|I_i - \mu_{k,t}^j| < 2.5\sigma_{k,t}^j$,其中 $j = \{Y, C_b, C_r\}$,则该点属于背景区域,标记为 0;否则属于前景运动区域,标记为 1。这样得到前景运动区域二值化掩模图像,即

$$B_i(x, y) = \begin{cases} 0, & |I_i - \mu_{k,t}^j| < 2.5\sigma_{k,t}^j, j = \{Y, C_b, C_r\} \\ 1, & \text{otherwise} \end{cases} \quad (12)$$

尽管高斯混合模型较以前的算法有了很大的进步,尤其在复杂环境下比简单的背景相减算法性能要好很多。但足,它仍然是基于像素级上进行的运算,这就使得它在检测目标时容易含有摄像机噪声导致的“伪目标”;或者由于背景和运动目标某些位置的像素值可能相差很小,最后得到的前景运动区域二值化掩模图像中可能存在着“空洞”现象。为了消除这些影响,本文用数学形态学和判断连通区域大小的方法进行后处理。首先对得到的二值图像 $B_i(x, y)$ 用适当的形态结构元素进行形态开-闭运算,消除孤立的噪声点,合并非连通的前景邻近区域,消除目标碎块。然后,对各个前景运动区域进行连通区域标记,填补区域内的“空洞”。最后,计算每个独立的前景运动区域的大小(即包含的像素点数目),排除小区域后得到较为精确的前景运动区域二值化掩模图像 $M_i(x, y)$ 。

3 阴影消除与运动目标检测

在上述得到的前景运动区域中,除了包含真实的运动目标区域以外,通常还包含运动目标的投射阴影区域。阴影是指由于物体表面未被光源直接照射而形成的暗区域。阴影可分为自身阴影和投射阴影。自身阴影是指物体自身没有被光线直接照射到的部分。而投射阴影则是指由于光线直接照射而致使物体投影得到的区域。通过对阴影的研究和观察发现,虽然投射阴影的形成取决于多种不同因素,但背景在投射阴影覆盖前后的纹理具有相似性。本文就是基于这种特性,利用阴影区域和已获取背景相应位置的 LBP(Local Binary Patterns, 局部二元图)纹理相似性来检测投射阴影。

3.1 局部二元图的纹理描述

局部二元图(LBP)纹理描述算子通过比较中心像素点与邻域像素点灰度值的大小来描述局部纹理信息^[18]。它对像素点与周围点的亮度差值进行阈值判断获得一串二进制码来表征像素点附近的局部纹理,本质上是一种局部的亮度梯度信息描述。

$$LBP_{P,R} = \sum_{p=0}^{P-1} S(g_p - g_c) \cdot 2^p \quad (13)$$

式中, g_c 表示中心像素点亮度值, g_p 表示距离中心像素点半径为 R 的等间隔的 P 个邻域像素点亮度值; $S(\cdot)$ 是一个判别函数, 定义为:

$$S(g_p - g_c) = \begin{cases} 1, & |g_p - g_c| > T_2 \\ 0, & |g_p - g_c| \leq T_2 \end{cases} \quad (14)$$

式中, T_2 为阈值。 T_2 值太小时易受到噪声影响, 太大时对纹理的描述区分性不够。对于相对平坦的背景图像可以取较小的值, 以增加其纹理描述性; 对于纹理复杂的背景图像, 则相应地取较大的值, 增加其抗噪声性能。本文取 $T_2 = 4$ 。

$LBP_{P,R}$ 可以输出 2^P 个不同值, 对应于由 P 个邻域像素点组成的可能的 2^P 种不同的二进制模板。图 1 给出了计算 $LBP_{8,1}$ 纹理特征的一个例子, 可以用生成的 8 位二进制码来表示该局部纹理。

86	76	58		1	0	1
73	77	76		0		0
56	66	80		1	1	0

$LBP_{8,1}$ 码: 01100101

图 1 8 邻域图像块的 LBP 纹理特征

3.2 阴影消除与运动目标检测

投射阴影区域是一个半透明区域, 背景区域在被阴影覆盖前后的局部纹理近似不变。对于提取出的前景运动区域, 本文采用 $LBP_{8,1}$ 算子提取当前帧像素点的局部纹理信息和背景相应位置点的局部纹理信息, 然后对两个 8 位 $LBP_{8,1}$ 码进行按位“异或”运算得到汉明距离。汉明距越大则意味着纹理差异越大, 如果汉明距离小于某一阈值 T_3 (如本文 $T_3 = 2$), 则判定该像素点为阴影像素点, 否则为运动目标像素点。

为了提高运动目标检测的精确度, 本文进一步结合阴影的空间几何属性来排除一些伪阴影点。如果检测到的阴影像素块在前景运动区域内部被前景像素点包围,那么该阴影像素块为伪阴影区域。

在检测到投射阴影后, 我们将它从前景运动区域中去除, 再进行一些后处理运算, 就可以得到运动目标区域。

4 实验结果与分析

本实验系统运行的主机 CPU 为 Intel Core(TM) 2, 主频 2.13 GHz, 内存 2 GB, Windows XP 操作系统, 软件环境为 Visual C ++ 6.0 平台。以 Carnegie Mellon University 提供的两个典型的阴影检测测试视频序列为对象, 包括室内场景的 Laboratory 视频序列和室外场景的 Campus 视频序列, 图像大小为 320×240 像素。图 2 为去除阴影前后的运动目标检测效果对比, 其中白色区域为检测出的目标, 灰色区域为检测出的阴影。

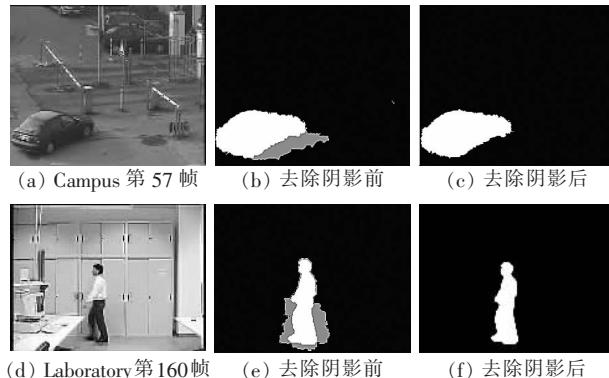


图 2 去除阴影前后的运动目标检测效果对比

为了定量评估阴影检测算法的性能, 引入阴影检测率 η 和阴影区分率 ξ , 其定义如下^[12]:

$$\eta = \frac{TP_S}{TP_S + FN_S} \quad (15)$$

$$\xi = \frac{\overline{TP}_F}{\overline{TP}_F + FN_F} \quad (16)$$

式中, TP_S 是指真正的阴影点的个数, FN_S 是被错误分类为目标点的阴影点的个数, TP_F 是指真正的目标点的个数, FN_F 是被错误分类为阴影点的目标点的个数, \overline{TP}_F 是指“真正的目标点的个数”减去“被检测为阴影点但属于前景目标点的个数”的值。为了获得相应的 η 和 ξ , 首先要取得某一视频序列的若干帧, 手工分割这些帧中的目标点和阴影点, 并对这些帧用阴影消除算法获得目标点和阴影点, 通过这些点计算 TP_S 、 FN_S 、 TP_F 、 \overline{TP}_F 和 FN_F , 最终得到 η 和 ξ 。

表 1 列出针对图 2(a) 和图 2(d) 的视频序列的阴影消除算法的性能, 与文献[12]中提到的 4 种方法 (SNP、SP、DNM1 和 DNM2) 相比, 本文提出的阴影消除算法在阴影检测率和阴影区分率上都有较大的提高。另外, 每帧的处理时间大约是 0.06 ~ 0.07 s, 若每隔 2 帧处理一次的, 则可以保证实时处理。

表 1 阴影消除算法性能比较^[12]

算法	Campus		Laboratory	
	$\eta/\%$	$\xi/\%$	$\eta/\%$	$\xi/\%$
SNP	80.58	69.37	84.03	92.35
SP	72.43	74.08	64.58	95.39
DNM1	82.87	86.65	76.26	89.87
DNM2	69.10	62.96	60.34	81.57
本文算法	81.40	92.61	84.51	93.17

5 结束语

本文提出的基于 YC_bC_r 颜色空间的背景建模及阴影消除方法, 能有效地处理外界光照条件变化、场景变化、背景扰动、阴影等带来的影响, 无论在室内还是室外都能够较好地检测出投射阴影和运动目标, 具有较高的实际应用价值。进一步的研究将主要集中于去除阴影后的运动目标识别及其行为分析。

参考文献:

- [1] YAMAMOTO S, MAE Y, SHIRAI Y, et al. Realtime multiple object tracking based on optical flows [C] // Proceedingss of the IEEE International Conference on Robotics and Automation. 21 ~ 27 May 1995, 3:2328 ~ 2333.
- [2] 初秀琴, 刘洋, 李玉山. 基于光流估计的运动目标检测系统设计实现 [J]. 系统工程与电子技术, 2007, 29(7):1174 ~ 1177.
CHU Xiuqin, LIU Yang, LI Yushan. Design of motion detection systems based on optical flow estimation [J]. Systems Engineering and Electronics, 2007, 29(7):1174 ~ 1177.
- [3] 周西汉, 刘勃, 周荷琴. 一种基于对称差分和背景消减的运动检测方法 [J]. 计算机仿真, 2005, 22(4):117 ~ 123.
ZHOU Xihan, LIU Bo, ZHOU Heqin. A Motion Detection Algorithm Based on Background Subtraction and Symmetrical Differencing [J]. Computer Simulation, 2005, 22(4):117 ~ 123.
- [4] PICCARDI M. Background subtraction techniques: a review [C] // Proc of 2004 IEEE International Conference on Systems, Man and Cybernetics. 10 ~ 13 Oct 2004, 4:3099 ~ 3104.
- [5] 代科学, 李国辉, 涂丹, 等. 监控视频运动目标检测减背景技术的研究现状和展望 [J]. 中国图像图形学报, 2006, 11(7):919 ~ 927.
DAI Kexue, LI Guohui, TU Dan, et al. Prospects and Current Studies on Background Subtraction Techniques for Moving Objects Detection from Surveillance Video [J]. Journal of Image and Graphics, 2006, 11(7):919 ~ 927.
- [6] 徐以美, 郭宝龙, 陈龙. 基于 RGB 颜色空间的减背景运动目标检测 [J]. 计算机仿真, 2008, 25(3):214 ~ 217.
XU Yimei, GUO Baolong, CHEN Long. Moving Objects Detection

- Based on RGB Color Space and Background Subtraction[J]. Computer Simulation, 2008, 25(3):214 - 217.
- [7] HARITAOGLU I, HARWOOD D, DAVIS L. W⁴: real-time surveillance of people and their activities[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8):809 - 830.
- [8] STAUFFER C, GRIMSON W E L. Adaptive background mixture models for real-time tracking[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Fort Collins, Colorado, USA, June 23 - 25, 1999, 2:246 - 252.
- [9] ZIVKOVIC Z. Improved adaptive Gaussian mixture model for background subtraction[C]// Proceedings of the International Conference on Pattern Recognition. Amsterdam, Netherlands, 2004: 23 - 26.
- [10] 陈振华,周锐锐,李光伟,等.一种改进的高斯混合背景模型算法及仿真[J].计算机仿真,2007,24(11):190 - 193.
CHEN Zhenhua, ZHOU Ruirui, LI Guangwei, et al. Simulation of an Improved Gaussian Mixture Model for Background Subtraction [J]. Computer Simulation, 2007, 24(11):190 - 193.
- [11] 焦波,李国辉,涂丹,等.一种用于运动目标检测的快速收敛混合高斯模型[J].中国图像图形学报,2008, 13(11):2139 - 2143.
JIAO Bo, LI Guohui, TU Dan, et al. A Fast Convergent Gaussian Mixture Model for Moving Object Detection[J]. Journal of Image and Graphics, 2008, 13(11):2139 - 2143.
- [12] PRATI A, MIKIC I, TRivedi M M, et al. Detecting moving shadows: algorithms and evaluation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2003, 25(7):918 - 923.
- [13] NADIMI S, BHANU B. Physical model for moving shadow and object detection in video[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004, 26(8):1079 - 1087.
- [14] 陈柏生,陈锻生.基于归一化rgb彩色模型的运动阴影检测[J].计算机应用,2006,26 (8):1879 - 1881.
CHEN Bosheng, CHEN Duansheng. Normalized rgb color model based shadow detection [J]. Journal of Computer Application, 2006, 26(8):1879 - 1881.
- [15] CUCCIARA R, GRANA C, PICCARDI M, et al. Detecting moving objects, ghosts and shadows in video streams[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2003, 25 (10):1337 - 1342.
- [16] 黄信想,刘秉瀚.基于HSV色彩空间的云模型车辆阴影检测[J].福州大学学报:自然科学版,2008,36(6):809 - 813.
HUANG Xinxiang, LIU Binghan. Vehicle shadow detection based on HSV color space and cloud model[J]. Journal of Fuzhou University:Natural Science, 2008, 36(6):809 - 813.
- [17] 朱周华.期望最大(EM)算法及其在混合高斯模型中的应用[J].现代电子技术,2003(24):88 - 90.
ZHU Zhouhua. EM Algorithm and Its Application in Mixture of Gaussian[J]. Modern Electronic Technique, 2003(24):88 - 90.
- [18] OJALA T, PIETIKÄINEN M, HARWOOD D. A comparative study of texture measures with classification based on feature distributions [J]. Pattern Recognition, 1996, 29 (1):51 - 59.

作者简介:



卢官明(1965-),男,浙江台州人。南京邮电大学通信与信息工程学院教授,博士,“信号与信息处理”学科学术带头人。1999年3月获上海交通大学通信与信息系统专业博士学位;2001年9月至2003年9月在东南大学博士后流动站从事博士后研究;2005年1月至2005年12月在瑞典于默奥(Umeå)大学作访问研究。目前主要的研究方向为图像处理与多媒体通信、计算机视觉、数字电视。

郎苏娟(1985-),女,江苏镇江人。南京邮电大学信号与信息处理专业硕士研究生。主要研究方向为图像处理与多媒体通信。

(上接第16页)

- [12] DUARTE M F, WAKIN M B, BARON D, et al. Universal distributed sensing via random projections[C]// Proc of 5th International Workshop on Inf Processing in Sensor Networks (IPSN '06). Nashville, TN, 2006: 177 - 185.
- [13] SWELDENS W. The lifting scheme:a new philosophy in biorthogonal wavelet constructions[J]. SPIE, 1995, 2569:68 - 79.
- [14] JINDAL A, PSOUNIS K. Modeling Spatially Correlated Data in Sensor Networks[J]. ACM Trans on Sensor Networks, 2006, 2 (4):466 - 499.
- [15] WAGNER R S, BARANIUK R G. An Architecture for Distributed Wavelet Analysis and Processing in Sensor Networks[C]// Information Processing in Sensor Networks(IPSN). Tennessee, USA, 2006:243 - 250.

作者简介:



胡海峰(1973-),男,安徽六安人。南京邮电大学通信与信息工程学院讲师,博士。主要研究方向为面向网络的信号处理和无线传感器网络中基于移动代理的关键技术等。

杨震(1961-),男,江苏苏州人。南京邮电大学校长,教授,博士生导师。(见本刊2009年第1期第14页)

基于 QoS 和 SLA 的网络计费系统设计

张登银¹, 吴超², 程春玲²

(1. 南京邮电大学 科技处, 江苏南京 210046
(2. 南京邮电大学 计算机学院, 江苏南京 210046))

摘要:传统网络计费一般采用单一计费原则,不能体现多级服务的差异性,也不能验证收取费用的合理性。为了解决这些问题,设计了一种基于 QoS(Quality of Service)和 SLA(Service Level Agreement)的计费策略,并在 Linux 环境下进行了系统实现,整个系统模块包括用户 SLA 签订、用户认证和配置、接纳控制、数据采集和计费处理。在实验室环境下,进行了系统功能验证和计费策略对比分析,结果表明基于 QoS 和 SLA 的计费系统可以较好地解决传统网络计费方式存在的不足,满足下一代网络应用发展的需要。

关键词:网络计费系统;服务质量;服务等级协定

中图分类号:TP393.07

文献标识码:A

文章编号:1673-5439(2009)06-0023-05

Design on Network Billing System Based on QoS and SLA

ZHANG Deng-yin¹, WU Chao², CHENG Chun-ling²

(1. Department of Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210046, China
(2. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210046, China))

Abstract: Traditional network billing, with a single billing method, could neither reflect the differentiation of the multi-level service nor validate rationality of billing. In order to solve these problems, we design a scheme of “QoS-SLA-based billing system”. And we implement the billing system under the circumstance of Linux, which includes SLA subscription, authentication and deployment, admit control, data collection, data disposal and billing. At last, the experiments of comparing the two billing schemes prove that our billing scheme could solve the disadvantages with traditional network billing scheme and could satisfy the demand of NGN’s development.

Key words: network billing system; quality of service; service level agreement

0 引言

网络计费系统对网络供应商(Network Provider, NP)来说极其重要,它不仅可以统计用户的使用费用,而且可以监督网络流量,优化网络资源分配。先进的网络计费系统是提供优质网络服务的重要保证^[1]。现有的计费方法主要有两种,即单一价格方法和基于使用量的线性定价方法^[2-3]。这些计费方法都采用单一的计费原则,制约了网络用户的数量和新业务的发展。根据业务类型、使用量和网络状态对业务进行计费,是解决这些问题的可行方法。

针对下一代网络(Next Generation Network ,

NGN)丰富的业务应用并具有开放业务平台的特点^[4],本文通过研究现有计费技术,设计了一种基于 QoS(Quality of Service)^[5]和 SLA(Service Level Agreement)^[6]的计费策略并进行了系统实现,以便网络供应商能够为用户提供差异化服务并验证收取费用的合理性,达到既能保证 QoS 和 SLA、缓解网络拥塞,又能合理分配网络资源的目的。我们在实验室环境下对系统功能进行了实验验证,并将本文设计的计费策略与传统计费方式进行对比分析。结果表明,本文设计的基于 QoS 和 SLA 的计费系统可以较好地解决传统网络计费方式存在的不足,满足下一代网络应用发展的需要。

1 基于 QoS 和 SLA 的计费策略

在基于 QoS 和 SLA 计费策略中,我们定义用户使用网络业务在一个周期 T 内,所要支付的总费用为 C ,它由用户的业务使用费用 $C_u(n)^j$ 、拥塞费用 $C_c(n)^j$ 和补偿费用 $C_p(n)^j$ 组成(其中 i 表示用户使用等级, j 表示业务类型):

$$C = \sum_j \sum_i [C_u(n)^j + C_c(n)^j - C_p(n)^j] \quad (1)$$

其中,业务使用费用 $C_u(n)^j = P_u^j \times R_T$ (P_u^j 等级为 i ,业务类型为 j 的使用费率; R_T 是周期 T 内业务的数据包流量),体现了区分业务定价的思想,不同业务以及相同业务的不同等级在单位带宽单位时间内收取的业务使用费用是不同的;拥塞费用 $C_c(n)^j = A \times P_c^j \times R_T$ (A 是拥塞费用的调节因子,由实时采集的 QoS 参数确定; P_c^j 是等级为 i ,业务类型为 j 的拥塞费率; R_T 是周期 T 内业务的数据包流量),体现了流量控制的思想,根据网络拥塞的实际情况制定拥塞价格,提高使用费用,利用价格杠杆引导用户使用资源;补偿费用 $C_p(n)^j = B \times (P_u^j - P_u^{j'}) \times R_T$ (B 是补偿费用的调节因子,由实时采集的 QoS 参数确定; $P_u^{j'}$ 是降级后等级为 i ,业务类型为 j' 的使用费率; R_T 是在周期 T 内业务的数据包流量),体现了 NP 对用户承诺的保证。综合计费体系克服了静态计费方法的不公平性,使得用户和 NP 两者的效益最大化。

基于 QoS 和 SLA 综合计费的处理流程,归纳如下:

- (1) 计费系统对接纳的业务数据进行处理,把采集的参数写入数据库,包括数据包流量、业务占用的带宽、端对端时延、丢包率等;
- (2) 计算业务使用费用 $C_u(n)^j = P_u^j \times R_T$;
- (3) 判断是否发生拥塞;若是,转(4);否则,转(6);
- (4) 根据采集到的各类 QoS 参数,计算拥塞调节因子 A ,计算拥塞费用 $C_c(n)^j = A \times P_c^j \times R_T$;
- (5) 再判断接纳网关在发生拥塞时是否对用户的业务进行了降级服务;若是,根据采集到的各类 QoS 参数,计算拥塞调节因子 B ,计算相应的补偿费用 $C_p(n)^j = B \times (P_u^j - P_u^{j'}) \times R_T$,转(6);否则,直接转(6);
- (6) 计算的总费用 C ,并进行实时扣费。

2 计费系统体系结构

基于 QoS 和 SLA 计费系统体系结构如图 1 所示,包括接入网关、目的网关、网络和计费中心 4 大部分。

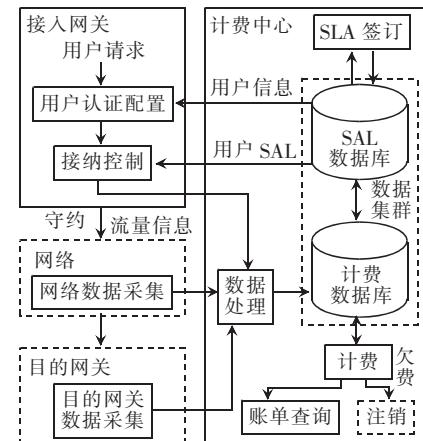


图 1 基于 QoS 和 SLA 计费系统体系结构

计费系统主要功能模块包括 SLA 签订、用户认证和配置、接纳控制、数据采集、数据处理和计费。其中,SLA 签订、用户认证和配置、接纳控制 3 个模块位于接入网关单元。本文所采用的 SLA 是静态或准静态的,即用户与 NP 事先签订,在系统中体现为 SLA 数据库;数据采集模块分散在接入网关、网络内部和目的网关上,完成网络相关信息采集;原始信息经数据处理模块分析处理后,以计费参数的形式存入计费数据库。

接入网关是计费系统的核心,我们将计费中心需要完成的功能,也融入到接入网关。接入网关中完整的业务流程,如图 2 所示。

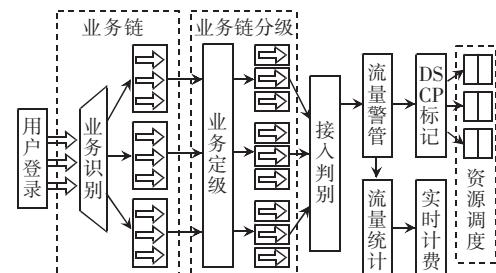


图 2 接入网关业务流程

用户登录后,发送或接收的业务流量经过业务识别模块进入各自的业务链;同一业务链中的流量经过业务定级模块后有了等级区分,进入不同等级的业务链;在接入判别模块,丢弃非法的业务流量,并对资源得不到满足的流量进行降级或丢弃处理;进入流量警管模块后,不同的业务流量根据 SLA 要求进行相应的限流,丢弃违约流量,将守约流量信息传递给流量统计模块,该模块将信息写入流量数据库,实时计费模块根据流量数据库信息实时计算业务费用并进行用户余额更新,这些费用信息将用于客户端的查询;从流量警管模块出来的流量还需打上 DS/CP 标记,最后对将要进入网络的流量进行队列调度和流量整形。

3 系统关键模块实现

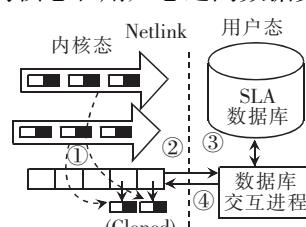
3.1 业务识别模块

业务识别的速度和精度是整个系统正常工作的前提。业务识别可以从 TCP/IP 协议栈各个层面展开。本系统采用 Linux 下最流行的 Netfilter/Iptables 包过滤及防火墙套件^[7], 建立相应规则, 对业务进行识别。分组在 Iptables 链(chain)中的遍历过程经历以下阶段^[8]: PREROUTING 链→FORWARD 链→POSTROUTING 链。

业务识别(分组过滤)通常在 FORWARD 链或 POSTROUTING 链中进行。本系统选择在 FORWARD 链中实现业务识别, 并设置了对应的新业务链, 在 FORWARD 链设定跳转到各个业务链的匹配规则, 与这些规则匹配的流量(已识别)将被分流到各自的新业务链中去, 实现流量的分类; 对于不能识别的流量, 则将它归为默认的业务(Default)链。新业务链的作用是作为分流后各种业务流的暂存区, 同时可对各个业务链进行流量统计。

3.2 业务定级模块

业务定级是实现 SLA 和 QoS 保证的关键, 业务定级的效率和准确性直接影响整个系统效率。我们提出了一种基于连接状态^[9]的 SLA 流量标记方案。基于连接状态是指通过对每个连接信息(包括套接字对、协议类型、协议连接状态和超时等)进行检测, 从而判断是否过滤数据包。此外, 这类连接还具有记忆连接状态以及在内存中为每个数据流建立上下文的能力。这些连接状态信息通常放在数据库中, 本文称之为连接状态数据库。图 3 为本方案实现的 Linux 内核态和用户态之间数据交互过程。



(3) 按顺序处理业务列表中的各个开展的业务;

(4) 流量统计模块从内核导出缓冲区流量,并根据用户的 IP 地址获取各类计费参数和费率,计算出各项费用和总费用,分别写入流量统计数据库和计费数据库。

(5) 检测缓冲区处流量是否抓取完毕,若是,转(6);否则,转(3)处理下一业务。

(6) NP 当前开展业务列表中的业务都处理完后,过周期 T 后跳转(1)。

4 实验结果与性能分析

实验环境为两台 CISCO 2950 交换机、4 台 IBM-PC 服务器(P4 2.66 G, 内存 512 M, 操作系统 Debian Linux 3.1)和两台客户机(P4 1.8 GHz, 内存 512 M, 操作系统 Windows XP)。PC1 作为接入网关, 部署 pppoe、Netfilter/Iptables 规则、Radius 和 MySQL 服务, 配置为软路由, 用户登录、业务识别、接纳判断、测量模块、资源管理、DSCP 标记、流量统计和业务计费模块均部署在 PC1; PC2 作为目的网关; PC3 用来模拟网络, 连接 PC1 和 PC2; PC4 上部署各种实验所需的应用类型业务, 如 Web、FTP、VoIP、P2P 等, 并通过交换机 1 与 PC2 相连; 客户机 A、B 通过交换机 2 连接到 PC1。

表 1 系统测试结果

用户名	IP 地址	业务类	服务等级	内部标识	DSCP	平均带宽/(kbit/s)	拥塞发生降级服务	流量/M	业务费用/元	拥塞费用/元	补偿费用/元	总费用/元
A 192.168.0.15		WEB	低	1030	000011	15	否	3	0.059	0	0	
		FTP	中	2023	000101	45	否	81	0.346	0	0	2.264
		BT	高	6013	010000	103	否	186	1.839	0	0	
B 192.168.0.21		WEB	低	1023	000001	16	是	4	0.081	0.163	0.053	
		FTP	低	2023	000100	27	是	49	0.177	0.358	0.113	3.511
		BT	中	6012	010000	81	是	146	1.273	2.478	0.853	

性能分析: 我们通过白盒测试方法对所有模块进行测试, 并通过黑盒测试方法对计费系统进行整体测试, 目的是为了验证本文提出的计费系统的正确性和可行性。测试结果比较客观的说明本文设计的计费系统是切实可行的, 已经在实验室环境下取得成功。实验中也发现一些问题, 例如: 业务识别模块的精度和速度与实际采用的识别方法和手段有很大关系, 随着技术的发展, 已经出现硬件级的识别设备和可以灵活扩充的特征库, 这也是近年来各大公司研究的热点问题; 业务定级采用的是我们自己开发的基于连接状态的 SLA 流量标记方案, 测试结果表明该模块可以很好地胜任本职工作, 但可进一步

4.1 实验一: 系统功能验证与性能分析

实验场景: 客户机 A、B 通过 PC1、PC2 和 PC3 访问 PC4 上部署的各类业务, 接入网关 PC1 对用户使用的业务进行识别、定级, 打上 DSCP 标识, 并从数据库中提取各类计费参数, 计算各项费用和总费用, 并把各费用数据写入数据库中。PC2 和 PC3 完成业务的 QoS 信息采集。针对用户开展的所有业务, 对业务的计费流程进行测试, 包括各模块之间的参数传递、模块组装后系统的整体功能、各项业务是否能正常合理地开展, 计费是否准确合理。

整体计费系统共测试 1 小时, 其中用户 A 的 WEB、FTP 以及 BT 业务同时开始, 持续测试了 30 分钟, 并设定期间网络正常状态; 然后用户 B 的 WEB、FTP 以及 BT 业务同时开始, 持续测试了 30 分钟, 并设定期间网络发生拥塞。整个计费系统部分测试结果如表 1 所示, 其中内部标识用于接纳网关表示用户签订业务的 SLA, 例如 2023(首部 2 表示业务类型, 这里标识为 ftp 业务, 中间 0 为保留位, 中间 1 表示用户签订的首选业务等级, 这里标识为中级, 尾部 3 表示用户签订的当网络拥塞时备选的业务等级, 这里标识为低级, 注意如果尾部标记为 0, 表示用户没有签订备选业务等级, 即当网络拥塞时用户不能得到服务); DSCP 用于标记允许进入网络的数据包。

改进算法, 提高其效率。

4.2 实验二: 计费策略对比与性能分析

实验场景: 针对 BT 业务, 对本系统的计费策略和现有计费方法中基于使用量的计费方法^[9]进行对比实验。在本系统的 BT 业务实验中, 首选业务等级为中级, 备选的业务等级为低级; 基于使用量计费的费率为 0.5 元/百兆。试验中采样周期为 6 s, 共测试了 1 小时内用户 BT 业务的总费用, 同时设定第 300 ~ 400 次周期(即 1 800 ~ 2 400 s)发生拥塞, 期间收取拥塞费用, 返还相应的补偿费用, 如图 4 所示。

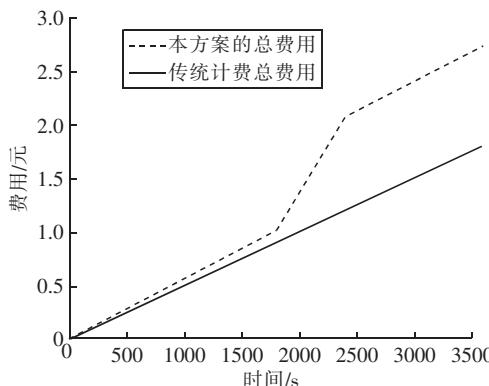


图 4 两类计费策略的比较

在实验中,本文提出的计费策略的总费用比传统计费方法的总费用要高一些,这主要是因为中级 BT 业务的费率略高于基于使用量计费的费率,并且当发生拥塞时,由于拥塞费用是以高价格定价的,且收取的拥塞费用高于因业务降级服务而补偿的费用。图 5 是图 4 中图线 1 所囊括的 3 种计费策略的费用在横坐标从 1 800 ~ 1 920 s 的局部放大图。由图 5 可见,网络拥塞时,由于业务的使用价格不变,只以流量为变量,所以使用费用呈现出平缓连续增加的状态;同时,接纳网关对该业务进行降级服务,但由于补偿费用小于拥塞费用,所以图 4 中 1 800 ~ 1 920 s 这段时间内总费用的斜率上升趋势较大。

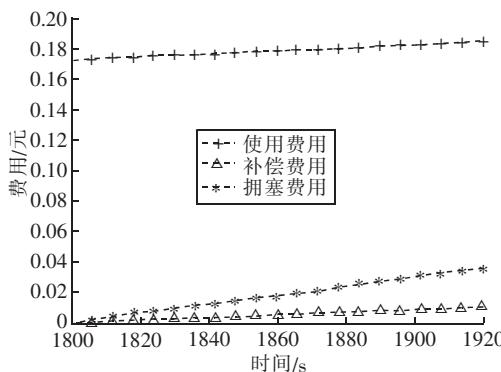


图 5 本方案中各种费用的局部比较

5 结束语

新的计费系统由于采用了精确到用户业务级的流量管理模式,大大细化了计费粒度;同时在计费策略中综合考虑 QoS 和 SLA,采用不同的计费方式,相对于现有的单一计费原则,使得计费的精确性和合理性大大提高。本文提出的计费系统既能保证 QoS 和 SLA,又能够合理的分配网络资源,克服了静态计费方法的不公平性,使得用户和 ISP 的效益都达到了最大化。

参考文献:

- [1] 张登银,王雪强. 基于服务级别和流量控制的网络计费[J]. 重庆邮电学院学报:自然科学版,2005,17(3):328~331.
ZHANG Dengyin, WANG Xueqiang. Multiservice and traffic-based network billing system[J]. Journal of Chongqing University of Posts and Telecommunications:Natural Science, 2005, 17(3):328~331.
- [2] 张登银,卢栋梁,王雪强. 基于数据包的下一代网络计费策略研究[J]. 电子与信息学报,2006,28(12):2382~2385.
ZHANG Dengyin, LU Dongliang, WANG Xueqiang. Research on Per-Packet Pricing Strategy for Next Generation Network[J]. Journal of Electronics & Information Technology, 2006, 28(12):2382~2385.
- [3] ZHANG Dengyin, ZHANG Li, TANG Zhiyun. A Novel Admission Control Algorithm Based on Negotiation and Price[J]. CHUPT, 2005, 12(1):76~79.
- [4] 杨晓艳,冯文江,刘海然. 基于软交换技术的下一代通信网络的应用研究[J]. 重庆大学学报:自然科学版,2009,17(5):33~36.
YANG Xiaoyan, FENG Wenjiang, LIU Hairan. The Researches and Applications of Next Generation Network Based on the Soft-switch Technology[J]. Journal of Chongqing University: Natural Science, 2009, 17(5):33~36.
- [5] 马秀芳,时和平. IP 网络中的 QoS 研究[J]. 现代有线传输,2003(3):48~54.
MA Xiufang, SHI Heping. Research on QoS for IP network[J]. Modern Wire Transmission, 2003(3):48~54.
- [6] VERMA D C. Service level agreements on IP networks[J]. Proceedings of the IEEE, 2004, 92(9):1382~1388.
- [7] HUBERT B. Linux Advanced Routing & Traffic Control HOWTO. [EB/OL]. <http://lartc.org/lartc.pdf>
- [8] GOUDA M G, LIU A X. A model of stateful firewalls and its properties[C]// Dependable Systems and Networks, Proceedings International Conference. 2005:128~137.
- [9] FALKNER M, DEVETSIKOTIS M. An overview of pricing concepts for broadband IP networks[C]// IEEE Communications Surveys. 2000:2~13.

作者简介:



张登银(1964-),男,江苏靖江人。南京邮电大学科技处副处长,研究员,博士。研究方向为 IP 网络技术,信号与信息处理。

吴超(1984-),男,浙江杭州人。南京邮电大学计算机学院硕士研究生。研究方向为网络服务质量。

程春玲(1972-),女,陕西西安人。南京邮电大学计算机学院副教授。研究方向为计算机网络,信息安全。

基于 BFGS 方法的拥塞速率控制算法

魏 涛^{1,2}, 张顺颐¹

(1. 南京邮电大学 信息网络技术研究所, 江苏南京 210003
2. 解放军理工大学 工程兵工程学院, 江苏南京 210007)

摘要:针对 TCP/AQM 对偶性模型采用梯度投影方法调整链路价格收敛速度慢的问题, 使用具有更快收敛速度的 BFGS 方法来进行链路价格的计算, 提出一种基于 BFGS 方法的拥塞速率控制算法。仿真结果证明, 利用 BFGS 方法所设计拥塞速率控制算法具有更快的收敛速度, 算法性能优于其它算法。

关键词:TCP/AQM 对偶性模型; 拥塞算法; BFGS 方法; 链路价格

中图分类号:TN915.04 文献标识码:A 文章编号:1673-5439(2009)06-0028-03

Congestion Rate Control Algorithm Based on BFGS Method

WEI Tao^{1,2}, ZHANG Shun-yi¹

(1. Institute of Information Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Engineering Institute of Corps of Engineers, PLA University of Science and Technology, Nanjing 210007, China)

Abstract: Aiming at the problem that gradient projection method for adjusting the link price is slower convergence in TCP/AQM dual model, the BFGS method with a faster convergence rate is used to calculate the link price and a congestion rate control algorithm by using BFGS method is presented in this paper. The simulation results successfully show that the proposed algorithm has the advantages over other algorithms of faster convergence rate and good performance.

Key words:TCP/AQM dual model; congestion control algorithm; BFGS method; link price

0 引言

随着 Internet 网络在最近几年中的迅猛发展, 拥塞问题成为制约网络应用和发展的一个瓶颈, 网络拥塞表现为数据分组时延增加、丢失概率增大、数据流吞吐量下降, 网络应用系统性能下降等状况。Internet 网络产生拥塞的根本原因在于用户(源端系统)提供给网络的负载大于网络资源和处理能力, 即用户数据发送速率和大于网络带宽。为了解决网络拥塞问题, 必须在 Internet 网络之上对用户的数据发送速率进行有效的拥塞控制。

现有的拥塞控制算法主要分为两类: 在源端系统上使用的 TCP 拥塞控制算法以及在路由器中使用的主动队列管理(Active Queue Management, AQM)算法。由于网络拓扑结构复杂, 规模庞大, 并

且处于不断地变化之中, 这造成拥塞控制问题的复杂性, 单纯基于算法参数调整的主观分析方法已经不能解决问题, 必须借助于数学上的非线性优化方法来分析、设计拥塞控制算法。S. H. Low^[1-2]等人基于优化理论和 F. P. Kelly^[3]等人用价格(Price)来表示网络拥塞度量的研究成果, 提出了完整的 TCP/AQM 对偶性模型。该模型把现有的 TCP 拥塞控制和 AQM 算法看作是求解具有适当效用函数的最优速率分配问题, 从而可从理论上分析网络在平衡状态时的性能, 如吞吐量、丢失率、时延和排队长度等。

TCP/AQM 对偶性模型提供了一种使整个网络各个用户数据流发送速率趋向于某个期望点的方法, 把用户合适拥塞速率的求取归结为一个非线性优化模型, 便于对拥塞控制机制的改进与设计, 然而, TCP/AQM 对偶性模型采用了梯度投影方法调整链路价

格,实际的网络流量是不断变化的,所以速率优化的目标也是变化的。因此要实现全局的最优化速率控制,关键在于提高拥塞速率控制算法的收敛速度,这样才能不断地逼近变化的优化目标。因此本文使用具有更快收敛速度的BFGS(Broyden Fletcher Goldfarb Shanno)方法^[4-7]来进行链路价格的计算,提出一种基于BFGS方法的拥塞速率控制算法。

1 TCP/AQM对偶性模型

TCP/AQM对偶性模型假设用户连接 r 所分配的速率为 x_r , x_r 对应的效用(收益)值为 $u_r(x_r)$, 显而易见, 网络带宽资源的分配集合 $\{x_r\}$ 应使 $\sum_r u_r(x_r)$ 取得最大值, 那么网络的优化模型为:

$$f = \max_{\{x_r\} \in S} \sum_r u_r(x_r) \quad (1)$$

约束条件为:

$$\sum_{r: l \in r} x_r \leq c_l, x_r \geq 0 \quad (2)$$

其中, $l \in L$, L 为网络中所有链路的集合, $r \in S$, S 为网络中所有用户的集合, 链路 l 的最大带宽资源为 c_l , $r: l \in r$ 为与链路 l 相关用户 r 的集合。

根据库恩-塔克条件, $\max_{\{x_r\} \in S} \sum_r u_r(x_r)$ 的解满足下列等式:

$$L(x, p) = \sum_{x_r \geq 0} u_r(x_r) - \sum p_r (\sum x_r - c_l) \quad (3)$$

用户 r 最优速率 x_r 的求取根据:

$$\begin{aligned} \frac{\partial L(x, p)}{\partial x_r} &= u'_r(x_r) - \sum p_r \\ &= u'_r(x_r) - q_r = 0 \end{aligned} \quad (4)$$

其中, $\sum p_r = q_r$ 为用户 r 路由上链路价格的和。

为了求解最优 $\{p_l\}$ 与 $\{x_r\}$, 使用下面的分布式求解算法:

在每个用户连接的发送方:

$$x_r = u_r'^{-1}(q_r) \quad (5)$$

在每个链路的路由器中:

$$p_l(k+1) = [p_l(k) + \gamma(y_l - c_l)]^+ \quad (6)$$

其中, y_l 为到达链路的所有用户的速率和, γ 为步长, $[z]^+$ 表示 $[z]^+ = \max\{z, 0\}$ 。

2 基于BFGS方法的拥塞速率控制算法

S. H. Low 采用式(6)的梯度投影方法调整链路价格, 利用 Lyapunov 稳定性理论可以证明其算法稳

定性, 但梯度投影方法收敛速度较慢, 然而 Internet 网络流量情况是瞬时变化, 实现全局的最优化速率控制, 必须提高拥塞控制算法的收敛速度, 因此, 必须使用具有更快收敛速度的二阶优化方法来求解链路价格。

在二阶优化方法中, Newton 法的优点是收敛速度快, 在这一点上, 梯度投影方法难以比拟。然而, Newton 法每次迭代都要计算目标函数的 Hessian 矩阵和它的逆矩阵, 当问题的维数 n 较大时, 计算量迅速增加。为此, 本文使用 BFGS 方法, BFGS 方法既保持了 Newton 法收敛速度快的优点, 又摆脱了 Hessian 矩阵的计算, 减少了计算量。

基于 BFGS 方法的拥塞速率控制算法描述如下:

已知目标函数 $L(x, p)$ 及其梯度 $\nabla L(x, p)$, 问题的维数 n , 终止限 ε 。

(1) 选取初始点 p_0 , 初始矩阵 $H_0 = I$, 给定终止限 $\varepsilon > 0$ 。

(2) 求初始梯度向量, 计算 $\nabla L(x, p_0)$, 若 $\|\nabla L(x, p_0)\| \leq \varepsilon$, 停止迭代, 输出 p_0 , 否则转(3)。

(3) 构造初始 BFGS 方向, 取 $u_0 = -H_0 \nabla L(x, p_0) = -\nabla L(x, p_0)$, 令 $k = 0$, 转(4)。

(4) 进行一维搜索, 由 $\frac{dL(x, p_k + t_k u_k)}{dt} = 0$ 求取

t_k , 令 $p_{k+1} = p_k + t_k u_k$, $s_k = p_{k+1} - p_k$, $y_k = \nabla L(x, p_{k+1}) - \nabla L(x, p_k)$, 转(5)。

(5) 求梯度向量计算 $\nabla L(x, p_{k+1})$, 若 $\|\nabla L(x, p_{k+1})\| \leq \varepsilon$, 停止迭代, 输出 p_{k+1} , 否则转(6)。

(6) 检验迭代次数, 若 $k+1 = n$, 令 $p_0 = p_n$ 转(3); 否则转(7)。

(7) 构造 BFGS 方向, 用 BFGS 公式 $H_{k+1} = H_k + \frac{1}{s_k^T y_k} \left[\left(1 + \frac{y_k^T H_k y_k}{s_k^T y_k} \right) s_k s_k^T - H_k y_k s_k^T - s_k y_k^T H_k \right]$ 计算 H_{k+1} , 取 $u_{k+1} = -H_{k+1} \nabla L(x, p_{k+1})$, 令 $k = k+1$, 转(4)。

在每个用户连接的发送方, 根据

$$x_r(k+1) = u_r'^{-1}(\sum p_l(k+1)) \quad (7)$$

来计算用户 r 的最优拥塞速率。

3 算法性能仿真

为了验证基于 BFGS 方法的拥塞速率控制算法的性能, 该算法在 NS2 网络仿真器^[8]中被实现, 并

与 REM(Random Exponential Marking)算法^[9]进行性能仿真对比,REM 算法是 S. H. Low 等人所提出 TCP/AQM 对偶性模型具体实现,该算法使用梯度投影方法来计算链路价格。仿真试验所使用的拓扑结构如图 1 所示。

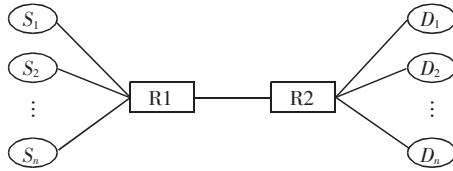


图 1 网络仿真实验拓扑结构

仿真所使用的网络结构中,在发送方 S_i 和接收方 D_i 之间存在数据流连接($i = 1, 2, \dots, 120$),分组长度为 1 024 字节。路由器 R1 与路由器 R2 之间的链路带宽为 100 Mbit/s,时延为 15 ms,缓冲区队列长度为 150 个分组。发送方与路由器 R1 之间的链路带宽为 40 Mbit/s,时延为 15 ms。接收方与路由器 R2 的链路带宽为 40 Mbit/s,时延为 15 ms。

仿真试验,REM 算法中, $\phi = 1.001, a = 0.1, \gamma = 0.001, b^* = 80$ 。基于 BFGS 方法的拥塞速率控制算法中, $\varepsilon = 0.001$ 。

在路由器 R1 与 R2 之间、发送方与路由器 R1 之间、接收方与路由器 R2 之间,分别配置 REM 算法以及本文所提出的基于 BFGS 方法的拥塞速率控制算法。

仿真试验 1 仿真时间 50 s,发送方和接收方之间为 120 个基于 TCP Reno 的 FTP 流。有效吞吐量(goodput)为全部数据流吞吐量与链路带宽之比。在路由器 R1 与 R2 之间,数据流在 2 种不同算法下,有效吞吐量随时间变化的过程如图 2 所示。仿真结果表明,使用基于 BFGS 方法的拥塞速率控制算法,数据流具有更高的有效吞吐量,更快的收敛速度,本文所提出的算法性能明显优于 REM 算法。

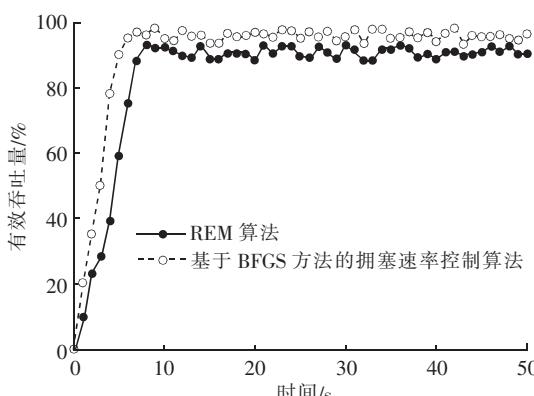


图 2 数据流有效吞吐量对比 1

仿真试验 2 仿真时间 50 s,仿真开始时有 60

个基于 TCP Reno 的 FTP 流,当 $t = 25$ s 时,另外 60 个基于 TCP Reno 的 FTP 流加入。在路由器 R1 与 R2 之间,数据流在 2 种不同算法下,有效吞吐量随时间变化的过程如图 3 所示。仿真结果表明,当数据流数目变化时,使用两种算法的数据流有效吞吐量都有所下降,但使用基于 BFGS 方法的拥塞速率控制算法,数据流的有效吞吐量高于 REM 算法数据流,具有更快的收敛速度,本文所提出的算法性能仍优于 REM 算法。

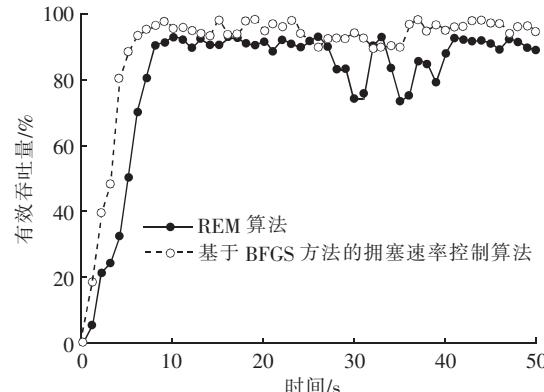


图 3 数据流有效吞吐量对比 2

4 结 论

网络流量情况是瞬时变化的,故速率优化的目标也是变化的,所以,拥塞控制算法的收敛速度非常重要。TCP/AQM 对偶性模型采用梯度投影方法调整链路价格,收敛速度较慢,本文使用具有更快收敛速度的 BFGS 方法来进行链路价格的计算,提出一种基于 BFGS 方法的拥塞速率控制算法,该算法具有更快的收敛速度,算法性能优于其它算法。

参考文献:

- [1] LOW S H, DAVID E. Lapsley, Optimization flow control I basic algorithm and convergence[J]. IEEE/ACM Trans on Networking, 1999, 7(6):861–874.
- [2] LOW S H. A duality model of TCP and queue management algorithms[J]. IEEE/ACM Trans on Networking, 2003, 11(4):525–536.
- [3] KELLY F P, MAULOO A, TAN D. Rate control for communication networks: Shadow prices, proportional fairness and stability[J]. Journal of Operations Research Society, 1998, 49(3):237–252.
- [4] BROYDEN C G. The convergence of a class of double rank minimization algorithms: the new algorithm[J]. Journal of the Institute of Mathematics and its Applications, 1970, 6:222–231.

(下转第 37 页)

基于无理数的DES加密算法

王 静¹, 蒋国平²

(1. 南京邮电大学 计算机学院, 江苏南京 210046)
(2. 南京邮电大学 自动化学院, 江苏南京 210046)

摘要:首先介绍数据加密标准(Data Encryption Standard, DES)和高级加密标准(Advanced Encryption Standard, AES),并对其安全性进行分析,然后提出基于无理数的DES加密方案。该方案利用无理数的伪混沌特性对密钥空间进行扩展,增加子密钥产生的随机性,使得每一组16次迭代所使用的子密钥各不相同,能够以和DES相同的时间开销,获得和AES相同的密钥空间。

关键词:数据加密标准;高级加密标准;无理数;伪混沌

中图分类号:TP309.7 文献标识码:A 文章编号:1673-5439(2009)06-0031-07

Irrational Numbers Based on DES Encryption Algorithm

WANG Jing¹, JIANG Guo-ping²

(1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)
(2. College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: DES and AES method for data encryption are discussed in this paper, and the analysis of security is also given, then an improved scheme based on irrational numbers is proposed. The encryption algorithm, which takes advantage of pseudo-chaos characteristics of irrational numbers, can expand key space, enhance the randomness of sub-keys and use different sub-keys in each group of 16 rounds. Therefore, the proposed scheme can obtain the same key space with AES algorithm by spending the same running time with DES algorithm.

Key words: DES; AES; irrational numbers; pseudo-chaos

0 引言

数据加密标准(DES)是美国国家标准局发布通告公开征求,由国际商业机器公司(IBM)推荐,并用于计算机数据保护的一种加密标准^[1-4]。它是使用最广泛的密钥系统,由于所需加密时间短,实现简单,所以DES算法在自动取款机(Automated Teller Machine, ATM)、磁卡及智能卡、加油站、高速公路收费站等领域被广泛应用,特别是在保护金融数据的安全中,自动取款机都使用DES。

随着计算机的发展,DES数据加密标准算法由于密钥长度较小(56 bit),已经不适应当今对数据加密安全性的要求。因此,1997年美国国家标准技术

研究院(National Institute of Standards and Technology, NIST)公开征集新的数据加密标准,即AES^[5-6]。此算法将成为美国新的数据加密标准而被广泛应用于各个领域中。尽管人们对AES还有不同的看法,但总体来说,AES作为新一代的数据加密标准汇聚了强安全性、高性能、高效率、易用和灵活等优点。AES设计有3个密钥长度:128, 192 和 256 bit。相对而言,AES的128密钥比DES的56密钥其安全性增强了1 021倍。当然,随着分组长度的增加,AES所用的加密时间也相应的比DES长。

本文提出基于无理数的DES算法,该算法是在原有DES算法的基础上增加了无理数的异或操作。无理数序列具有无限不循环特性,类似伪混沌序列。

收稿日期:2009-03-20;修回日期:2009-09-25

基金项目:国家教育部新世纪优秀人才支持计划(NCET-06-0510)、江苏省“六大人才高峰”高层次人才计划(SJ209006)、江苏省自然科学基础研究计划(08KJD510022)、南京邮电大学“青蓝计划”(NY207113)资助项目

通讯作者:蒋国平 电话:(025)83492256 E-mail:jianggp@njupt.edu.cn

64 bit 密钥在产生子密钥之前就先进行与无理数的异或处理,使得密钥本身并不直接参与子密钥的产生,并且异或处理在大多数 CPU 上都可以高速执行,即:利用无理数的伪混沌特性对密钥空间进行扩展,增加子密钥产生的随机性,并且呈现伪随机变化,在几乎不增加时间开销的基础上扩展其密钥空间,密钥空间由原来的 2^{64} ,扩展成为 2^{128} ,增强信息的保密性,达到和 AES 相同的密钥长度,可以很好的避免穷举法攻击。

1 DES 和 AES 算法

(1) DES 算法实现过程中,16 次加密变换为最重要的部分。64 bit 的密钥经过固定的置换和移位产生 16 个子密钥,并且每个分组使用相同的子密钥进行加密,具体情况如图 1 所示。

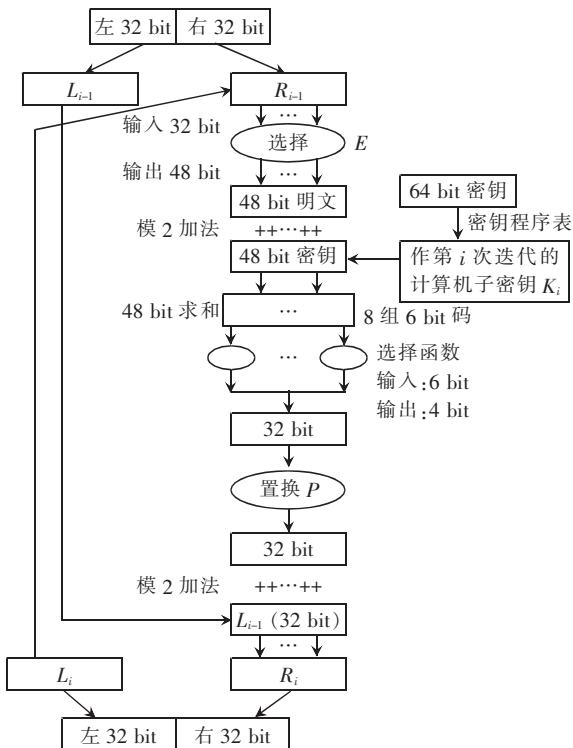


图 1 16 次加密变换过程的第一次迭代

参见图 1, 其中第 i 次迭代为 $L_i = R_{i-1}$ 和 $R_i = L_{i-1} \oplus f(R_{i-1}, K_i)$, 并且 $f(R_{i-1}, K_i) = P(S(E(R_{i-1}) \oplus K_i))$ 。 E 是一个固定的扩展置换, 将 R_{i-1} 从 32 bit 映射成 48 bit(有些位映射一次, 有些位映射两次)。 P 是另一个在 32 bit 上的固定置换, 解密涉及相同的密钥和算法, 但内部每一轮使用的子密钥的顺序正好相反。

子密钥的产生过程如图 2 所示。

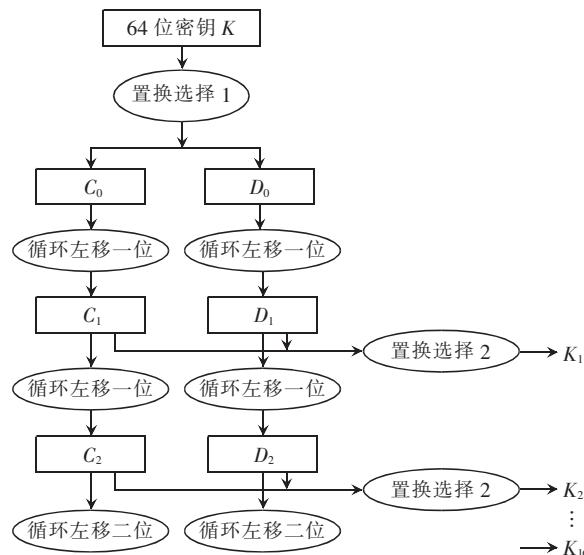


图 2 子密钥的产生过程

(2) AES 算法的加密和解密流程如图 3 所示。

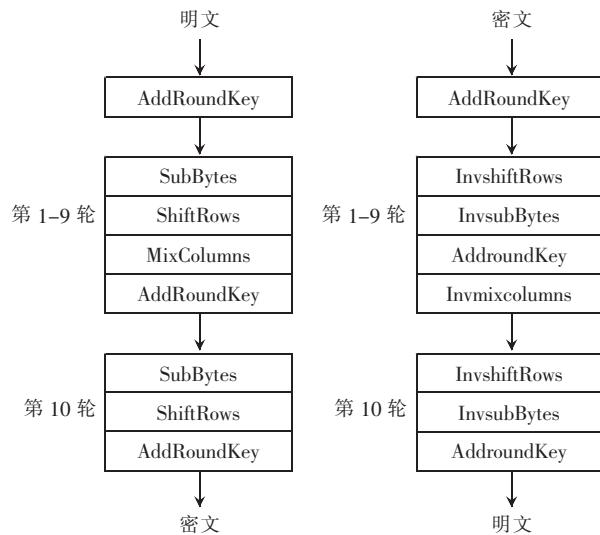


图 3 AES 算法的加密和解密原理图

加密算法的输入分组和解密算法的输出分组均为 128 位。输入分组用一个以字节为单位的 4×4 的矩阵来描述。

字节代换变换(SubBytes): AES 定义了一个 S 盒(SBox), 字节代换的正向变换就是一个在 SBox 中的查表操作。字节代换的逆向变换相应地在一个逆 S 盒(InvSBox)中进行查表操作。

行移位变换(ShiftRows): 行移位的操作比较简单, 在正向行移位变换中, State 的第 1 行保持不变, 第 2 行、第 3 行和第 4 行分别循环左移 1 个字节、2 个字节和 3 个字节。逆向行移位变换执行相反方向的移位操作。

列混淆变换(MixColumns): 列混淆变换较为复杂, 正向列混淆变换对每列独立地进行操作, 每列中的每个字节被映射为一个新值, 这个变换基于 State

的矩阵乘法。逆向列混淆变换(InvMixColumns)正矩阵变换的逆变换。

轮密钥加变换(AddRoundKey):轮密钥加变换的正向和逆向变换相同,都是用 128 位的 State 按位与 128 位的密钥异或。

密钥扩展算法(KeyExpansion):AES 密钥扩展算法输入 4 字(16 字节)密钥,输出值是一个 44 字(156 字节)的一个线性数组,每 4 个字组成一组用于算法中初始轮密钥加变换和另外 10 轮中每轮的密钥加操作。

在理论和实践基础上,AES 被认为是“安全的”,因为要破解它的话,唯一有效的方法是强行生成所有可能的密钥。

2 基于无理数的 DES 算法

2.1 DES 算法分析

一次一密密码体制的无条件安全性,是在 Shannon 提出完善保密的理论之后,才得到严格的数学证明的^[7]。Shannon 指出,仅当密钥至少和明文一样长时,才能达到无条件安全。也就是说除了一次一密方案外,再无其他加密方案是无条件安全的。一个密码体制(M,C,K,E,D)叫做完善保密的,如果对所有的明文 $m \in M$,密文 $c \in C$,满足条件概率 $prob(m/c) = prob(m)$ 。(M,C,K,E,D)分别对应明文空间、密文空间、密钥空间、加密算法和解密算法。那么对于 DES 算法,密钥空间为 2^{64} ,当明文也达到 2^{64} 时,密钥就会重复使用,因此,应该尽量的扩大密钥空间,才能增加信息的安全性^[8-11]。

对于 DES 密码, $K = 2^{56} \approx 7 \times 10^{16}$,即使使用每秒种可以计算 100 万个密钥的大型计算机,也需要算 10^6 天才能求得所使用的密钥,这意味着如果一台计算机的速度是每一秒钟检测一百万个密钥,则它搜索完全部密钥就需要将近 2^{285} 年的时间,因此看来是很安全的。但是,随着科学技术的发展,1998 年在由美国 RSA 公司发起的一项竞赛中,数万名志愿者使用普通 PC 机,利用业余时间,在 59 天后破译了 56 位的 DES 密钥。

Diffie 和 Hellman 指出,如果设计一种 $1 \mu\text{s}$ 可以核算一个密钥的超大规模集成片,那么它在一天内可以核算 8.64×10^{10} 个密钥。如果由一个百万个这样的集成片构成专用机,那么它可以在不到一天的时间内用穷举法破译 DES 密码。他们曾于 1977 年估计:这种专用机的造价约为 2 000 万美元。

目前最强大的超级计算机是 IBM 的蓝色基因,IBM 耗费了 5 年的时间以及 1 亿美元的成本开发出了 Blue Gene 超级计算机系统,它拥有 65 536 个双核心处理器,峰值运算性能达到 367 TFlops。即 367×10^{12} 的计算能力,能以接近每秒千万亿次运算的速度连续运行,该系统的计算能力超过家用电脑的 10 万倍,售价 150 万美元。那么它可以在 3 分钟时间内用穷举法破译 DES 密码。因此,当出现超高速计算机后,人们可考虑把 DES 密钥的长度再增长一些,以此来达到更高的保密程度。因此出现了 3DES 算法。

3DES 算法,无论从理论上还是实践上,都是一种比较安全的加密算法。虽然利用穷举攻击,该算法最终可以破译,但需要 $2^{112} \approx 5 \times 10^{33}$ 次穷举,使用当今最强大的超级计算机,也需要 1.5×10^{14} 天,因此,所花费的代价实在太大,可以说得不偿失。相对来说,3DES 算法是一种比较安全的加密算法,但是时间开销将是 DES 算法的 3 倍,本文提出的基于无理数 DES 算法,将以 DES 的时间开销达到和 3DES 相同的安全性。

2.2 基于无理数 DES 算法介绍

混沌系统是一种高度复杂的非线性系统,同时系统的输出具有不可预测性及区间上的遍历性。这些特性决定了混沌非常适合于信息的安全保密。由于混沌保密通信具有实时性强、保密性高等优点,因此在保密通信领域显示出了强大的生命力。因为对初始值的极端敏感,所以混沌保密通信一般需要很严格的同步系统,这限制了它在保密通信领域的应用。将混沌转变成数字序列后,我们发现该序列类似于无理数,是一种非周期无限长的数字序列^[12-15]。因此,利用具有伪混沌特性的无理数进行数据加密将不会面临要求严格同步的问题。

基于无理数的 DES 算法仍是一种分组密码,将明文序列划分成等长的分组,对每一组用相同的加密算法和不同的密钥进行加密。分组密码有其自身的特点,首先分组密码容易被标准化,因为现代的通信通常是分成块的处理和传输的;其次分组密码很容易实现同步,因为一个分组的传输错误不会影响到其它的分组,同时丢失一个分组也不会对随后分组的正确解密产生影响,即传输错误不会扩散。对于分组加密来说,对密钥长度的各种要求是相互矛盾的。安全性要求长的密钥,特别是多轮迭代分组密码需要很长的密钥。而简洁性和密钥本身的保密性要求密钥尽可能的短。

图4中, K_1, K_2, \dots, K_{16} 为16轮迭代所需要的子密钥, 置换选择1和置换选择2为两个固定的置换。在密钥程序表产生子密钥的过程中, 64位密钥在产生子密钥之前就先进行了无理数控制的随机异或处理再进行固定的置换选择1。参加每轮迭代的密钥都呈现伪随机变化, 从而增加了密钥的保密性, 保护了原密钥。控制异或所使用的随机数 b 乃由无理数发生器产生, 收发双方有相应的同步系统进行同步。因为该算法是分组密码, 一个分组的传输错误不会影响到其它的分组, 即传输错误不会扩散, 所以很容易实现同步。

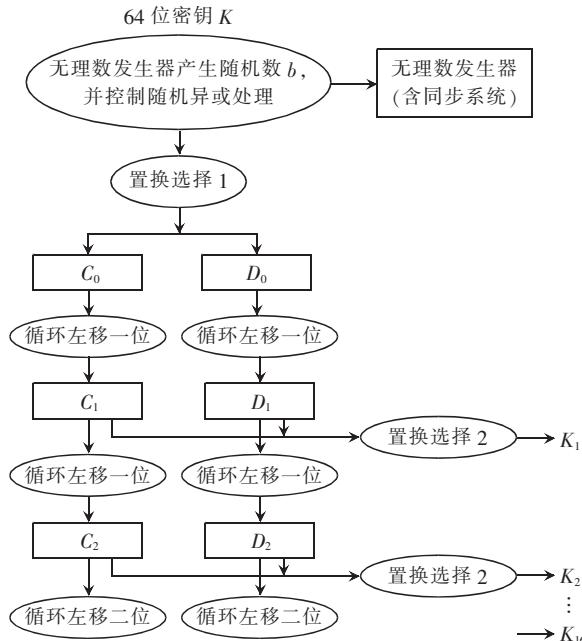


图4 基于无理数DES算法的子密钥产生原理

由图4可见, 64位密钥在产生子密钥之前就先进行了无理数控制的异或处理, 使得密钥本身并不直接参加子密钥的产生。该算法利用无理数的伪混沌特性对密钥空间进行了扩展, 增加了子密钥产生的随机性, 使得每一组的16次迭代所使用的子密钥都不相同, 并且呈现伪随机变化。在几乎不增加时间开销的基础上扩展了其密钥空间, 若无理数发生器产生的伪随机数为64位, 则密钥空间为 2^{128} , 增强了数据的保密性, 使其可以避免穷举法攻击。

举个例子说明, DES算法的密钥长度为64位, 假设原密钥为

$K = 11111111\ 11111111\ 11111111\ 11111111$

$11111111\ 11111111\ 11111111\ 11111111$,

那么现在假设产生一个随机数 b , 随机数 b 的第一位有两种取值0或者1, 则

$b = 01111111\ 11111111\ 11111111\ 11111111$

或者为 $11111111\ 11111111\ 11111111\ 11111111$, $11111111\ 11111111\ 11111111\ 11111111$, 那么将 K 和 b 进行异或, 相应得出 K_1 和 K_2 , 变成2个密钥; 再假设 b 的第二位也有两种取值, 0或者1, 那么将 K 和 b 进行异或, 相应可以得出 $K_1K_2K_3K_4$, 产生4个密钥, 即 2^2 个 K ……以此类推, 有64位的 b 就有 2^{64} 个 K , 而 K 本身的密钥空间就是 2^{64} , 因此, 在此基础上扩大了 2^{64} 倍, 就变成 $2^{64} \times 2^{64} = 2^{128}$ 。

基于无理数的DES算法是在DES算法基础上增加无理数的异或操作, 其目的有3个:

首先, 无理数的异或操作在大多数CPU上都可以高速执行, 所以该算法和DES算法的运行时间也是基本相同的。

其次, 能够增加子密钥产生的随机性, 使得每一组的16次迭代所使用的子密钥都不相同, 即: 能够扩展密钥空间, 若无理数发生器产生的伪随机数为64位, 则密钥空间为 2^{128} 。

最后, 基于无理数的DES算法仍是分组加密算法, 一个分组的传输错误不会影响到其它的分组, 即: 传输错误不会扩散; 另外, 具有伪混沌特性的无理数进行数据加密时不需要很严格的同步系统。因此, 基于无理数的DES算法容易实现同步。

3 仿真结果

以下所有的仿真都是在相同的仿真环境MATLAB7.1中完成的, 每种算法都做了多次仿真, 由于篇幅受限, 只从中截取了一部分数据。

(1) 表1是DES算法和基于无理数DES算法在相同的明文和密钥的条件下仿真的结果。其中, DES算法的密钥空间为 2^{64} , 而基于无理数的DES算法的密钥空间为 2^{128} 。

表1 DES与基于无理数DES算法加密和解密时间的比较

时间/s

DES 加密时间	改进 DES 加密时间	DES 解密时间	改进 DES 解密时间
0.667 936	0.693 657	0.702 563	0.682 190
0.662 521	0.694 369	0.687 896	0.687 268
0.683 996	0.703 038	0.666 201	0.693 584
0.671 693	0.699 149	0.708 764	0.703 066
0.660 216	0.703 030	0.676 635	0.687 979
0.667 709	0.699 707	0.690 962	0.705 175
0.666 820	0.698 257	0.687 276	0.693 584
0.668 871	0.699 144	0.693 611	0.698 993
0.666 329	0.699 229	0.669 503	0.699 294

为了更加直观的观察 DES 算法与基于无理数的 DES 算法的所用加密和解密时间的区别,根据表 1 中的数据,做仿真图 5 和图 6。由图 5、图 6 可知,DES 算法和基于无理数 DES 算法的加密和解密的时间几乎相同,但是 DES 算法的密钥空间为 2^{64} ,而基于无理数的 DES 算法的密钥空间扩展为 2^{128} 。

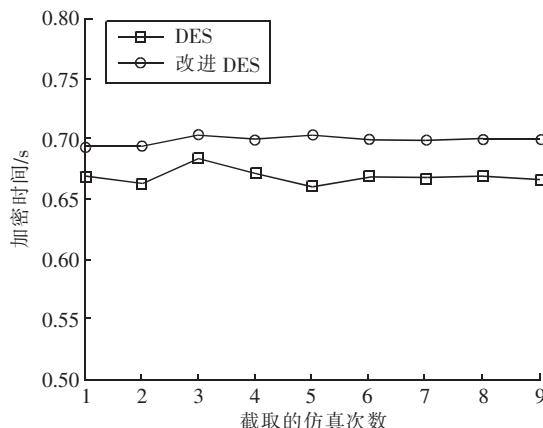


图 5 DES 算法与基于无理数 DES 算法加密时间的比较

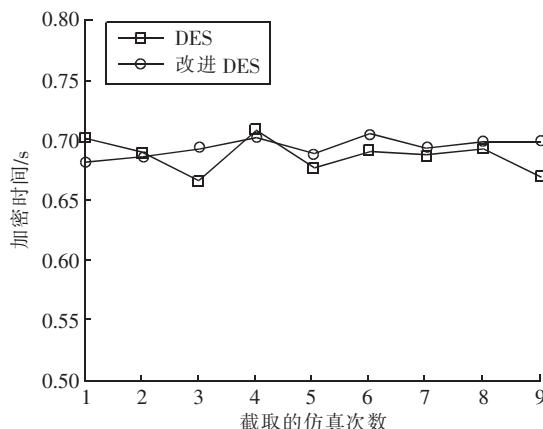


图 6 DES 算法与基于无理数 DES 算法解密时间的比较

参见图 5 和图 6, 基于无理数的 DES 算法时间开销比 DES 算法的时间开销稍微大 0.03 s, 但是密钥空间扩大了 2^{64} 倍, 所以用很少的时间开销获得了较大的密钥空间。

表 2 介绍 DES 算法和基于无理数 DES 算法在相同明文和密钥下的密文比较。由表 2 可以看出:

① 基于无理数异或的 DES 算法可以用很少的时间开销, 获得比 DES 算法大 2^{64} 倍的密钥空间, 达到和 AES 相同的密钥长度。

② 在相同的明文和密钥前提下, 经多次仿真后 DES 算法产生的密文是完全相同的, 而对于基于无理数的 DES 算法, 密文则是随机的, 即通过增加子密钥产生的随机性增加了密文的随机性。

表 2 DES 和基于无理数 DES 算法在相同明文和密钥下的密文比较

DES 密文	改进 DES 密文	随机数 b
0F67F40FDAD62857	BD2442BA9704E0AC	7950288419716939
0F67F40FDAD62857	EEEFA82E3022A20F	4338327950288419
0F67F40FDAD62857	1C04F139FB996D47	7169399375105820
0F67F40FDAD62857	06B6A0CE02C4354B	3589793238462643
0F67F40FDAD62857	D3B148544BB40CED	6433832795028841
0F67F40FDAD62857	2C0796602D064617	3832795028841971
0F67F40FDAD62857	45EA457B1494B38E	9502884197169399
0F67F40FDAD62857	1C509A157C5452BB	7932384626433832
0F67F40FDAD62857	C0690DFE3D7F45A9	2643383279502884

其中, 明文为 1234567891234567, 密钥为 1234567891234567。

(2) 表 3 为高级加密标准 AES 和数据加密标准 DES 在加密 128 位数字序列时的时间开销。

表 3 AES 和 DES 在加密 128 位数字序列时的时间比较

AES 加密时间	DES 加密时间	AES 解密时间	DES 解密时间	时间/s
3.740 846	1.335 876	3.707 661	1.455 126	
3.726 647	1.365 942	3.714 293	1.375 992	
3.720 749	1.367 884	3.694 422	1.334 402	
3.719 963	1.330 032	3.705 473	1.417 128	
3.721 440	1.363 516	3.698 905	1.343 571	
3.724 454	1.333 614	3.702 909	1.381 724	
3.730 541	1.337 645	3.700 726	1.375 344	
3.758 997	1.332 656	3.703 571	1.387 013	
3.737 737	1.357 322	3.699 166	1.339 606	

可见, 数据加密标准 DES 在加密 128 位二进制数字序列时所消耗的时间要比 AES 少, 但是其密钥空间为 2^{64} , 远小于 AES, 从而抵挡不了穷举法攻击。

从仿真图 7 和图 8 中, 可以更加直观的看出, DES 算法在加密和解密 128 位数字序列时的时间开销要比 AES 算法少 2.5 s, 同时密钥空间小 2^{64} 倍。

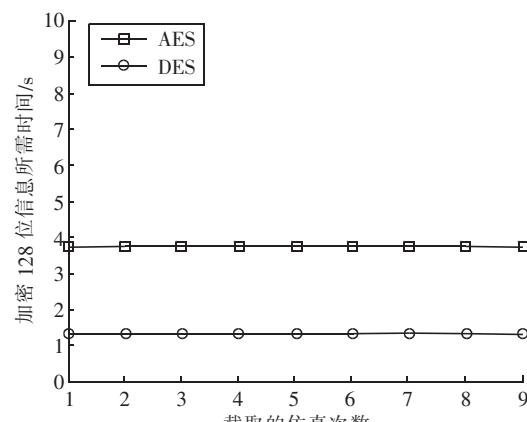


图 7 AES 和 DES 在加密 128 位数字序列时的时间比较

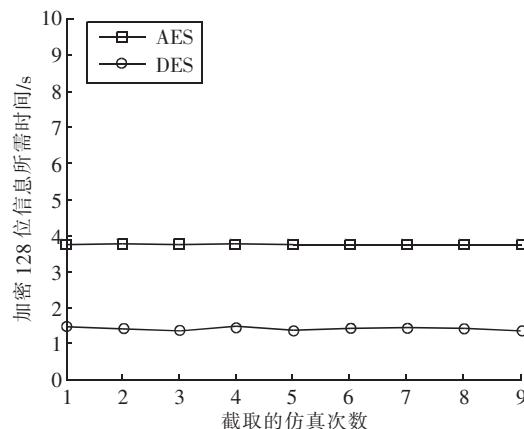


图 8 AES 和 DES 在解密 128 位数字序列时的时间比较

(3) 表 4 为高级加密标准 AES 和基于无理数 DES 算法在相同密钥空间 2^{128} 下的加密和解密时间的比较。

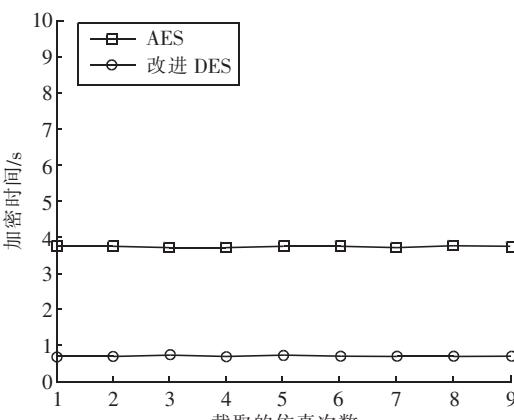
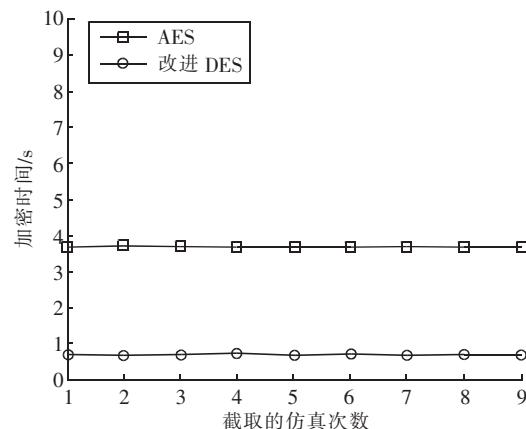
表 4 AES 和基于无理数 DES 算法比较

AES 加密时间	改进 DES 加密时间	AES 解密时间	改进 DES 解密时间	时间/s
3.740 846	0.693 657	3.707 661	0.682 190	
3.726 647	0.694 369	3.714 293	0.687 268	
3.720 749	0.703 038	3.694 422	0.693 584	
3.719 963	0.699 149	3.705 473	0.703 066	
3.721 440	0.703 030	3.698 905	0.687 979	
3.724 454	0.699 707	3.702 909	0.705 175	
3.730 541	0.698 257	3.700 726	0.693 584	
3.758 997	0.699 144	3.703 571	0.698 993	
3.737 737	0.699 229	3.699 166	0.699 294	

其中,对于 AES 加密标准有

```
Plaintext:00 04 08 02      key:00 04 08 02      Cipher:98 dc a7 5d
    01 05 09 03              01 05 09 03          fe e3 56 15
    02 06 00 04              02 06 00 04          e6 f7 e3 7c
    03 07 01 05              03 07 01 05          d8 49 bd c4
```

从仿真图 9 和图 10 中,可以更加直观的看出,基于无理数 DES 算法加密和解密的时间开销要比 AES 算法少 3 s,但是密钥空间却和 AES 算法相同,可以将密钥空间扩展为 2^{128} 。

图 9 AES 算法和基于无理数 DES 算法在相同密钥空间 2^{128} 下的加密时间比较图 10 AES 算法和基于无理数 DES 算法在相同密钥空间 2^{128} 下的解密时间比较

4 结束语

DES 算法是一种常用的加密方法,但因密钥空间较小,使其应用受到很大限制。本文提出了一种基于无理数的 DES 加密算法,该算法大大扩展了密钥空间,以 DES 算法所用时间达到 AES 的密钥空间,即为 2^{128} ,显著提高了数据加密的速率和安全性,并且无理数的同步也易于实现。基于这些优点,相信该算法将有着广泛的应用前景。

参考文献:

- [1] 杨波.现代密码学[M].北京:清华大学出版社,2003.
YANG Bo. Modern cryptography[M]. Beijing: QingHua University Press, 2003.
- [2] 马银华,刘明生.提高 DES 加密强度的密钥选位方法研究[J].计算机工程,2000,26(3):68-69.
MA Yinhua, LIU Mingsheng. The study of cryptographic key permutation choose for enhancing DES Encryption Intensity[J]. Computer Engineering, 2000, 26(3):68-69.
- [3] 左涛,蒋国平.基于混沌序列的数据加密标准 DES 算法研究[J].信息安全与保密通信,2004(2):33-34.
ZUO Tao, JIANG Guoping. Research on chaotic-sequence-based DES algorithm[J]. China Information Security, 2004(2):33-34.
- [4] 顾超.动态 DES 算法[J].计算机应用与软件,2007,24(7):164-166.
GU Chao. Dynamic DES algorithm[J]. Computer Applications and Software, 2007, 24(7):164-166.
- [5] 王先培.新一代数据加密标准——AES[J].计算机工程,2003,29(3):69-70.
WANG Xianpei. A new advanced encryption standard—AES[J]. Computer Engineering, 2003, 29(3):69-70.
- [6] PARIKH C, PATEL P. Performance evaluation of AES algorithm on various development platforms[C]//IEEE International Symposium

- on Consumer Electronics. Irving, TX, June 2007;1 – 6.
- [7] SHANNON C E. Communication theory of secrecy system[J]. Bell Systems Technical Journal, 1949, 28(4):656 – 715.
- [8] LIN Z. A study and analysis on a high intensity public data encryption algorithm[C]// Proceedings of the 3d World Congress on Intelligent Control and Automation. Hefei China, 2000;2492 – 2494.
- [9] 王立胜. 数据加密标准 DES 分析及其攻击研究[J]. 计算机工程, 2003, 29(13):130 – 132.
- WANG Lisheng. Analysis and attack research of DES cipher[J]. Computer Engineering, 2003, 29(13):130 – 132.
- [10] SPILLMAN R. Classical and contemporary cryptology[M]. Upper Saddle River, NJ: Pearson Education, 2005.
- [11] 冯登国. 国内外密码学研究现状及发展趋势[J]. 通信学报, 2002, 23(5):18 – 26.
- FENG Dengguo. Status quo and trend of cryptography[J]. Journal of China Institute of Communications, 2002, 23(5):18 – 26.
- [12] WANG J, JIANG G P, YANG H. Improved DES algorithm based on irrational numbers[C]// IEEE International Conference Neural Networks and Signal Processing. 2008;629 – 631.
- [13] YANG H, JIANG G P. Irrational-based time-hopping modulation for UWB[J]. IEEE Trans on Communications Circuits and Systems II: Express Briefs, 2008, 55(4):364 – 368.
- [14] 罗启彬, 张健. 一种新的混沌伪随机序列生成方式[J]. 电子与信息学报, 2006, 28(7):1262 – 1265.

(上接第 30 页)

- [5] FLETCHER R. A new approach to variable metric algorithms[J]. Computer Journal, 1970, 13:317 – 322.
- [6] GOLDFARB D. A family of variable-metric methods derived by variational means[J]. Mathematics of Computation, 1970, 24:23 – 26.
- [7] SHANNO D F. Conditioning of quasi-Newton methods for function minimization[J]. Mathematics of Computation, 1970, 24: 647 – 650.
- [8] Ns-2. Network Simulator [EB/OL]. <http://www.isi.edu/nsnam/ns>.
- [9] SATHURALIY A, VH L, SH L, et al. REM: Active Queue Management[J]. IEEE Network, 2001, 15(3):48 – 53.

LUO Q B, ZHANG J. A new approach to generate chaotic pseudorandom sequence[J]. Journal of Electronics & Information Technology, 2006, 28(7):1262 – 1265.

- [15] XIANG T, LIAO X F. A novel block cryptosystem based on iterating a chaotic map[J]. Physics Letters A, 2006, 349(2):109 – 115.

作者简介:



王 静(1982 –), 女, 山东临沂人。南京邮电大学计算机学院博士研究生。主要研究方向为复杂动态网络与信息安全技术。

蒋国平(1966 –), 男, 江苏扬中人。南京邮电大学研究生部主任兼学科建设办公室主任, 教授, 博士生导师。(见本刊 2009 年第 2 期第 68 页)

作者简介:



魏 涛(1974 –), 男, 山东泰安人。南京邮电大学信息网络技术研究所博士研究生, 解放军理工大学工程兵工程学院讲师。主要研究方向为: 计算机网络服务质量、拥塞控制、网络建模与优化。

张顺颐(1944 –), 男, 江苏南京人。南京邮电大学信息网络技术研究所教授, 博士生导师。(见本刊 2009 年第 1 期第 5 页)

三维 MIMO 信道物理模型的统计特征

海 淩, 张业荣

(南京邮电大学 电子科学与工程学院, 江苏南京 210046)

摘要: 提出以散射矩阵来概括并简化电波传播过程中散射体对电磁波的影响, 并对传统的距离分集和极化分集多入多出(MIMO)信道模型进行扩展, 建立了一种可以应用于各种分集情况的三维通用MIMO物理信道模型。再以此模型为基础, 对三维环境中的MIMO信道统计特征进行了研究, 推导出任意情况下交叉极化鉴别度(XPD)和子信道间相关性的计算方法, 并进行了仿真验证。仿真结果表明从文中所提出信道模型提取得到的统计特征与理论分析的结果是相符的。

关键词: MIMO; 混合分集; 物理信道模型; XPD; 相关性

中图分类号: TN011 文献标识码: A 文章编号: 1673-5439(2009)06-0038-05

Statistical Characteristics of 3-D Physical MIMO Channel Model

HAI Lin, ZHANG Ye-rong

(College of Electronic Science and Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: A scattering matrix is proposed to describe and simplify the influence of scatters during propagation. Then by extending traditional physical multi-input multi-output (MIMO) channel models of space-diversity and polarization-diversity, and with the help of the scattering matrix mentioned above, we established a 3-D common MIMO physical channel model, which is applicable for MIMO systems of any antenna diversity, and based on which calculating method of statistical characteristics of 3-D MIMO channels, such as correlation coefficient and cross-polar discriminations (XPD), are discussed and deduced. Simulation results show that the statistical characteristics obtained from our new model matches the theoretical results well.

Key words: MIMO; mix-diversity; physical MIMO channel model; XPD; correlation

0 引言

MIMO(Multiple-Input Multiple-Output)技术通过多天线形成的多个并行子信道进行空分复用, 能够在不增加系统带宽和天线发送总功率的情况下, 有效对抗无线信道衰落的影响, 大大提高系统的频谱利用率和信道容量^[1-3]。基于这些优点, MIMO技术已经在许多无线通信系统中得到了广泛的应用。

MIMO无线系统的信道容量取决于子信道之间的相关特性, 低的相关特性将最大限度地发挥多天线的传输潜力, 提高信道容量。以往的研究中, MIMO系统多采用距离分集, 为了获得较低的子信道相关

性, 天线间距要求有数个波长。这样就使得系统尺寸变大, 在常规移动载体和便携式设备中很难有足够的空间安放这种天线系统, 此时极化分集就可以显示出其优越性: 正交极化的线极化天线即使共点安装也可形成低相关性的子信道, 因此适合体积受限的情况; 而且在提高信道容量方面, 极化分集可以达到和距离分集相同的效果^[4-5]。然而由于平行极化波的地面损耗较为严重, 极化分集在提升系统性能方面也具有一定的局限性。为了同时研究使用距离分集和极化分集的系统性能, 本文旨在建立同时应用了这两种分集方式的通用信道模型, 并将单纯的距离分集或极化分集看作此模型的特殊情况。

当前世界上对极化分集的研究主要集中在二维分集的情况。虽然三维极化分集的有效性已经被证明^[6-7],但对三维 MIMO 信道模型的研究尚不太多,其主要原因之一就是难以对此类模型的统计特性进行研究,而 MIMO 信道的统计特征在描述信道特性、建立统计随机信道模型等方面是举足轻重的。研究三维物理信道特性一般需要对电磁散射过程进行繁琐的计算,而且由于去极化的复杂影响,研究此类模型的统计特征在理论上有非常大的困难。基于这个原因,本文提出以散射矩阵来概括散射体对电磁波的影响,并依此建立三维 MIMO 物理信道模型,使建模所需的计算更加简单,并且能够较方便地对其统计特性进行研究,而散射矩阵的参数更可以依照具体环境进行设计,能够灵活满足不同环境下的要求。

1 三维混合分集物理信道模型

考虑一个由 n_r 副发送天线、 n_t 副接收天线组成的 MIMO 系统模型。第 m 副发送天线到第 n 副接收天线之间的信道衰落系数用 h_{mn} 表示,则信道衰落系数矩阵可以表示为

$$\mathbf{H}(t) = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n_t} \\ h_{21} & h_{22} & \cdots & h_{2n_t} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n_r 1} & h_{n_r 2} & \cdots & h_{n_r n_t} \end{pmatrix} \in C^{n_r \times n_t} \quad (1)$$

建立物理信道模型,即通过设计天线与散射体的相关参数来模拟实际环境,再经由相关计算以获得信道衰落系数矩阵。为了研究信号经由散射体后幅度和极化方向的变化^[8-9],使用了一个散射公式,但这个公式的计算比较繁琐,并且利用此方法建立的信道,其统计特征的研究较为困难。因此引入一个散射矩阵,用它来描述散射体对信号带来的影响:

$$\mathbf{A}_l = \begin{pmatrix} A_{xx}^{(l)} & A_{xy}^{(l)} & A_{xz}^{(l)} \\ A_{yx}^{(l)} & A_{yy}^{(l)} & A_{yz}^{(l)} \\ A_{zx}^{(l)} & A_{zy}^{(l)} & A_{zz}^{(l)} \end{pmatrix} \quad (2)$$

根据电磁散射理论,发送信号在散射体上发生散射,产生一个新的辐射场,如图 1 所示。

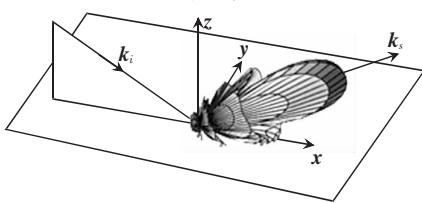


图 1 电磁散射示意图

而散射后到达接收端的信号仅仅是这个辐射场中朝向接收端的那一部分,矩阵 \mathbf{A}_l 所描述的就是到达接收端的信号相对于到发送信号的极化方向和功率的变化。所以通过适当地设计 \mathbf{A}_l 的各个元素,可以对各种类型的散射体进行模拟,灵活地对任何环境下散射体带来的影响进行仿真分析,而且计算较为简单。

考虑 M 个发送天线和 N 个接收天线组成的系统,可以是利用不同天线位置和不同极化方向形成的 MIMO 混合分集。设第 m 个发送天线上发送信号为 $x_m(t)$,极化方向为 $(\alpha_{mT}, \beta_{mT})$,天线方向图为 G_{mT} 。为了方便分析,将 $x_m(t)$ 沿坐标轴分解为 3 个方向的分量:

$$\begin{pmatrix} T_{mx} \\ T_{my} \\ T_{mz} \end{pmatrix} = \begin{pmatrix} \sin\alpha_{mT}\cos\beta_{mT} \\ \sin\alpha_{mT}\sin\beta_{mT} \\ \cos\alpha_{mT} \end{pmatrix} x_m(t) \quad (3)$$

假设收发端之间存在着 L 个散射体。第 m 个发送天线到达第 l 个散射体时有相移 $\frac{2\pi|\vec{r}_{ml}|}{\lambda}$,而这段路径的传播损耗为 g_{ml} ,那么经由第 l 个散射体后向接收端传播的那部分信号在 3 个方向的分量为

$$\begin{pmatrix} R_{mx}^{(l)} \\ R_{my}^{(l)} \\ R_{mz}^{(l)} \end{pmatrix} = G_{mT}(\vec{r}_{ml}) g_{ml} \mathbf{A}_l \exp\left(-j2\pi\frac{|\vec{r}_{ml}|}{\lambda}\right) \begin{pmatrix} T_{mx} \\ T_{my} \\ T_{mz} \end{pmatrix} \quad (4)$$

第 l 个散射体到达第 n 个接收天线时有相移 $\frac{2\pi|\vec{r}_{nl}|}{\lambda}$,这段路径的传播损耗为 g_{nl} 。如果第 n 个接收天线的极化方向为 $(\alpha_{nR}, \beta_{nR})$,方向图为 G_{nR} ,那么第 n 个接收天线上接收到的信号为

$$y_n(t) = \sum_{l=1}^L G_{nR}(\vec{r}_{nl}) g_{nl} \exp\left(-j2\pi\frac{|\vec{r}_{nl}|}{\lambda}\right) \begin{pmatrix} \sin\alpha_{nR}\cos\beta_{nR} \\ \sin\alpha_{nR}\sin\beta_{nR} \\ \cos\alpha_{nR} \end{pmatrix} \begin{pmatrix} R_{mx}^{(l)} \\ R_{my}^{(l)} \\ R_{mz}^{(l)} \end{pmatrix} \quad (5)$$

$$\text{令 } \begin{pmatrix} r_1^{(n)} \\ r_2^{(n)} \\ r_3^{(n)} \end{pmatrix} = \begin{pmatrix} \sin\alpha_{rn}\cos\beta_{rn} \\ \sin\alpha_{rn}\sin\beta_{rn} \\ \cos\alpha_{rn} \end{pmatrix}, \begin{pmatrix} t_1^{(m)} \\ t_2^{(m)} \\ t_3^{(m)} \end{pmatrix} = \begin{pmatrix} \sin\alpha_{lm}\cos\beta_{lm} \\ \sin\alpha_{lm}\sin\beta_{lm} \\ \cos\alpha_{lm} \end{pmatrix},$$

$A_{ij}^{(l)}$ 代表矩阵 \mathbf{A}_l 的各个元素,如果发送天线间距与接收天线间距相对收发端距离可以忽略,那么路径损耗可近似为 $g_{ml}g_{nl}$,则第 m 个发送天线到第 n 个接收天线之间的信道系数为

$$h_{mn} = \sum_{l=1}^L G_{mT}(\vec{r}_{ml}) G_{nR}(\vec{r}_{nl}) g_{l} \cdot$$

$$\exp\left[\frac{-j2\pi(|\vec{r}_{nl}|+|\vec{r}_{ml}|)}{\lambda}\right] \sum_{i=1,j=1}^3 r_i^{(n)} t_j^{(m)} A_{ij}^{(l)} \quad (6)$$

至此通过几何方法为三维情况下的信道向量建立了模型。通过此物理信道模型求解信道衰落系数矩阵,计算量仅为^[8-9]所用方法的一半左右。在下一部分将利用这个模型对统计模型所需的参数进行探讨。

$$\rho_{mn,pq} = \frac{E(h_{mn} h_{pq}^*) - E(h_{mn})E(h_{pq}^*)}{\sqrt{E(h_{mn} h_{mn}^*) - E(h_{mn})E(h_{mn}^*)} \sqrt{E(h_{pq} h_{pq}^*) - E(h_{pq})E(h_{pq}^*)}} \quad (7)$$

一般假设各条子信道的冲击响应均满足均值为零的高斯分布,则有

$$\rho_{mn,pq} = \frac{E(h_{mn} h_{pq}^*)}{\sqrt{E(h_{mn} h_{mn}^*)} \sqrt{E(h_{pq} h_{pq}^*)}} \quad (8)$$

传统的一维信道模型,即仅利用距离分集来降低相关性,而天线均采用相同极化方向(一般都考虑垂直极化)这一情况下的信道模型。此时子信道向量的计算公式为^[11]:

$$h_{mn} = \sum_{l=1}^L g_l \exp\left[\frac{-j2\pi(|\vec{r}_{nl}|+|\vec{r}_{ml}|)}{\lambda}\right] \quad (9)$$

将式(9)代入(8),并假设不同散射体之间相互独立,则有

$$E(h_{mn} h_{pq}^*) = \sum_{l=1}^L E(g_l^2) \cdot \exp\left[\frac{-j2\pi(|\vec{r}_{nl}|+|\vec{r}_{ml}|+|\vec{r}_{ql}|+|\vec{r}_{pl}|)}{\lambda}\right] \quad (10)$$

一般假设 g_l 为独立同分布的高斯变量,则 $E(g_l^2)$ 为定值。将式(10)代入式(8),可得到

$$\rho_{mn,pq} = \sum_{l=1}^L \exp\left[\frac{-j2\pi(|\vec{r}_{nl}|+|\vec{r}_{ml}|+|\vec{r}_{ql}|+|\vec{r}_{pl}|)}{\lambda}\right] \quad (11)$$

即传统距离分集信道模型子信道间相关系数的计算公式。

2.2 三维信道模型相关系数计算

三维混合分集信道模型,即既考虑距离分集情

$$\rho_{mn,pq} = \frac{\sum_{l=1}^L \exp\left[\frac{-j2\pi(|\vec{r}_{nl}|+|\vec{r}_{ml}|+|\vec{r}_{ql}|+|\vec{r}_{pl}|)}{\lambda}\right]}{\sqrt{\sum_{l=1}^L \sum_{i=1,j=1}^{3,3} \sum_{u=1,v=1}^{3,3} r_i^{(n)} t_j^{(m)} r_u^{(q)} t_v^{(p)} E(A_{ij}^{(l)}) E(A_{uv}^{(l)})}} \sqrt{\sum_{l=1}^L \sum_{i=1,j=1}^{3,3} \sum_{u=1,v=1}^{3,3} r_i^{(n)} t_j^{(m)} r_u^{(q)} t_v^{(p)} E(A_{ij}^{(l)}) E(A_{uv}^{(l)})} \quad (15)$$

2.3 XPD 计算

电波的极化方向在经过散射体之后很可能发生改变,不同极化方向的接收天线将各接收到一部分功率的电波,XPD(cross-polar discrimination)就是从功率方面描述极化方向在传播过程中的变化趋势的参

2 统计参数计算和推导

2.1 传统一维信道模型相关系数计算

任意两条子信道之间的相关性可以通过下式求得^[10]:

况又考虑极化分集情况的信道模型。采用第1节所述的信道模型,以式(6)作为子信道向量的计算公式,并且为简化计算假设收发端均采用理想点源天线组成的阵列,此时有

$$E(h_{mn} h_{pq}^*) = \sum_{l=1}^L \exp\left[\frac{-j2\pi(|\vec{r}_{nl}|+|\vec{r}_{ml}|+|\vec{r}_{ql}|+|\vec{r}_{pl}|)}{\lambda}\right].$$

$$E(g_l^2 \sum_{i=1,j=1}^{3,3} r_i^{(n)} t_j^{(m)} A_{ij}^{(l)} \sum_{i=1,j=1}^{3,3} r_i^{(q)} t_j^{(p)} A_{ij}^{(l)}) \quad (12)$$

决定 A_l 各元素之变化的并不是散射体本身,而是散射体相对收发端的位置和方向。由这些参数所引起的 A_l 之变化,各元素之间并没有确定的相互关系。因此可以认为接收信号在某个极化方向上发生的相对变化,和在与其正交的极化方向上发生的变化无关,即矩阵 A_l 的各个元素互不相关。令

$$E(A_{ij}^{(l)} A_{uv}^{(l)}) = E(A_{ij}^{(l)}) E(A_{uv}^{(l)}) \quad (13)$$

则上式变为

$$E(h_{mn} h_{pq}^*) = \sum_{l=1}^L E(g_l^2) \exp\left[\frac{-j2\pi(|\vec{r}_{nl}|+|\vec{r}_{ml}|+|\vec{r}_{ql}|+|\vec{r}_{pl}|)}{\lambda}\right].$$

$$\sum_{i=1,j=1}^{3,3} \sum_{u=1,v=1}^{3,3} r_i^{(n)} t_j^{(m)} r_u^{(q)} t_v^{(p)} E(A_{ij}^{(l)}) E(A_{uv}^{(l)}) \quad (14)$$

如果 $E(g_l^2)$ 为定值,则三维情况下子信道间的相关系数为

$$\rho_{mn,pq} = \frac{\sum_{l=1}^L \sum_{i=1,j=1}^{3,3} \sum_{u=1,v=1}^{3,3} r_i^{(n)} t_j^{(m)} r_u^{(q)} t_v^{(p)} E(A_{ij}^{(l)}) E(A_{uv}^{(l)})}{\sqrt{\sum_{l=1}^L \sum_{i=1,j=1}^{3,3} \sum_{u=1,v=1}^{3,3} r_i^{(n)} t_j^{(m)} r_u^{(q)} t_v^{(p)} E(A_{ij}^{(l)}) E(A_{uv}^{(l)})}} \quad (15)$$

数,其定义为与发送端相同极化方向的天线接收到的功率同与发送端极化方向相垂直的天线接收到的功率之比。对于传统的二维极化分集理论^[12-13]来说,

水平极化的XPD为 $XPD_H = \frac{D(h_{HH})}{D(h_{HV})}$,垂直极化的

XPD 为 $XPD_V = \frac{D(h_{VV})}{D(h_{VH})}$, 其中 $D(h) = E(hh^*)$ 。对于三维模型来说, 只考虑垂直与水平的 XPD 是不够的, 于是将三维情况下的 XPD 定义为一组数值:

$$\left\{ \begin{array}{l} XPD_{xy} = D(h_{xx})/D(h_{xy}) \\ XPD_{xz} = D(h_{xx})/D(h_{xz}) \\ XPD_{yx} = D(h_{yy})/D(h_{yx}) \\ XPD_{yz} = D(h_{yy})/D(h_{yz}) \\ XPD_{zx} = D(h_{zz})/D(h_{zx}) \\ XPD_{zy} = D(h_{zz})/D(h_{zy}) \end{array} \right. \quad (16)$$

其中, $D(h) = E(hh^*)$ 而 x, y, z 表示平行于三维空间各坐标轴的极化方向。将收发天线的极化方向选为和各坐标轴平行, 并分别带入式(6), 计算结果即 $h_{xx}, h_{xy}, h_{xz}, h_{yx}, h_{yy}, h_{yz}, h_{zx}, h_{zy}, h_{zz}$ 。再利用式(16)即可得到全部 XPD。

由于信号在到达接收端时, 其幅度和极化方向都会随着收发端和散射体的位置改变而产生很大的变化, XPD 也随之变化。由于天线的分集距离相对收发端和散射体之间的距离来说可以忽略, 所以可近似认为所有的天线可以使用同一组 XPD 数值。

3 仿真验证

本文采用单环模型模拟传输环境: 假设散射体的水平位置在接收端周围呈圆环状分布, 散射体的高度服从正态分布, 圆环半径为 200 m, 收发端间距为 1 000 m, 载波频率为 3 000 MHz。收发端可以是多个处于任意位置并且具有任意极化方向的天线组成的阵列, 为简便起见, 讨论收发端均为理想线极化天线或点源天线的情况。

首先假设收发端所有天线均为垂直极化, 即仅考虑垂直极化的一维情况, 此时由传播中的去极化效应造成的、处于非垂直方向的那部分场强, 将不会被接收。此时本文第 3 节所提出的混合分集物理信道模型可以看做是等同于纯粹的距离分集模型^[11]的一种特殊情况。假设接收端的两根天线完全重合, 发送端的两根天线位置如图 2 所示。

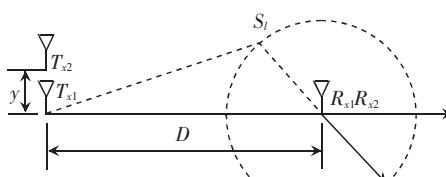


图 2 单环模型

对本文所提出的三维模型采用一维特殊情况的天线及环境参数, 并在与收发端连线相垂直的方向 (y 轴) 上进行距离分集, 与传统一维物理信道模型^[11]的结论进行比较的结果见图 3。

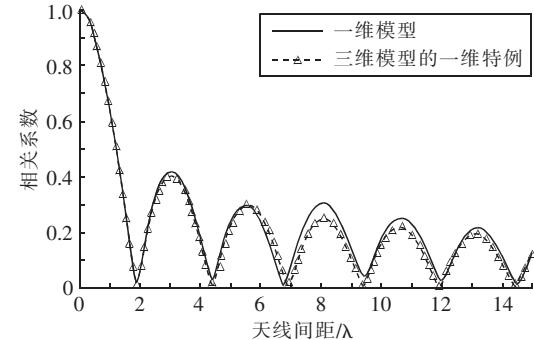


图 3 一维情况下相关系数与发送天线间距的关系

从图 3 可以看出两条曲线基本吻合, 所以仅考虑距离分集的一维信道模型可以看作本文所提出三维混合分集模型的特例。

令收发端的天线间距均在与收发端连线相垂直的方向上发生改变, 则发送端天线间距和接收端天线在 y 轴上间距的联合分布与相关系数的关系见图 4。

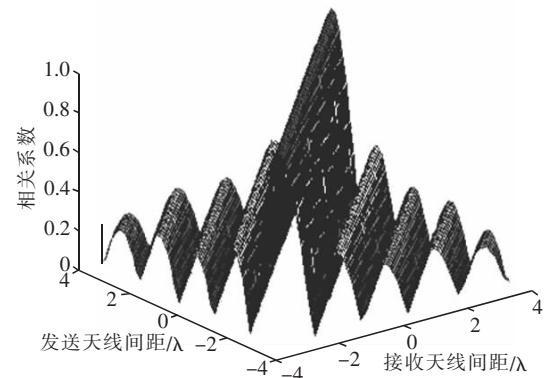


图 4 收发端天线间距与相关系数的关系

上述仿真结果验证了本文所提出的三维模型在天线仅存在距离分集的一维特殊情况下, 与传统距离分集一维模型的一致性, 接下来演示此模型在三维混合分集情况下的性能。

图 5 显示了发送端的两根天线极化方向的夹角分别为 $0, \frac{\pi}{4}, \frac{\pi}{3}$ 和 $\frac{\pi}{2}$ 时, 天线距离与相关系数的关系。

由图 5 可以推出: 两根天线的极化方向相差越多, 子信道之间的相关性越低。与当天线极化方向相互垂直时, 即使天线之间的距离很近, 子信道间的相关性也很低。这就验证了即使是在天线没有间距的情况下, 采用极化分集也能明显降低子信道之间的相关性, 获得良好的分集效果, 从而证明了本文所

提出物理信道模型可以合理地对各种混合分集情况进行模拟。

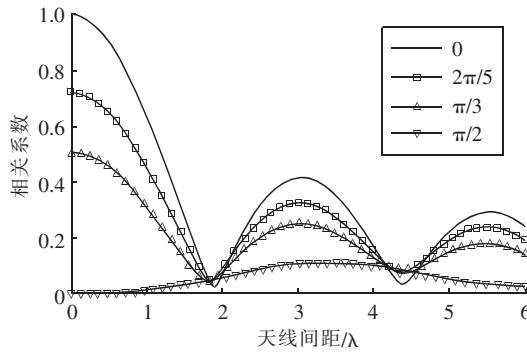


图5 天线极化方向与相关系数的关系

4 结束语

当前二维的MIMO极化分集被研究的比较多,而三维极化分集MIMO系统的有效性已经被证实,此类系统需要合理的三维信道模型的支持。本文针对这一现状,通过几何方法建立了一个基于散射矩阵的三维混合分集物理信道模型,并对从物理信道提取统计参数的方法进行了研究。通过此模型能够合理地提取出相关性等统计参数,而且仿真结果也证明了此模型在对各种混合分集情况进行模拟时的正确性。

参考文献:

- [1] TELATAR I E. Capacity of Multi-Antenna Gaussian Channels[R]. AT&T Bell Laboratories Internal Technical Memorandum, June 1995.
- [2] FOSCHINI G J, GANS M J. On limits of wireless communications in a fading environment when using multiple antennas[J]. Wireless Personal Communications, 1998 (6):311 - 335.
- [3] FOSCHINI G J. Layered space-time architecture for wireless communication in fading environment when using multiple antennas[J]. Bell Labs Technical Journal, 1996,1(2):41 - 59.
- [4] ANDREWS M A, MITRA P P, CARVALHO R. Tripling the Capacity of Wireless Communications Using Electromagnetic Polarization [J]. Nature, 2001,409:316 - 318.

- [5] NABAR R U, BÖLCSKEI H, ERCEG V. Performance of multi-antenna signaling techniques in the presence of polarization diversity[J]. IEEE Trans on Signal Processing, 2002,50(10):2553 - 2562.
- [6] CHIU C Y, YAN J B, MURCH R D. Compact three-port orthogonally polarized MIMO antennas[J]. IEEE Antennas Wireless Propag Lett, 2007,6:619 - 622.
- [7] MTUMBUKA M C, EDWARDS D J. Investigation of tri-polarised MIMO technique[J]. IEE Electronics Letters, 2005,41:137 - 138.
- [8] SVANTESSON T. A double-bounce channel model for multi-polarized mimo systems[C]//56th Proceedings of Vehicular Technology Conference. IEEE Espoo, Finland, Sept, 2002:691 - 695.
- [9] WANG Xuedong, LI Jiandong. A Generic Channel Model For MIMO Systems[C]//ICCT'06. 27 - 30 Nov, 2006:1 - 4.
- [10] SVANTESSON T. A Physical MIMO Radio Channel Model for Multi-element Multi-polarized Antenna systems[C]// Proc IEEE VTC 2001 Fall. Atlantic City, NY, October 2001:1083 - 1087.
- [11] STEGE M, JELITTO J, BRONZEL M. A multiple input-multiple output channel model for simulation of TX- and RX-diversity wireless systems[C]// Proc IEEE Vehic Technol Conf. Boston, MA, 2000:833 - 839.
- [12] ANREDDY V R, INGRAM M A. Capacity of measured Ricean and Rayleigh indoor MIMO channels at 2.4GHz with polarization and spatial diversity[C]// Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '06). Las Vegas, Nev, USA, April 2006,2:946 - 951.
- [13] ALATOSSAVA M, HENTILÄ L, HOLAPPA V M. Comparison of Outdoor-to-Indoor and Indoor-to-Outdoor MIMO Propagation Characteristics at 5.25 GHz[C]// Proc of VTC2007-Spring, IEEE 65th Vehicular Technology Conference. Dublin, Ireland, 2007.

作者简介:



海 凛(1982-),男,安徽和县人。南京邮电大学电子科学与工程学院博士研究生。研究方向为数字电视移动接收、多域协同环境中的多天线链路信道模型等。

张业荣(1963-),男,安徽和县人。南京邮电大学电子科学与工程学院副院长,教授,博士生导师。(见本刊2009年第4期第68页)

一种新型的网络带宽最优分配机制

冯慧斌, 张顺颐, 刘超, 王攀

(南京邮电大学信息网络技术研究所, 江苏南京 210003)

摘要:研究了新一代网络中的网络带宽最优分配机制。应用经济学中的社会福利函数思想和根据流量的优先级构造了基于公平意识的流量效用函数,证明了网络带宽最优分配模型的纳什均衡解存在性,提出了一种基于优先级的网络带宽最优分配机制。仿真结果表明提出的带宽分配机制既能使得流量根据优先级获得区分服务又能保证带宽分配相对公平,从而保证网络中不同业务流的服务质量(QoS)。

关键词:带宽分配; 纳什均衡; 弹性流

中图分类号:TP393 文献标识码:A 文章编号:1673-5439(2009)06-0043-05

A Novel Network Bandwidth Optimal Allocation Mechanism

FENG Hui-bin, ZHANG Shun-yi, LIU Chao, WANG Pan

(Institute of Information Networks Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The network bandwidth optimal allocation mechanism in new generation network is investigated in this paper. The fairness aware flow utility function is constructed by using the social welfare function and the elastic flow's priority, the Nash Equilibrium of the network bandwidth optimal allocation model is proved, and the bandwidth optimal allocation mechanism on flow's priority is proposed. Simulation result shows the proposed mechanism not only can achieve differentiated services according to the priority, but ensures the relative fairness of the bandwidth allocation as well, thus guarantees the QoS of different service flow.

Key words: bandwidth allocation; Nash equilibrium; elastic flow

0 引言

新一代网络是以IP技术为核心可同时支持语音、数据和多媒体等各种业务的融合网络,由于网络中的业务具有不同的业务特征和服务质量需求,因此研究如何为承载的不同业务类型的新一代网络公平合理地分配网络带宽资源,对提升网络整体性能和效用具有重要意义。

国内外研究人员对传统网络的带宽分配问题进行了大量的研究。文献[1]提出了一种动态带宽分配模型来确保网络中用户的服务质量(QoS)需求同时最大化网络整体效用,并且理论分析和仿真实验

验证了模型的正确性。文献[2]利用议价方法来研究区分服务网络中的带宽分配问题,提出了基于网络效用最大化的议价数学模型,给出了求解网络带宽最优分配算法。文献[3]研究了利用非合作博弈理论建立无线蜂窝网络带宽模型,证明了模型的纳什均衡解的存在性和唯一性。文献[4]利用排队算法设计了新型的网络带宽管理方法给不同的业务分配带宽,其核心思想是利用排队调度来保证用户带宽需求和数据包的时延约束。文献[5]利用控制理论来研究网络中自适应QoS需求的动态带宽分配问题,控制器根据流量的丢包率来动态调整用户带宽,并提出了基于AIMD(Additive Increase Multiplic-

ative Decrease)控制模型来获得最优带宽分配。文献[6]提出了基于QoC(Quality of Control)动态带宽分配模型及算法,能有效地提高传统静态分配策略的效率。文献[7]利用博弈论中的VCG(Vickrey-Clarke-Groves)算法规划P2P网络中的带宽分配和计费使得P2P网络中能提供有效的带宽分配和计费方案,从而在服务开销和用户收益两个方面达到最优。文献[8]提出一种短流优先的公平带宽分配机制FPIP(Fair Proportional Integral based series compensation and Position feedback compensation),通过区别处理短流和长流的报文,FPIP能够将带宽优先分配给短流,然后将剩余的带宽在长流之间公平分配。

已有的带宽分配模型主要研究了流量最优带宽分配,其并未考虑流量间的优先级特性和流量之间带宽分配的相对公平性对网络性能的影响。本文根据流量的优先级特性构造了基于公平意识的流量效用函数,证明了最优网络带宽分配的纳什均衡,在此基础上提出了基于优先级的网络带宽最优分配机制。仿真验证了提出的带宽分配机制既能使得流量根据优先级获得区分服务又能保证各流量的带宽分配相对公平。

1 博弈论及纳什均衡

博弈论是研究具有理性的不同主体在“策略相互依存”情形下相互作用的数学工具^[9]。博弈模型可以用 $G = \{I, S, U(\cdot)\}$ 描述:其中 I 是博弈参与者的集合, S 是博弈所有参与者的策略空间,它是一个非空的、紧闭凸集, $U(\cdot)$ 是博弈模型的参与者效用函数。在博弈模型 G 中,博弈参与者 i 采用策略 s_i^* 的收益是在其它参与者策略组合 $s_{-i}^* = (s_1^*, s_2^*, s_{i-1}^*, \dots, s_{i+1}^*, s_{m-1}^*, s_m^*)$ 确定下获得的,如果参与者 i 的效用函数满足不等式: $u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*)$, $\forall s_i \in S_i$,则称 s_i^* 为博弈模型 G 的一个Nash均衡。

2 基于Max-Min的新一代网络带宽分配模型

假设网络中有 L 个网络链路,网络中有 N 个根据其路由路径标识的弹性流量竞争网络带宽 C ,令 r_i 表示保证流量*i*最低服务质量所需的带宽,向量 $r = (r_1, r_2, \dots, r_N)$ 表示流量所需最小带宽向量, p_i 表

示流量*i*获得最佳服务质量所需的最大带宽,向量 $p = (p_1, p_2, \dots, p_N)$ 表示流量所需最大带宽向量,则 $0 < r_i < p_i$ 。定义一个 $L \times N$ 维流量矩阵 $A = (a_{i,l})_{N,L}$,如果流量*i*使用链路*l*则 $a_{i,l} = 1$,否则 $a_{i,l} = 0$,由于所有流量的速率之和必须小于链路的带宽约束关系因此可得 $(Ax)_l \leq (C)_l$ 。根据上述定义可得对于任一流量*i*其分配带宽 x_i 必须满足下式:

$$x_i = \{x_i \in R^N \mid r_i \leq x_i \leq p_i \text{ and } (Ax)_i \leq (C)_i\} \quad (1)$$

由于每个流量都希望获得最大带宽来获得最佳服务质量从而最大化自己效用,定义流量所获带宽 x_i 与期望所获最大带宽 p_i 的之差的倒数为用户的效用,则流量效用函数表示如下:

$$U(x_i) = \frac{1}{a(p_i - x_i)^\alpha} \quad (2)$$

其中,参数 $0 < \alpha \leq 1$ 表示流量对带宽的敏感度, α 的值越小表示该流量对带宽越敏感,反之则不敏感。由式(2)可知当流量所获带宽 x_i 与期望所获最大带宽 p_i 接近时,其获得服务质量也越高,流量的收益 $U(x_i)$ 越接近 $+\infty$,当流量所获带宽 x_i 偏离期望所获最大带宽 p_i 时,流量的收益急剧下降,其获得服务质量也越来越低。

为了建立网络最优带宽分配模型,根据文献[10]首先定义Harsanyi类型社会福利函数:

$$U(x_i, x_{-i}) = \sum_{i=1}^N w_i U(x_i) \quad (3)$$

其中, $x_{-i} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$, $U(x_i)$ 为式(2)所定义的效用函数,由于社会福利是由个体效用加权的线性组成构成,定义 w_i 为流量在社会福利中的加权因子,考虑业务不同的服务质量需求和带宽分配需求,可定义社会福利函数加权因子为 $w_i = k_i(p_i - r_i)^{\alpha+1}$,其中 $k_i = 1, 2, 3, 4$ 表示流量优先级, k_i 的值越小表示流量*i*的优先级越高。

上述定义的社会福利函数只考虑了单个流量的效用水平,对于整个网络来说必须考虑所有流量之间的均衡,为了保证各流量之间公平地共享链路带宽,可以最大最小每个流量的效用,即:

$$(S) \quad \max_{x_i} \min_{x_{-i}} U(x_i, x_{-i}), i = 1, 2, \dots, N \quad (4)$$

且

$$(Ax)_l \leq (C)_l, l = 1, 2, \dots, L$$

问题S是根据经济学中的最大最小原则下的社会福利函数形式。假定人们都是风险规避者,在“无知的面纱”下进行选择博弈策略。一个社会的富裕程度要由其处境最差的人来决定。只要状态X1中处境最差的人要好于状态X2中处境最差的

人,那么前者就好于后者。如果在状态 X_3 中处境最差的人与其他状态中处境最差的人一样好,则状态 X_3 是最大最小最优的。

满足式(4)求得的任何带宽分配都能满足 Max-Min 公平性。假设所有流量所需分配最小带宽总是小于链路容量即满足 $(Ar)_l < (C)_l$ 。当经过链路 l 的流量满足 $(Ap)_l < (C)_l$ 时,由于所有流量所获总带宽小于链路容量,因此所有流量都能获得最佳服务质量,任何一种带宽分配都是最优的。当 $(Ap)_l > (C)_l$ 时,由于流量所需分配总带宽大于链路容量,因此流量所分配带宽必须满足 $(Ax)_l = (C)_l$ 。当 $(Ax)_l < (C)_l$ 时,如果 $a_{i,l} = 0$,则表示流量 i 不参与链路带宽分配。如果 $a_{i,l} = 1$,则可把 x_i 写成 $x_i = (C)_l - \sum_{j \neq i} a_{j,l} x_j$,无论 $a_{i,l}$ 取值都可以把式(4)的最大最小化模型转换成一个最小化模型 $\min_x U(x)$,其表达式如下:

$$(D) \quad \min_x U(x) = \min_x \sum_{i=1}^N w_i U(x_i) \quad (5)$$

$$\text{s. t.} \quad (Ax)_l = (C)_l, l = 1, 2, \dots, L \quad (6)$$

$$x_i \leq p_i, i = 1, 2, \dots, N \quad (7)$$

$$x_i \geq r_i, i = 1, 2, \dots, N \quad (8)$$

其中,式(6)表示带宽分配过程中的链路容量约束,式(7)和式(8)表示流量的带宽需求。

3 模型纳什均衡分析

假设模型满足式(8)的约束条件,则问题 D 的可行域是非空的、凸的且是完备的集合。根据式(3)定义的效用函数可求得 $\frac{\partial f}{\partial x_i x_j} = 0 (i \neq j, i = 1, 2, \dots, N, j = 1, 2, \dots, N)$,其二阶导数 $\frac{\partial^2 f}{\partial x_i^2} > 0 (i = 1, 2, \dots, N)$,从而可得其 Hessian 矩阵 $\nabla^2 U(x)$ 是正定矩阵。由于效用函数 $U(x)$ 的 Hessian 矩阵是正定的,所以 $U(x)$ 是严格的凸函数,根据最优化理论可知问题 D 在可行域内存在唯一的最优解。

假设 $L(x, \lambda, \mu)$ 为 Lagrange 函数,其中 λ 为与链路 l 容量约束 Lagrange 乘子, $\mu_i \geq 0, i = 1, 2, \dots, N$ 为与式(7)关联的 Lagrange 乘子,则 $L(x, \lambda, \mu)$ 表达式如下:

$$L(x, \lambda, \mu) = U(x) - \lambda((Ax)_l - C_l) - \sum_{i=1}^N \mu_i(x_i - p_i) \quad (9)$$

根据 Kuhn-Tucker 一阶条件,可得如下所示的表达式:

$$w_i(p_i - x_i^{*})^{-1-\alpha} - \lambda(a_{i,l}) - \mu_i = 0, i = 1, 2, \dots, N \quad (10)$$

$$\lambda((Ax)_l - (C)_l) = 0 \quad (11)$$

$$\mu_i(x_i - p_i) = 0, i = 1, 2, \dots, N \quad (12)$$

由于假设模型满足式(8)的约束条件且经过链路 l 的流量才参与带宽分配,因此可令 $a_{i,l} = 1$,构造 Lagrange 函数只需构造式(11)和式(12)满足式(6)和式(7)即可。由于 $x_i \leq p_i$,因此对所有流量 $i = 1, 2, \dots, N$ 要满足式(12)条件必须是 $\mu_i = 0$,整理式(10)可得问题 D 的最优解:

$$x_i^{*} = p_i - [w_i/\lambda]^{1/(1+\alpha)}, i = 1, 2, \dots, N \quad (13)$$

引理 1 问题 D 网络带宽最优分配解是 $x_i^{*} = p_i - [w_i/\lambda]^{1/(1+\alpha)}, i = 1, 2, \dots, N$,该解能保证网络中流量获得 Max-Min 公平带宽分配。

由于式(13)是问题 D 的最优解,问题 S 是问题 D 等价问题,因此式(13)也是原问题 S 的最优解,由此可得定理 1。

定理 1 假设 $r_i \leq x_i$,网络带宽分配模型 S 的纳什均衡是 $x_i^{*} = p_i - [w_i/\lambda]^{1/(1+\alpha)}$,其中 $(Ax)_l = (C)_l, r_i \leq x_i \leq p_i, \lambda \geq 0, l = 1, 2, \dots, L, i = 1, 2, \dots, N$ 。

证明 假设 $x^{*} = (x_i^{*}, x_{-i}^{*})$ 是满足约束式(6)、式(7)和式(8)的模型的最优解,令流量 i 分配带宽为 x_i 且 $x_i \neq x_i^{*}$,则带宽分配向量 (x_i, x_{-i}^{*}) 必须满足 $(Ax)_l \leq (C)_l$ 。如果 $a_{i,l} = 0$,则 $x_i = 0$,如果 $a_{i,l} = 1$,则 $x_i + \sum_{j \neq i} a_{j,l} x_j^{*} \leq (C)_l$,而 $x_i^{*} + \sum_{j \neq i} a_{j,l} x_j^{*} = (C)_l$,比较两式可得 $x_i \leq x_i^{*}$ 。由于 $U(x_i, x_{-i}^{*})$ 关于 x_i 的严格递增函数,因此可得如下表达式:

$$U(x_i, x_{-i}^{*}) \leq U(x_i^{*}, x_{-i}^{*}), i = 1, 2, \dots, N \quad (14)$$

由纳什均衡定义可得 $x^{*} = (x_i^{*}, x_{-i}^{*})$ 即网络带宽分配模型 S 的纳什均衡。证毕。

4 基于优先级的最优网络带宽分配机制

由微观经济学可知,上一节定义的拉格朗日乘子 λ 可理解带宽分配的边际成本。由于流量的最小带宽需求为 r_i ,结合式(13)可得:

$$x_i^{*} = p_i - [w_i/\lambda]^{1/(1+\alpha)} \geq r_i \quad (15)$$

整理上式得: $w_i \leq \lambda(p_i - r_i)^{(1+\alpha)}$,根据第 1 节定义效用函数的加权因子 $w_i = k_i(p_i - r_i)^{a+1}$ 可得:

$$k_i \leq \lambda, i = 1, 2, \dots, N \quad (16)$$

由式(16)可清楚看出只要流量的优先级取值不大于带宽分配的边际成本,流量就可以分配到所需的小带宽。把 w_i 代入式(15)可把最优带宽分配表达式写成如下式子:

$$x_i^* = p_i - (p_i - r_i)(k_i/\lambda)^{1/(1+\alpha)} \quad (17)$$

从式(17)可知拉格朗日乘子 λ 是所有流量优先级的函数即 $\lambda = f(k_1, k_2, \dots, k_N)$,由于带宽分配必须满足式 $\sum x^* = C$,将式(17)代入可把最优带宽分配写成 p_i 和 r_i 加权向量表达式:

$$x_i^* = (1 - \beta_i)p_i + \beta_i r_i \quad (18)$$

$$\sum_{j=1}^N p_j = c$$

$$\text{其中}, \beta_i = \frac{\sum_{j=1}^N (p_j - r_j)(k_j/k_i)^{1/(1+\alpha)}}{\sum_{j=1}^N (p_j - r_j)(k_j/k_i)^{1/(1+\alpha)}}.$$

从式(18)可看出,只要 $0 \leq \beta_i \leq 1$,则流量 i 的可分配的最佳带宽可写为 $(1 - \beta_i)p_i + \beta_i r_i$,且能保证其所分配带宽在最大带宽 p_i 和最小带宽 r_i 之间。由上述分析可以给出基于优先级的最优带宽分配机制。

基于优先级的网络最优带宽分配机制:

For ($i = 1$ to N), 计算 $\sum r_i$ 和 $\sum p_i$

If ($\sum p_i \leq c$), 则给所有流量分配其最大带宽 p_i

If ($\sum r_i \leq c < \sum p_i$)

For ($i = 1$ to N) 计算流量 i 的值 β_i

If ($0 \leq \beta_i \leq 1$), 则分配给流量 i 的最佳带宽为 $x_i^* = (1 - \beta_i)p_i + \beta_i r_i$

For ($i = 1$ to N) 返回流量 i 的带宽 x_i^*

5 数值仿真

本节将通过仿真实验来验证提出的带宽分配机制的正确性及有效性。

实验一 假设网络中有 4 个弹性流量 1、2、3 和 4,其对应的最小带宽向量为 $\bar{r} = [0.5, 0.75, 1, 1.25]$,所需最大带宽向量为 $\bar{p} = [2, 3, 4, 5]$,在实验一中假设网络中所有流量有相同的优先级 $\bar{k} = [1, 1, 1, 1]$ 。

从图 1 可以看出由于流量的优先级相等,因此随着网络带宽容量的增加各流量获得的带宽也是线性增长的,当带宽增长到流量最大需求之和时各流量都能获得最佳的服务质量。

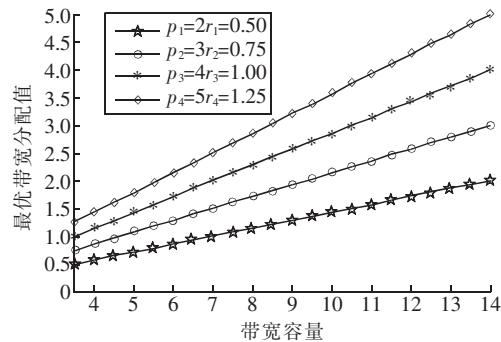


图 1 优先级相等情况下的各流量带宽分配示意图

实验二 假设网络中有 4 个弹性流量 1、2、3、4,其对应的最小带宽向量为 $\bar{r} = [0.5, 1, 0.5, 1]$,所需最大带宽向量为 $\bar{p} = [1, 3, 1, 3]$,假设前 2 个流量有相同的优先级且 $\bar{k} = [2, 2]$,后 2 个流量有相同的优先级但高于前 2 个流量优先级且 $\bar{k} = [1, 1]$ 。从图 2 可以看出虽然流量 1 和流量 3 的带宽需求是相同的,但是由于流量 3 的优先级高于流量 1 因此在带宽分配中高优先级的流量分配到更多的带宽。随着带宽容量逐渐增长,优先级高的流量 3 也不能立即分配所需最大带宽,当带宽增长到流量最大需求之各流量之间带宽分配差异消除了,每个流量都分配到了最大带宽而达到最佳服务质量,保证了各流量带宽分配的相对公平。从图同样地可以看出流量 2 和流量 4 也是如此。而且由于两个流量的带宽需求大于流量 1 和 3,从而导致两个流量分配的带宽也同样大于流量 1 和 3。

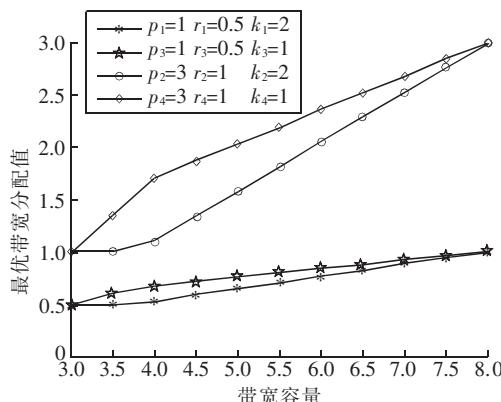


图 2 优先级两两相等情况下的各流量带宽分配示意图

实验三 假设网络中有 4 个弹性流量 1、2、3 和 4,其对应的最小带宽向量为 $\bar{r} = [0.5, 0.5, 0.5, 0.5]$,流量所需最大带宽向量为 $\bar{p} = [1, 1, 1, 1]$,且假设 4 个流量的优先级向量为 $\bar{k} = [1, 2, 3, 4]$ 。从图 3 可以看出虽然各流量的带宽需求都是相同的,但随着带宽容量不断增长优先级高的流量分配到的带宽大于优先级低的流量,说明带宽分配机制能根据流量优先级提供区分服务。随着带宽容量不断增

大,各流量分配带宽差距不断减少,当带宽增长到所有流量的最大带宽需求之和时每个流量都分配到了最大带宽而达到最佳服务质量,从而保证了各流量带宽分配的相对公平。

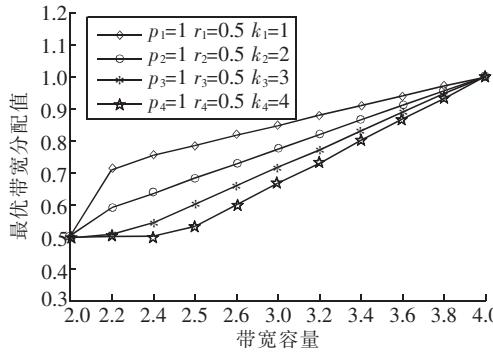


图3 优先级不一致情况下的各流量带宽分配示意图

6 结束语

本文针对新一代网络的弹性流业务特征,研究了基于区分服务和公平意识的新一代网络带宽分配机制。借鉴 Harsanyi 类型社会福利函数和根据流量的优先级构造了基于 Max-Min 的网络流量效用函数。根据弹性流的效用函数,证明了提出的网络带宽分配模型纳什均衡解的存在性并给出其表达式。在此基础上给出了基于优先级的网络带宽最优分配机制。仿真验证了提出的带宽分配机制既能使得流量根据优先级获得区分服务又能保证带宽分配相对公平,表明了提出的机制能根据不同流量的特征有效地分配网络带宽,从而能提升网络的整体效用。

参考文献:

- [1] ELIAS J, MARTIGNON F, CAPONE A, et al. A new approach to dynamic bandwidth allocation in Quality of Service networks performance and bounds[J]. Computer Networks, 2007, 51: 2833 – 2853.
- [2] GUAN Yongpei, WEIL G, OWEN Henry, et al. A pricing approach for bandwidth allocation in differentiated service networks[J]. Computers & Operations Research, 2008, 35: 3769 – 3786.
- [3] SAHASRABUDHE A, KOUSHIK K. Bandwidth Allocation Games under Budget and Access Constraints[C] // Proc of CISS. March 2008: 761 – 769.
- [4] QADEER M A, SHAHID H, YAZDAN J A. Differential Allocation of

Bandwidth to Services based on Priority[C] // ICI. September 23, 2008: 761 – 769.

- [5] RABBY M, RAVINDRAN K. Dynamics of End-to-End Bandwidth Allocations in QoS-adaptive Data Connections[C] // Proc of LAN-MAN. September 2007: 96 – 101.
- [6] ZHAO Weiquan, LI Di. Performance Optimization Based On Dynamic Bandwidth Allocation for Networked Motion Control Systems[C] // Proc of PEITS. 2008: 83 – 87.
- [7] 黄冠尧, 洪佩琳, 李津. P2P-VCG: 一种基于博弈论的带宽分配方案[J]. 计算机研究与发展, 2007, 44(1): 78 – 84.
HUANG Guanyao, HONG Peilin, LI Jin. P2P-VCG: A game theory Proposal for bandwidth allocation[J]. Journal of Computer Research and Development, 2007, 44(1): 78 – 84.
- [8] 张鹤颖, 蒋杰, 窦文华. 一种短流优先的公平带宽分配机制[J]. 软件学报, 2007, 18(3): 765 – 774.
ZHANG Heying, JIANG Jie, DOU Wenhua. Fair bandwidth allocation mechanism with Preference to short flows[J]. Journal of Software, 2007, 18(3): 765 – 774.
- [9] FUDENBERG D, TIROLE J. Game Theory[M]. Cambridge, MA: The MIT Press, 1991: 10 – 29.
- [10] HARSANYI J C. Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility[J]. The Journal of Political Economy, 1955, 63(4): 309 – 321.

作者简介:



冯慧斌(1980 -),男,江西遂川人。南京邮电大学信息网络技术研究所博士研究生。主要研究方向为无线通信与计算机通信网。

张顺颐(1944 -),男,江苏南京人。南京邮电大学信息网络技术研究所教授,博士生导师。(见本刊2009年第1期第5页)

刘超(1977 -),男,江苏镇江人。南京邮电大学信息网络技术研究所博士研究生。主要研究方向为无线传感器网络。

王攀(1977 -),男,新疆阿克苏人。南京邮电大学信息网络技术所助理研究员。主要研究方向为网络流量建模与预测。

基于 CSP 的进程行为取证方法研究

孙国梓^{1,2}, 俞超^{1,2}, 陈丹伟^{1,2}

(1. 南京邮电大学 计算机技术研究所, 江苏南京 210003)
(2. 南京邮电大学 计算机学院, 江苏南京 210046)

摘要:针对取证过程中所获取的进程异常行为,提出进程行为事件重建犯罪过程的方法。该方法使用CSP(通信顺序进程)理论来形式化描述具有威胁乃至破坏性的进程操作及进程间的通信,根据系统保存的进程行为记录建立进程通信状态模型,使用基于路径搜索的进程行为解释算法分析模型内所有可能的进程通信序列,形成进程通信行为规则,在排除不符合规则的通信序列的基础上,找到能够形成合理证据链的通信序列。通过案例分析进行了证据的形式化及CSP建模,给出了进程行为的具体分析解释和原型系统,验证该方法的可行性和有效性。

关键词:计算机取证;CSP;进程行为;通信序列

中图分类号:TP309 文献标识码:A 文章编号:1673-5439(2009)06-0048-06

Research on Forensic Methods of the Process Behavior Based on CSP

SUN Guo-zi^{1,2}, YU Chao^{1,2}, CHEN Dan-wei^{1,2}

(1. Institute of Computer Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)
(2. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: For the abnormal process behaviors got in digital forensic process, the paper gives a method of reconstructing the crime process with the process events. In the method, it firstly formally describes dangerous process operations and process communications by CSP theory, and builds a communication state model of process using the process records in system to get the process communication rules, then finds the communication sequence which can form reasonable evidence link after analyzing all possible communication sequences in the model with interpretation algorithm based on path search and excluding communication sequences out of rules. In the last, the paper gives a specific analysis and explanation of process behaviors by formalizing evidences and constructing CSP model, and validates the feasibility and effectiveness of the method with the developed simulation software.

Key words: computer forensic; CSP; process behavior; communication sequence

0 引言

计算机取证是运用计算机及其相关学科的原理与方法来获取与计算机相关的证据以证实某个客观事实存在的过程^[1]。计算机证据分析是计算机取证的核心和关键步骤。计算机证据分析就是通过分析基本数据的含义、属性、相互关系以查证信息的存在、信息的来源以及信息的传播途径,进而重构某事

件的真实情况(如犯罪事件或嫌疑人的特征、犯罪行为、犯罪动机等)。如何对现有证据进行智能分析、推理,以最接近已发生事实的方式来重建整个犯罪过程,成为计算机取证的工作重点。事实上,单凭原始证据无法满足诉讼的要求,所以取证调查过程中必须对原始电子数据证据的有效性进行检查,并且使用特定的规则对证据进行正确的分析处理,使得证据在逻辑上能够串联起来形成有效的证据链,

从而使犯罪事实得以重现并作为呈堂证据。

Pavel Gladyshev 和 Ahmed Patel 提出运用有限状态机来进行犯罪事件重建^[2-3],从理论上表明自动推理分析或重建犯罪事件具有可行性。许多实际数字系统使用有限状态机建模,或者可以用有限状态机来描述。但很多现实问题使用该方法建立模型不够直观方便,不能直接表达系统中的异步、并发、冲突等行为。除此之外,计算机自动取证分析方面的研究成果比较少。

通信顺序进程(CSP)是一种数学框架,可以用它来描述与分析由多个组件或者过程组成的系统,CSP 中针对并发系统中进程的特点定义了多种进程间的复合操作,如并发进程、或进程、选择进程、顺序进程、屏蔽进程、混合进程等,可以对系统中进程的非法行为作详细准确的形式化描述,更加真实地反应系统中发生的客观事件^[4-6]。CSP 的数学基础是进程代数,其自身提供了比较丰富的演算手段,在 CSP 模型的基础上,对系统中进程行为进行形式化描述,并通过规则作严格的数学推理,可以对计算机上的进程犯罪行为进行有效的验证和解释。

CSP 是著名计算机科学家 C. A. R. Hoare^[7-8]在 1978 年提出的代数理论,它是一种并发、分布式程序设计语言模型。CSP 具有基于进程代数的特点,长于描述事件的发生和进程之间的关系,在分布计算环境中具有较为完整的代数演算能力,因此在协议描述技术中得到了很好的应用。目前,CSP 的描述范围已得到了多种扩展,如计数器模型、迹模型等。同时,CSP 模型在时间特性描述方面进行了扩展,如 CSP-R,赋时 CSP,通信共享资源 CSR 和通信实时状态机 CRSM。

1 CSP 基本运算定义

CSP 使用数学化的符号来描述进程及其操作,基本运算符如下所示:

(1) STOP 表示一个中断的进程,该进程不可能与外部发生通信,实际中,可表示死锁和进程不收敛;

(2) SKIP 表示一个进程除终止外不做任何事情;

(3) $P;Q$ 为进程间的顺序组合运算,在 P 结束之前都是执行 P 的事件,然后才执行 Q 的事件;

(4) $P \square Q$ 为进程间的选择运算,由外部环境决定进入进程 P 或者进程 Q 的状态;

(5) $P \Pi Q$ 为进程间的或运算,它是内部选择,外部环境对进入哪个状态不起任何作用;

(6) $P \parallel Q$ 表示进程间的并发运算,在进程 P 和 Q 中相同的事件会同时执行,其他事件交替执行;

(7) $P \backslash X$ 表示进程的屏蔽运算,对于进程 P 和事件集合 X ,得到屏蔽进程 P 中所有 X 中事件的进程;

(8) $P \mathbb{M} Q$ 表示进程间的混合运算,混合之后的进程中所执行的每个事件为进程 P 或进程 Q 中的一个事件;

(9) $a \rightarrow P$ 是进程的前缀表示法, a 是进程 Q 中的第一个事件,且执行事件 a 后, Q 的剩余部分是 P ,那么进程 P 可以表示为 $a \rightarrow P$;

(10) $x:A \rightarrow P(x)$ 是进程的选择表示法,表达式给出了从多个前缀事件中选择一个事件执行的描述,其中 A 是进程全体事件集合的一个子集; $x:A$ 表示 A 中任一事件均可为该进程的第一个事件; $P(x)$ 表示执行事件 x 后的剩余进程部分;

(11) $P = F(P)$ 是进程的递归表示法。

2 基于 CSP 的证据分析

2.1 迹模型

迹是进程过程中可见事件的有限序列。traces(P) 表示 P 可以执行的所有迹的集合。例如进程 $P_1 = a \rightarrow \text{STOP}$ 的踪迹为:traces(P_1) = {<>, < a > }。在递归进程中,迹的集合可能是无穷的。例如进程 $P_2 = a \rightarrow P_2$ 的踪迹为 a 的所有可能序列:traces(P_2) = {<>, < a >, < a,a >, < a,a,a >, ... }。

2.2 证据的形式化

为便于计算机证据的自动化分析和事件重建,对计算机犯罪调查中所搜集到的进程行为相关的计算机证据做出形式化的划分,定义了进程行为记录、进程行为序列、进程通信序列和进程入侵证据等概念,为事件重建的实现奠定了基础。在整个过程中使用了基于 CSP 的证据的形式化表示方法,利用 CSP 形式化方法对基于事件的数学模型进行数字证据的分析。图 1 给出了形式化定义的关系。

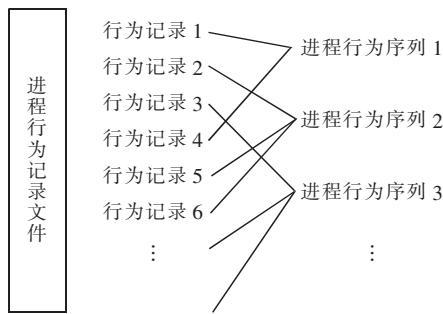


图1 形式化定义关系图

(1) 进程行为记录文件。

是包含所有进程行为记录的文件。利用进程行为监视工具对系统中进程的危险行为(包括进程间的注入、获取密码、非法终止、修改注册表关键信息等)进行记录,并形成文件保存在系统中。

(2) 进程行为记录。

是一条包含了某个进程在确定时间的行为事件记录。每个进程行为记录代表了系统在某时间段所表现出的可记录到的特性。为了更好地表示系统进程通信和非通信的行为记录,我们定义了两种形式化的进程行为记录:进程系统操作记录和进程通信操纵记录。

进程系统操作记录的路径上包含了系统时间、单个进程标识和非通信事件记录(例如进程对注册表关键信息的修改和特殊文件的读取保存等),表示进程在某个系统时间进行的非通信操纵。记作: $pbr = (t, p, e)$,其中 t 代表系统时间, p 代表系统中唯一存在的进程,用PID标示, e 代表非通信事件。

进程通信操纵记录的路径上包含了系统时间、通信发起方进程、通信接收方进程和通信事件记录(例如进程之间的代码注入等),表示通信发起方进程在某系统时间发起与通信接收方进程的通信事件。记作: $pcbr = (t, ps, pd, e)$,其中 t 代表系统时间, ps 代表系统中发起通信的进程, pd 代表系统中接受通信的进程, e 代表通信事件。

(3) 进程行为序列。

是同一进程相关的所有行为记录的集合,由于系统记录无法还原进程程序的整体结构,所以进程行为序列只是进程在众多分支循环路径中运行的一条子路径,在通信顺序进程中称为迹。记作: $pbs = < a.pbr_1, a.pcbr_1, a.pbr_2, \dots >$ 。 $a.pbr_1$ 表示进程 a 的系统操作记录。

(4) 进程通信网。

是整个系统中所有的进程行为序列的集合,不同进程间由通信过程相互关联,它是整个系统进程

活动的映射。记作: $pcn = (a.pbs, b.pbs, c.pbs, \dots)$ 。

(5) 进程通信序列。

是由一个或多个进程通信事件相关联的两个或多个进程行为记录的集合,记作: $pcs = < a.pbr_1, a.pcbr_1, b.pcbr_3, b.pcbr_4, \dots >$ 。由于通信需要两个进程,所以进程通信序列是由两个以上的进程行为记录组成的,单个进程行为记录无法构成进程通信序列。进程通信序列是进程通信网中的一条有效路径,即系统中一种可能的入侵过程。

(6) 进程入侵证据。

使用制定的进程通信操作规则对进程通信序列进行排查,对无法正确解释的通信过程和异常进程行为予以排除,剩余的进程通信序列则形成进程入侵证据集合,为调查人员分析整个犯罪现场的入侵过程提供了有效的解释。记作: $pie = (pcs_2, pcs_4, pcs_9, \dots)$ 。

3 基于路径搜索的进程行为解释算法

基于路径搜索的进程行为解释算法的主要思想是:整个进程入侵过程即可表示为模型中通过进程间通信所联系起来的不同进程行为事件的组合。首先根据被监视系统的进程行为记录为最小单位构建通信顺序进程中迹的网状模型;然后从系统的某个最初状态开始按照迹网状模型的路径进行遍历,生成所有具有有效路径的进程通信序列;最后通过进程通信行为规则对所生成的进程通信序列进行判断,排除无效的序列,最后得到符合规则的有效进程行为解释序列。下面给出了算法及规则的详细描述。

(1) 算法输入。

进程行为记录($a.pbr_1, b.pcbr_2, \dots$)和进程数目 n 。

(2) 算法输出。

符合规则的能够被合理解释的进程通信序列集合 M 。

(3) 算法步骤。

步骤1:根据同一进程中行为记录的时间顺序,对同一进程的行为记录进行排序连接,形成进程行为序列 $< a.pbr_1, a.pcbr_1, a.pbr_2, \dots >$ 。并依次生成所有进程的行为序列。

步骤2:检查所有进程行为序列中是否存在通信操作记录 $pcbr(t, ps, pd, e)$,如果有则将通信操作前的发起方进程 ps 状态节点链接到接收方进程 pd 的最接近 t 时刻的状态节点上。重复步骤2直到所

有的通信连接被标识。

步骤3:选取某一进程作为起始进程,根据进程通信状态图遍历所有路径分支,生成该进程的进程通信序列。重复步骤3直到所有进程都被遍历。

程序的伪代码如下:

```
i = 0;
while( i < n ) // 检查所有进程
{
    a = 未被检查过的进程;
    p = a;
    while( a 中的所有通信序列没有都被枚举过 )
    {
        while( 记录 p.r 不是该进程 p 最后一条记录 )
        {
            if( p.r 是通信操作记录 )
                从两条路径中选择一条没被标记的路径,
                做标记并 p.r 入栈操作;
            else 做标记并 p.r 入栈操作;
        }
        从栈底到栈顶顺序输出记录到 x;
        if( 进程通信序列 x 中包含两个或者 \
            多个进程的行为记录 )
            加入进程通信序列集合 M;
        else 舍弃 x;
        while( p.r 以下的路径都已被标记 )
            p.r = 出栈;
    }
    i ++; // 下一个进程
}
```

步骤4:根据进程通信行为规则对集合M中的进程通信序列进行合理的解释,并排除不可能的或者无意义的进程通讯序列。剩下的集合M中包含的序列既是作为进程入侵证据。

4 进程通信行为规则

根据进程入侵行为的特点和进程间通信的基本准则建立了进程通信行为规则,使得进程行为解释算法可以通过该规则来判断进程通信序列是否符合潜在入侵证据的条件。同时对每条规则进行了相应的形式化描述,为规则提供严密的语法论证,从而增强所获取的计算机证据可信性和严密性。

4.1 规则的语言描述

规则1:通信双方使用的通信协议应该相同。例如:通信双方都使用信号量、互斥量、临界区、管道等进程同步中的一种方式。

规则2:通信请求事件与通信响应事件应当成

对出现才算完成整个通信过程,所以下面两种情况不符合通信顺序逻辑:(1)两个进程进行通信时出现只有通信请求事件而无通信响应。(2)两个进程进行通信时出现只有通信响应事件而无通信请求。

规则3:在三方或以上通信序列中,只有出现特殊的通信方式(如代码植入技术,远程线程等),通信序列才算有效。

规则4:在通信序列中必须有一个或者多个犯罪行为(如对私密数据的操作,对计算机系统的破坏等)。

规则5:被定义为安全行为的进程所产生的进程通信序列中存在入侵行为的,应当予以排除。

4.2 规则的形式化定义

数据集定义:

$A \subseteq D \times D$ 是进程通信序列生成函数;

$\text{Number}(P_i)$ 是计算进程通信序列中进程数的函数。

$\text{Method}(pcbr.e)$ 是获取通信事件中使用的通信方式的函数。

D 是进程行为序列集合。

E 是进程通信方式集合。

E' 是进程特殊通信方式集合。

R 是进程危害行为集合。

Q 是安全进程集合。

B 是进程通信序列的集合,用来表示进程入侵的潜在证据集合, $B \subseteq D \times D$ 。

P 是不同进程行为序列的组合,用来表示进程通信序列。

进程行为序列集合 $D = \{pbs_0, pbs_1, pbs_2, \dots, pbs_m\}$ 。

进程通信方式集合 $E = \{e_0, e_1, e_2, \dots, e_n\}$ 。

进程特殊通信方式集合 $E' = \{e'_0, e'_1, e'_2, \dots, e'_n\}$ 。

进程危害行为集合 $R = \{r_0, r_1, r_2, \dots, r_i\}$ 。

安全进程集合 $Q = \{q_0, q_1, q_2, \dots, q_i\}$ 。

进程通信序列集合 $B = \{P_0, P_1, P_2, \dots, P_i\}$ 。

进程通信序列 $P = \{pbr_0, pbr_1, pbr_2, \dots, pbr_i, pcbr_0, pcbr_1, pcbr_2, \dots, pcbr_j\}$ 。

规则1': $\exists P_i \subset B, \forall pcbr_j \in P_i$, 若 $\text{Method}(pcbr_j.e) \in E$, 则 P_i 有效。

规则2': $\exists P_i \subset B, \forall pcbr_m \in P_i, \forall pcbr_n \in P_i (m \neq n)$, 若 $\text{Method}(pcbr_m.e) = \text{Method}(pcbr_n.e) \in E \wedge pcbr_m.p_s \neq pcbr_n.p_s$, 则 P_i 有效。

规则3': $\exists P_i \subset B, \exists pcbr_j \in P_i$, 当 $\text{Number}(P_i) > 2$ 且 $\text{Method}(pcbr_j.e) \in E'$ 时, 则 P_i 有效。

≥ 3 时, $p_{cbr_j} \in E'$, 则 P_i 有效。

规则 4': $\exists P_i \subset B, \exists p_{br_j} \in P_i$, 若 $p_{br_j}, e \in R$, 则 P_i 有效。

规则 5': $\exists P_i \subset B, p_{br_0} \in P_i, \exists p_{br_m} \in P_i, \exists p_{cbr_n} \in P_i$, 若 $p_{br_0}, p \in Q \wedge (p_{br_m} \in R \vee p_{cbr_n} \in E')$, 则 P_i 无效。

5 案例分析

5.1 案例说明

某用户通过网上银行对其账户中的现金进行支配使用,但是之后出现使用用户名和密码无法登陆的情况,随后到银行查询后,发现该账户的现金在网络银行交易系统中被转账,随后调查人员在该用户使用的计算机上进行调查取证,根据计算机中系统进程记录工具的对进程的记录,了解到系统中进程行为记录如表 1 所示。

表 1 进程行为记录

时间	进程	事件	事件序号
10:25:21	a	创建进程	a_1
10:26:43	a	向进程 b 内写入代码, 并执行	a_2
10:27:11	a	等待 b 中代码执行	a_3
10:28:18	a	进程自我终止	a_4
10:20:22	b	创建进程	b_1
10:28:17	b	向进程 a 发送代码成功运行的信号	b_2
10:28:28	b	写注册表, 加入开机自启动	b_3
10:28:20	b	运用消息机制, 获得进程 c 密码框里的密码	b_4
10:28:24	b	用户名, 密码等关键信息写入文件 t	b_5
10:28:59	b	通过电子邮件将文件 t 发送至账户 x	b_6
11:12:02	b	进程终止	b_7
08:09:20	c	创建进程	c_1
10:28:21	c	发送密码至进程 b	c_2
11:12:21	c	用户终止进程	c_3

5.2 证据的形式化及 CSP 建模

根据以上对系统进程相关操作的描述记录, 将同一进程的行为记录根据时间先后顺序连接, 构成进程行为序列 $a = < a_1, a_2, a_3, a_4 >$, $b = < b_1, b_2, b_3, b_4, b_5, b_6, b_7 >$, $c = < c_1, c_2, c_3 >$, 使用进程前缀表示法表示为: $a = a_1 \rightarrow a_2 \rightarrow a_3, b = b_1 \rightarrow b_2 \rightarrow b_3 \rightarrow b_4 \rightarrow b_5 \rightarrow b_6 \rightarrow b_7, c = c_1 \rightarrow c_2 \rightarrow c_3$ 。其中包含通信事件 a_2, b_2, b_4, c_2 , 根据这 4 个通信事件链接 a, b, c 三个进程, 形成完整的进程通信网络, 根据案件建立的进程通信状态图如图 2 所示, 通过搜索进程通信路径查找进

程通信序列并使用前缀表示法表示, 如表 2 所示。

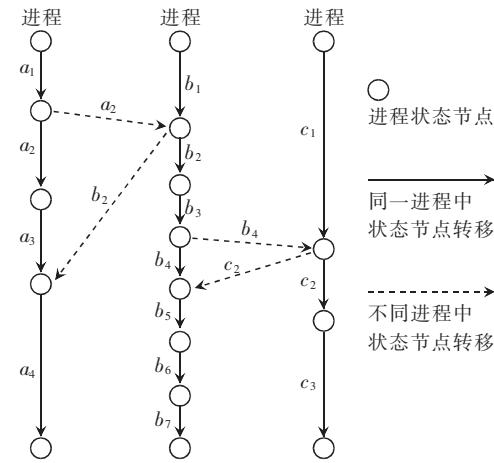


图 2 进程通信状态图

表 2 进程通信序列的 CSP 前缀表示法

序列序号	起始进程	进程通信序列
1	a	$a_1 \rightarrow a_2 \rightarrow b_2 \rightarrow b_3 \rightarrow b_4 \rightarrow b_5 \rightarrow b_6 \rightarrow b_7$
2	a	$a_1 \rightarrow a_2 \rightarrow b_2 \rightarrow a_4$
3	a	$a_1 \rightarrow a_2 \rightarrow b_2 \rightarrow b_3 \rightarrow b_4 \rightarrow c_2 \rightarrow c_3$
4	a	$a_1 \rightarrow a_2 \rightarrow b_2 \rightarrow b_3 \rightarrow b_4 \rightarrow c_2 \rightarrow b_5 \rightarrow b_6 \rightarrow b_7$
5	b	$b_1 \rightarrow b_2 \rightarrow a_4$
6	b	$b_1 \rightarrow b_2 \rightarrow b_3 \rightarrow b_4 \rightarrow c_2 \rightarrow c_3$
7	b	$b_1 \rightarrow b_2 \rightarrow b_3 \rightarrow b_4 \rightarrow c_2 \rightarrow b_5 \rightarrow b_6 \rightarrow b_7$
8	c	$c_1 \rightarrow c_2 \rightarrow b_5 \rightarrow b_6 \rightarrow b_7$

5.3 进程行为分析解释

分析序列 1,5,8: 由于进程 b 与进程 c 进行通信时进程 c 未对通信行为 b_4 做响应, 所以根据通信行为规则 2 排除序列 1, 同理可以排除序列 5, 8。

分析序列 2: 进程通信序列 3 中没有出现对计算机具有威胁的行为, 根据通信行为规则 4 排除序列 2。

分析序列 6,7: 起始进程为 b 的进程通信序列中出现了入侵行为, 但是由于进程 b 被定义为安全行为进程, 如果没有其他进程入侵该安全进程的话, 则无法解释安全进程出现入侵行为, 所以根据规则 5 排除序列 6, 同理排除 7。

分析序列 3,4: 进程通信序列 3 和 4 符合所有规则, 可通过调查人员分析判断, 序列 4 有效的说明的整个入侵的过程, 该序列即是所要求得的进程入侵证据。根据该入侵序列的起始进程为 a, 则可通过查找进程 a 的源头, 查到入侵者的相关信息。

5.4 原型系统实现

在 CSP 模型形式化取证方法的理论基础之上, 在 Windows 平台上实现了进程行为取证工具。该原

型系统通过读取监控软件的事件记录数据库,对进程相关数据进行归类、分析,并对进程操作进行形式化,使用进程行为解释算法构建进程通信序列,最后通过形式化的进程通信规则对进程通信序列进行筛选,得到图3所示的潜在入侵证据。

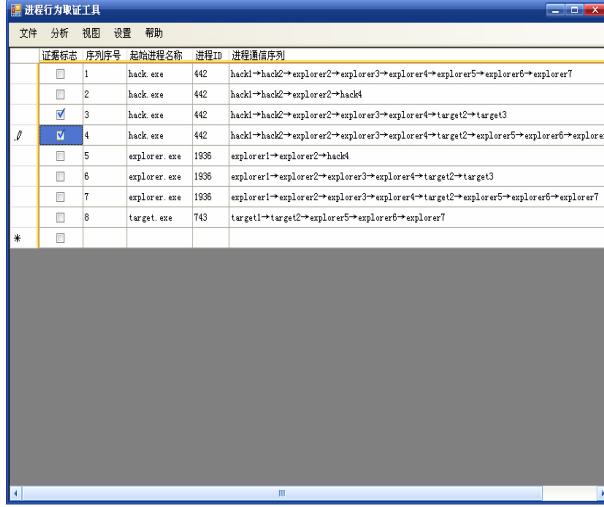


图3 进程行为取证分析结果

6 结束语

对计算机取证结果有效性和可信性加以研究是一个复杂的系统工程。文章从操作系统中软件进程行为角度出发,应用CSP理论建立取证模型,为分析解释进程行为做了较为完整的形式化描述,针对所描述系统的实时性加入了时间概念,使得模型能够真实地反应系统事件。提出了使用规则筛选的进程行为解释算法,通过对案例加以分析,给出了分析结果,验证了该行为解释算法的取证效果。

参考文献:

- [1] TEELINK S, ERBACHER R F. Improving the computer forensic analysis process through visualization [J]. Communications of the ACM, 2006, 49(2): 71-75.

- [2] GLADYSHEV P. Finite state machine analysis of a blackmail investigation [J]. International Journal of Digital Evidence, 2005, 4(1): 1-13.
- [3] GLADYSHEV P, PATEL A. Finite state machine approach to digital event reconstruction [J]. Digital Investigation, 2004, 1(2): 130-149.
- [4] XU D, PHILBERT N, LIU Z T, et al. Towards formalizing UML activity diagrams in CSP [C]// Proc of IEEE International Symposium on Computer Science and Computational Technology. 2008: 450-453.
- [5] BAO T, LIU S F, YAN F, et al. Network data collection formal analysis based on communication sequential process [C]// Proc of IEEE 11th International Conference on Computer Supported Cooperative Work in Design. 2007: 42-46.
- [6] SAKELLARIOU I, VLAHAVASA I, FUTOB I, et al. Communicating sequential processes for distributed constraint satisfaction [J]. Information Sciences, 2006, 176(5): 490-521.
- [7] HOARE C A R. Communicating sequential processes [M]. New Jersey: Prentice Hall International, 1985.
- [8] 韩志耕,罗军舟,王良民.不可否认协议分析的增广CSP方法 [J].通信学报,2008,29(10):8-18.
HAN Z G, LUO J Z, WANG L M. Extended-CSP based analysis of non-repudiation protocols [J]. Journal on Communications, 2008, 29(10):8-18.

作者简介:



孙国梓(1972-),男,安徽天长人。南京邮电大学计算机学院副教授,博士。主要研究方向为计算机取证技术,计算机通信网与安全。

俞超(1983-),男,江苏宜兴人。南京邮电大学计算机学院硕士研究生。主要研究方向为计算机取证技术。

陈丹伟(1970-),男,陕西商洛人。南京邮电大学计算机学院副教授,博士。主要研究方向为计算机通信网与安全,计算机取证技术。

基于 RSSI 的无线传感器网络环境参数分析与修正方案

凡高娟, 王汝传, 孙力娟

(南京邮电大学 计算机学院, 江苏南京 210046)

摘要:定位技术是无线传感器网络数据采集的基础服务,而定位精度的高低在很大程度上取决于距离测量的精度。基于 RSSI(接收信号强度)测距技术无须添加任何硬件设施、用较少的通信开销和较低的实现复杂度,十分适应于能量受限的无线传感器网络。通过对 RSSI 测距模型进行分析,并提出一种针对室内环境的参数修正方案。通过自行研发的传感器节点 UbiCell 上进行验证分析,实验表明,采用环境参数修正方案后,明显提高了测距的精度。

关键词:无线传感器网络;接收信号强度;距离估计

中图分类号:TN911 文献标识码:A 文章编号:1673-5439(2009)06-0054-04

Distance-assistant Node Coverage Identification Model for Wireless Sensor Networks

FAN Gao-juan, WANG Ru-chuan, SUN Li-juan

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: Localization is one of the basic services for data collection in wireless sensor networks. The localization accuracy to some extent depends on the accuracy of distance estimation. The advantage of employing the RSSI(Received signal strength indications)-based distance estimation have some metric that no extra hardware and fewer communication cost and lower implementation complex which can adapt to energy-limit wireless sensor network. In this paper, we studied and analyzed RSSI-based distance estimated model, and provide a parameter calibration method which in specific to indoor environmental applications. Experience on our prototype node—UbiCell and shows that the proposed calibration method performs well in distance estimate accuracy.

Key words: wireless sensor networks; received signal strength indication; distance estimation

0 引言

无线传感器网络由大量感知节点组成,是以数据为中心的采集平台,可广泛用于环境监测、应急通信、健康护理、安全监控等应用领域^[1]。在这些应用中,事件发生的位置及其获取信息节点的位置是无线传感器网络应用的基础。如何对发生事件或目标节点的定位成为当前研究热点之一。基于距离的定位必须首先测量节点间的距离,当前常用测量距

离的方法有 GPS^[2]、红外线^[3]、超声波^[4] 和 RS-SI^[5-6]等。由于无线传感器网络应用于一些无人可达领域,节点配备 GPS 方案代价太高,其红外线、超声波都需要额外的硬件设备支持,增加了节点的硬件成本。

RSSI 测距方法是利用电磁波在传输过程中^[7],接收信号功率与传输距离存在着某种变化关系,从而根据特定环境推导出这一特定关系,实现实时系统定位。利用 RSSI 定位不需要增加任何额外的硬

收稿日期:2009-06-30

基金项目:国家自然科学基金(60573141, 60773041, 60973139), 江苏省自然科学基金(BK2008451), 国家高技术研究发展计划(863 计划)(2006AA01Z219, 2007AA01Z404, 2007AA01Z478), 2006 江苏省软件专项, 南京市高科技项目(2007 软资 106), 现代通信国家重点实验室基金(9140C1105040805), 江苏省博士后基金(0801019C), 江苏高校科技创新计划(CX08B-085Z, CX08B-086Z) 和江苏省六大高峰人才资助项目

通讯作者:王汝传 电话:(025)83492867 E-mail:wangrc@njupt.edu.cn

件设备,成本低廉,只需较少的通信开销和较低的实现复杂度,这在能量有限的传感器网络是非常重要的。由于 RF 信号受环境因素影响很大^[8],在其应用中,受到地板、墙壁和人体等各种物体等障碍物的阻拦,电磁波会存在着反射、绕射及衍射,使得 RSSI 值随机变化较大,无法利用 RSSI 来获得节点间的准确距离。因此,对环境关系的设定是达到精确定位的基础。

在无线传感器网络定位应用中,RSSI 受环境的影响很大。相同的节点对相同相对位置,在不同的环境下,其 RSSI 可能有很大的不同。此外,在同一环境下,在不同的区域或节点不同的方向,尽管距离相同,其 RSSI 值也可能会不相同。所以在基于 RSSI 定位技术中,首先要对环境因素进行分析。

虽然基于 RSSI 环境设定都有很多研究,如文献[9—10]根据曲线拟合法求得接收功率与距离之间的关系,但没有考虑到室内无线信号的传输受多种因素影响,不能真实表达出环境因素影响下的真实距离。根据实际环境进行采样,建立环境数据库,这样虽然提高了距离估计的精度,但这样占用空间大,且没有考虑到单个数据值不能反映出环境的变化。本文通过对无线传播模型进行分析,并根据采集的数据进行分析与研究,提出一种基于 RSSI 的无线传感器网络环境参数修正方法。

1 无线信道模型

无线传感器网络中常采用的无线传播模型主要有 3 种^[11—12]:Free-Space 模型、Two-Ray Ground Reflection 模型和 Shadowing 模型。其中,Shadowing 模型充分考虑了其环境因素的变化情况。其 Friis 自由空间方程为

$$\overline{PL(d_0)}[\text{dB}] = -10\log_{10}\frac{G_t G_r \lambda^2}{(4\pi)^2 d_0^2 L} \quad (1)$$

其中, G_t, G_r 分别为发射、接收天线的增益, λ 为电磁波在自由空间传播时的波长, d_0 为收发节点天线间的距离, $L(L \geq 1)$ 为系统损失。通常情况下, $G_t = G_r = 1, L = 1$ 。

现实应用中,在一定的距离下接收到的信号强度是一个随机量,信号通过多径传播,产生路径损失,其路径损失表达为:

$$PL(d)[\text{dB}] = \overline{PL(d_0)} + 10\eta\log_{10}(d/d_0) \quad (2)$$

其中, d 为发射节点与接收节点之间的距离, d_0 表示发射节点和参考节点之间的距离,一般取 1 m。 η

为路径衰减因子,一般取值为 2~5。 $\overline{PL(d_0)}$ 为距离 d_0 处信号强度的测量值。

那么其节点经过路径损失后,得到的接收功率 P_r 为:

$$P_r = P_t - PL(d) \quad (3)$$

其中, P_t 为发射节点的发送信号功率。

然而,在实际应用环境中,由于多径、绕射、障碍物等因素,用对数-常态分布模型将更加合理,节点收到信号时的路径损耗为:

$$PL(d)[\text{dB}] = \overline{PL(d_0)} + 10\eta\log_{10}(d/d_0) + X_\sigma \quad (4)$$

其中, $PL(d)$ 为经过距离 d 后的路径损耗, X_σ 为平均值为 0 的高斯随机变量,其标准差一般为 4~10。

在此情况下,距离 d 后的接收功率为:

$$P_r(d)[\text{dB}] = P_t[\text{dB}] - PL(d)[\text{dB}] = \\ P_t - (\overline{PL(d_0)} + 10\eta\log_{10}(d/d_0) + X_\sigma) \quad (5)$$

在实际定位应用过程中,首先需要估算网络运行环境下的路径衰减因子 η 和阴影 X_σ ,一般是多次测量取均值的方法得到这两个参数。在第 2 节中对通过数据进行分析,并针对当前方法中的不足之处,提出一种网络环境参数修正模型。

2 基于 RSSI 的无线传感器网络环境参数分析与修正

2.1 实验环境及参数

本文的测试程序是基于 MantisOS 系统,通过我们自行研发的 Ubicell^[13] 节点为实验平台,该平台包括 ATmega128 处理器/CC1000 通信模块组成的 Ubicell 节点若干,及 MIB510 型号的编程板组成。通过 MantisOS 对 Ubicell 节点烧写收/发程序,基站不但负责数据信息的收发,还将接收到的信息通过串口显示在主机上。其应用环境和节点参数如图 1 所示。

Radio frequency	903.845 MHz
Time	2 s
Transmission power	-9 dBm
Antenna	Omni-antenna

图 1 Ubicell 节点参数及部署环境

为了充分体现无线传感器网络具体应用环境,我们通过对室外、室内环境下接收信号进行采集,分析影响节点接收信号的环境参数,并通过数据分析,得出适合具体环境的参数。

2.2 网络环境建立步骤

基于 RSSI 的无线传感器网络环境参数需要通过对数据进行采集,并在此基础上分析之间关系,得到具体的因子,并对未知接收信号进行距离估计。其建立步骤见下。

2.2.1 RSSI 信息采集与修正

从 CC1000 接收的数据包中读出的 RSSI 值是芯片寄存器的值,需要通过 A/D 转换为接收节点的 RF 管脚的功耗值。其方法如式(6)、式(7)所示。

$$V_{\text{RSSI}} [V] = \text{ADC} \times V_{\text{Bat}} / 1024 \quad (6)$$

$$P_r [\text{dBm}] = -50 \times 3V_{\text{RSSI}} - 45.5 \quad (7)$$

其中, V_{RSSI} 是 A/D 转换前的 RSSI 电压值,单位为 V, V_{Bat} 为电源电压,该值为一个常数 3 V。

对 RSSI 的采集分为室外环境及室内环境下数据的采集。

室外环境是在空旷的草坪上进行,采用两个 Ubicell 节点放置高度为 0.5 m 的方凳上。固定 1 号节点为接收信息节点,2 号节点为发射节点,每隔 2 m 向 1 号节点发送信息,并移动 2 号节点,在距离 1 号节点 1 m, 2 m, ..., 11 m 等多个位置,每次取 100 个数据包。其不同距离下测量到室外 RSSI 值如图 2 所示。

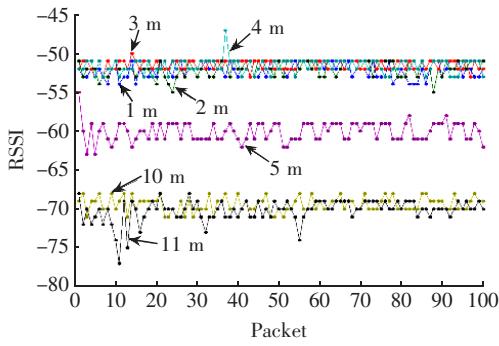


图 2 室外 RSSI 数据

在室外环境中,在距离一定的情况下,其 RSSI 值在一定范围内跳动,而并不是一个固定的值,且随着距离的增大,其 RSSI 变小。从图 2 中可以看出,根据不同位置下的测量信号强度,可以得到各组 RSSI $\sim d$ 之间的数据对。通过式(2),可以在不同距离下的路径损失因子:

$$\eta_i = \frac{P_t - PL(d_0) - RSSI_i}{10 \log_{10}(d_i/d_0)} \quad (8)$$

其中, $RSSI_i$ 表示在 d_i 距离下的接收信号强度值。

根据式(8)得到如表 1 中不同参数距离下的 η_i 。

表 1 不同参考距离下的 η_i

d_i/m	η_i
2	3.822 3
3	2.296 3
4	1.834 8
5	2.786 4
10	2.883 6
11	2.845 8

通过表 1 可以看到,不同的参考距离求得的路径损失因子并不相同,具有一定的差异,其差异部分是由于节点之间的随机噪声,当拿这些值作为参数来估计距离时,会产生很大的误差,所以我们提出一种修正模型求得 η :

$$\eta = \left| \frac{\sum_{i=1}^N P_t - PL(d_0) - RSSI_i}{\sum_{i=1}^N 10 \log_{10}(d_i/d_0)} \right| \quad (9)$$

N 表示采集的距离个数,用修正模型求得 $\eta = 2.744 9$ 。

其室内环境同图 1,数据采集规则与室外相同,通过对 1 ~ 7 m 距离下数据采集,在采集过程中,同时记录在改变节点方向及当有障碍物进入感知区域时的 RSSI 值,其数据如图 3 所示。

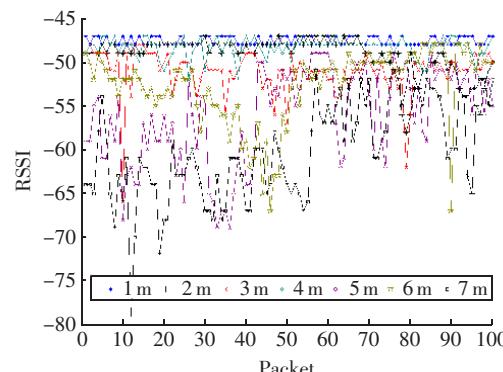


图 3 室内 RSSI 数据

从图 3 中可以看出,受室内环境因素影响,其 RSSI 值呈现一种不确定分布,当发射节点与接收节点距离比较近时,RSSI 值比较稳定,但随着距离的增加,其波动性比较大。

根据式(5)及式(8),可以求出其阴影参数为:

$$X_{\sigma i} = -P_t + PL(d_i) + RSSI_i \quad (10)$$

根据式(10),可得出不同参考距离下的 $X_{\sigma i}$ 如表2所示。

表2 不同参考距离下的 $X_{\sigma i}$

d_i/m	$X_{\sigma i}$
2	0.1763
3	2.4902
4	8.7496
5	3.6497
6	9.2432
7	3.9708

为避免距离估计中产生误差,针对 X_{σ} ,提出的修正模型为

$$X_{\sigma} = \frac{\sum_{i=1}^N (-P_t + PL(d_i) + RSSI_i)}{\sum_{i=1}^N i} \quad (11)$$

同理,用修正模型求得 $X_{\sigma} = 4.7134$ 。通过这两个参数修正,就可以根据接收信号强度估算出节点之间的距离。

2.2.2 距离估计

通过对上述室内外模型评估与参数修正,根据 RSSI 值,得到对未知距离节点的距离估算公式为:

$$d = d_0 \times 10^{\frac{-P_t - PL(d) - X_{\sigma} + RSSI}{10\eta}} \quad (12)$$

3 实验分析与验证

为了验证本修正模型的有效性,本节对数据进行重新采集,分别用修正前与修正后两种方案根据接收到的信息对距离进行估计,从而对两种情况下的估计距离及误差进行分析与比较。

图4描述了真实距离与估计距离之间的关系。从图4中可以看出,相对于修正前的方案,采用修正后的网络环境参数后,其估计的距离更接近于真实距离。

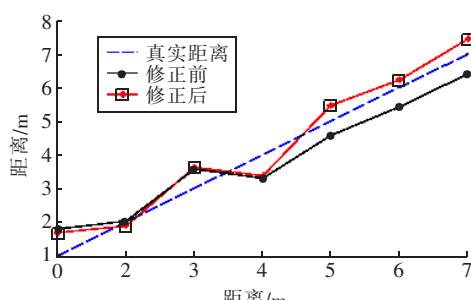


图4 真实距离与估计距离

图5描述基于 RSSI 的无线传感器网络环境参

数修正前与修正后的测距误差情况。从图5中可以看出,在1~7 m范围内,其修正前的最大测距误差为0.7709 m,最小测距误差为0.1314 m。使用修正后的方案,其最大测距误差为0.6638 m,最小误差为0.0225 m,明显提高了测距的精度。

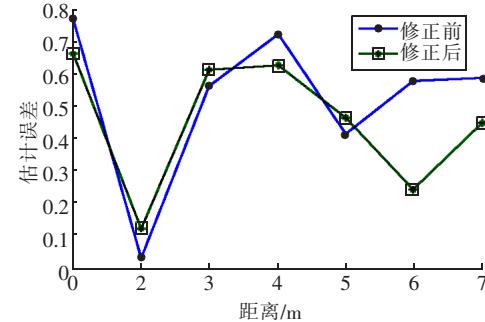


图5 两种情况下的测距误差

4 结论

本文对无线信道模型进行分析及当前 RSSI 测距存在问题的基础上,提出一种针对具体应用的无线传感器网络环境参数修正方案,以提高 RSSI 测距的精度,为高精确定位打下基础。实验结果表明,其修正后的方案明显提高了节点距离估计的精度,此修正方案可用于不同环境下基于 RSSI 定位应用中。

参考文献:

- [1] AKYILDIZ I F, SU W, SANKARASUBDAM Y, et al. Wireless Sensor Networks: A Survey [J]. Comp Networks, 2002, 38(4):393-422.
- [2] AZUMA R. Tracking Requirements for Augmented Reality [J]. Communication of the ACM, 1993, 36(7):50-51.
- [3] BULUSU N, HEIDEMANN J, ESTRIN D. GPS-less Low Cost Outdoor Localization For Very Small Devices [J]. IEEE Personal Communications Magazine, 2000, 7(5):28-34.
- [4] GIROD L, EST RIN D. Robust Range Estimation Using Acoustic and Multimodal Sensing [C]// Proc of the IEEE/ RSJ Int'l Conf on Intelligent Robots and Systems. Maui: IEEE Robotics and Automation Society, 2001, 3:1312-1320.
- [5] GIROD L, BYCHOVSKIY V, ELSON J, et al. Locating Tiny Sensors in Time and Space: A Case Study [C]// WERNER B. Proc of the 2002 IEEE Int'l Conf on Computer Design: VLSI in Computers and Processors. Freiburg: IEEE Computer Society, 2002:214-219.
- [6] ZHOU G, HE T, KRISHNAMURTHY S, et al. Models and solutions for radio irregularity in wireless sensor networks [J]. ACM Trans on Sensor Networks, 2006, 2(2):221-262.

(下转第63页)

基于鲁棒性的移动自组织网络路由选择算法

徐占洋^{1,2}, 张顺颐¹

(1. 南京邮电大学 信息网络技术研究所, 江苏南京 210003
2. 南京信息工程大学 计算机系, 江苏南京 210044)

摘要: 移动 Ad hoc 网络中传输流时, 通常受到节点的移动性引起的路由中断影响。当传输实时数据流时, 必须提供具有鲁棒性的路由。提出了基于鲁棒吞吐量的路由选择方法进行路由选择, 可以极大地提高 MANET 网络系统的吞吐量速率。按需鲁棒路由算法(distributed on-demand routing and flow admission, DRFA) 方案选择性的发现路由, 保证准备传输的流或文件不中断(包括不重新选择路由)的传输。

关键词: MANET; 吞吐量; 鲁棒性; 路由选择

中图分类号: TP393 文献标识码: A 文章编号: 1673-5439(2009)06-0058-06

Route Selection Based on Robustness for Mobile Ad hoc Network

XU Zhan-yang^{1,2}, ZHANG Shun-yi¹

(1. Institute of Information Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Department of Computer Sciences, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Flows transported across mobile ad hoc wireless network (MANET) were often suffered from the route breakups, which was caused by the nodal mobility. The robust route should be provided for the real time flow transported in MANET. In the paper, in order to improve the robust throughput of the MANET, the route selection scheme based on robustness was proposed to select the route. The robust route selection DRFA (distributed on-demand routing and flow admission) was used to select the route to guarantee the flows (or file date) transported without the breakups (or without re-selecting the route).

Key words: MANET; throughput; robustness; route selection

0 引言

无线移动 Ad hoc 网络 MANETs (mobile ad hoc networks) 是一种无固定设施的、复杂的分布式系统, 由无线自由移动的节点组成。可以广泛应用于作战、救灾等紧急情况。无线自组织网络与有线网络、蜂窝无线网络相比具有动态拓扑、链路带宽受限、容量时变等特点。

由于移动 Ad hoc 无线网络自身的这种特点, 一个移动 Ad hoc 无线网络的传输性能由它的传输链路的状态决定, 而它的链路状态具有一个动态随机过程的特性(包括与链路性能相关的噪声干扰、数

据速率、无线信号传输范围), 且受到节点的特性(如节点的移动模型和资源状态)、网络拓扑结构的基本图连通性以及由传输负载进程要求的 QoS(服务质量)目标影响。在典型的按需 Ad hoc 路由选择算法下, 一个源节点发起一个路由发现进程来查找到目标节点的路由, 位于所选择的路由上的节点使用路由转发表去转发由源节点流产生的分组。而在主动路由选择算法里, 所有节点周期性的交换链路状态数据, 且在其路由表中保存到网络中所有目标节点的转发目录。尽管这两种方案被证明在许多情况下, 具有很好的效果, 但在这两种情况下, 路由的鲁棒性并没有作为它的选择而考虑^[1-2]。而在移动

ad hoc 网络中,由于节点的移动和/或节点和链路故障以及通信传输质量的波动,路由中断经常发生。结果,用于传输一个文件或一组分组的流路由在完成该流传输之前而被中断。

当一个活动的路由被中断,无论是按需路由或主动路由算法都开始去发现一个替代路由。在源节点发现新的至目标节点的路由期间,累积在前路由的节点(这些节点的转发链路已经失败)缓存里的分组可能失效,因为产生的排队时延可能对于基本流的有效传输来说高的不可接受。对于有些应用产生的程序(或文件),需要按照严格实时方法对其作为一个整体通过网络进行传输。因此,对这些传输,经常需要确保按照完整的流,即在传输路由没有中断(包括重新查找路由)的情况下传输。

在本文中,研究一种路由选择算法,为被接纳的流提供可接受的 QoS 水平。路由选择中,每个节点的选择的主要依据是该节点的鲁棒性能,该鲁棒性能最大概率的保证传输的流(如程序或文件)完整地在一个路由中完成传输。该算法既可以在按需路由选择方案中使用,也可以在主动路由发现方案中应用。

1 系统鲁棒吞吐量计算

现假设随机变量 N_t 表示的是在时间 t 时刻,存在于系统中的会话数目。另外假设在时间 T_n 到来之前,第 n 个会话包括在系统里,且在时间 T_n 处,该会话终止,离开了系统。对于该会话,离开的原因通常有两个:要么是传输被中断;要么是因为该会话成功完成传输后的离开。为了更好的度量系统性能,我们假定,一旦一个会话离开,则提供给该会话(或流)的服务参数可以充分确定。而离开的会话事件代表一个与其相关联的传输任务成功完成(如果该会话传输的是一个程序文件至目标节点,则只有在它离开时整个程序文件已经准确传递完毕,这时才可以认为成功完成传输),据此可以提供一个完全信任的系统,表明该系统成功完成基本的传输任务。当传输的数据文件只有部分完成传输时,其会话保持的周期时间很短,然后被丢弃,在此情况下,任何后来消息传输(该消息属于基本的原始程序或文档)的恢复必须涉及重传在失败前已经发送的部分数据。如果发生这种情况的原因,是一个由中断事件触发的会话离开(比如,节点的移动性或链路传输的退化,或节点失败产生的导致沿该会话路由出

现的故障从而引起的中断事件),则与这个会话相关的基本传输的任务可以认为已经失败。所以,对于传输行为只是部分完成的会话,可以没有任何回报值(回报)授予该系统。当然,有时按照成功传输要求的条件定义,部分回报值也是合适的。

设 $N = \{N_t; t \geq 0\}$ 是系统的会话窗口尺寸的过程,它是一个在状态空间 $E = \{0, 1, 2, \dots\}$ 上的随机过程,表示的是在时间上的系统支持的流(或会话)总数的变量。相应地,设随机变量 R_n 表示,在时间 T_n 处,对于正离开的会话 n ,系统获得的回报(rewards)。相应的回报过程可以表示为 $R = (R_n; n = 1, 2, \dots)$, 该回报值假设在区间 $[0, +\infty)$ 内。参见图 1, N_t 和 R_n 之间的关系^[3]。

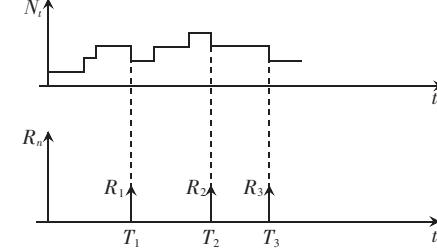


图 1 网络系统中会话的随机过程及其离开后的相关回报
在运行时间为 $[0, t]$ 上,网络系统获得的整个回

报可以用随机变量 $\eta = \sum_{n=1}^{N_t^d} R_n$ 来表示,这里的 N_t^d 表示在时间周期 $(0, t]$ 上,发生地会话离开的全部个数。则系统在每个单位时间获得的平均回报表示为 ξ ,可以用它来指示一个 MANET 系统的运行效率。

$$\xi = \lim_{t \rightarrow \infty} \frac{1}{t} E(\eta_t) \quad (1)$$

在对 η 进一步分析前,使用以下定理:

Wald 等式定理^[4]: 设 $\{\tau_k\}$ 为存在期望的独立同分布序列,又设正整数随机变量 N 与 $\{\tau_k\}$ 的发展一致,意即对 $\forall n, \{N \leq n\}$ 只依赖 $\tau_1, \tau_2, \dots, \tau_n$ (从而与 τ_{n+1}, τ_{n+2} 独立),并且再设 $EN < \infty$,那么

$$E\left(\sum_{k=1}^N \tau_k\right) = EN \cdot E\tau_1 \quad (2)$$

由于获得的回报过程 $R = (R_n)$ 是一个独立同分布的随机变量序列,另该回报值为有限的均值,且会话离开点过程也是统计独立的。同时,回报过程 R 与计数过程相关(或更一般地可以假设变量 N_t^d 是一个与序列 R 有关的停止时间),因此,在 $(0, t]$ 上发生的离开平均数也是一个有限值,可以应用上述 Wald 等式定理,对式(1)分析,有:

$$\xi = \lim_{t \rightarrow \infty} \frac{1}{t} E\left(\sum_{n=1}^{N_t^d} R_n\right) = \lim_{t \rightarrow \infty} \frac{1}{t} E(N_t^d) E(R) \quad (3)$$

此时会话离开速率可被定义为流吞吐量,如果度量单位为 flows/s,则定义 λ_d 为网络系统的吞吐量,有:

$$\lambda_d = \lim_{t \rightarrow \infty} \frac{1}{t} E(N_t^d)$$

因此,

$$\xi = \lambda_d E(R) \quad (4)$$

由于在网络上传输的消息种类的差别,不同类型的流对网络系统要求也不尽相同,所以不同类型的流对应的会话离开后,系统所得到的回报也不同。为了对传输的不同类型流所获得的回报值进行比较,将网络系统中传输的流分为 K 类,即: $F_1, F_2, \dots, F_k, \dots, F_K$ 。对应 k 类型的流提供的负载流量(回报)表示为 $\lambda_0^k = \lambda_0 f(k)$ (单位为 flows/s),这里 $f(k) \in [0, 1]$ 表示所有 k 类流速率的规定相对载荷的能力,因此, $\sum_{k=1}^K f(k) = 1$, λ_0 (flows/s) 表示系统提供的整个流负载速率能力。

对于 k 类的流,用 $R_n(k)$ 表示由第 n 个离开的 k 类的流(在时间 $R_n(k)$,该时间表示对应的流离开的时间)获得的回报。相应的回报过程可以表示为 $R(k) = \{R_n(k), n \geq 1\}$ 。有关的回报随机变量均假定具有有限均值。变量 N_t^d 表示发生在 $(0, t]$ 上的离开的 k 类的流的数目,相应的计数过程表示为 $N_t^d(k) = \{N_t^d(k), t \geq 0\}$ 。则相应的网络系统长期获得的回报均值可以定义为:

$$\xi = \lim_{t \rightarrow \infty} \frac{1}{t} E\left(\sum_k \sum_{n=1}^{N_t^d(k)} R_n(k)\right) \quad (5)$$

而在 $(0, t]$ 时间上,网络系统对于所有 k 类型流提供的总的回报均值 ξ_k 为:

$$\xi_k = \lim_{t \rightarrow \infty} \frac{1}{t} E\left(\sum_{n=1}^{N_t^d(k)} R_n(k)\right) \quad (6)$$

由于该回报变量是独立同分布的随机变量,具有的均值是 $E(R(k))$,且计数变量 N_t^d 是一个与序列 $\{R(k)\}$ 有关的有限均值停止时间,根据 Wald 引理得到(注意到,类型 k 是一个有限值的数,因此,基本的期望值和求和的操作可以互换):

$$\xi = \lim_{t \rightarrow \infty} \frac{1}{t} E\left(\sum_k \sum_{n=1}^{N_t^d(k)} R_n(k)\right) = \sum_k \lambda_d(k) E(R(k)) \quad (7)$$

$$\xi_k = \lim_{t \rightarrow \infty} \frac{1}{t} E\left(\sum_{n=1}^{N_t^d(k)} R_n(k)\right) = \lambda_d(k) E(R(k)) \quad (8)$$

这里的 $\lambda_d(k) = \lim_{t \rightarrow \infty} E(N_t^d(k))/t$ 表示类型为 k 的流离开速率,单位为 flows/s。

每个节点在其活动期间,按照网络系统给定接纳控制方法对每个流实现接纳控制。在这个过程期间,源节点发起一个路由发现过程,如果没有检测到可接受的路由,那么,源节点锁定该流。

1.1 网络的吞吐量速率

为了更好的描述各种类型流的传输特性,现设 $F(k)$ 为稳态分布的流速率、 $H(k)$ 为每个流传输时会话的保持(或传输)时间,用以刻画 k 类型流的传输特性。

另假设, $S(k)$ 是一个随机变量,该变量表示 k 类型的流所选择路由的存活时间的分布,该路由是用于传输 k 类型的流。当检查一个待选路由时,通过考虑具有路由链路和节点生存期和移动模型的随机过程来计算它的生存时间的分布(该模型的计算方法详见文献[5])。

定义 $T(k)$ 是一个随机变量,该变量表示一个被接纳的 k 类型的流的会话通过所选择的路由的传输时间。如果路由是在会话预期完成传输时间 $H(k)$ 的结束之前被中断,可以认为该传输没有完成。因为 $S(k)$ 表示的是 k 类型的流的路由存活时间,如果 $S(k) \geq H(k)$,则该 k 类型的流的传输将成功完成。在本文中, $T(k) = H(k)$ 表示会话传输已经成功完成,相应地,当 $S(k) < H(k)$,则意味着会话传输被提前终止,所以得到:

$$T(k) = \min\{H(k), S(k)\} \quad (9)$$

网络系统吞吐量可以另外定义为:单位时间内,由网络传送到预期目的节点的信息单元的(有限的)平均值。为了利用 $F(k)$ 、 $H(k)$ 和 $T(k)$ 来计算网络系统的吞吐量速率,设 $M(k)$ 表示累积的信息单元总数(也被定义为所有涉及的传输进程获得的回报),这些信息是由 k 类会话在它的传输时间内成功地传送到它的目的节点。用变量 $M(k) = F(k)T(k)$ 表示由 k 类会话传输完成后获得的回报值。所以系统的吞吐量速率用 λ_d (单位为 flows/s)也可以表示为:

$$\lambda_d = \xi = \sum_k \lambda_d(k) E(F(k)T(k)) \quad (10)$$

1.2 网络的鲁棒吞吐量速率

按照上文所述,对于某个具体应用,最根本要求是成功实现流会话,并从信息的基本传输中获得充分受益。基于这种情况,如果一个会话在其通信期间中断或提前终止,都会明显地降低传送分组的回报值。任何这样的崩溃很可能会严重的降低完整传输数据文件的及时性。且在这种情况下很难及时计算分组的回报值。

通常的情况是传送到目的节点的信息单位(比如,消息或分组)的有效值是在传输会话不出现中断的情况下而获得的。比如,一个源节点传送一个可执行的计算机程序文件至目标节点,如果该流的传输进程过早地被终止,那么到达目的节点的部分分组将没有任何意义,尽管这部分分组部分传送已经成功完成。在这种情况下,由分组传送实现的网络吞吐量表示的仅是没有任何信任度的交易(事务)的一部分。对于此类应用,网络管理系统赋予它可以是完全的吞吐量信任(如果没有中断成功传输),或没有任何回报值(没有成功传输)。

因此,通过引入鲁棒吞吐量(RT)度量,用来说明由相关的移动节点从接收的完整或部分传输的交易中获得的收益。

设 $\delta(k) \in [0, 1]$ 是一个随机变量,它表示与 k 类型会话传输事件关联的部分成功情况。设信任度 $R(k)$ (一个 k 类型会话在它终止的传输上所获得的,依赖于它的成功传输等级): $R(k) = F(k) \cdot \delta(k)$ 。在这种方法里, $R(k)$ 指出了相关传输的消息比率(对于一个 k 类会话在终止的传输上获得的信任度)。连续成功的 k 类变量序列组成进程 $\delta(k) = \{\delta_n(k); n=1, 2, \dots\}$ 。

随机变量 $\delta(k)$ 的分布通常由变量 $H(k)$ 、 $L(k)$ 和指定定义的回报函数的分布确定。

定义与一个回报函数相关的有效因在作为系统的鲁棒吞吐量,单位为 bits/s。也就是:

$$f_r = \xi = \sum_{k=1}^K \lambda_d(k) E(F(k) \cdot T(k) \cdot \delta(k)) \quad (11)$$

鲁棒吞吐量速率(单位为 flows/s) 定义为:

$$\lambda_r = \sum_{k=1}^K \lambda_d(k) E(\delta(k)) \quad (12)$$

对于所有 n 和 k ,当 $\delta_n(k)=1$ 时,在路由的生存期内,每个会话对接收的所有已经传输的数据完全信任,是独立于会话终止的原因和被终止的会话的已经持续实现的时间。在这种情况下,鲁棒吞吐量 $f_r(\lambda_r)$ 就是网络系统的吞吐量水平 $f_r(\lambda_d)$ 。为了简化计算,本文考虑的主要基于流传输完成的鲁棒吞吐量,即在此条件下, $\delta_n(k)=1$ 。

1.3 移动节点的鲁棒性指标计算

设节点 X_i 是移动 ad hoc 网络中的一个活动节点,该节点具有的总的流吞吐量为 η_i ,另设 $\eta_i(k)$ 为节点 X_i 对应的 k 类型的流的吞吐量。所以有 $\eta_i = \sum_k \eta_i(k) \leq \xi_k$ 。按照前文可知,对于网络系统中的任意节点 i ,必有 $\eta_i \leq \xi$,同时对于任意类型 k 类的

流,必有 $\eta_i(k) \leq \xi_k$ 。因此,对于每个移动节点,其基于完成的鲁棒吞吐量由与其没有中断的(或重新发现路由)情况下完成的会话(或流)传输确定。

因此引入路由链路的吞吐量评价指标来估算候选路由的鲁棒性。

定义 SRVI(Source Robustness vulnerability Index) 为源节点至目标节点的路由链路脆弱性指标,该指标的定义如下:

节点 i, j 为两个相邻节点,且节点 i 为路由查找阶段的前一个节点(即 $j = (i+1), i=0, 1, 2, \dots$),其中, $i=0$ 时,表示目标节点为源节点自身, $N_{\text{hop}}(j)$ 同时也表示源节点至当前节点 j 的跳数。节点 j 执行以下的计算方法来计算从源节点至当前自身所累积的路由脆弱性指标 $SRVI(j)$:

$$SRVI(j) = (SRVI(i) \times N_{\text{hop}}(i) + \eta_j) \div N_{\text{hop}}(j) \quad (13)$$

相应的如果需要传输的是特定 k 类型流,则至节点 j 用于传输 k 类型流的链路累计的路由脆弱性指标 $SRVI_k(j)$ 为:

$$SRVI_k(j) = (SRVI_k(i) \times N_{\text{hop}}(i) + \eta_j(k)) \div N_{\text{hop}}(j) \quad (14)$$

如果节点 i 为源节点,则其是 $SRVI(i)$ 和 $SRVI_k(i)$ 均按照 0 来计算。该值越大,则意味着所查找的路由的鲁棒性越好。

2 鲁棒路由选择算法描述及性能分析

与其它网络系统不同,在移动 ad hoc 网络中,由于节点的移动经常会导致链路中断,且这种故障会引起路由的经常性中断。特别是当传输流用于实时应用时,这种故障的影响更大。所以期望实现一个路由发现机制去发现一个网络路由,通过该网络路由有足够的资源来保证一个鲁棒性目标水平(因此,路由的生存时间具有更高的概率足够长用来传输,而不发生早期的传输的中断)。在本文中主要考虑和计算的是分布式按需路由和流接纳方案 DRFA(distributed on-demand routing and flow admission) (该方案通过对 AODV^[6] 方案的扩展操作可以方便地实现),用以保证实现目标应用。

在路由发现期间,路由请求(RREQ)分组洪泛式的在整个移动 ad hoc 网络中广播(见文献[7—9])。使用 DRFA 方案,一个节点接收这样一个请求分组后,洪泛式转发到它的相邻节点(如果它没有转发过该分组)。

在接收到 RREQ 分组情况下,预期的目标节点在一定周期内等待并收集一些 RREQ 分组(如果有的话)。然后它进行检查被携带在这些分组头部的累积鲁棒性状态指标(与其它指标一样,比如遍历的路由跳数)。然后,目标节点然后选择一个最佳 RREQ 分组(该 RREQ 分组的 SRVI 指标在所有分组中,值最大,如果有相同的,就选择跳数最小者)。算法具体描述如下:

DRFA——处理 RREQ 分组

(1) 为了收集一些 RREQ 分组(如果有的话),等待一个规定的周期时间;

- (2) 检查接收到得 RREQ 分组;
- (3) 选择一个最佳 RREQ 分组;
- (4) 如果是目标节点,那么,产生 RREP 分组
- (5) 否则,修改 RREQ 分组;
- (6) 向前转发 RREQ 分组;
- (7) End。

图 2 给出了在 NS-2.31 系统下,MANET 系统由 60 个移动节点组成,发送的是一个应用程序文档。发送数据的移动节点从 4 个开始变化到 14 个,模拟时间为 1 200 s。分别使用 DRFA 方案和非 DRFA 方案(如 AODV 协议)来模拟系统鲁棒吞吐量。

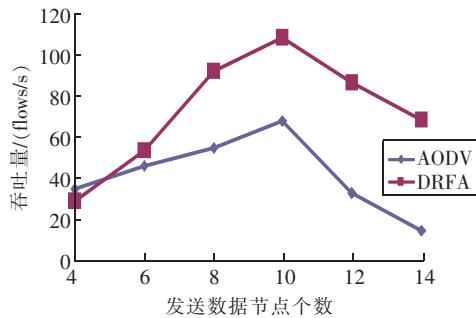


图 2 MANET 吞吐量性能

可以发现,在相同的条件下,当网络系统的负载速率较低时,非 DRFA 方案获得的吞吐量要比在 DRFA 方案下获得的要高,这是由于在 DRFA 方案下,路由的选择主要是依据路由能提供的鲁棒性性能,所以可能造成所选的路由时延稍大。在这种情况下,这些方案获得 MANET 传输性能之间的差别应当是非常小的。

但当 MANET 网络系统的负载较高时,使用 DRFA 方案,可以产生一个改善很多的鲁棒吞吐量性能。因为当负载速率增加时,在非 DRFA 方案下,提供给流的通信资源路径具有较高提前中断的机会,这样造成传输该类流经常重发,最后,可能由于缺乏链路容量资源,流被阻塞。而在 DRFA 方案里,

将被接纳的流分配至那些能提供满足要鲁棒性要求的路由,必然地,通信容量资源可以更好地利用,从而网络系统的鲁棒吞吐量性能获得改善。由于网络链路故障率不受负载速率影响,但是用于接纳新的流的容量资源的有效性受网络负载影响较大,所以在一个较高负载率的情况下,至关重要的是使用一个结合的鲁棒性和面向流接纳的方案。

3 结束语

我们提出了鲁棒吞吐量和鲁棒吞吐量容量度量来刻画一个 MANET 的传输流的能力,且为一些特应用提供支持的服务要求(这些应用通常要求流传输没有任何中断地进行)。使用新的方案 DRFA 进行路由选择,可以极大地提高 MANET 网络系统的吞吐量速率。按需鲁棒路由算法 DRFA 方案选择性的发现路由,保证准备传输的基本会话或文件不中断(包括不重新选择路由)的传输。

还有更多的适应机制可以被集成到本文所描述的方案中来,用以提高移动 ad hoc 无线网络的鲁棒性能。比如,交叉层方法的应用^[10],在这种方法下,通过配置基本的参数和方案(如软件配置的调制/编码方案、相关的业务数据速率和转发范围)用来改善网络链路和路由的生存期。此外,为了改善一个 MANET 网络布局的鲁棒性,尽量使用位置稳定的和有能力的节点作为中继节点^[10],这些中继节点包括,无人地面车辆和无人飞行器,这些节点的位置的稳定性减少了节点的移动对于所选路由的稳定性,因此改善了网络的鲁棒吞吐量。

参考文献:

- [1] CAVALCANTI D, KUMAR A, AGRAWAL D P. Wireless Ad hoc Networking[M]. Florida: Auerbach Publications, 2007.
- [2] QUAN Letrung, KOTSIS G. Reducing Problems in Providing Internet Connectivity for Mobile Ad Hoc Networks[C]// Proc of Wireless and Mobility. 2008:113 - 127.
- [3] RUBIN I, ZHANG Runhe. Robust throughput and routing for mobile Ad hoc wireless networks[J]. Ad hoc Networks, 2009, 7(2):265 - 280.
- [4] 钱敏平. 应用随机过程[M]. 北京:北京大学出版社, 2004.
- [5] ZHANG Runhe, RUBIN I. Mobility induced robust throughput behavior in mobile ad hoc networks[C]// Proc of IEEE 60th Vehicular Technology Conference. Los Angeles, 2004, 4:2708 - 2711.
- [6] PERKINS C, BELDING-ROYER E, DAS S. Ad hoc on Demand Distance Vector (AODV) Routing Protocol[S]. IETF Internet RFC

3561, July 2003.

- [7] RUBIN I, BEHZAD A, JU H, et al. Ad hoc wireless networks with mobile backbones[C]// Proc of the 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications. Barcelona, Spain, 2004, 1: 566 – 573.
- [8] HUANG X, RUBIN I, JU H. A mobile backbone network routing protocol with flow control[C]// Proc of Military Communications Conference. 2004, 2: 1086 – 1092.
- [9] JU H, RUBIN I. Backbone topology synthesis for multi-radio meshed wireless LANs[C]// Proc of IEEE INFOCOM 2006 Conference. Barcelona, Spain, 2006.
- [10] HSU J, RUBIN I. Performance analysis of multi-rate capable random access MAC protocols in wireless multi-hop networks[C]// Proc of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications. Helsinki, Finland, 2006.

作者简介:



徐占洋(1975 –),男,江苏灌云人。南京邮电大学信息网络技术研究所博士研究生,南京信息工程大学计算机系讲师。研究方向为无线自组织网络技术,网络服务分析等。

张顺颐(1944 –),男,江苏南京人。南京邮电大学信息网络技术研究所教授,博士生导师。(见本刊2009年第1期第5页)

(上接第57页)

- [7] REGHELIN R. RF-based location system using cooperative calibration[C]// Proceedings of the 3rd IEEE International Workshop on Wireless Ad-hoc and Sensor Networks. New York, 2006.
- [8] ELNAHRAWY E, MARTIN R P. The limits of localization using signal strength:a comparative study[C]// IEEE SECON. Santa Clara, CA, USA. 2004.
- [9] LI Xinrong. RSS-Based Location Estimation with Unknown Path loss Model[J]. IEEE Trans on Wireless Communication, 2006, 5: 3626 – 3633.
- [10] ARIAS J, LAZARO J, ZULOAGA A. GPS-less location algorithm for wireless sensor networks[J]. Computer Communications, 2007, 30(14/15): 2904 – 2916.
- [11] HOMAYOUN N, HOMAYOUN H. Phase Modeling of Indoor Radio Propagation Channels[J]. IEEE Trans on Vehicular Technology, 2000, 49(2): 220 – 231.
- [12] THEODORE S R. Wireless Communications: Principles and Practice (2nd Edition)[M]. New Jersey: Prentice Hall Press, 2001, 69 – 138.
- [13] Ubicell://www.wsns.net.cn

作者简介:



凡高娟(1983 –),女,河南周口人。南京邮电大学计算机学院博士研究生。主要研究方向为无线传感器网络拓扑控制。

王汝传(1943 –),男,安徽合肥人。南京邮电大学计算机学院教授,博士生导师。(见本刊2009年第1期第68页)

孙力娟(1963 –),女,江苏南京人。南京邮电大学计算机学院院长,教授,博士生导师。主要研究方向为无线传感器网络、智能优化算法。

不同掺杂浓度的 CdS:Mn/SiO₂ 核壳纳米结构的光致发光

薛洪涛

(南京邮电大学理学院,江苏南京 210046)

摘要:通过反胶束法合成了分散性较好的 Mn²⁺掺杂的 CdS/SiO₂核壳纳米结构,在合成过程中,没有添加任何偶联剂。利用高分辨透射电镜和电子衍射仪器对合成的纳米颗粒的结构进行了表征。进一步研究了这些纳米颗粒的光致发光谱、光致发光激发谱和电子自旋共振谱,对于不同的 Mn²⁺掺杂的 CdS/SiO₂核壳纳米结构的发光特性和机制进行了详细的分析。这些稳定的荧光纳米颗粒可望在生物、医学等方面以及与材料相关的领域内有广泛的应用。

关键词:CdS:Mn²⁺/SiO₂;反胶束法;核壳纳米结构;光致发光

中图分类号:O469 文献标识码:A 文章编号:1673-5439(2009)06-0064-04

Photoluminescence of CdS/SiO₂ Core-Shell Nanostructures with Different Manganese Concentration

XUE Hong-tao

(College of Science, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: Mono-dispersed Mn-doped CdS/SiO₂ core-shell structures were synthesized via reverse-micelle method without adding a coupling agent. High resolution transmission electron microscopy and selected area electron diffraction were used to characterize the products. The photoluminescence spectra and photoluminescence excitation spectra along with electron paramagnetic resonance measurement were studied. The properties and mechanism of photoluminescence of these nano-particles were analyzed in detail. It is hoped that the steady and fluorescent CdS:Mn²⁺/SiO₂ core-shell nanostructures would be useful material in biological and medicinal systems, as well as materials related fields.

Key words:CdS:Mn²⁺/SiO₂; reverse-micelle method; core-shell nanostructures; photoluminescence

0 引言

近些年来,半导体纳米颗粒由于在光电方面有着不同于相应的体材料的广泛应用而倍受关注。其中,CdS 纳米颗粒和硅土包裹的 CdS 纳米颗粒就引起了人们相当大的兴趣。CdS 是非常重要的Ⅱ—VI 族化合物半导体材料之一,由于较宽的带隙(2.42 eV)和明显的非线性光学特性,CdS 纳米材料在荧光探测、传感器、太阳能电池、发光二极管等光电装置方面有着重要的应用^[1-4]。此外,目前也出现了大量的磁性离子掺杂的 CdS 纳米材料,即稀磁

半导体^[5-6]。但是,CdS 纳米材料的光电特性和其稀磁半导体的磁光特性受其表面态和分散性的影响较大,所以改善其表面结构和提高纳米颗粒之间的分散性能是至关重要的。那么对纳米颗粒进行包裹是值得推荐的方法之一,一方面对纳米颗粒的表面进行修饰可以降低其表面能,提高粒子的抗氧化能力;另一方面适当的修饰也可以调节纳米颗粒与其它材料之间的相容性。作为包层材料,SiO₂ 具有以下优异性能^[7-8]:可以屏蔽纳米颗粒之间的相互作用,防止粒子团聚;具有良好的生物相容性、亲水性以及非常好的稳定性;并且 SiO₂ 微球的制备技术已

经相当成熟。因此,近些年来,核壳结构二氧化硅/纳米颗粒或磁性纳米颗粒的制备和应用成为研究的热点之一。但是,大部分报道^[9-11]的这种核壳结构合成过程比较复杂,颗粒尺寸较大,而且有些在合成过程中需要添加偶联剂。本文利用反胶束法,通过改变实验温度和合成时间以及样品清洗过程,在没有添加任何偶联剂的情况下,成功地合成了单分散性较好的,尺寸较小的、不同 Mn²⁺ 浓度掺杂的 CdS/SiO₂核壳纳米结构,利用高分辨透射电镜、电子衍射、荧光光谱仪和电子自旋共振仪器,对合成的纳米颗粒的结构和具有不同 Mn²⁺ 掺杂浓度的样品的发光特性和机制进行了详细的分析。研究结果表明,良好的包裹使得 Mn²⁺ 在纳米颗粒中的光致发光得到了有效的改善,而恰当的掺杂浓度也是提高 Mn²⁺ 发光效率的一个重要因素。

1 实验部分

通过反胶束法合成了 Mn²⁺ 掺杂的 CdS/SiO₂核壳纳米颗粒,SiO₂壳层利用正硅酸乙酯水解而成,详细的实验过程类似 Yang Heesun^[12]等人的报道。本实验过程简单地描述如下:反应物是 Cd(CH₃COO)₂ · 2H₂O,Mn(CH₃COO)₂ · 4H₂O 和 Na₂S · 9H₂O。一定浓度的 S²⁻ 和(Cd²⁺ + Mn²⁺) 的水溶液先分别和表面活性剂(AOT)的正庚烷溶液混合形成胶束溶液,并持续搅拌 15 分钟,然后将其中一种胶束液缓慢地加入另一种胶束液中,在混合过程中用磁力搅拌器不停地搅拌形成 CdS:Mn²⁺ 量子点。少量的正硅酸乙酯(TEOS)加入上述混合胶束液中,在一定质量百分比(~18%)的 NH₄OH 的催化作用下进行水解凝结反应生成 silica,最后的混合溶液被持续搅拌 24 小时。实验过程中,所有混合溶液的搅拌过程都在接近 0 ℃ 下进行。反应结束后,在超声震荡下,用甲醇和酒精对样品进行反复地离心和清洗以便除去没有反应的杂质。在反胶束溶液中,Cd²⁺ 和 S²⁻ 在水溶液中的浓度分别为 0.2 和 0.4 M,而水和 AOT 在正庚烷溶液中的浓度分别为 0.6 和 0.1 M,即水和 AOT 的摩尔比 W = 6。我们制备了掺杂浓度不同的样品,其中 3 个典型的掺杂样品 a,b,c 中,Mn²⁺ 的掺杂浓度大约分别为:0.8%、2%、4%(Mn²⁺ 的掺杂浓度是指 Mn²⁺ 对 CdS 纳米颗粒的摩尔比)。后面图 2 和图 3 中的曲线 a,b 和 c 分别代表相应的样品 a,b 和 c 的特性。

合成样品的结构和形貌用型号为 FEI Tecnai G²20 S-TWIN,并且自带电子衍射装置(SAED)的高分辨透射电镜(HRTEM)来表征;EPR 谱线在操作频率为 9.8 GHz,型号为 Bruker EMX-10/12 的电子顺磁共振谱仪上记录;光致发光谱线从 Jobin Yvon 公司生产的型号为 FluoroMax-2 荧光光谱仪上进行测量而获得。

2 结果和讨论

图 1 显示了 silica 包裹的 CdS 的纳米颗粒(样品 b,掺杂浓度为 2%)是在合成一周后的电镜图。从图 1(a)中可以看出,颗粒分散性较好,且呈现球形或椭球形,而且半径 r 大部分在 10 nm 左右,比根据公式^[13]:

$$\left(\frac{r+1.5}{r}\right)^3 - 1 = \frac{27.5}{W}$$

计算出的半径要大些(本实验中水和 AOT 的摩尔比 W = 6),主要是因为表面有 silica 包裹的缘故。图 1(b)是图 1(a)中样品的高分辨电镜图,图 1(a)中清楚地展现了这种核壳结构,壳层的厚度大约有 5 ~ 7 nm,核的直径在 10 nm 左右,大于其体材料的激子波尔直径(大约 5.8 nm)^[14],晶格间距为 0.364 nm 的清晰的晶格指纹可归属于六角状的 CdS 纳米晶体的 {100} 面^[15],图 1(b)中右上角的选区电子衍射图(SAED)也表明样品具有良好的结晶性。此外,在壳层里面,也观察到了结晶相,推断可能是多孔 SiO₂。

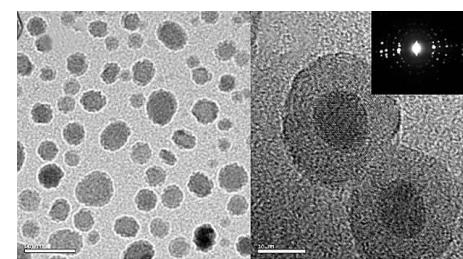


图 1 Mn²⁺ 掺杂的 CdS/SiO₂ 核壳纳米结构图

为了探索 Mn²⁺ 在 CdS 纳米颗粒中的局域分布,我们对样品 a,b,c 分别进行了 EPR 谱的测量,如图 2 所示。从图 2 中可知,样品 a,b 的 EPR 谱有明显的 6 条线,在这两套六线谱中,g 因子分别为 2.033 和 2.002。谱线来源于 Mn²⁺ 中允许的磁偶极跃迁($\Delta m_s = \pm 1$; $\Delta m_l = 0$),而磁偶极跃迁是发生在 Mn²⁺ 的 3d 电子基态能级⁹S_{5/2} 的超精细劈裂的 Zeeman 能级之间,这种超精细结构起源于 Mn²⁺ 中未配

对的3d电子的自旋 $S = 5/2$ 和核自旋 $I = 5/2$ 之间的相互作用。当掺杂的 Mn^{2+} 浓度较低时,无论CdS纳米颗粒是具有立方相还是六角相结构,这种超精细劈裂的特征是在临近的允许跃迁之间有 $\Delta B_{HFS} = 7.0$ mT,样品a基本上符合这个劈裂间距。同时也观察到,在样品a中,除了允许跃迁外,在每条超精细线的低磁场侧有禁阻跃迁出现($\Delta m_s = \pm 1; \Delta m_l = \pm 1$)。当主晶格是立方相时,禁阻跃迁的出现表明 Mn^{2+} 处于四面体晶体场中的 Cd^{2+} 的取代位置;当主晶格是六角相时,这些禁阻跃迁线会更加明显,并且随着 Mn^{2+} 浓度的增加和允许跃迁的超精细线叠加在一起,这一现象在样品b、c中均可观察到,只是在样品c中很难再区分超精细线和禁阻跃迁线(也称精细线)。同时可推断掺杂 Mn^{2+} 会改变II—VI族纳米晶的结晶相,一般认为,只有掺杂 Mn^{2+} 浓度很低时,才会出现立方相,随着 Mn^{2+} 浓度的增加,II—VI族纳米晶变为六角相^[16]。这一结论和我们对样品b的高分辨电镜结果是一致的。另外,在样品b中还有一个附加的超精细结构就是大部分超精细线之间的劈裂间距大于7.0 mT,最大的有12.0 mT,这一超精细结构表明独立的 Mn^{2+} 在主纳米晶的表面或间隙位置^[14]。研究表明:当 Mn^{2+} 浓度较低时, Mn^{2+} 更容易进入主晶格的取代位置;随着 Mn^{2+} 浓度的增加,一些 Mn^{2+} 就局域在主纳米颗粒的表面或间隙位置,或 Mn^{2+} 之间发生偶极相互作用,或进一步形成 Mn 团簇;当 Mn^{2+} 浓度较高时,偶极相互作用和交换耦合作用合并在一起使得超精细结构变成一个宽的共振峰。显然, Mn^{2+} 在纳米颗粒中的分布强烈地受到 Mn^{2+} 的掺杂浓度的影响。

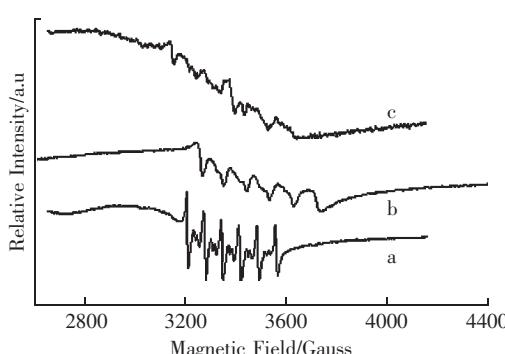


图2 CdS:Mn/SiO₂ 纳米颗粒的EPR谱线,曲线

a、b、c中[Mn²⁺]分别为0.8%、2%、4%

下面对样品a、b和c分别进行了光谱测量。图3显示了不同 Mn^{2+} 掺杂浓度的样品扩散在水溶液中的归一化的光致发光谱(PL)和光致发光激发谱(PLE)。其中,样品a的掺杂浓度约为0.8%,激发

波长为420 nm。样品b和c的掺杂浓度约为2%和4%,激发波长为360 nm。图4是实际的与 Mn^{2+} 相关的特征辐射峰的强度随着 Mn^{2+} 浓度的变化趋势。随着 Mn^{2+} 浓度的增加,其特征辐射峰的强度先是增加,然后又下降,原因是 Mn^{2+} 浓度太低时,对总的发光强度也是有影响的。从图3中可看出与 Mn^{2+} 相关的辐射峰值随着 Mn^{2+} 浓度的增加并没有发生明显的移动,只是随着掺杂 Mn^{2+} 浓度增加,与 Mn^{2+} 相关的辐射峰有少许的红移。样品a的辐射峰值在580 nm,样品c的辐射峰值在588 nm,其原因可能和 Mn^{2+} 在晶体场中的3d⁵能级分布结构有关^[17]。但是,随着 Mn^{2+} 浓度的增加,PL谱的谱线明显展宽,半高宽由75 nm增加到137 nm,而且在样品c中,除了在588 nm处的与 Mn^{2+} 相关的特征辐射峰外,在630 nm处也出现了一肩峰。而且随着掺杂 Mn^{2+} 浓度的进一步增加,与 Mn^{2+} 相关的特征辐射峰(~580 nm)逐渐消失。同时我们也测量了3个样品的PLE谱。当监控波长为580 nm时,样品a中的PLE谱的峰值主要在415 nm左右,同时在370 nm处有一肩峰;样品b、c中有3个峰,分别位于415、398和370 nm;在样品c中,当监控波长为630 nm时,PLE谱的峰与监控波长为580 nm时的PLE谱基本相同。3个样品的PLE谱在370 nm处都有峰值,而对没有掺杂的CdS纳米颗粒的PL谱测量表明PLE峰值也在370 nm(文章中没出示意图),所以370 nm处的峰值可认为是对CdS纳米颗粒的带边激发;而位于415和398 nm处的峰可认为是对 Mn^{2+} 内部能级跃迁的激发。虽然 Mn^{2+} 的辐射峰一般起源于 $^4T_1 \rightarrow ^6A_1$ 的跃迁,但其激发可有多个能级,而且随着 Mn^{2+} 浓度的增加, Mn - Mn 之间的相互作用使得多个跃迁能级之间可能产生交叠。所以随着掺杂 Mn^{2+} 浓度的增加,580 nm处的峰强逐渐减弱,在大于600 nm的区域出现了发光峰,但是更高的 Mn^{2+} 浓度会导致发光的猝灭。我们也发现掺杂后CdS纳米颗粒本身的PL谱基本消失,可能是由于CdS纳米颗粒的s-p轨道和 Mn^{2+} 的d轨道杂化所致,硅土壳层的完好包裹使得这种杂化程度加强,而杂化为能量转移提供了一条有效的途径。由此可推断在CdS纳米颗粒与 Mn^{2+} 的局域能级之间发生了能量转移,这种能量转移也加强了 Mn^{2+} 的特征辐射,而这种能量转移的有效性与 Mn^{2+} 在纳米颗粒中的局域分布有很大关系。所以很好地调节和控制 Mn^{2+} 在纳米颗粒中的浓度应该是改善其发光性能的有效途径之一。

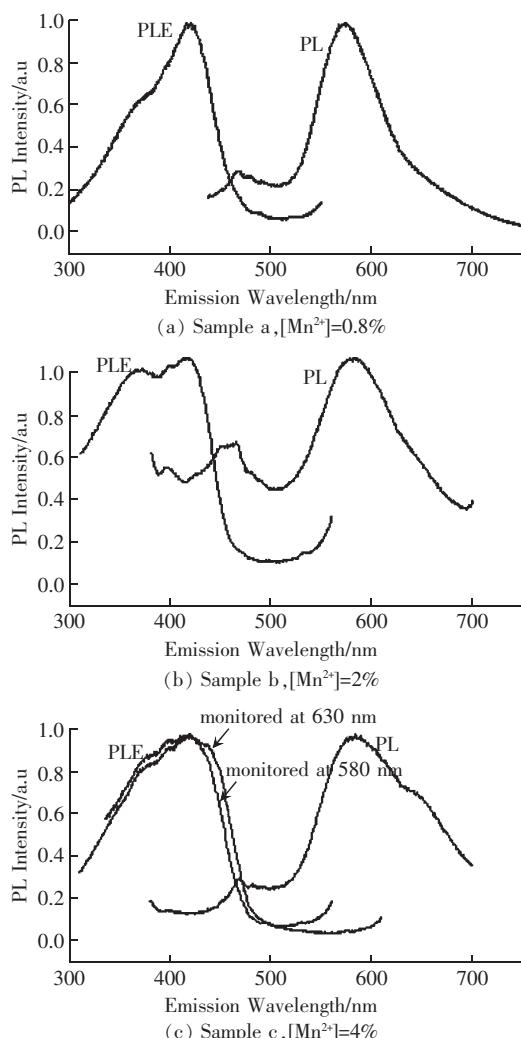


图 3 不同 Mn²⁺ 浓度掺杂的 CdS/SiO₂ 纳米颗粒的光致发光和光致发光激发谱

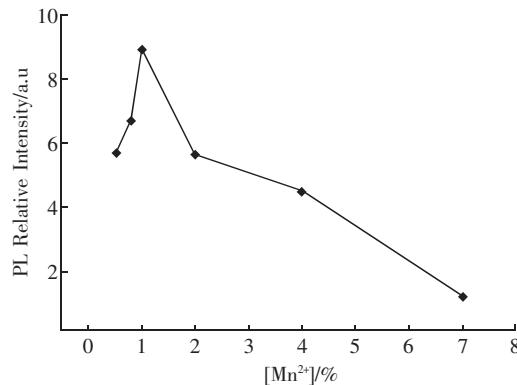


图 4 Mn²⁺ 的 580 nm 特征辐射峰强度随其浓度的变化趋势

3 结 论

在本文中用反胶束法合成了分散性较好的掺杂浓度不同的 CdS:Mn/SiO₂ 核壳纳米颗粒, 并研究了其光致发光谱、光致发光激发谱及电子顺磁共振谱。

通过 EPR 谱的测量确认了不同掺杂浓度的 Mn²⁺ 在核壳纳米颗粒中的局域分布情况。不同掺杂浓度的 CdS:Mn/SiO₂ 纳米颗粒有不同的发光特性。浓度较低而 Mn²⁺ 处于 Cd²⁺ 的晶格取代位置时 Mn²⁺ 的特征辐射较强, 而且发光峰半高宽较窄。同时光致发光谱和光致发光激发谱的测量结果表明 CdS 主纳米晶体和 Mn²⁺ 之间有明显的能量转移, 这种能量转移和 Mn²⁺ 的局域分布及壳层的包裹有很大的关系, 这也是 Mn²⁺ 的特征辐射加强的主要原因。

参 考 文 献:

- [1] HENGLEIN A. Small-particle research: physicochemical properties of extremely small colloidal metal and semiconductor particles[J]. Chem Rev, 1989, 89: 1861 – 1863.
- [2] ROSSETTI R, NAKAHARA S, BRUS L E. Quantum size effects in the redox potentials, resonance Raman spectra, and electronic spectra of CdS crystallites in aqueous solution[J]. J Chem Phys, 1983, 79: 1086 – 1088.
- [3] BARRELET C J, WU Y, DAVID C B, et al. Synthesis of CdS and ZnS nanowires using single-source molecular precursors[J]. J Am Chem Soc, 2003, 125: 11498 – 11499.
- [4] BERMAN A, CHARYCH D. Uniaxial alignment of cadmium sulfide on polymerized films: Electron microscopy and diffraction study[J]. Adv Mater, 1999, 11: 296 – 298.
- [5] ZHOU H J, HOFMANN D M, ALVES H R, et al. Correlation of Mn local structure and photoluminescence from CdS: Mn nanoparticles[J]. J Appl Phys, 2006, 99: 103502.
- [6] BHARGAVA R N. Doped nanocrystalline materials-physics and applications[J]. J Lumin, 1996, 70: 85 – 94.
- [7] TARTAJ P, GONZALEZ-CARRE N O T, SERANA C. Magnetic behavior of γ-Fe₂O₃ nanocrystals dispersed in colloidal silica particles[J]. J Phys Chem B, 2003, 107(1): 20 – 24.
- [8] YANG H H, ZHANG S Q, CHEN X L, et al. Magnetite-containing spherical silica nanoparticles for biocatalysis and bioseparations[J]. Anal Chem, 2004, 76(5): 1316 – 1337.
- [9] YANG Y H, GAO M Y. Preparation of fluorescent SiO₂ particles with single CdTe nanocrystals core by the reverse microemulsion method[J]. Adv Mater, 2005, 17: 2354 – 2357.
- [10] TENG F, TIAN Z J. Preparation of CdS-SiO₂ core-shell particles and hollow SiO₂ spheres ranging from nanometers to microns in the nonionic reverse microemulsions[J]. Catal Today, 2004, 93/95: 651 – 657.
- [11] WANG Z X, CHEN J F. Sythesis of monodispersed CdS nanoballs through-irradiation route and building core-shell structure CdS@SiO₂[J]. Mater Res Bull, 2007, 42: 2211 – 2218.
- [12] YANG H S, HOLLOWAY P H. Water-soluble silica-overcoated CdS: Mn/ZnS semiconductor quantum dots[J]. J Chem Phys, 2004, 121: 7421 – 7426.

(下转第 74 页)

基于 K 近邻分类间隔的特征选择方法研究

李 云¹, 张腾飞², 杨文杰¹

(1. 南京邮电大学 计算机学院, 江苏南京 210046
(2. 南京邮电大学 自动化学院, 江苏南京 210046)

摘要: 特征选择是机器学习和模式识别领域的一个关键问题。文中详细分析研究一类基于 K 近邻分类间隔的特征选择算法, 并着重讨论当 $K > 1$ 时, 特征选择的评价准则和搜索策略的设计, 同时在多个数据集上验证其性能。

关键词: 特征选择; K 近邻; 分类间隔

中图分类号: O235; TP273+.22

文献标识码: A

文章编号: 1673-5439(2009)06-0068-07

Feature Selection Based on Margin of K -Nearest Neighbors

LI Yun¹, ZHANG Teng-fei², YANG Wen-jie¹

(1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210046, China
(2. College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: Feature selection is one of key problems in machine learning and pattern recognition. In this paper, a type of feature selection methods based on Margin of K -nearest neighbors is discussed. Furthermore, the feature selection evaluation criterion and search strategy is introduced when the value of k is more than 1. Meanwhile, the experimental results on different data sets are presented.

Key words: feature selection; K -nearest neighbors; classification margin

0 引言

从原始特征集中选出最优特征子集是模式识别和机器学习等领域的一个关键问题, 同时也是一个棘手问题, 它通常包括两个关键环节: 特征子集的评价准则和搜索策略。而特征选择算法从模型上一般可以分为两类: 过滤器和封装器^[1]。前者是将特征选择作为一个预处理过程, 独立于学习算法, 而后者则将学习算法的结果作为特征子集评价准则的一部分。一般过滤器模型的时间复杂度较低, 效果不佳, 而封装器模型的时间复杂度较高, 效果较好。随着数据量和特征维数的不断增长, 过滤器模型更具有现实意义。

分类间隔 Margin 目前是机器学习领域的一个研究热点, 它通常用来测量分界面与被预测样本间

的距离, 描述分类器性能的置信度, 并且它已经成为许多算法设计的理论基础和依据。目前对支持向量机(SVM)^[2]的分类间隔研究比较深入和透彻, 因此基于 SVM 分类间隔的特征选择方法也比较多, 如 SVM-RFE^[3], $R^2W^{2,4}$ 以及文献[5]中所介绍的算法等。而 K 近邻分类器^[6]是一种最简单的分类器, 应用非常广。 K 近邻分类就是取未知样本的 K 个近邻, 看这 K 个近邻中多数属于哪一类, 就把未知样本分为哪一类。当 $K = 1$ 时, 称为最近邻分类, 对其分类间隔研究已有一定的成果。当 $K > 1$ 时, 对其分类间隔研究比较少, 而基于 K 近邻分类间隔的特征选择算法则更少。本文将介绍一类基于最近邻分类间隔的特征选择算法, 并特别介绍当 K 的取值大于 1 时所构建的特征选择算法。

1 最近邻分类间隔特征选择算法

1.1 两类间隔 Margin

Crammer^[7]指出目前存在两类间隔 Margin,一种是样本间隔 Sample-margin 从一个样本到由分类规则推导出的决策边界的距离,如图 1(a)所示。其中支持向量机 SVM 就是利用一组特殊的称为支持向量的样本,代替原始样本对整个训练数据集进行描述,并寻找一个最优超平面使得分类的两类的分类间隔最大,这个间隔即 Sample margin. 此外假设间隔 Hypothesis-margin 是在不改变任何样本点分类结果的条件下分类器可以移动的距离。需要注意的是,这种 Margin 要求对分类器之间的距离进行度量。它被应用在 AdaBoost 方法^[8]中,如图 1(b)所示。在本文中,我们更关注 K 近邻分类间隔,且当 K = 1 时已经有相关结论^[7]:

(1) 如果找到一个具有最大 Hypothesis-margin 的类别集,那么它同样具有最大的 Sample-margin;

(2) 在样本集 S 中,对于某个样本 x_i ,很容易得到它的 Hypothesis-margin:

$$\theta_s(x_i) = \frac{1}{2} (\|x_i - \text{nearmiss}(x_i)\| - \|x_i - \text{nearhit}(x_i)\|) \quad (1)$$

其中, $\text{nearmiss}(x_i)$ 和 $\text{nearhit}(x_i)$ 分别表示与 x_i 不同类和同类的最近邻样本。

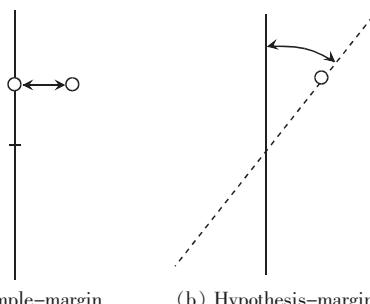


图 1 两类 Margin 示意图

Sample-margin 在分类识别中具有最自然的特点,为此它使 SVM 具有很好的分类性能,而 Hypothesis-margin 则在计算上更容易,具有比 Sample-margin 更低的误差界。

1.2 相关特征选择算法

对于最近邻分类,其 Hypothesis-margin 可以明确地通过式(1)来计算。而基于最近邻分类间隔的经典特征选择算法主要有以下两种。

1.2.1 Relief

Relief 算法^[9]通过寻找每个样本的最近邻同类

样本和不同类样本,然后根据一定规则计算出各个特征与类的关联程度,也就是给每个特征赋予相应的权值,并降序排列,从中选取其值大于选定门限值或者前 m 个特征构成最后的特征子集。该算法容易实现,且能有效地消除不相关特征,具体描述如下:

```

    算法 Relief
    输入: 每个训练样本的特征矢量和类标记
    输出: 特征的关联程度矢量  $\omega$ 
    第一步: 初始化:  $w = (0, 0, \dots, 0)$ ;
    第二步: FOR  $i = 1, 2, \dots, N$  //  $N$  为训练样本数
        DO BEGIN
            (a) 随机选择样本  $x_i$ ;
            (b) 寻找与  $x_i$  的  $\text{nearhit}(x_i)$  和  $\text{nearmiss}(x_i)$ ;
            (c) FOR  $f = 1, 2, \dots, n$  计算
                 $w_f = w_f - \text{diff}(f, \text{nearhit}(x_i), \text{nearmiss}(x_i)) / N +$ 
                 $\text{diff}(f, \text{nearhit}(x_i), \text{nearmiss}(x_i)) / N$ 
        END;
    END

```

其中, n 表示原始特征数, 函数 $\text{diff}(f, x_1, x_2)$ 是计算两个样本在某个特征 f 上的差异, 对于离散特征, 可以定义为:

$$\text{diff}(f, x_1, x_2) = \begin{cases} 0; & x_{1f} = x_{2f} \\ 1; & \text{otherwise} \end{cases} \quad (2)$$

对于连续特征,可以定义为:

$$\text{diff}(f, x_1, x_2) = \frac{|x_{1f} - x_{2f}|}{\max(f) - \min(f)} \quad (3)$$

其中, x_1, x_2 为两个样本, x_{1f} 表示 x_1 在特征 f 上的取值。 $\max(f)$ 和 $\min(f)$ 分别表示特征 f 的最大值与最小值。

Relief 算法归根到底是利用最近邻分类间隔,求得特征的权重,然后对特征进行排序。但是其不能处理多类情况、且无法消除冗余特征,算法的时间复杂度为 $O(N^2n)$ 。

为处理多类情况,将 Relief 扩展到 ReliefF^[10],从随机选取的样本所在的类中选取 k 个近邻样本 $\text{nearhit}(x_i)^j (j=1, 2, \dots, k)$, 并从每个不同类中各选取 k 个近邻样本 $\text{nearmiss}(x_i)^j (j=1, 2, \dots, k)$, 取它们的平均值,从而可以应付多类和噪声数据。则计算某个特征 f 的权值 w_f 的公式修改为:

$$w_f = w_f - \frac{\left(\sum_{j=1}^k \text{diff}(f, x_i, \text{nearhit}(x_i)^j) \right)}{kN} + \frac{\sum_{c \neq \text{class}(x_i)} [p(c) \times \left(\sum_{j=1}^k \text{diff}(f, x_i, \text{nearmiss}(x_i)^j) \right)]}{kN} \quad (4)$$

其中, $P(c)$ 表示类概率, $\text{class}(x_i)$ 表示样本 x_i 所在的类。

ReliefF 仍然无法消除冗余特征, 并且对异常值也缺少处理机制, 后来又提出了 I-Relief 算法^[11-12], 主要采用期望最大化算法(Expectation-Maximization EM)来最大化目标函数。但该算法时间开支很大, 需要多次迭代, 每次迭代的时间复杂度是 $O(N^2n)$, 且每次迭代需要存储和计算 $3 \times N$ 的矩阵。因此 I-Relief 不太适合大规模数据的处理。

1.2.2 Simba

Simba^[13-14]也是一个基于分类间隔的特征选择算法, 它能很好地评价特征的性能。它的评价准则是直接最大化最近邻($K=1$)分类器的 Hypothesis 间隔。Simba 算法流程如下:

算法 Simba

第一步: 初始化

第二步: FOR $i = 1, 2, \dots, N$

- (a) 随机选择样本 x_i
- (b) 找到 $\text{nearmiss}(x_i)$ 和 $\text{nearhit}(x_i)$
- (c) FOR $f = 1, 2, \dots, n$ 计算

$$\Delta_f = \frac{1}{2} \left(\frac{(x_{if} - \text{nearmiss}(x_i)_f)}{\|x_i - \text{nearmiss}(x_i)\|} - \frac{(x_{if} - \text{nearhit}(x_i)_f)}{\|x_i - \text{nearhit}(x_i)\|} \right) w_f$$

$w = w + \Delta$

$$\text{第三步: } w \leftarrow \frac{w^2}{\|w^2\|_\infty}$$

其时间复杂度为 $O(N^2n)$ 。实验结果表明它能获得比 Relief 更好的性能, 但 Simba 对噪声数据的处理能力不强。因此作者将其进行扩展, 设计基于 K ($K > 1$) 近邻分类间隔的特征选择算法, 相关研究成果见文献[15]。

2 K 近邻分类间隔特征选择算法-Lmba

2.1 评价准则

假设存在训练集 S 含有 N 个样本 $\{x_i, y_i\}_{i=1}^N$, 并且每个样本由 n 维特征矢量描述 $x_i = (x_{i1}, x_{i2}, \dots, x_{in}) \in \Re^n$, 其标记 y_i 是离散的。我们可以构建二值矩阵 \mathbf{B} , 其元素 $b_{ij} \in \{0, 1\}$ 表示标记 y_i 和 y_j 是否相同。此外 k 个与 x_i 标记相同的最近邻同类样本称为目标近邻, 也就是说它们与 x_i 的距离最短且标记相同, 这样可以定义二值矩阵 \mathbf{T} , 其元素 $t_{ij} \in \{0, 1\}$ 表示 x_j 是否是 x_i 的目标近邻。矩阵 \mathbf{B} 和 \mathbf{T} 都是固定的, 在特征选择的过程中是不变的, 而距离度量采用的是欧氏距离。

损失函数是机器学习中常用来寻找分类误差与

分类间隔之间平衡的方法。一旦选定损失函数, 学习算法就该考虑如何最小化损失函数以便得到最大的分类间隔^[7]。对于输入样本 x_I , 其在 K 近邻分类中的损失函数可以定义如下:

定义 1 设 S 为训练集, x_i 为样本, 则 x_i 的损失函数为:

$$\begin{aligned} L_S(x_i) &= \sum_j t_{ij} \|x_i - x_j\|^2 + c \sum_{jp} t_{ij}(1 - b_{ip}) h_{jp}(x_i) \\ h_{jp}(x_i) &= \theta_i + \|x_i - x_j\|^2 - \|x_i - x_p\|^2 \downarrow_+ \end{aligned} \quad (5)$$

其中, $[z]_+ = \max(z, 0)$ 表示 hinge 损失, c 为正常数, 通常通过交叉验证得到。

值得注意的是在损失函数的第一项中只是惩罚那些与 x_i 较远的目标近邻样本, 而不是所有与 x_i 具有相同标记的样本。而第二项则惩罚那些与 x_i 的目标近邻距离较近且与 x_i 标记不同的样本^[16-17]。而作者尤其关注那些位于特定区域内具有与 x_i 不同标记的样本, 其与 x_i 的距离小于 x_i 到其所有目标近邻的距离再加上一个邻域 θ_i , 定义为:

$$\theta_i = |\|x_i - \text{nearmiss}(x_i)\|^2 - \|x_i - \text{nearhit}(x_i)\|^2| \quad (6)$$

此邻域的定义可以保证在损失函数中, 至少有一个不同类的近邻被考虑到, 从而确保损失函数的完整性。

由于选择不同的特征子空间可以影响样本间的距离, 从而影响 K 近邻分类的损失函数。因此可以通过选择特征子空间使得损失函数最小, 从而可以将损失函数作为特征选择的评价准则。如果许多样本有着低损失和大的分类间隔, 那么就可以保证算法有着好的泛化性能。这样就可以设计基于 K 近邻分类间隔的特征选择评价准则, 首先将 K 近邻分类的损失函数转化为所选择的特征子集的函数。

定义 2 假设 S 为训练集, x_i 为样本, w 为特征集中每个特征权重构成的权重矢量, 则样本 x_i 的基于特征权重的损失函数定义为:

$$\begin{aligned} L_S(w, x_i) &= \sum_j t_{ij} \|x_i - x_j\|_w^2 + \\ &\quad c \sum_{jp} t_{ij}(1 - b_{ip}) h_{jp}(w, x_i) \\ h_{jp}(w, x_i) &= \theta_i + \|x_i - x_j\|_w^2 - \|x_i - x_p\|_w^2 \downarrow_+ \end{aligned} \quad (7)$$

其中, $\|z\|_w = \sqrt{\sum_j w_j^2 z_j^2}$, $w_j \in [0, 1]$ 。定义 2 在计算样本距离时考虑了特征的权重。我们可以通过特征的权重来对特征进行排序, 再选择重要特征。

这样特征选择的评价准则就可以定义为所有样本的损失函数之和。

定义3 训练集 S , 权重矢量 w , 则评价函数为:

$$e(w) = \sum_i L_s(w, x_i) \quad (8)$$

损失函数中融入了分类间隔的思想。尤其是式(5)中的第二项, hinge 损失是由那些与 x_i 的标记不同且与 x_i 的距离小于 x_i 与其所有目标近邻的距离再加上一个预先定义的邻域 θ_i 。这样评价函数就会选择特征子集,使得与 x_i 标记不同的样本在所选择的特征空间里距离 x_i 和其目标近邻比较远,这样它们不会影响到 x_i 的目标近邻,从而获得大的分类间隔,保证分类的准确性。

此外,邻域 θ_i 的定义包含了最近邻分类中样本 x_i 的 Hypothesis 间隔的思想。

2.2 特征选择算法 Lmba

为了寻找使得评价函数最小化的特征子集,许多搜索策略可以使用,如顺序前向和反向搜索、增 l 减 r 、顺序浮动搜索、遗传算法和分支定界等^[1]。但是,它们只是将特征权重赋为 1 或者 0, 分别表示特征被选或者没被选,且它们的时间复杂度至少为 $O(N^2n^2)$, 其中 N 为训练集的大小, n 为特征数。因此我们想寻找一种更柔性的、时间开支较少的搜索策略。由于 $e(w)$ 几乎是平滑的, 可以考虑利用梯度下降去寻找权重矢量 w , 使得评价函数最小, 从而对特征进行排序。评价准则的梯度定义如下:

$$\begin{aligned} \frac{\partial e(w)}{\partial w_f} &= \sum_i \frac{\partial L_s(w, x_i)}{\partial w_f} = \\ &\sum_i \left(2w_f \sum_j t_{ij} (x_{if} - x_{jf})^2 + c \sum_{jp} t_{ij} (1 - b_{ip}) \frac{\partial h_{jp}(w, x_i)}{\partial w_f} \right) \end{aligned} \quad (9)$$

而 hinge 损失的梯度定义如下:

$$\begin{aligned} \frac{\partial h_{jp}(w, x_i)}{\partial w_f} &= \begin{cases} 0; & \text{if } \theta_i + \|x_i - x_j\|^2 \leq \|x_i - x_p\|^2 \\ g(w_f); & \text{otherwise} \end{cases} \\ g(w_f) &= 2w_f ((x_{if} - x_{jf})^2 - (x_{if} - x_{pf})^2) \end{aligned} \quad (10)$$

算法 Lmba 的步骤如下:

算法 Lmba

第一步: 初始化

第二步: 计算矩阵 B 和 T

第三步: FOR $i = 1, 2, \dots, N$

- (a) 随机选择样本 x_i
- (b) 找到 $\text{nearmiss}(x_i)$ 和 $\text{nearhit}(x_i)$, 并得到 θ_i 的值
- (c) FOR $f = 1, 2, \dots, n$ 计算

$$\nabla_f = 2w_f \sum_j t_{ij} (x_{if} - x_{jf})^2 + c \sum_{jp} t_{ij} (1 - b_{ip}) \frac{\partial h_{jp}(w, x_i)}{\partial w_f}$$

$$(d) w = w - \frac{\nabla}{\|\nabla\|}$$

第四步: 根据特征权重 w 排序

2.3 算法分析

在算法 Lmba 的每次循环中, 通过一个样本 x_i 来修改 w 。由于特征的权重在不断增加, 因此 ∇ 的作用在相对减少, 这样算法就会典型收敛。Lmba 的时间开支主要是计算 B 和 T 以及 w , 它们各自的复杂度为 $O(N^2)$, $O(N^2)$ 和 $O(N^2kn)$ 。而由于 k 值通常比较小, 因此总的时间复杂度为 $2O(N^2) + O(N^2n) \approx O(N^2n)$ 。

Lmba 搜索方法与 Simba 类似, 只是梯度的方向不一样。但是 Lmba 是通过最小化 K 近邻分类器的损失函数来间接优化分类间隔, 而 Simba 是直接优化最近邻分类器的 Hypothesis 间隔。此外, Lmba 中 k 值是大于 1 的, 因此它能更好地处理噪声数据。而在 θ_i 的定义中也融入了最近邻分类的 Hypothesis 间隔的思想。因此 Lmba 比 Simba 更鲁棒, 且可以看作是 Simba 的扩展。

3 实验

我们将在不同的数据集上验证所提出的特征选择方法。实验分成 3 个部分:

首先在基准和合成数据集上验证评价函数的正确性, 看是否能将重要特征排在前面。使用 Matlab 中的 random 函数去产生合成数据。两个合成数据 S1 和 Multi-class 具有不同的类别数和特征数, 但是都含有 100 个样本。对于 S1, 一些特征作为重要特征且呈高斯分布, 不重要特征的取值是随机的。对于 Multi-class 数据集, 样本的取值是随机的 $X = \{x_1, x_2, \dots, x_{100}\}$, 而样本的标记是由下面的 Matlab 函数产生的 $Y = \text{bin2dec}(\text{num2str}(X(:, 1:2) > 0))$ 。基准和现实数据集都来自 UCI 机器学习资料库。这些数据集的描述如表 1 所示。对于 Iris 和 Monk 数据集, 基于一定的先验知识, 它们的重要特征都是已知的, 重要特征的序号如表 1 中第 4 列(从左到右)所示。

表 1 基准和合成数据集描述

数据集	特征数	类别数	重要特征	排序结果
Iris	4	3	3,4	{3,4}, 2, 1
Monk	6	2	2,4,5	{5,4,2}, 1, 6, 3
Multi-class	10	4	1,2	{2,1}, 9, 6, ...
S1	22	6	1-6	{2,3,6,1,4,5}, 8, ...

然后在基准数据集 Multi-features 上进行实验,该数据集是荷兰公共事业地图上的手写字母(0~9)的特征构成,含有2 000 个样本,649 个特征以及 10 类。

最后是基于人脸图像的性别分类实验。在性别识别中,用来训练的样本包括男女人脸图像各 500 张,特征是 1 584 维的 gabor 小波过滤器。测试集包括 15 类性别数据,它们对应不同的姿态、表情、背景等。具体描述见表 2 和图 2。

表 2 性别分类的人脸图像测试集

编号	人脸描述	样本数	编号	人脸描述	样本数
1	正面 1	1 278	9	正面戴眼镜	813
2	正面 2	1 066	10	向右 10°	814
3	向下 10°	820	11	向右 20°	815
4	向下 20°	819	12	向右 30°	805
5	向下 30°	816	13	向上 10°	819
6	笑	805	14	向上 20°	816
7	张嘴	815	15	向上 30°	816
8	闭眼	805			

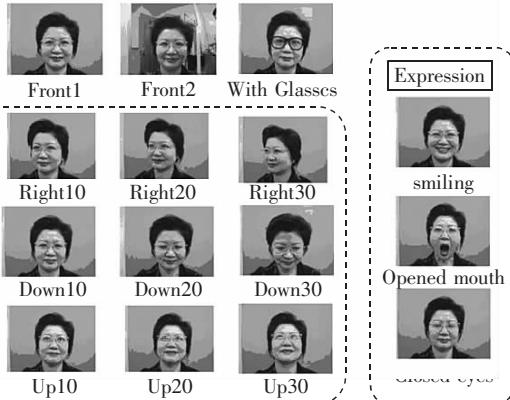


图 2 测试人脸样例图像

3.1 实验结果

使用 Lmba 对基准数据集中的特征排序,排序结果如表 1 中最后一列所示。从表 1 中可以看到,Lmba 能够将重要特征排在最前面。对于 Multi-features 数据集,我们采用 5 次交叉验证(Cross-Validation)和最近邻分类器的分类准确率去评价所选择的不同大小特征子集的性能。由于已有实验结果表明 Simba 算法的性能是优于 Relief^[13~14],因此我们仅将所提出的算法 Lmba 与 Simba 和 Mitra's 算法^[18]进行比较。Mitra's 是一个经典的特征选择算法,它是基于 K 近邻聚类,由于很难确定其参数 k 的确切值,因此在实验中只是给出其近似值,使得所选择的特征子集的维数近似等于或稍微大于 Lmba 和

Simba 所选择的特征数。在 Lmba 算法中,参数 c 和 k 被设为 1 和 3,后面的实验中也采用相同的设置。实验结果如表 3 所示。

表 3 Multi-features 数据集实验结果

选择特征数	算法	准确率/方差
13	Lmba	86.93/0.13
	Simba	89.96/0.10
	Mitra's	85.47/0.14
26	Lmba	90.15/0.09
	Simba	92.60/0.07
	Mitra's	87.51/0.11
52	Lmba	93.81/0.06
	Simba	92.92/0.07
	Mitra's	90.19/0.09
104	Lmba	95.14/0.04
	Simba	94.12/0.05
	Mitra's	92.86/0.07
208	Lmba	95.68/0.04
	Simba	94.32/0.05
	Mitra's	93.41/0.06
416	Lmba	95.76/0.04
	Simba	94.40/0.05
	Mitra's	94.43/0.05

3.2 性别分类

在本节中,当所选择的特征子集维数是已知的,比较特征选择算法 Lmba、Simba 和 Mitra's 在人脸图像上的性别分类结果。所选择的特征子集维数为 254、508 和 1 016。分类器采用的是支持向量机^[2,19],其中参数 C 设为 1。在所有测试集上的平均准确率如表 4 所示。不同算法在 15 个测试集上的结果如图 3 所示。X 轴为测试集的编号,而 Y 轴为性别分类准确率。

表 4 性别分类平均准确率

选择特征数	算法	平均准确率/%
254	Lmba	76.23
	Simba	75.26
	Mitra's	73.19
508	Lmba	77.72
	Simba	75.78
	Mitra's	75.00
1016	Lmba	79.08
	Simba	77.58
	Mitra's	76.59

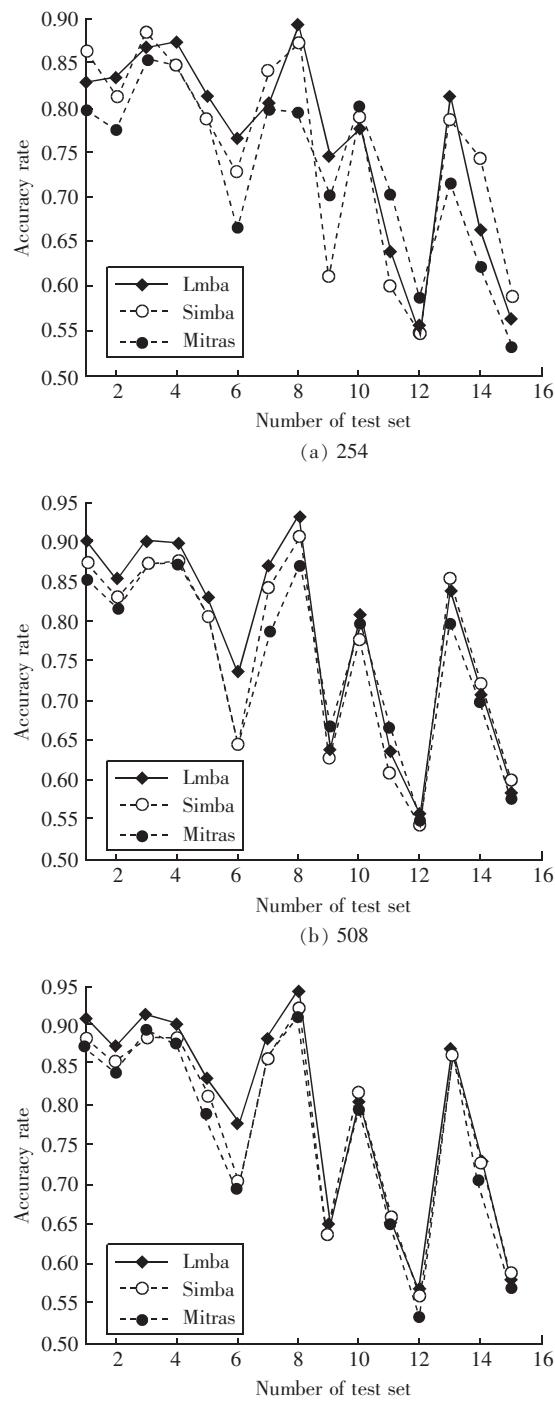


图3 在15个测试集上的准确率,分别对应不同特征数

3.3 观察评论

(1) 所提出的评价准则能够准确地排序重要特征。

(2) 对不同的数据集,Lmba 和 Simba 所选择的特征子集通常能在不同分类器上获得比 Mitra's 更好的性能。同时 Lmba 在大部分情况下分类性能最好,且对于性别分类,Lmba 获得最好性能的测试集数目比 Simba 多。

(3) 随着所选择特征数量的增加,Lmba 能获得

更好的性能,且在性别分类中 Lmba 获得最高准确率的测试集数量也在增加。

对于实验结果,Simba 是基于最近邻分类的 Hypothesis 间隔,它在没有噪声的数据集上和采用最近邻分类器时效果最好。因此我们只需要在这样的环境下对 Lmba 和 Simba 进行比较。当然如果采用其它分类器,如 K 近邻分类器($K > 1$),并且数据集里含有噪声,则 Lmba 肯定能获得比 Simba 更好的性能。这主要是因为 Lmba 中所使用的 K 近邻规则能更好地处理噪声数据^[4],而且评价准则是基于 K 近邻分类间隔的。众所周知,如果一个算法能够选择具有大分类间隔的特征子集,则肯定有好的泛化能力^[13]。

4 结论

本文首先介绍了两类间隔 Margin 的定义,然后分析了各种经典的基于最近邻分类间隔的特征选择算法。最后完整地介绍了 $K > 1$ 时所构建的基于 K 近邻分类间隔的特征选择算法,并给出相关的理论分析和实验结果,从中可以看出作者所提出的算法 Lmba 能够获得比 Simba 和 Mitra's 更好的性能。虽然评价准则是基于 K 近邻分类,但在其它分类器上,如 SVM,仍能获得比 Simba 和 Mitra's 更好的性能。

参考文献:

- [1] LIU H, YU L. Toward integrating feature selection algorithms for classification and clustering[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(3):1–12.
- [2] CORTES C, VAPNIK V. Support vector networks[J]. Machine Learning, 1995, 20(3):273–297.
- [3] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1/3):389–422.
- [4] WESTON J, MUKHERJEE S, CHAPELLE O, et al. Feature selection for SVMs[C]// Advances in Neural Information Processing Systems. 2001, 13:668–674.
- [5] 任双桥,傅耀文,黎湘,等. 基于分类间隔的特征选择算法[J]. 软件学报, 2008, 19(4):842–850.
- REN S Q, FU Y W, LI X, et al. Feature Selection based on classes margin[J]. Journal of Software, 2008, 19(4):842–850.
- [6] COVER T, HART P. Nearest neighbor pattern recognition[J]. IEEE Trans on Information Theory, 1967, 13:21–27.
- [7] CRAMMER K, BACHRACH R G, NAVOT A, et al. Margin analysis of the lvq algorithm[C]// Advances in Neural Information Process-

- ing System. La Jolla CA, 2002, 14: 462 – 469.
- [8] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of online learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119 – 139.
- [9] KIRA K, RENDELL L. A Practical approach to feature selection[C] // Proc Ninth Int'l Workshop Machine Learning(ICML). 1994: 249 – 256.
- [10] KONONERKO I. Estimating attributes: analysis and extension of RELIEF[C] // Proc of European Conf on Machine Learning. 1994: 171 – 182.
- [11] SUN Y J. Iterative Relief for feature weighting: algorithms, theories, and applications[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 26(6): 1035 – 1051.
- [12] SUN Y J, TODOROVIC S, GOODISON S. Local learning based feature selection for high dimensional data analysis[J/OL]. IEEE Trans on Pattern Analysis and Machine Intelligence, http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.190.
- [13] BACHRACH R G, NAVOT A, TISHBY N. Margin based feature selection-theory and algorithm[C] // Proc of the 21th Int'l Conf on Machine Learning(ICML). Banff Canada, 2004.
- [14] GUYON I, GUNN S, NIKRAVESH M. Feature extraction, foundations and applications [M]. New York: Springer Physica-Verlag, 2006.
- [15] LI Y, LU B L. Feature selection based on loss-margin of nearest neighbor classification[J]. Pattern Recognition, 2009, 42: 1914 – 1921.
- [16] WEINBERGER K Q, BLITZER J, SAUL L K. Distance metric learning for large margin nearest neighbor classification[C] // Advances in Neural Information Processing Systems(NIPS). 2005, 18: 1473 – 1480.
- [17] WEINBERGER K Q, SAUL L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research, 2009, 10: 207 – 244.
- [18] MITRA P, MURTHY C A, PAL S K. Unsupervised feature selection using feature similarity[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301 – 312.
- [19] CHANG C C, LIN C J. LIBSVM: a library for support vector machines. [EB/OL]. <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.ps.gz>

作者简介:



李云(1974-),男,安徽望江人。南京邮电大学计算机学院副教授。2005年毕业于重庆大学计算机学院,获工学博士学位。2007年9月从上海交通大学计算机科学与技术博士后流动站出站。目前的主要研究方向是机器学习、数据挖掘和隐私保护等。

张腾飞(1981-),男,河南夏邑人。南京邮电大学自动化学院讲师。2007年毕业于上海海事大学电气自动化系,获工学博士学位。目前的主要研究方向是智能信息处理、模式识别与智能控制等。

杨文杰(1977-),女,山东泰安人。南京邮电大学计算机学院讲师。2005年毕业于南京理工大学计算机科学与技术学院,获工学博士学位。目前的主要研究方向是计算机视觉和信息隐藏。

(上接第67页)

- [13] HOENER C F, ALLAN K A, BARD A J, et al. Demonstration of a shell-core structure in layered cadmium selenide-zinc selenide small particles by x-ray photoelectron and Auger spectroscopies [J]. J Phys Chem, 1992, 96: 3812 – 3817.
- [14] CAO L, MIAO Y M. Exciton interactions in CdS nanocrystal aggregates in reverse micelle[J]. J Chem Phys, 2005, 123: 024702.
- [15] PINNA N, WEISS K, URBAN J, et al. Triangular CdS nanocrystals: Structural and optical studies[J]. Adv Mater, 2001, 13: 261 – 264.
- [16] BRIELER F J, GRUNDMANN P. Formation of Zn_{1-x}Mn_xS nanowires within mesoporous silica of different pore sizes[J]. J Am Chem Soc, 2004, 126: 797 – 807.
- [17] FURDYNA J K. Diluted magnetic semiconductors[J]. J Appl Phys, 1988, 64: R29.

作者简介:



薛洪涛(1976-),男,江西赣州人。南京邮电大学理学院物理实验中心讲师。2006年在复旦大学获理学硕士学位。现主要从事微/纳米制备及其应用研究。

保单驱动索赔离散风险模型的精算量分布

徐小阳¹, 唐加山²

(1. 南京邮电大学 通信与信息工程学院, 江苏南京 210003
2. 南京邮电大学 理学院, 江苏南京 210046)

摘要:研究了一类索赔是由保单驱动的带随机利率的离散时间非寿险风险保险模型,证明了该模型的盈余首次超过给定水平的时间、破产前最大盈余、破产持续时间以及盈余首次回复为正后的瞬间值等精算量的分布都可以由一类积分方程的唯一解给出。

关键词:离散时间; 风险模型; 精算量分布

中图分类号:O211.6 文献标识码:A 文章编号:1673-5439(2009)06-0075-04

Distributions of Actuarial Variables of Discrete Risk Model with Policies Driving Claims

XU Xiao-yang¹, TANG Jia-shan²

(1. College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
(2. College of Science, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: We consider one class of discrete non-life insurance risk model with random interest force and the claims being driven by the insured policies, and prove that the distributions of the first hitting time for a given level of the surplus, the maximum surplus before ruin, the duration of ruin, the surplus immediately after the risk process turns back to positive and etc can be given by the unique solution of one class of integral equations.

Key words: discrete time; risk model; distribution of actuarial variable

0 引言

近几十年来,随着金融市场的不断繁荣和发展,风险理论一直是保险精算学领域最热门的研究课题之一,对于风险保险模型破产概率等问题的研究不但具有重要的理论意义,而且具有重要的实际应用价值。

经典风险模型在学术界已经得到了深入的研究^[1-2],取得了非常丰富的成果,例如破产概率、破产持续时间、盈余回复为正后的瞬间的盈余分布、破产前最大盈余的分布、盈余首次越过水平 x 的分布以及破产前盈余、破产后赤字与破产前最大盈余的联合分布^[3-4]等等。根据市场的实际需要,在经典模型的基础上,学者们提出了很多与实际更加接近

的保险模型^[5-6]。近年来,根据索赔是由保单驱动的思想,学者们又提出了一种基于进入过程的风险保险模型^[7-9]。本文的目的是对于离散化的一类基于进入过程的风险保险模型,通过引入随机利率而建立一类新的风险模型,对于新建立的模型,给出了其若干精算量的分布,结果表明这些分布都可以由一类积分方程的唯一解给出。

1 模型描述

设保险公司的初始准备金用 $u > 0$ 表示,对于时刻 $n (n = 1, 2, 3, \dots)$,假设在时间区间 $(n-1, n]$ 中只有两种可能,一是有且只有一个投保客户到达,该事件以概率 $p (0 < p < 1)$ 发生,二是没有客户到达,

此时从 $n-1$ 时刻到 n 时刻, 公司没有额外的保费收入, 因此可以不妨假设客户的到达是服从参数为 p 的二项随机序列。假设保险公司为客户提供 K 种类型的保单, 不同的保单有不同的有效期 $a_1 < a_2 < \dots < a_K$, 到达的客户可以以一定的概率选择其中一种保单购买。若 n 时刻到达的客户选择的保单有效期为 C_n , 则 C_n 是取值于 $\{a_1, a_2, \dots, a_K\}$ 的随机变量, $P(C_n = a_k) = p_k$ ($1 \leq k \leq K$), 且 $\sum_{k=1}^K p_k = 1$ 。不同时刻到达的客户选择保单是相互独立的, 保险公司将根据保单有效期的不同收取不同的保费, 即 n 时刻到达的客户需要缴纳的保费为 $f(C_n)$, 其中 $f(\cdot)$ 是一个严格单增的确定性函数, 假设 n 时刻购买保单的客户的不幸事件将在 S_n 时间以后发生, 若该不幸事件发生在保单的有效期之内, 则客户向保险公司要求索赔, 索赔额为 V_n , 否则保单失效, 保险公司不对客户负任何责任。

为了给出风险模型的盈余过程, 再假设随机变量序列 $\{U_n, n=1, 2, \dots\}$ 独立同分布于 $U \sim b(1, p)$, 并假设 n 时刻的随机利率为 $i_n > 0$, 记 $Z_n = 1 + i_n$, 则本文所讨论的风险模型在 n 时刻的盈余可以由下面的递推公式给出

$$R_n = \begin{cases} u, & n=0 \\ (R_{n-1} + U_n \cdot f(C_n)) \cdot Z_n - U_n \cdot V_n \cdot I_{\{S_n \leq C_n\}}, & n=1, 2, 3, \dots \end{cases} \quad (1)$$

其中, $I_{\{\cdot\}}$ 是示性函数。式(1)表明, 在时间区间 $(n-1, n]$ 中, 可能的保费在 $n-1$ 时刻收取, 可能的赔付在 n 时刻兑现, $n=1, 2, \dots$ 。

对于上述风险保险模型, 作出如下假设: (1) U_n, C_n, Z_n, V_n, S_n ($n=1, 2, 3, \dots$) 相互独立; (2) 随机变量序列 Z_n 同分布于 Z , 并记它的分布函数为 $F_Z(\cdot)$, 对 V_n, S_n 亦作类似假设, 并记相应的分布为 $F_V(\cdot)$ 以及 $F_S(\cdot)$; (3) 为了保证保险公司能够盈利, 需要假设 $E(f(C_n)) > E(V_n \cdot I_{\{S_n \leq C_n\}})$ 。

在上述模型中若假设 $X_n = U_n \cdot f(C_n)$ 且 $Y_n = U_n \cdot V_n \cdot I_{\{S_n \leq C_n\}}$, 则对于 X_n 与 Y_n 相互独立的情形, 文献[10]已经进行了详细的讨论并给出若干精算量的分布。在本文中, 将对模型(1)讨论它的若干精算量的分布问题, 尽管本文模型中 X_n 与 Y_n 不是相互独立的, 但仍然可以利用文献[10]中条件数学期望的方法讨论有关精算量的分布问题。

为了下文讨论的方便, 定义保险公司的破产时刻 T_u 为 $T_u = \inf\{n; n > 0 | R_n < 0 | R_0 = u\}$, 若对所有的 $n > 0$, 都有 $R_n \geq 0$, 则定义 $T_u = \infty$, 另外, 随机变

量 $R_1 = (u + U_1 \cdot f(C_1)) \cdot Z_1 - U_1 \cdot V_1 \cdot I_{\{S_1 \leq C_1\}}$ 的分布将在本文的研究中起着重要的作用, 下面的引理 1 给出了它的分布 $F_u(x) = P(R_1 \leq x | R_0 = u)$, 显然, 该分布是由初始资本 u 以及随机变量 U_1, C_1, Z_1, V_1 和 S_1 的分布来确定。

引理 1 时刻 1 的盈余 R_1 的分布函数 $F_u(x)$ 为

$$F_u(x) = q \cdot F_Z\left(\frac{x}{u}\right) + p \sum_{k=1}^K p_k \left(F_S(a_k) \cdot \int_0^\infty F_Z\left(\frac{v+x}{u+f(a_k)}\right) dF_V(v) + (1 - F_S(a_k)) F_Z\left(\frac{x}{u+f(a_k)}\right) \right)$$

其中, $-\infty < x < \infty$ 。

证明 设 $q = 1 - p$, 则由上文中关于模型的假定, 对于任意实数 x , 有

$$\begin{aligned} F_u(x) &= P(R_1 \leq x | R_0 = u) \\ &= P((u + U_1 \cdot f(C_1)) \cdot Z_1 - U_1 \cdot V_1 \cdot I_{\{S_1 \leq C_1\}} < x) \\ &= P(uZ_1 < x) \cdot P(U_1 = 0) + P((u + f(C_1)) \cdot \\ &\quad Z_1 - V_1 \cdot I_{\{S_1 \leq C_1\}} < x) \cdot P(U_1 = 1) \\ &= q \cdot F_Z\left(\frac{x}{u}\right) + p \sum_{k=1}^K P((u + f(a_k)) \cdot Z_1 - \\ &\quad V_1 \cdot I_{\{S_1 \leq C_1\}} < x) \cdot P(C_1 = a_k) \\ &= q \cdot F_Z\left(\frac{x}{u}\right) + p \sum_{k=1}^K p_k P(S_1 \leq a_k) P((u + f(a_k)) \cdot \\ &\quad Z_1 - V_1 < x) + P(S_1 > a_k) P((u + f(a_k)) \cdot Z_1 < x) \\ &= q \cdot F_Z\left(\frac{x}{u}\right) + p \sum_{k=1}^K p_k \left(F_S(a_k) \cdot \int_0^\infty F_Z\left(\frac{v+x}{u+f(a_k)}\right) dF_V(v) + \right. \\ &\quad \left. (1 - F_S(a_k)) F_Z\left(\frac{x}{u+f(a_k)}\right) \right) \end{aligned}$$

2 若干精算量的分布

在本节中, 将对第 1 节中定义的风险模型式(1)讨论其若干精算量的分布问题, 由于从大数定律的角度来看, 随着时间的推移, 保险公司的盈余终将趋于无穷大, 因此, 首先讨论保险公司的盈余首次达到一个给定水平的时间的分布问题, 为此, 给出下面的定义。

定义 1 对于任意给定 $x > 0$, 定义盈余过程首次越过水平 x 的时刻 T_u^x 为

$$T_u^x = \inf\{n; R_n \geq x | R_0 = u\}$$

显然当 $x \leq u$ 时 $T_u^x = 0$, 因此只讨论 $x > u$ 的情形。

定理 1 设 T_u^x 的分布为 $\gamma_u^x(n) = P(T_u^x = n)$, 则 $\gamma_u^x(1) = 1 - F_u(x)$, 且

$$\gamma_u^x(n) = \int_{-\infty}^x \gamma_u^x(n-1) dF_u(s), n = 2, 3, \dots \quad (2)$$

进一步,该模型在有限时间内越过水平 x 的概率为

$$\mathbf{P}(T_u^x < \infty) = \sum_{n=1}^{\infty} \gamma_u^x(n) = 1.$$

证明 式(2)可以由类似于文献[10]中定理6.1的证明方法得到。为了证明 $\mathbf{P}(T_u^x < \infty) = 1$, 证当 $m \rightarrow \infty$ 时, 有 $\mathbf{P}(T_u^x > m) \rightarrow 0$ 成立, 为此在模型式(1)的基础上, 定义如下的盈余过程

$$\tilde{R}_n = \begin{cases} u, \\ \tilde{R}_{n-1} + U_n \cdot f(C_n) - U_n \cdot V_n \cdot I_{\{S_n \leq C_n\}}, \end{cases} n = 1, 2, 3, \dots \quad (3)$$

式(3)中各个参数的假设同模型式(1), 比较式(1)

和式(3)可知, 对于任意的 $n \geq 0$ 都有 $R_n \geq \tilde{R}_n$, 令 $W_n = U_n \cdot f(C_n) - U_n \cdot V_n \cdot I_{\{S_n \leq C_n\}}$, 则由前面的假设可知 $\{W_n, n \geq 1\}$ 是独立同分布的随机变量序列, 设它们与 W 同分布, 则 $\mathbf{E}W > 0$, 取 ε 满足 $0 < \varepsilon < \mathbf{E}W$, 对于定理中的 u 和 x , 取 m 足够大使得 $x - u < m(\mathbf{E}W - \varepsilon)$, 则

$$\begin{aligned} \mathbf{P}(T_u^x > m) &= \mathbf{P}(R_1 < x, R_2 < x, \dots, R_m < x) \\ &\leq \mathbf{P}(R_m < x) \\ &\leq \mathbf{P}(\tilde{R}_m < x) \\ &\leq \mathbf{P}(W_1 + W_2 + \dots + W_m < x - u) \\ &\leq \mathbf{P}(W_1 + W_2 + \dots + W_m < m(\mathbf{E}W - \varepsilon)) \\ &\leq \mathbf{P}\left(\frac{W_1 + W_2 + \dots + W_m}{m} - \mathbf{E}W < -\varepsilon\right) \\ &\leq \mathbf{P}\left(\frac{W_1 + W_2 + \dots + W_m}{m} - \mathbf{E}W < -\varepsilon\right) + \\ &\quad \mathbf{P}\left(\frac{W_1 + W_2 + \dots + W_m}{m} - \mathbf{E}W \geq \varepsilon\right) \\ &= \mathbf{P}\left(\left|\frac{W_1 + W_2 + \dots + W_m}{m} - \mathbf{E}W\right| \geq \varepsilon\right) \end{aligned}$$

由辛钦大数定律(文献[11]中定理5.2.5)可知定理1结论成立, 证毕。

上述精算量并未考虑保险公司是否破产, 实际上, 可以考虑保险公司在破产前的盈余情况, 为此, 设 $L_u(x) = \mathbf{P}(\sup_{0 \leq k \leq T_u} R_k \leq x)$, 它表示风险模型式(1)当初始准备金为 u 时破产前最大盈余的分布。由于当 $x < u$ 时, $L_u(x) = 0$, 因此只讨论 $x \geq u$ 的情形。类似于文献[10]中定理4.1的证明可得。

定理2 对于 $x \geq u$, $L_u(x)$ 是下面积分方程的唯一解,

$$L_u(x) = F_u(0) + \int_0^x L_s(x) dF_u(s)$$

保险公司是否破产是一个不确定事件, 公司破产也并不等价于倒闭, 对于正在运作中的保险公司而言, 破产可能只是暂时性的问题, 直观上, 保险公司能够运行的越久, 则破产的概率就越小, 因此考虑保险公司第一次破产持续的时间将是一个有意义的问题, 为此, 定义保险公司破产后盈余首次回复为非负值的时刻为 τ_u , $\tau_u = \inf\{n: n > T_u, R_n \geq 0\}$ 。从而公司的首次破产持续时间可以定义为

$$\hat{T}_u = \begin{cases} \tau_u - T_u, & T_u < \infty \\ 0, & T_u = \infty \end{cases}$$

记其概率分布为: $\varphi_n(u) = \mathbf{P}(\hat{T}_u = n), n \geq 1$ 。

为了给出破产持续时间概率分布以及下文中的其他结果, 先介绍一个引理, 为此, 令

$M_n(u, x) = \mathbf{P}(R_1 < 0, R_2 < 0, \dots, R_n < 0, R_{n+1} \geq x), x \geq 0$ 它表示 1 时刻破产且持续 n 个时间点后保险公司在 $n+1$ 时刻的盈余不小于 x 的概率, 则有

引理2

$$M_1(u, x) = \int_{-\infty}^0 (1 - F_S(x)) dF_u(s)$$

$$M_n(u, x) = \int_{-\infty}^0 M_{n-1}(s, x) dF_u(s), n = 2, 3, \dots$$

进一步, 若记 $M(u, x) = \sum_{n=1}^{\infty} M_n(u, x)$, 则

$$M(u, x) = M_1(u, x) + \int_{-\infty}^0 M(s, x) dF_u(s)$$

证明 类似于文献[10]中式(3.5)以及式(3.7)的推导可得。

定理3 对于初始准备金为 u 的风险模型式(1), 破产持续 n 个时间点的概率为

$$\varphi_n(u) = M_n(u, 0) + \int_0^{+\infty} \varphi_n(s) dF_u(s)$$

证明 类似于文献[10]中定理2.1的证明可得。

下面来讨论首次赤字结束后的瞬间保险公司的盈余分布问题。对于任意的 $x > 0$, 初始准备金为 u , 且破产持续 n 个时间点时盈余回复为正后的瞬间盈余不小于 x 的概率定义为:

$$N_n(u, x) = \mathbf{P}(R_{\tau_u} \geq x, \hat{T}_u = n), n \geq 1, x > 0$$

另外, 盈余首次由负值回复为正后的瞬间盈余不小于 x 的概率定义为 $N(u, x) = \mathbf{P}(R_{\tau_u} \geq x)$, 则类似于文献[10]中定理3.1和定理3.2的推导可得

定理4 对于 $N_n(u, x)$, 有

$$N_n(u, x) = M_n(u, x) + \int_0^{+\infty} N_n(s, x) dF_u(s), n \geq 1$$

对上式关于 n 从 1 到无穷求和得

$$N(u, x) = M(u, x) + \int_0^{+\infty} N(s, x) dF_u(s)$$

其中, $M(u, x)$ 由引理 2 给出。

最后来讨论破产时刻前后正负盈余与破产前最大盈余的联合分布, 实际上, 若定义破产前盈余、破产后赤字与破产前最大盈余的联合分布为

$$Q_u(x, y, z) = P(R_{T_{u-1}} \leq x, R_{T_u} \geq -y, \sup_{0 \leq k \leq T_u} R_k \leq z), \\ x > 0, y > 0, z > u > 0$$

则有

定理 5 对于上述 3 个精算量的联合分布, 有

$$Q_u(x, y, z) = (F_u(0) - F_u(-y)) I_{\{u \leq x\}} + \\ \int_0^z Q_s(x, y, z) dF_u(s)$$

证明 类似于文献[10]中定理 5.1 的证明可得。

注记: 本文主要结果中的分布都满足某个积分方程, 这些方程都具有相似的形式, 例如

$$\psi(u, x) = \varphi(u, x) + \int_a^b \psi(s, x) dF_u(s)$$

该类积分方程解的存在唯一性问题可以通过泛函分析中的压缩映射原理得到, 具体参见文献[10]中定理 7.1 的证明, 详情略。

3 结束语

本文考虑了一类离散时间的非寿险风险模型, 模型中的索赔是由保单驱动的, 而且带有随机利率。通过研究, 我们利用文献中的条件数学期望的方法证明了该模型的若干精算量的分布都可以由一类积分方程的唯一解给出。

参考文献:

- [1] 成世学. 破产论研究综述[J]. 数学进展, 2002, 31(5):403–422.
CHENG S X. The survey for researches of ruin theory[J]. Advances in Mathematics, 2002, 31(5):403–422.
- [2] 成世学, 伍彪. 完全离散的经典风险模型[J]. 运筹学学报, 1998, 2(3):42–54.
CHENG S X, WU B. Classical risk model in fully discrete setting [J]. OR Transactions, 1998, 2(3):42–54.
- [3] 孙立娟, 顾岗. 离散时间保险风险模型的破产问题[J]. 应用概率统计, 2002, 18(3):293–299.
SUN L J, GU L. Ruin problems for the discrete time insurance risk model[J]. Chinese Journal of Applied Probability and Statistics, 2002, 18(3):293–299.
- [4] 蒲冰远, 唐应辉, 刘燕. 离散风险模型破产问题的进一步研究[J]. 电子科技大学学报: 自然科学版, 2007, 36(2):382–383.
PU B Y, TANG Y, LIU Y. Further analysis about ruin in discrete risk model[J]. Journal of University of Electronic Science and Technology of China: Natural Science, 2007, 36(2):382–383.
- [5] 杨娟, 陈圣滔. 两个离散风险模型破产问题的研究[J]. 长春大学学报: 自然科学版, 2006, 16(5):6–8.
YANG J, CHEN S T. Ruin problem for the discrete time insurance model with random rates of interest[J]. Journal of Changchun University: Natural Science, 2006, 16(5):6–8.
- [6] 钟朝艳, 黑韶敏. 带随机利率离散时间风险模型的破产问题[J]. 大理学院学报: 自然科学版, 2007, 6(2):55–57.
ZHONG C Y, HEI S M. Bankruptcy problem in discrete time risk model with random rates of interest[J]. Journal of Dali University: Natural Science, 2007, 6(2):55–57.
- [7] LI Z H, ZHU J X, CHEN F. Study of a risk model based on the entrance processes[J]. Statistic & Probability Letters, 2005, 72: 1–10.
- [8] 唐加山, 徐小阳. 混合双险种风险模型的破产概率[J]. 南京邮电大学学报: 自然科学版, 2009, 29(1):43–45.
TANG J S, XU X Y. Ruin probabilities of a mixed double-line insurance risk mode[J]. Journal of Nanjing University of Posts and Telecommunications: Natural Science, 2009, 29(1):43–45.
- [9] XIAO H M, TANG J S. Ruin probability of one kind of entrance processes based insurance risk model[R]. Nanjing University of Posts and Telecommunications, 2009.
- [10] 俞雪梨, 肖纲景. 随机利率离散时间风险模型的破产问题[J]. 应用概率统计, 2009, 25(2):38–46.
YU X L, XIAO G J. Ruin problems for the discrete time risk model under stochastic rates of interest[J]. Chinese Journal of Applied Probability and Statistics, 2009, 25(2):38–46.
- [11] 夏乐天. 应用概率统计[M]. 北京: 机械工业出版社, 2008.
XIA L T. Applied probability and statistics[M]. Beijing: China Machine Press, 2008.

作者简介:

徐小阳(1985–), 女, 贵州安顺人。南京邮电大学通信与信息工程学院硕士生。(见本刊 2009 年第 1 期第 45 页)。

唐加山(1968–), 男, 安徽天长人。南京邮电大学理学院统计系主任, 教授, 博士。(见本刊 2009 年第 1 期第 45 页)

应用层组播的效率优化技术研究

饶 翔¹, 张顺颐¹, 许建真², 陈 涛¹, 周 筠¹

(1. 南京邮电大学 信息网络技术研究所, 江苏南京 210003
2. 南京邮电大学 实验室建设与设备管理处, 江苏南京 210046)

摘要:应用层多播(ALM)作为 IP 多播的替代在互联网中得到广泛的应用。提出一个基于优先级的动态分层应用层多播模型 CDMP(The Classified Dynamic Model Based on the Priority in Application-Level Multicast)。该模型通过对不同性质的结点的分层来搭建整个架构, 通过分域将性质接近的结点放在一起以保证数据传输的效率和数据共享的公正性, 并且减少了结点寻找路由的代价, 加快了当结点失败时的路由修复速度。通过分析和仿真证明, CDMP 协议具有良好的上述性质。

关键词:应用层多播; 优先级; 效率

中图分类号: TP393 文献标识码: B 文章编号: 1673-5439(2009)06-0079-06

Study of the Efficiency Optimization Technology of ALM

RAO Xiang¹, ZHANG Shun-yi¹, XU Jian-zhen², CHEN Tao¹, ZHOU Jun¹

(1. Institute of Information Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
(2. Department of Laboratory Construction and Equipment Management, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: ALM has been widely deployed in internet these years. We present an ALM model called The Classified Dynamic Model Based on the Priority in ALM (CDMP). We classify hosts into different levels and areas according to node's priority. This architecture guarantees network stability and justness, as well as less traffic during the dynamic routing discovery process, it also ensures broken routes could be repaired locally without rediscovery. By the analysis and simulation, CDMP has achieved good performance.

Key words: application layer multicast; priority; efficiency

0 引言

多播是一种点到多点(或多点到多点)的通信方式, 即多个接收者同时接收源地址发送的相同信息。目前有两种多播方式, 一种工作在网络层, 被称为 IP 多播, 另一种则工作在应用层, 称为应用层多播(ALM, Application Layer Multicast)。IP 多播是对互联网“单播、尽力发送”模型的重要扩充, 多播的主要功能在路由器上实现, 通过合并重复信息传输来减少带宽浪费和降低服务器的负担。由于 IP 多播在传输技术和管理上存在很多问题^[1-5], 目前没有在互联网中普遍采用。

ALM 的主要思想为使各项应用自我组织到一个逻辑覆盖网络, 通过该网络进行点到点的单播传输。相比于 IP 多播, ALM 显然在效率上低一些, 但 ALM 更加容易部署, 不需要网络层的多播支持, 比如支持多播的路由器, 也不需要一个全局的认证号(比如 IP 地址)。另外由于 ALM 传输时为点对点单播, 单播的优点(诸如流量控制, 拥塞控制, 信任传输)都可以在 ALM 中研究并利用。但是另外一方面, 由于数据在终端之间传输, 端与端之间的延迟将会较大, 转发速率则会较小。此外, 当覆盖层的多条边映射到一条相同的网络层链路时, 则会导致相同数据的多个拷贝在同一条链路传输, 致使带宽使用

率降低。所以,在与IP多播比较时,端对端的延迟和带宽的利用是两个比较重要的衡量标准。

1 CDMP模型

1.1 CDMP的设计理念

ALM假定多播成员都主动且友好地贡献资源用于转发,且转发资源相对丰富。而实际情况的网络中有各种性能不同的终端,甚至有恶意的结点,因此为使不同的个人、组织和资源间实现安全、协调地数据传输,对成员采用区别对待来保证整个模型的公正性及有效性是有必要的。本文主要采取的对结点的衡量标准为:

(1) 结点本身性能,如:内存、CPU等;

(2) 结点拥有的带宽,本模型中带宽与结点的出度成正比;一个拥有高带宽的结点负责向更多的结点传输数据。对于一个终端而言,它的出度数,即它在应用层多播中的子结点的数量,要低于某个限定值 d_{\max} ,否则终端所具有的计算资源和网络资源将无法满足复制、分发多播信息的要求,从而造成多播通信的性能下降。当结点的出度达到这个限定值时,称为结点达到饱和,否则称为未饱和。又设某结点已使用出度为 d_u ,则该结点的饱和度 η 定义为:

$$\eta = \frac{d_u}{d_{\max}} \times 100\% \quad (1)$$

(3) 结点在模型中所处的时间。在P2P网络中,结点的分布是遵循power-law^[6],由PLOD(power_law out_degree)^[7]算法产生。Power-law的含义是指节点的度数为 m 的概率与 m 的 $-\lambda$ 次幂成比例,即 $P(d=m)=c \cdot m^{-\lambda}$,其中 c 为一常数, $\lambda \in [2.2, 2.4]$ 。

它的另一个含义的数学公式表达为:

$$P\{X > T\} = T^{-\alpha} \quad (2A)$$

其中, X 为一随机变量,代表结点的生命均值, T 为指代时间。

由于它的记忆性,它的条件概率为:

$$P\{T > b | T' = a\} = \left(\frac{b}{a}\right)^{-\alpha} \quad (2B)$$

其中, T 为指代结点生命的时间变量, T' 即指现在的生命长度。该公式表明,结点将来存在时间的长度与已经存在的生命长度成正比。即当某结点在ALM中存在时间越长,它继续存在的可能性越大,无疑这样的结点具较高的稳定性。

1.2 CDMP的拓扑模型描述及其分析

(1) 考核层(testing-nodes layer)。

考核层(testing-nodes layer)为最底层,层中结点皆为叶子结点,即无其它结点从它接收数据。本模型中加入一个考核层的原理在于防止恶意结点的频繁加入退出,影响周边结点的链接关系,从而导致模型的内部调整,增大模型的维护开销。故每一个新加入的结点进入考核层,只承担叶子结点的角色,在一定的时间内,若没有退出,则进入上一层。CDMP拓扑模型见图1。

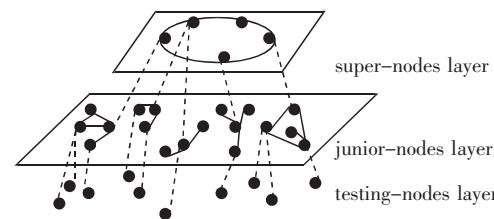


图1 CDMP 拓扑模型

设到达系统的结点数目 $X(t)$ 符合强度为 $\lambda > 0$ 的泊松过程,即: $X(0)=0$; $X(t)$ 是独立增量过程;并且在长度区间 t 内:

$$P\{X(t+s)-X(t)=n\}=e^{-\lambda t} \frac{(\lambda t)^n}{n!}, n=0,1,\dots \quad (3)$$

由简单的证明可知,假设考核层的各个结点继续进入普通层的概率为 p , $p \in (0,1)$,那么考核层结点进入普通层的结点数 $Y(t)$ 符合强度为 λp 的泊松过程,即:

$$P\{Y(t+s)-Y(t)=n\}=e^{-\lambda p t} \frac{(\lambda p t)^n}{n!}, n=0,1,\dots \quad (4)$$

故考核层的增加不仅对恶意结点具备有效的剔除能力,同时对CDMP模型的整体架构不产生任何动荡,恰恰是一个平稳的过程。

(2) 普通层(junior-nodes layer)。

普通层(junior-nodes layer)是一个二维的平面层,层中结点从底层经考核进入,层中分为几个独立的域,一个域的定义如下:

1) 每个域有一个leader结点,负责从上层结点接受数据,并且向本域内结点传输。leader结点保存域中其他结点的所有信息。

2) 每个结点有唯一的坐标(talent, life), talent为结点的性能经评估后的值,life为结点在模型中已经存在的生命时间。每个坐标对应唯一的一个结点。

3) 域内任何两个结点间是连通的。

由其结点的坐标定义可见,层中结点只要不退

出,其稳定性是随时间的推移而增长的,层中所有的结点呈向右“漂移”的状态。

另外设一个域中结点集为 $V = \{\text{leader}, N_1, N_2, \dots, N_m\}$,各结点的最大出度和已使用出度为 $\{(d_u, d_{\max}), (d_{u1}, d_{\max1}), \dots, (d_{um}, d_{\maxm})\}$,我们给出域的饱和度 η 的定义:

$$\eta = \frac{\alpha \cdot d_u + (1 - \alpha) \cdot \sum_{i=1}^m d_{u_i}}{\alpha \cdot d_{\max} + (1 - \alpha) \cdot \sum_{i=1}^m d_{\max_i}} \times 100\% \quad (5)$$

其中, $\alpha \in (0, 1)$ 为该域的系统常量,平衡 leader 结点和普通成员结点间的权值。当 $\eta = 1$ 时,称该域已经达饱和,当 $\eta \geq m$ (其中 $m \in (0, 1)$ 为一系统常量)时,称该域已经达到准饱和的状态而当新结点加入普通结点层时不接受该新结点作为本域的成员结点。

(3) 超级结点层(super-nodes layer)。

超级结点层(super-nodes layer)其中结点皆为性能较高,状态稳定的结点,负责管理整个拓扑模型。层中结点从 normal-nodes 中选拔进入。各个超级结点间环状互联。每个结点保存有自己在本层中的唯一 ID、前驱结点 ID(pred)、后继结点 ID(succ)及其本结点和后继结点相对应的普通层的 leader 结点。采取环状的理由为:

1) 一个环的度为常数 1,即 $O(1)$,这大大减少了拓扑复杂度,例如在秘钥分布机制中;

2) 环形结构没有路由选择,不会出现拥塞和死锁,易于实现分布控制和高速通信;

3) 通过携带预定和流控制信息的“令牌”(token),可以有效地实现安全、可靠、完全有序的消息传输。

由于环状结构中某一结点一旦发生故障,将导致整个环路不通,故为保证更可靠的传输,可在本层结点中让每个结点保存 m 位它的后继结点。鉴于 super-node 的稳定性,某一结点成为失败结点的情况为小概率事件,故给定一个合理的 m 值, m 个超级结点同时失效的事件可视为不可能事件。

由其拓扑描述可知:

任何一个结点只属于其中的某一特定的层;

特定的层代表结点的性能及其稳定性和在整个拓扑模型中的级别;

在同层中,只有普通层中的结点根据其坐标有相应的不同的优先级;

任何一个上层结点都是由下自上渐进式的。

1.3 CDMP 协议的结点维护

(1) 结点加入。

新结点 N 加入时发送 request 信息,请求信息中将包含自己的位置信息,终端的性能以及带宽信息。引导程序将根据它的地理位置给出最近的 normal-node 层的某一结点 M 的信息, N 连于 M ,在特定时间 $\Delta = \mu / \text{talent}$ (μ 为一常数)之后,则再由引导程序将其映射到第二层,通过一个简单的映射函数:

$$f(talent, life) = talent \cdot i + life \cdot j \quad (6)$$

映射到第二层中点(talent, life),若发现坐标位置(talent, life)已被使用时,则采取退避策略,对 life 进行 $\pm \varepsilon$ (ε 为给定的一较小的值)修正,成功映射后结点计算与周边 leader 结点的逻辑距离,逻辑距离的计算公式:

$$d = \sqrt{\lambda(T_i - talent)^2 + (1 - \lambda)(L_i - life)^2} \quad (7A)$$

其中, $\lambda \in (0, 1)$, 为一系统常数,最后得到其中的最小值所对应的 leader 结点作为将来自己所在域的 leader 结点,亦即计算出:

$$d_{\min} = \min_{T_i, L_i \in D} \left\{ \sqrt{\lambda(T_i - talent)^2 + (1 - \lambda)(L_i - life)^2} \right\} \quad (7B)$$

其中, D 为以(talent, life)为圆心的圆,如图 2a 所示,即得到和自己性能最接近的周边的 leader 结点的信息 P 并向 P 发送 JOIN 信息,若 P 所在的域未达到准饱和的状态,则 P 给出 N 的可能父结点的信息,考虑到实际网络中的链路传输压力,在本模型中 P 负责给出本域中离 N 物理距离最近的结点 Q_1 的信息,若 Q_1 未饱和,则 N 连于 Q_1 。否则, P 再给出离 N 结点次近的结点 Q_2 ,若 Q_2 未饱和,则 N 连于 Q_2 , N 向 P 发送 JOIN_OK 信息,否则, P 再给出离 N 结点次近的结点 Q_3 ,依次类推,直至 N 成功连至某一父结点(见图 2b)。若 P 所在的域已经达到准饱和状态,则 N 扩大起先的搜索半径,选择距 N 结点逻辑距离次近的 leader 结点,如图 2 所示, R 结点,再依次类推,直至连至 R 域中某未饱和结点。至此, N 成功进入第二层。为保证模型的可靠性传输, N 结点可同时保存自己的父结点的父结点的信息为自己的第二父结点。

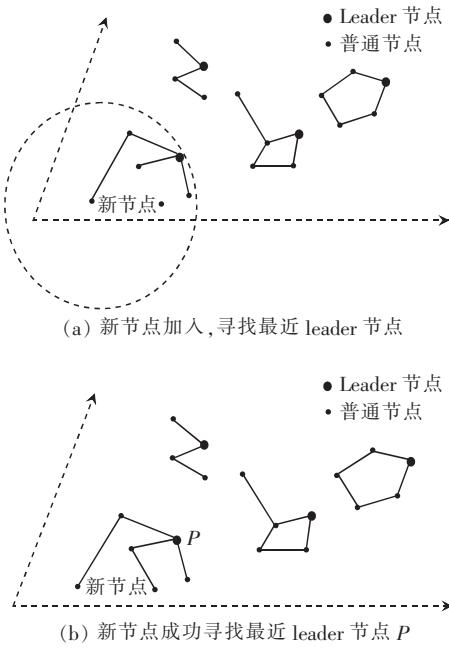


图2 新节点加入过程

(2) 结点退出。

结点退出分为友好退出和不正常退出。在友好退出下,退出的结点发送 Remove 信息给自己的父结点并自己所在域的 leader 结点。leader 结点接收到信息并在自己的信息中去除该结点的信息。但若为不友好退出,即退出结点无法发送 Remove 信息,则由其周边结点通过 heartbeat 测出它已不在而自动删除其相关信息并上报 leader 结点,若它有连于它的子结点,其子结点自动连至其第二父结点。若某结点的父结点和第二父结点同时失效,则该结点重新向 leader 结点发送 JOIN 信息,并且该结点的 life 值不变。

若失败结点为 leader 结点,则当域中其他结点发现其退出后,则根据结点的性质选举新的 leader 结点,其性质表现为这个性能的终端在时间 life 下在系统中的累积表现值,数学表达式如下:

$$p = \text{talent}^2 \int_0^{life} \ln(t+1) dt \quad (8)$$

由于对数函数 $\ln t$ 平滑递增的曲线,并且其一阶导 $\ln t' = 1/t$,即随着时间 t 的增长其函数值 $\ln t$ 增长的幅度越来越小,故式中选择函数 $\ln t$ 作为被积函数。当选择出最大的 p 值时,leader 结点亦诞生了,诞生后新的 leader 结点采用原有 leader 结点的坐标,以便原 leader 结点的 super_father 结点能够依旧正常地往下传数据。

若失败结点为 super-node 结点 u ,则由 u 的前驱结点 p 通过 heartbeat 测出,之后 p 采取两个步骤。首先 p 发送 LINK_REPAIR 给 u 的后继结点 s , s 收

到后回复 LINK_REPAIR_ACK 信息。同时 p 又发送修复信息于 u 对应的普通层的 leader 结点 l , l 回复。至此,修复成功。

(3) 模型的结点分布控制及其调整。

首先,给定域空间的相关定义如下:设一个域中结点集 $V = \{leader, N_1, N_2, \dots, N_m\}$,对应的坐标为 $\{(talent, life), (talent_1, life_1), \dots, (talent_m, life_m)\}$,各结点的最大出度和已使用出度为 $\{(d_u, d_{max}), (d_{u1}, d_{max1}), \dots, (d_{um}, d_{maxm})\}$,域的直径 d 定义为 leader 结点与其它域内结点间的最大逻辑距离,即:

$$d = \max_{i \in [1, 2, \dots, m]} \{ \sqrt{\lambda(talent - talent_i)^2 + (1 - \lambda)(life - life_i)^2} \} \quad (9)$$

域的空间大小 Θ 定义为以 d 为直径的圆的面积:

$$\Theta = \pi d^2 \quad (10)$$

域的密度 ρ 的定义:

$$\rho = \frac{m+1}{\Theta} = \frac{m+1}{4\pi d^2} \quad (11)$$

域的性质 P 定义为域中各结点的性质的加权平均,即:

$$P = \frac{\beta \cdot p + (1 - \beta) \sum_{i=1}^m p_i}{m+1} = \frac{\beta \cdot talent^2 \int_0^{life} \ln(t+1) dt + (1 - \beta) \sum_{i=1}^m talent_i^2 \int_0^{life} \ln(t+1) dt}{m+1} \quad (12)$$

其中, $\beta \in (0, 1)$ 为一系统常量,用来权衡 leader 结点和成员结点的权值。

1) 域的合并和拆分。

由模型的定义及其描述,不难看出,随着时间的增长,模型中的整个第二层中的各个域向拓扑图的右边,即 x 轴正方向推进。且随着个别结点的退出,距 y 轴越远的域,其中的结点数目将越小。模型中为防止某一域的组员过多过少,或分布区域的太大而导致系统的不稳定,每个域的 leader 结点定期地检查本域的情况,并且根据检查的结果做出合理的拆分或者合并的工作。

故此,在模型中设定当某个域符合某特定条件:空间大小 Θ 大于某特定值 Θ_{max} 后,或者密度 ρ 小于特定值 ρ_{min} 时,与周边的域进行合并。

合并时受以下条件限制:合并后的域空间大小不超过 Θ_{max} ,或者拆分后的域的密度不低于 ρ_{min} 。

2) super-node 结点的选取。

Super-nodes 层的结点为整个拓扑模型的骨干结点,它们的出度应足以应付底下层的 leader 结点的需要,故当 super-nodes 层的饱和度达到某一特定值 η_{\max} 时,normal-nodes 层将有部分性能较高的结点进入超级结点层,为降低计算复杂度,在本文中将选取集体性能较优的域,即根据式(12),选取最大的 P 值,得到的相应的域即作为候补进入超级结点层。

2 仿 真

2.1 仿真环境及实现

所有实验都在一台 PC 机上完成,PC 机的配置为 AMD Athlon (TM) XP 1800+, CPU 频率 1150.134 MHz, 内存 512 Mb, 操作系统内核为: Linux 2.6.20-15-generic i686 GNU/Linux。仿真工具为 NS-2 (The Network Simulator)^[8], 该仿真器对 TCP 协议、路由协议和在无线网络或者有线网络环境中的多播协议都有重要的支持。

仿真拓扑结构图通过 NS2 内置的 GT-ITM^[9]来生成。初始拓扑中的结点按指数分布(Exponential (On/Off) model)动态进入 CDMP 环境, Exponential (On/Off) model 有 4 个参数:[start time], up interval, down interval, [finish time], 仿真过程皆采用默认值:<start time> 默认值是从仿真开始后 0.5 s; <finish time> 默认值是仿真结束的时刻; <up interval>、<down interval> 指明了指数分布的意义, 定义了节点或链接将要各自 up 和 down 的时间。up 和 down 间隔值默认分别为 10 s 和 1 s。每个网络节都是采用 DropTail queue 的方式。我们在结点与结点之间建立 UDP 的联机,并在其上架构 CBR (Constant Bit Rate) 应用程序, CBR 流的源结点能以固定速率产生数据。CBR 的传送速度为 1 Mb, 每一个封包大小为 1 000 Bytes。结点的处理延迟分布在 10 到 90 ms 之间, 结点间的带宽分为 10 MP 或者 100 MP。

仿真最后的结果通过 gawk 对流量跟踪文件(traffic trace file)的分析得出,由绘图工具 Gnuplot 对二进制数据文件作图。

2.2 仿真结果

图 3 为当恶意节点从 0% 逐渐增加到 50% 的情况下,对系统的开销的变化。“×”为没有考核层的情况,“+”为系统加上考核层的情况。由图 3 可

见,当没有 testing_level 时,恶意节点可轻易的成为非叶子节点,进而其频繁的加入退出对系统的震荡就大,而当存在 testing_level 时,恶意节点的加入退出开销仅为针对叶子节点的开销,虽也呈线性增长,但显然斜率小很多,当恶意节点数目增多时,考核层的考核效果相当明显。

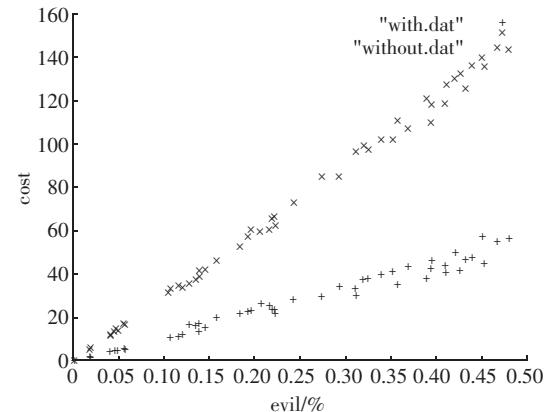


图 3 恶意节点变化对系统开销影响

图 4 显示为在普通层节点间 100 条链路的情况下,各链路上的端对端延迟(End-to-End Delay)的情况下(单位为 10 ms),包括了处理时延和传输时延。“×”为不分域情况下的分布,“+”为分域情况下的时延分布。由于不分域情况下,处理时延较小的节点会以一定概率连接到时延较大的,从而导致两者间的链路上的总时延较大,但在分域情况下,相接近性能的节点是在同一个域内的,不存在太大牵连的情况。

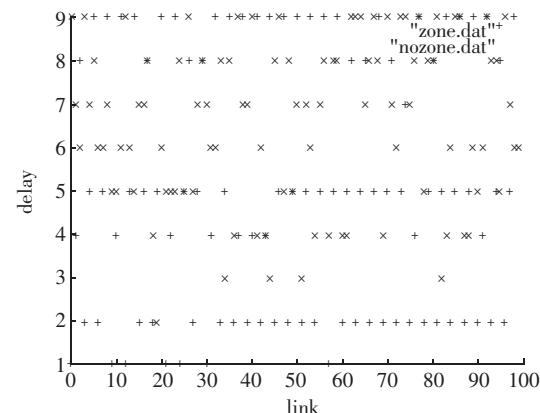


图 4 普通层链路上端对端时延分布

图 5 为对图 4 的两种情况下的值取平均后的值(平均时延),为求其统计效果,我们对此进行了 25 次的仿真,可以看到,分域后的各链路的时延接近在 50 ms 左右,而不分域的情况下则是 65 ms 左右,分域后平均时延降低了 30%。

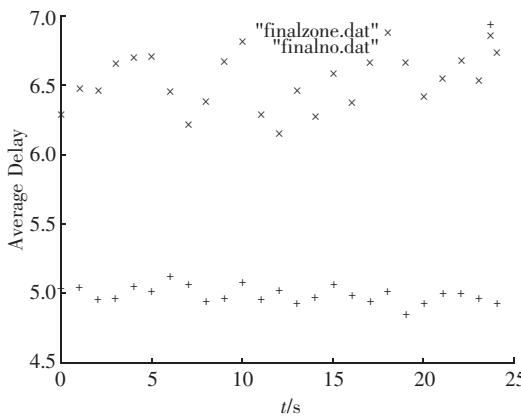


图5 普通层链路上端对端时延分布均值

3 结束语

整体上,CDMP的3层模型结构把不同性质的结点分布在不同的层中,有利于整个模型的传输性能的稳定和效率;在CDMP的第二层中,把性能优良和稳定性较高的成员聚集在一个域,通过保证友好的成员受到友好的待遇来提高模型的公正性,并且由于结点的路由查找和修复皆在域内进行而使结点的路由查找和修复的代价都得到减少;当发生拥塞时,通过及地的剔除优先级较低的成员以使整个系统稳定,同时也继续使得优良结点在系统中接收和传输数据,具有一定的激励性质;另外由于结点的出度与结点的带宽成正比,这极大地提高了网络的带宽利用率。仿真证明,通过分级和分域有利于整个系统的稳定和公正并进行快速的数据分发。

参考文献:

- [1] JANNOTTI J, GIFFORD D K, JOHNSON K L, et al. Overcast: Reliable Multicasting with an Overlay Network[C]// Proc of OSDI. October 2000:197–212.
- [2] CHU Y H, RAO S G, ZHANG H. A case for end system multicast [C]// Proc of ACM Sigmetrics. June 2000:1–12.
- [3] ZHUANG S, ZHAO B, JOSEPH A, et al. Bayeux : An Architecture

for Scalable and Fault-tolerant Wide-Area Data Dissemination[C]// Proc of NOSSDAV. June 2001.

- [4] RATNASAMY S, HANDLEY M, KARP R. Application-level multicast using content-addressable networks[C]// Proc of NGC. 2001.
- [5] ROWSTRON A, KERMARREC A M, CASTRO M, et al. The design of a large-scale event notification infrastructure[C]// Proc of NGC. 2001.
- [6] BUSTAMANTE F E, QIAO Y. Peer lifespan and its role in p2p protocols[C]// Proc of the International Workshop on Web Content Caching and Distribution. Sept 2003.
- [7] RIPEANU M. Peer-to-Peer architecture case study: Gnutella network [R]. Chicago: University of Chicago, 2001.
- [8] BRESLAU L, ESTRIN D, FALL K, et al. Advances in Network Simulation[J]. IEEE Computer, 2000(5):59–67.
- [9] GT – ITM. Georgia tech internetwork topology models[EB/OL]. <http://www.cc.gatech.edu/projects/gitm/>

作者简介:



饶翔(1977-),男,贵州印江人。南京邮电大学信息网络技术研究所博士生。主要研究方向为下一代网络性能监测与优化、网络QoS监测控制、网络业务感知技术。

张顺颐(1944-),男,江苏南京人。南京邮电大学信息网络技术研究所教授,博士生导师。(见本刊2009年第1期第5页)

许建真(1966-),男,安徽砀山人。南京邮电大学实验室建设与设备管理处副处长,副教授。研究方向为计算机网络与通信技术。

陈涛(1985-),女,浙江宁波人。南京邮电大学计算机专业硕士研究生。研究方向为计算机通信与网间互连技术。

周筠(1985-),女,江苏南京人。南京邮电大学计算机专业硕士研究生。研究方向为计算机通信与网间互连技术。

随机工艺变化下互连线 ABCD 参数建模与仿真

张瑛^{1,2}, 王志功², 方承志¹, 杨恒新¹

(1. 南京邮电大学 电子科学与工程学院, 江苏南京 210046
2. 东南大学 射频与光电集成电路研究所, 江苏南京 210096)

摘要:集成电路的不断发展使得互连线的随机工艺变化问题已经成为影响集成电路设计与制造的重要因素。基于电报方程建立了工艺变化下互连线的分布参数随机模型,推导出互连线 ABCD 参数满足的随机微分方程组,并提出了基于蒙特卡洛法的互连线 ABCD 参数统计分析方法,通过对 ABCD 参数各参量系数的正态性进行偏度-峰度检验,给出了最差情况估计。实验结果表明所提出的互连线随机模型及统计分析方法可以对工艺变化下的互连线传输性能进行有效的评估。

关键词:互连线;电报方程;随机建模;随机微分方程;蒙特卡洛法

中图分类号:TN47 文献标识码:B 文章编号:1673-5439(2009)06-0085-06

Modeling and Simulating the ABCD Parameters of Interconnects in the Presence of Random Process Variations

ZHANG Ying^{1,2}, WANG Zhi-gong², FANG Cheng-zhi¹, YANG Heng-xin¹

(1. College of Electronics Science and Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210046, China
2. Institute of RF&OE-ICs, Southeast University, Nanjing, 210096, China)

Abstract: With the great development of Integrated Circuits(IC), random process variations of interconnects has been an important factor which impacts the IC design and manufacture. On the basis of telegraph equation, the stochastic interconnect model with distributed parameters is proposed in the presence of process variations. The stochastic differential equation of the ABCD parameters is derived, and Monte Carlo method based statistical analysis method for the ABCD parameters is presented. Jarque-Bera test is made for the normality, and the worst-case estimation is given. Experimental results demonstrate that the proposed stochastic model and the statistical analysis method can evaluate the transmission performance of interconnects in the presence of process variations effectively.

Key words: interconnects; telegrapher's equation; stochastic modeling; stochastic differential equation; Monte Carlo method

0 引言

随着超大规模集成电路迅速向高速度与大规模方向进展,作为连接装置的互连线在集成电路设计与制造中正在扮演着越来越重要的角色。互连线寄生耦合效应的存在使得电路系统性能更加难以预测,其对信号产生的反射、散射及串扰等影响可能导

致整体电路设计的失败。而随着半导体制造工艺和技术的提高,器件尺寸持续减小,而刻蚀、注入等工艺步骤的扰动并没有相应的减小,因而集成电路特性对工艺扰动的灵敏度在增加^[1]。同样的问题也存在于互连线的制造过程中,电路设计者需要知道制造工艺参数产生的变化对互连线传输性能的影响,因此有必要对其进行估计和预测^[2-5]。目前射

频与微波集成电路中互连线的长度已经与信号波长可以相比拟,甚至大于信号波长,此时应采用具有分布参数的传输线模型及理论对互连线的工作情况进行分析。

针对上述问题,本文考虑工艺扰动导致的工艺参数随机变化对互连线电气分布参数的影响,采用数值仿真及拟合方法得到电气分布参数的近似表达式,并引入高斯宽平稳随机过程建立了互连线分布参数的随机模型;推导出互连线ABCD参数满足的复随机微分方程组,并提出了基于蒙特卡洛法的互连线ABCD参数统计分析方法,通过对ABCD参数各参量系数的正态性进行偏度-峰度检验,给出了最差情况估计。

1 互连线的分布参数随机模型

随着集成电路的工作频率和集成度的不断提高,其互连线的分布参数效应愈加严重,需采用传输线模型进行分析。设互连线的电容、电感、电导和电阻等电气分布参数分别为 $C(z), L(z), G(z)$ 和 $R(z)$,其中 z 为互连线长度坐标,于是有相应的传输线时域电报方程

$$\begin{cases} -\frac{\partial u(z,t)}{\partial z} = R(z)i(z,t) + L(z)\frac{\partial i(z,t)}{\partial t} \\ -\frac{\partial i(z,t)}{\partial z} = G(z)u(z,t) + C(z)\frac{\partial u(z,t)}{\partial t} \end{cases} \quad (1)$$

其复频域形式为:

$$\begin{cases} -\frac{\partial U(z,s)}{\partial z} = R(z)I(z,s) + sL(z)I(z,s) \\ -\frac{\partial I(z,s)}{\partial z} = G(z)U(z,s) + sC(z)U(z,s) \end{cases} \quad (2)$$

互连线的电气分布参数由其内部导体形状与尺寸以及介质的材料与分布等工艺参数决定,在生产过程中这些工艺参数由于工艺扰动等因素的影响,不可避免地会产生一些随机变化,目前所采用的经典模型都将这些随机变化所引起的偏差忽略掉了。但随着互连线尺寸越来越小,而工艺步骤的扰动并没有相应的减小,这些偏差就有必要给予足够的重视。

在互连线的设计与制造中会有多个主要的制造工艺参数,如导体的宽度、导体的高度和介质厚度等等,但为讨论方便,首先考虑一个制造工艺参数的情况。

令 $\gamma(z)$ 表示工艺参数,考虑工艺变化的影响,互连线的电阻等电气分布参数分别为:

$$\begin{aligned} \tilde{R}(z) &= H_r[\gamma(z)] = H_r[\gamma_0(z) + \lambda(z)] \\ \tilde{G}(z) &= H_g[\gamma(z)] = H_g[\gamma_0(z) + \lambda(z)] \\ \tilde{L}(z) &= H_l[\gamma(z)] = H_l[\gamma_0(z) + \lambda(z)] \end{aligned} \quad (3)$$

$$\tilde{C}(z) = H_c[\gamma(z)] = H_c[\gamma_0(z) + \lambda(z)]$$

其中, $\gamma_0(z)$ 为工艺参数设计的标称值, $\lambda(z)$ 表示由工艺参数的随机变化量,函数 H_r, H_g, H_l, H_c 表示相同的工艺参数对不同电气分布参数的不同影响。

工艺参数的随机变化量 $\lambda(z)$ 与其标称值 $\gamma_0(z)$ 相比是相对较小的量,因此对式(3)进行泰勒展开且只保留一阶项得到^[2,5]:

$$\begin{aligned} \tilde{R}(z) &\approx H_r[\gamma_0(z)] + \left. \frac{\partial H_r}{\partial \gamma} \right|_{\gamma=\gamma_0} \cdot \lambda(z) \\ &= R(z) + h_r(z) \cdot \lambda(z) \\ \tilde{G}(z) &\approx H_g[\gamma_0(z)] + \left. \frac{\partial H_g}{\partial \gamma} \right|_{\gamma=\gamma_0} \cdot \lambda(z) \\ &= G(z) + h_g(z) \cdot \lambda(z) \end{aligned} \quad (4)$$

$$\begin{aligned} \tilde{L}(z) &\approx H_l[\gamma_0(z)] + \left. \frac{\partial H_l}{\partial \gamma} \right|_{\gamma=\gamma_0} \cdot \lambda(z) \\ &= L(z) + h_l(z) \cdot \lambda(z) \end{aligned}$$

$$\begin{aligned} \tilde{C}(z) &\approx H_c[\gamma_0(z)] + \left. \frac{\partial H_c}{\partial \gamma} \right|_{\gamma=\gamma_0} \cdot \lambda(z) \\ &= C(z) + h_c(z) \cdot \lambda(z) \end{aligned}$$

其中, $R(z), G(z), L(z), C(z)$ 是与互连线工艺参数设计的标称值对应的电阻等电气分布参数的标称值。

鉴于目前绝大多数互连线结构的复杂性,互连线寄生效应的完全建模非常困难,其分布参数的解析解一般很难得到,通常采用数值方法进行参数提取。此外,VLSI(Very Large Scaled Integrations)中的布线现虽已达10层以上,但邻近层间的相互影响以及与接地平面相耦合的寄生效应仍是其主要部分。下面举例说明如何采用数值仿真和曲线拟合方法来计算 $h_r(z), h_g(z), h_l(z), h_c(z)$ 。

图1所示为一互连线截面,x轴为接地平面,当导体2的宽度和高度发生变化时,导体2的自电容分布参数 C_{22} 也随之变化。通过数值计算^[6-8]得到其变化规律如图2和图3所示。由图2和图3,当导体2的宽度及高度在一定范围内变化时,采用曲线拟合方法可以得到导体2的自电容参数与它们之间的近似表达式为

$$\begin{aligned} C_{22} &= C + a \cdot \Delta W \\ C_{22} &= C + b \cdot \Delta H \end{aligned} \quad (5)$$

其中, C 为互连线电容的标称值, a, b 为常量, $\Delta W, \Delta H$ 分别为导体 2 宽度与高度的变化量。比较式(4)与式(5), 显然当工艺变化量较小时, $h_r(z), h_g(z), h_l(z), h_c(z)$ 可分别取为常量 h_r, h_g, h_l, h_c 。

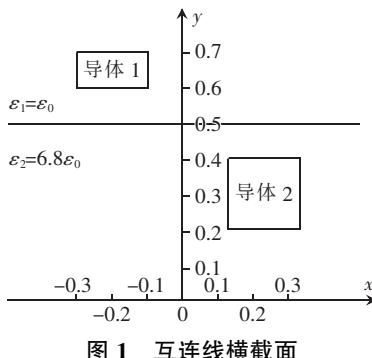


图 1 互连线横截面

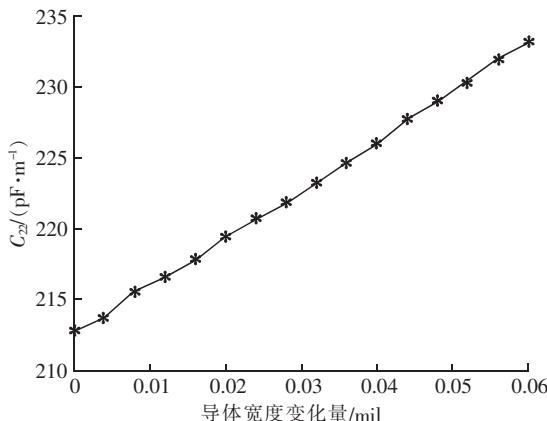


图 2 导体 2 宽度变化引起的变化曲线

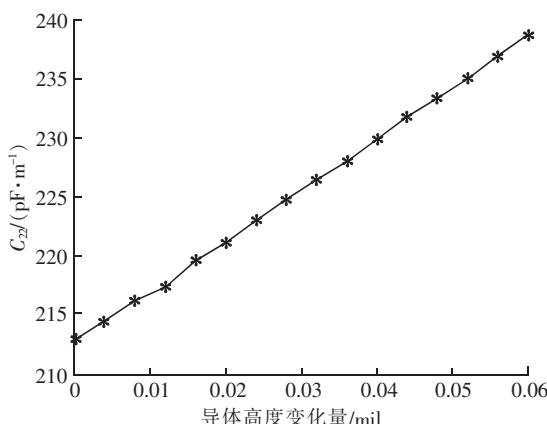


图 3 导体 2 高度变化引起的变化曲线

因此由式(2)和式(4)得到电报方程的复频域随机形式:

$$\begin{cases} -\frac{\partial U(z,s)}{\partial z} = \tilde{R}(z)I(z,s) + s\tilde{L}(z)I(z,s) \\ -\frac{\partial I(z,s)}{\partial z} = \tilde{G}(z)U(z,s) + s\tilde{C}(z)U(z,s) \end{cases} \quad (6)$$

其中, $\tilde{R}(z) = R(z) + h_r\lambda(z)$, $\tilde{G}(z) = G(z) + h_g\lambda(z)$, $\tilde{L}(z) = L(z) + h_l\lambda(z)$, $\tilde{C}(z) = C(z) + h_c\lambda(z)$, 均为沿互连线的随机过程。考虑工程实际, 将工艺参数沿互连线的随机变化量 $\lambda(z)$ 取为宽平稳的高斯随机过程。

式(6)即为互连线的复频域随机模型。若为多互连线系统, 则模型式(6)中各随机变量相应改为向量形式即可。下面基于互连线的随机模型式(6)对互连线的 ABCD 参数矩阵进行分析。

2 互连线的 ABCD 参数

2.1 ABCD 参数的随机模型

根据 ABCD 参数矩阵定义, 互连线始端与 z 处的电流与电压满足如下关系式:

$$\begin{bmatrix} U(z,s) \\ I(z,s) \end{bmatrix} = A(z,s) \begin{bmatrix} U(0,s) \\ I(0,s) \end{bmatrix} = \begin{bmatrix} A(z,s) & B(z,s) \\ C(z,s) & D(z,s) \end{bmatrix} \begin{bmatrix} U(0,s) \\ I(0,s) \end{bmatrix} \quad (7)$$

当考虑互连线的工艺变化时, 将式(7)两端对 z 取一阶偏导, 并联立式(6)得到 ABCD 参数各参量所满足的微分方程组为

$$\begin{cases} \frac{\partial A(z,s)}{\partial z} = [\tilde{G}(z) + s\tilde{C}(z)]B(z,s) \\ \frac{\partial B(z,s)}{\partial z} = [\tilde{R}(z) + s\tilde{L}(z)]A(z,s) \\ \frac{\partial C(z,s)}{\partial z} = [\tilde{G}(z) + s\tilde{C}(z)]D(z,s) \\ \frac{\partial D(z,s)}{\partial z} = [\tilde{R}(z) + s\tilde{L}(z)]C(z,s) \end{cases} \quad (8)$$

并且满足初始边界条件

$$\begin{cases} A(0,s) = D(0,s) = 1 \\ B(0,s) = C(0,s) = 0 \end{cases} \quad (9)$$

式(8)即工艺参数随机变化下互连线 ABCD 参数的随机模型。根据式(8), 由于 $\tilde{G}(z), \tilde{R}(z), \tilde{C}(z), \tilde{L}(z)$ 为随机过程, 因此 $A(z,s), B(z,s), C(z,s), D(z,s)$ 相应成为复随机过程。为研究其统计特性, 采用蒙特卡洛法与数值计算相结合的方法进行研究, 并与经典的未考虑工艺参数随机变化的互连线 ABCD 参数进行比较分析。

2.2 蒙特卡洛法分析随机模型的 ABCD 参数

蒙特卡洛法也称为随机模拟(Random Simulation)法, 有时也称为随机抽样(Random Sampling)技术或统计实验(Statistical Testing)法, 是统计数学的一个分支, 利用随机数进行统计实验, 将求得的统计

特征值作为被研究问题的近似解^[9]。采用蒙特卡洛法对互连线随机模型进行 ABCD 参数的数值仿真与分析分为 4 个步骤:

- (1) 产生互连线随机样本;
- (2) 每个互连线样本的 ABCD 参数数值求解;
- (3) 对互连线样本的 ABCD 参数数值结果进行统计分析;
- (4) 利用统计分析结果对互连线性能进行评估。

2.3 互连线样本的 ABCD 参数数值求解

设第 i 个互连线样本的电气分布参数分别 $\tilde{G}_i(z), \tilde{R}_i(z), \tilde{C}_i(z), \tilde{L}_i(z)$, 代入式(8)得到其对应的 ABCD 参数各参量满足的微分方程组为

$$\begin{cases} \frac{\partial A(z,s)}{\partial z} = [\tilde{G}_i(z) + s\tilde{C}_i(z)]B(z,s) \\ \frac{\partial B(z,s)}{\partial z} = [\tilde{R}_i(z) + s\tilde{L}_i(z)]A(z,s) \\ \frac{\partial C(z,s)}{\partial z} = [\tilde{G}_i(z) + s\tilde{C}_i(z)]D(z,s) \\ \frac{\partial D(z,s)}{\partial z} = [\tilde{R}_i(z) + s\tilde{L}_i(z)]C(z,s) \end{cases} \quad (10)$$

式(10)为复微分方程组, 直接求解相当困难, 因此对 ABCD 参数各参量进行一阶麦克劳林展开^[10]

$$\begin{cases} A(z,s) \approx a_0(z) + a_1(z)s \\ B(z,s) \approx b_0(z) + b_1(z)s \\ C(z,s) \approx c_0(z) + c_1(z)s \\ D(z,s) \approx d_0(z) + d_1(z)s \end{cases} \quad (11)$$

将式(11)代入式(10)并令方程左右两边 s^i ($i=0,1$) 系数相等, 得到 ABCD 参数各参量的系数所满足的常微分方程组为

$$\begin{cases} a'_0(z) = \tilde{G}_i(z)b_0(z) \\ a'_1(z) = \tilde{G}_i(z)b_1(z) + \tilde{C}_i(z)b_0(z) \\ b'_0(z) = \tilde{R}_i(z)a_0(z) \\ b'_1(z) = \tilde{R}_i(z)a_1(z) + \tilde{L}_i(z)a_0(z) \\ c'_0(z) = \tilde{G}_i(z)d_0(z) \\ c'_1(z) = \tilde{G}_i(z)d_1(z) + \tilde{C}_i(z)d_0(z) \\ d'_0(z) = \tilde{R}_i(z)c_0(z) \\ d'_1(z) = \tilde{R}_i(z)c_1(z) + \tilde{L}_i(z)c_0(z) \end{cases} \quad (12)$$

而由 $A(z,s), B(z,s), C(z,s), D(z,s)$ 的初始条件式(9)可以得到各参量系数所满足的初始边界条件为

$$\begin{aligned} a_0(0) &= d_0(0) = 1, b_0(0) = c_0(0) = a_1(0) = b_1(0) \\ &= c_1(0) = d_1(0) = 0. \end{aligned}$$

式(12)可通过 4 阶龙哥-库塔法^[11]进行数值求解。

3 仿真实验结果及分析

例 1: 均匀互连线。设互连线长 $l = 0.01$ m, 根据其设计的工艺参数计算得到的电阻、电感等电气分布参数的标称值分别为 $R(z) = 8.24 \Omega/m, G(z) = 905 \text{ nS/m}, L(z) = 309 \text{ nH/m}, C(z) = 144 \text{ pF/m}$ 。

由于工艺变化导致电气分布参数发生变化, 设这些变化量服从高斯分布, 其均值为 0, 均方差取为标称值的 10%。Monte Carlo 仿真实验的互连线样本个数取 2 000 个, 得到式(11)中 ABCD 参数各参量系数的仿真结果如表 1 所示。由表 1 可以看出 ABCD 参数各参量在未考虑工艺变化时的数值和其考虑工艺变化时的均值相等, 这表示服从均值为 0 的高斯分布的工艺变化不会对 ABCD 参数各参量产生系统偏差, 但会产生随机的变化, 尤其对 A 参量和 D 参量的一阶系数 $a_1(0.01)$ 和 $d_1(0.01)$ 影响较大, 它们的直方图如图 4 和图 5 所示。对 $a_1(0.01)$ 和 $d_1(0.01)$ 的正态性进行显著性水平 $\alpha = 0.01$ 的偏度-峰度(Jarque-Bera)检验, 结果接受假设, 利用 3σ 值可知它们的变化量最差情况可达到 $7.41E-15$ 和 $8.19E-15$, 均超过其理想情况下取值的 10%, 因此工艺变化的影响是不可忽略的。

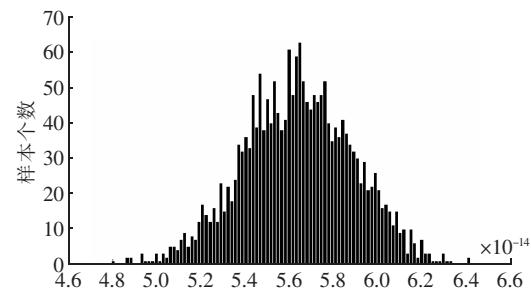


图 4 均匀互连线 $a_1(0.01)$ 的直方图

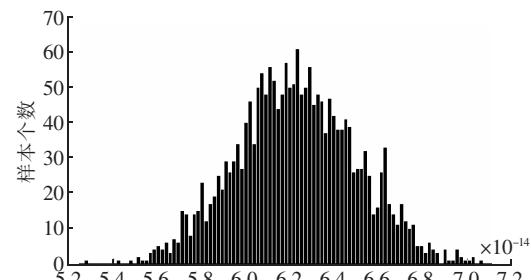


图 5 均匀互连线 $d_1(0.01)$ 的直方图

表1 均匀互连线ABCD参数各参量系数的仿真结果

仿真条件	$a_0(0.01)$	$a_1(0.01)$	$b_0(0.01)$	$b_1(0.01)$	$c_0(0.01)$	$c_1(0.01)$	$d_0(0.01)$	$d_1(0.01)$
未考虑工艺变化的理想情况	1.00	5.64e-14	8.24e-2	3.09e-9	9.05e-9	1.44e-12	1.00	6.23e-14
Monte Carlo 仿真得到的均值	1.00	5.64e-14	8.24e-2	3.09e-9	9.05e-9	1.44e-12	1.00	6.24e-14
Monte Carlo 仿真得到的均方差	1.56e-11	2.47e-15	1.81e-3	6.78e-11	1.99e-10	3.16e-14	1.72e-11	2.73e-15

例2:非均匀互连线。设互连线长 $l = 0.03$ m, 根据其设计的工艺参数计算得到的电阻、电感等电气分布参数的标称值分别为

$$L(z) = 387/[1 + k(z)] \text{ nH/m}$$

$$C(z) = 104.13/[1 - k(z)] \text{ pF/m}$$

$$R(z) = 1.2 \Omega/\text{m}$$

$$G(z) = 98 \text{ nS/m}$$

其中, $k(z) = 0.25[1 + 0.6\sin(\pi z + \pi/4)]$ 。

同样设工艺参数导致的电气分布参数高斯变化量的均值为0, 均方差取为标称值的10%。Monte

Carlo 仿真实验的互连线样本个数取2000个, 得到式(11)中ABCD参数各参量系数的仿真结果如表2所示。由于工艺变化对A参量和D参量的一阶系数 $a_1(0.03)$ 和 $d_1(0.03)$ 影响较大, 这里给出其直方图如图6和图7所示, 并且对 $a_1(0.03)$ 和 $d_1(0.03)$ 的正态性进行显著性水平 $\alpha = 0.01$ 的偏度-峰度(Jarque-Bera)检验, 结果接受假设, 利用 3σ 值可知它们的变化量最差情况可达到 $1.13e-14$ 和 $1.25e-14$, 同样超过了其理想情况下取值的10%。

表2 非均匀互连线ABCD参数各参量系数的仿真结果

仿真条件	$a_0(0.03)$	$a_1(0.03)$	$b_0(0.03)$	$b_1(0.03)$	$c_0(0.03)$	$c_1(0.03)$	$d_0(0.03)$	$d_1(0.03)$
未考虑工艺变化的理想情况	1.00	8.47e-14	3.60e-2	8.50e-9	2.94e-9	4.93e-12	1.00	9.27e-14
Monte Carlo 仿真得到的均值	1.00	8.46e-14	3.60e-2	8.50e-9	2.94e-9	4.92e-12	1.00	9.27e-14
Monte Carlo 仿真得到的均方差	2.24e-12	3.78e-15	8.00e-4	1.88e-10	6.53e-11	1.11e-13	2.47e-12	4.16e-15

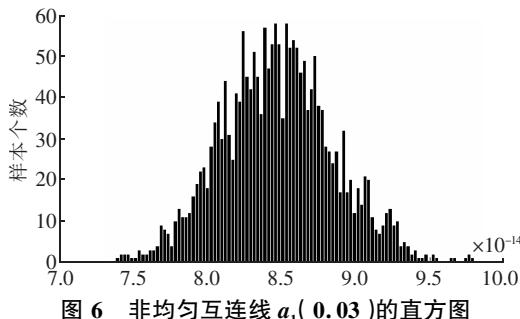
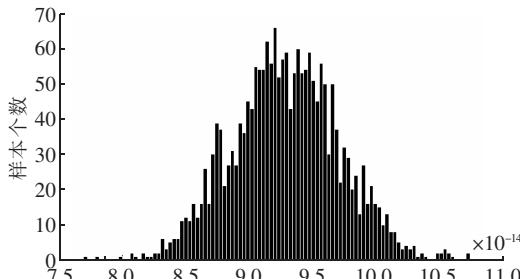
4 结论

目前集成电路已经开始进入纳米工艺阶段, 互连线的工艺变化问题日益成为影响集成电路性能和成品率的重要因素。本文考虑在生产过程中工艺参数的随机变化对互连线电气分布参数的影响, 建立了工艺变化下互连线的分布参数随机模型, 推导出工艺变化下互连线ABCD参数满足的随机微分方程组, 并采用蒙特卡洛法对ABCD参数进行统计实验仿真与分析。实验结果表明该互连线的随机模型可以对互连线的传输性能进行有效的评估, 对于采用统计的方法进行互连线制造过程的控制及优化有着重要意义, 在以后的工作中将对本文提出的模型及其分析方法在工程中的应用做进一步的研究。

参考文献:

- [1] 郝跃,荆明娥,马佩军. VLSI集成电路参数成品率及优化研究进展[J]. 电子学报,2003,31(12A):1971-1974.
HAO Yue, JING Minge, MA Pejun. State of the Art on Study of Parametric Yield and Its Optimization for VLSI[J]. Acta Electronica Sinica, 2003, 31(12A):1971-1974.
- [2] WANG J M, GHANTA P, VRUDHULA S. Stochastic analysis of interconnect performance in the presence of process variations[C]// IEEE/ACM International Conference on Computer Aided Design.

由实验结果可以看出, 互连线工艺参数的随机变化对其ABCD参数各参量的一阶系数影响较大, 尤其是A参量和D参量, 而一阶系数往往是影响互连线时域响应特性(譬如延迟)的主要因子, 因此互连线的工艺变化对其传输性能的影响在某些情况下可能会相当严重, 尤其随着电路工作频率的提高和互连线分布参数效应的愈加严重, 互连线工艺变化的影响可能造成整个电路系统的失效。

图6 非均匀互连线 $a_1(0.03)$ 的直方图图7 非均匀互连线 $d_1(0.03)$ 的直方图

- San Jose, CA USA, 2004:880–886.
- [3] 张瑛, WANG J M, 肖亮, 等. 工艺参数随机扰动下的传输线建模与分析新方法[J]. 电子学报, 2005, 33(11): 1959–1964.
ZHANG Ying, WANG J M, XIAO Liang, et al. A New Stochastic Modeling and Analysis Method for Transmission Lines in the Presence of Random Process Variations[J]. Acta Electronica Sinica, 2005, 33(11): 1959–1964.
- [4] 骆祖莹. 芯片功耗与工艺参数变化——下一代集成电路设计的两大挑战[J]. 计算机学报, 2007, 30(7): 1054–1063.
LUO Zuying. Power Consumption and Process Variations: Two Challenges to Design of Next-generation Ics[J]. 2007, 30(7): 1054–1063.
- [5] LIN Yan, HE Lei, HUTTON M. Stochastic Physical Synthesis Considering Prerouting Interconnect Uncertainty and Process Variation for FPGAs[J]. IEEE Trans on Very Large Scale Integration Systems, 2008, 16(2): 124–133.
- [6] MEI K. Measured Equation of Invariance: Anew Concept in Field Computations[J]. IEEE Trans on Antennas and Propagation, 1994, 42(3): 320–328.
- [7] 张瑛, 肖亮, 吴慧中, 等. 互连线分布电容偏差计算的非均匀有限差分法[J]. 南京理工大学学报: 自然科学版, 2005, 29(6): 740–744.
ZHANG Ying, XIAO Liang, WU Huizhong, et al. Non-uniform Finite Difference Method for Computing Distributed Capacitance Deviation of Interconnects[J]. Journal of Nanjing University of Science and Technology: Natural Science, 2005, 29(6): 740–744.
- [8] 张瑛, WANG J M, 肖亮, 等. 传输线电容参数提取中的奇异性处理方法[J]. 微波学报, 2005, 21(6): 14–18, 42.
ZHANG Ying, WANG J M, XIAO Liang, et al. Singularity Treatment Approach for Capacitance Parameters Extraction of Transmission Lines[J]. Journal of Microwares, 2005, 21(6): 14–18, 42.
- [9] 殷显安. 试验模拟的蒙特卡洛方法[J]. 测试技术学报, 1994, 8(2): 49–51.
- YIN xianan. Monte Carlo Method for test simulation[J]. Journal of Test and Measurement Technology, 1994, 8(2): 49–51.
- [10] LI Xiaochun, MAO Junfa, HUANG Huifen. Accurate analysis of interconnect trees with distributed RLC model and moment matching [J]. IEEE Trans on Microwave Theory and Techniques, 2004, 52(9): 2199–2206.
- [11] 戴嘉尊, 邱建贤. 微分方程数值解法[M]. 南京: 东南大学出版社, 2002: 16–21.
DAI Jiazhun, QIU Jianxian. Numerical Solution for Differential Equations[M]. Nanjing: South East University Publication, 2002: 16–21.

作者简介:



张瑛(1980-),男,安徽黄山人。南京邮电大学电子科学与工程学院讲师,博士,东南大学博士后流动站在职博士后。研究方向为射频与微波集成电路设计。

王志功(1954-),男,河南荥阳人。东南大学射频与光电集成电路研究所所长,教授,博士生导师。从事超高速、微波和毫米波集成电路、光电集成电路设计,2000年荣获教育部“长江学者特聘教授”。

方承志(1976-),男,湖南安化人。南京邮电大学电子科学与工程学院讲师,博士。研究方向为信号与信息处理。

杨恒新(1971-),男,山东郯城人。南京邮电大学电子科学与工程学院副教授。现从事电子电路理论教学及EDA、微机应用、单片机应用研究。

声 明

为适应我国信息化建设的需要,扩大作者学术交流渠道,实现期刊编辑、出版工作的网络化,本刊已加入《中国学术期刊(光盘版)》和《中国期刊网》全文数据库、《万方数据——数字化期刊群》、《中文科技期刊数据库》,作者著作权使用费随本刊稿酬一次性给付。如不同意将文章编入相关数据库,请在来稿时声明,本刊将做适当处理。

嵌入式通信服务器 E1 网络驱动程序设计与实现

高建国,戴海鸿

(南京邮电大学 通信与信息工程学院,江苏南京 210003)

摘要:提出了采用 MPC8270 嵌入式处理器和嵌入式 Linux 操作系统组成网络通信服务器以实现在 E1 传输网络上进行 IP 网络通信的设计方案,着重论述了实现该方案的关键技术之一 E1 网络驱动程序的设计。给出了基于 MPC8270 FCC 通信控制器的 HDLC(高级数据链路协议)通信模式进行 E1 网络通信的网络驱动程序设计方法。配置 E1 网络驱动程序的网络通信服务器已在实际的通信系统中得到应用。

关键词:MPC8270; FCC; HDLC; E1; 嵌入式 Linux; 网络驱动程序

中图分类号:TN915.04 文献标识码:B 文章编号:1673-5439(2009)06-0091-05

Design and Implementation of E1 Network Driver for Embedded Communication Server

GAO Jian-guo, DAI Hai-hong

(College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In this paper, a design scheme of using MPC8270 embedded processor and embedded Linux operating system to composite the network communication server to achieve at IP network communication on E1 transmission network is proposed. Focused on discussing one of the keys to implement the scheme, which is the design of E1 network driver. The design method of the driver for realizing the E1 network communication under HDLC communication mode based on the MPC8270 FCC Communication Controller is given. The network communication server with E1 network driver has been applied in real communication systems.

Key words: MPC8270; FCC; HDLC; E1; embedded Linux; network driver

0 引言

E1 是我国电信传输网一次群使用的传输标准, E1 被广泛应用于基站、交换局间、广域网的数据传输链路。经过信息基础设施的长期建设和发展, 我国已经拥有了丰富的 E1 信道资源。近年来, 随着 IP 网的迅猛发展, 基于 IP 包的分组交换在通信中所占的比重日益增大, 实现在 E1 上的分组传输方式, 可使 E1 通信电路资源得到充分地利用。

HDLC(High-Level Data Link Control, 高级数据链路协议)是一个在同步网上传输数据并面向比特的数据链路层协议。HDLC 被广泛应用于点对点或多点的数据链路通信中。将 IP 包承载在 HDLC 上

在 E1 电路上进行传输的技术(简称 IP over HDLC over E1)是目前成熟的技术。

MPC8270 是美国飞思卡尔公司推出的一种功能强大的双核 RISC(精简指令集计算机)处理器, 它具有一个高性能的 PowerPC 内核和一个能独立处理与多种外围设备通信的通信处理模块(CPM)。CPM 支持多个通信控制器, 如 FCC(快速串行通信控制器)、SCC(串行通信控制器)、SMC(串行管理通信控制器)、MCC(多信道通信控制器)等。CPM 的 FCC 控制器含有执行 HDLC 通信协议的底层微代码, FCC 控制器还具有接口复用功能, 与 E1 外围接口通信功能芯片结合后, 可简便地组成在 E1 上进行 HDLC 协议通信(HDLC over E1)的硬件平台。

嵌入式 Linux 2.6.10 操作系统是一个内核精简、稳定、高效的操作系统,它的网络通信子系统功能强大,并提供了许多广域网通信的网络驱动模型架构,其中就含有通用的 HDLC 网络驱动模型架构层,用户开发自己的 HDLC 网络驱动程序时只需在驱动程序中针对自己实际使用的硬件通信控制器进行设置,其余按照 Linux 网络驱动模型和通用的 HDLC 网络驱动模型架构进行编程,即可实现高层的 IP 网络通信应用程序使用标准的 Linux Socket 网络通信接口方式与底层的 HDLC 网络设备进行 IP 网络数据包的交换功能(IP over HDLC),最终实现 IP 在 E1 上的网络通信传输。

综上所述,设计基于 MPC8270 嵌入式处理器的 E1 网络通信硬件平台、在硬件平台上配置嵌入式 Linux 操作系统、开发自己的 E1 网络驱动程序和 IP 网络通信应用程序来组成 E1 网络通信服务器。在 E1 传输网中,依靠 E1 网络通信服务器点对点间的通信,可以完成 E1 网络上的 IP 网络通信功能(IP over HDLC over E1)。

本文重点论述了在嵌入式 Linux 2.6.10 操作系统内核的基础上开发 E1 网络驱动程序的方法。

1 E1 网络通信服务器硬件结构

E1 网络通信服务器是 E1 网络通信系统的硬件基础,通信服务器采用了 128 MB 内存、16 MB 闪存的存储设计。16 MB 的闪存用来存储 Linux 操作系统的内核、根文件系统 RAMDISK 映像、保存用户应用程序生成的数据;128 MB 的内存用来运行操作系统程序、应用程序、存储运行过程中的数据。

MPC8270 的 CPM 通信处理器可以控制 3 个 FCC 通信控制器对外围通信接口进行控制管理,提供初始化配置、启动执行通信协议的微代码、传输控制、中断处理等通信服务功能,FCC 可以配置成 HDLC、异步 HDLC、透明、以太网等传输模式,FCC 可通过 TSA(时隙分配器)对多达 4 个 TDM(时分复用)的信号接口提供支持。在我们的网络通信服务器硬件设计中采用了 FCC2 和 FCC3 分别和 2 个 E1 接口电路连接,并配置成 HDLC 传输模式,通过 E1 传输线路分别和两个对端的网络通信服务器连接,实现广域网的通信功能。而 FCC1 则配置成以太网传输模式,实现以太局域网的通信功能。用 SMC1 和串行接口连接,提供控制台终端功能^[1]。图 1 简要描述了 E1 网络通信服务器结构。

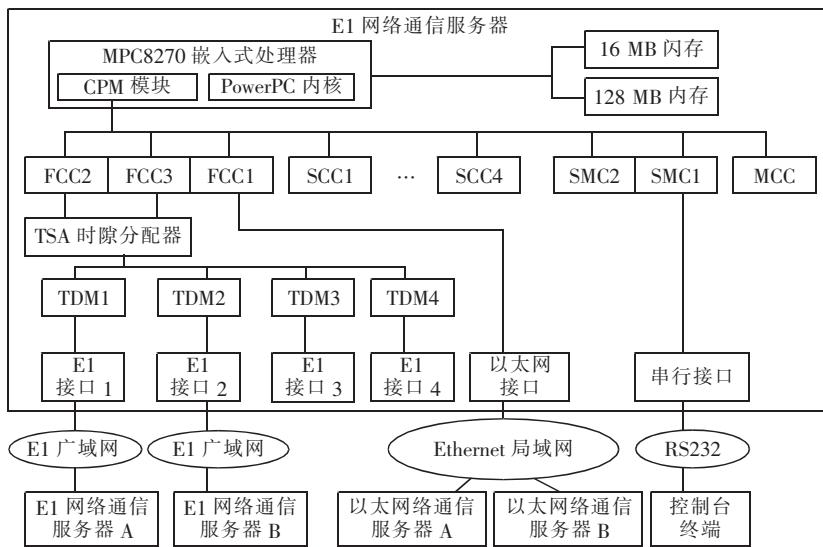


图 1 E1 网络通信服务器结构

2 E1 网络通信服务器软件结构

E1 网络通信服务器的设计目标是满足本端服务器同时能与两个对端服务器进行 IP 网络通信的功能,网络的拓扑结构是一点对两点。通信服务器采用嵌入式 Linux 操作系统,通信软件由 IP 网络通信应用程序、Linux2.6.10 通用 HDLC 驱动层内核模

块、E1 网络驱动程序组成。

IP 网络通信应用程序使用标准的 Linux 的 TCP/UDP 套接字方式与 Linux 网络接口进行 IP 通信。

E1 网络驱动程序通过 Linux 的网络接口和应用层交换 IP 数据包;驱动程序将 MPC8270 处理器的 FCC2、FCC3 通信控制器配置成 HDLC 模式,驱动

E1 接口电路,由 FCC 底层的微代码自动完成 E1 上的 HDLC 通信。

嵌入式 Linux2.6.10 针对使用 HDLC 协议的网络通信设备设计了广域网(WAN)通信的通用 HDLC 驱动层内核模块。E1 网络驱动程序调用通用 HDLC 驱动层模块的函数来创建 HDLC 网络设备接口,处理 HDLC 通信。

为了实现本端服务器与两个对端服务器进行点对点的 E1 网络通信,在本端服务器的 Linux 操作系统中需要创建两个广域网络接口,网络设备名称为

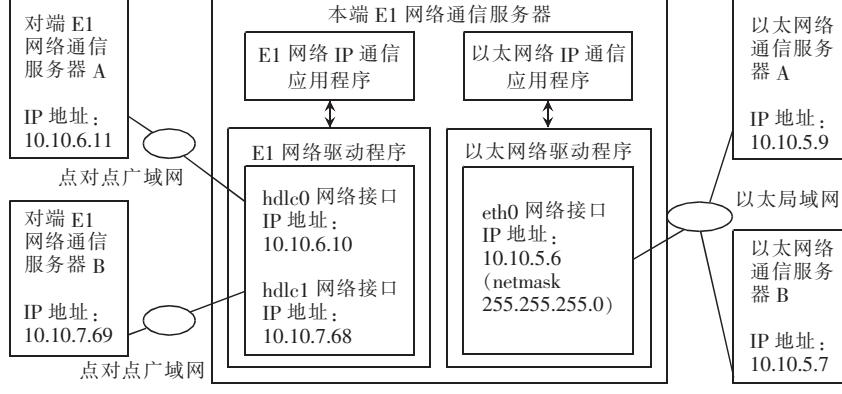


图 2 E1 网络驱动程序绑定的网络接口

在嵌入式 Linux 系统中创建 hdle0 和 hdle1 网络设备需要 sethdle 实用程序来帮助,创建网络接口则使用 Linux 系统的 ifconfig 命令。

3 E1 网络驱动程序设计

E1 网络驱动程序模块按照 Linux 的网络驱动^[2]和通用的 HDLC 驱动层^[3]框架设计。驱动程序的处理函数如图 3 中白色文本框所示。

当驱动程序装入系统后首先运行的是驱动模块的初始化函数。在初始化函数中,需要创建两个 hdle 网络设备数据结构,在设备数据结构中设置有保存未发送完成的上层应用程序传送来的 skb (socket buffer) 数据结构指针的循环队列数组。初始化函数所做的重要工作是对 FCC2、FCC3 进行初始化,通过创建、初始化收/发描述符队列;设置 FCC 参数寄存器;初始化 FCC 模式寄存器;初始化 SI2 RAM;设置 FCC 收发参数初始化命令;配置 CPM 复用逻辑 SI2 时钟路由寄存器;启用 TDM 相关同步及收发数据引脚的对外连接;使能 TDM;启用 TDM 相关时钟的对外连接;初始化 FCC 中断寄存器等具体步骤使 FCC 设置为 hdle 传输模式并控制 E1 接口电路的传输^[4]。在初始化函数中,通过中断安装步

hdle0 和 hdle1。分配两个 IP 地址和服务器的 E1 网络驱动程序绑定,这两个 IP 地址应分属在不同的子网上。参见图 2 举例,一个 IP 地址在 10.10.6 子网上,另一个 IP 地址在 10.10.7 子网上。为避免 E1 网络 IP 通信应用程序的 IP 数据包传送到原 Linux 系统的以太网络驱动程序中,这两个 IP 地址不能与服务器局域网通信的以太网络 IP 地址在同一子网(10.10.5)上。对端 E1 网络通信服务器也需要配置同样的 E1 网络驱动程序以及创建广域网络接口,接口的 IP 地址必须与本端的在同一子网上。

骤,使得 FCC 的中断与中断处理例程连接。通过初始化 hdle 设备数据结构和注册 hdle 设备步骤建立设备打开、设备连接、设备关闭、设备控制、数据传输、传输超时函数的回调连接。

当上层的 IP 网络通信应用程序通过 Socket 向 hdle0 网络接口(目的 IP 地址与 hdle0 网络接口的 IP 地址在同一子网上)发送 IP 数据包时,系统生成 skb 调用驱动程序的设备数据传输函数。在该函数中,将 skb 结构指针保存到循环队列数组中并将 skb 中的 IP 数据包送到 FCC2 的发送描述符队列中,设置发送描述符状态未就绪等待 FCC2 从 E1 传输线路上用 HDLC 协议发送出去。当发送完成,FCC2 产生发送完成中断,由发送完成中断处理函数删除发送刚开始时保存的 skb。这个删除工作必须要在驱动程序中做,否则,上层的 IP 应用程序的 socket 发送完 256 个数据包后就不能继续工作。当 FCC2 产生发送超时问题时则调用传输超时函数,在该函数中通常是做停止系统的发送数据传输队列以暂停应用程序的发送工作。当 FCC2 通过 HDLC 协议从 E1 传输线路上接收到 IP 数据包时保存到 FCC2 的接收描述符队列中并产生接收中断,由接收数据中断处理函数从接收描述符中取出数据包,进行校验后生成 skb 上传到 IP 应用程序的 socket 中。hdle1 网

络接口的数据收发处理也是如此(参见图3)。

当在Linux系统上执行ifconfig命令建立网络接口时系统调用驱动程序的设备打开函数,在该函数中,进行打开hdlc设备、开启系统的发送数据传

输队列处理。当使用sethdlc实用程序时系统调用设备连接函数,在该函数中,对编码和CRC校验规则进行选择。当Linux系统卸载驱动模块时调用驱动模块卸载函数,在该函数中释放已创建的资源。

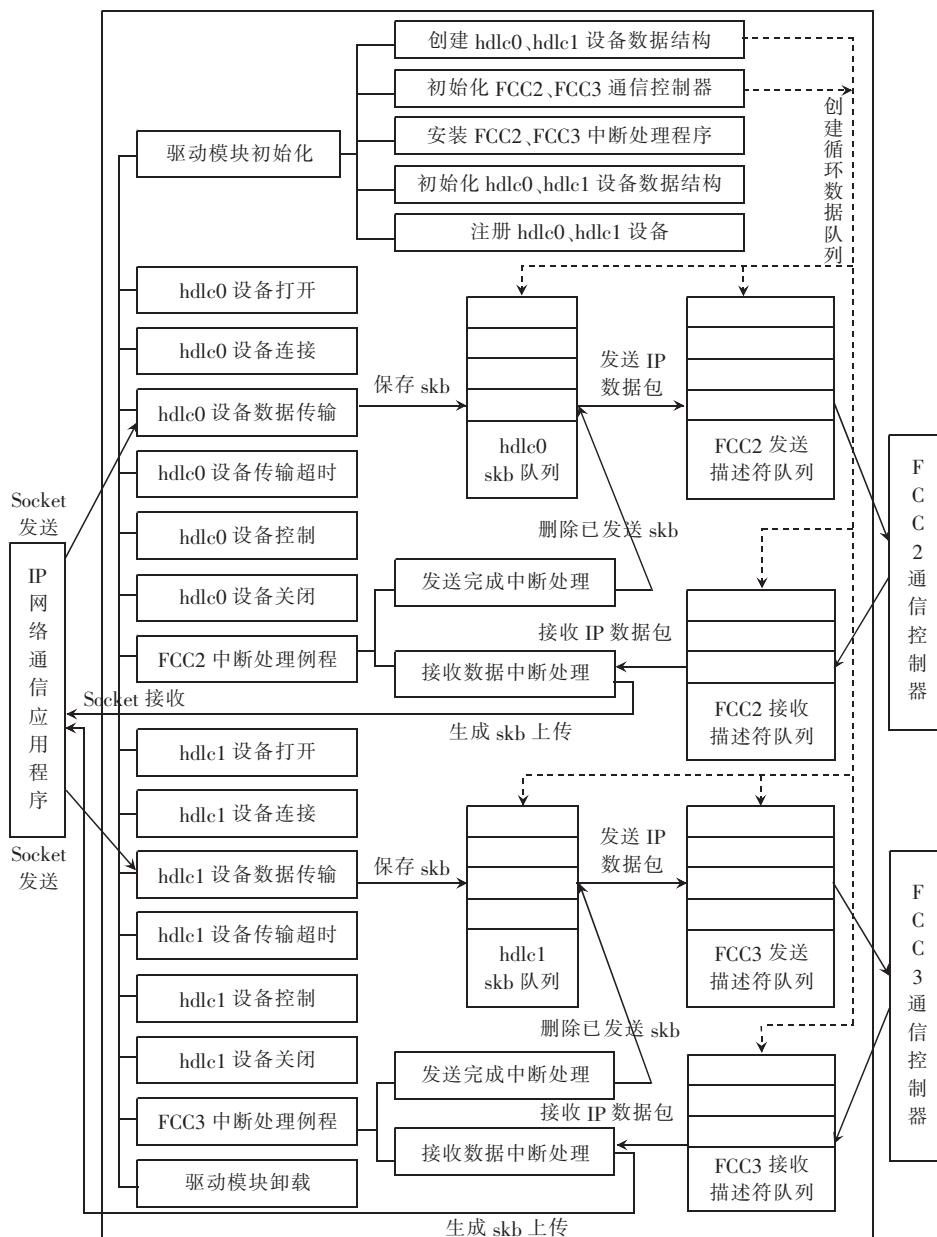


图3 E1 网络驱动程序处理函数和数据收发处理流程

4 E1 网络驱动程序开发工具、运行环境

E1 网络驱动程序需要在 PowerPC 处理器的环境下运行,同时还需要修改 Linux2.6.10 内核进行重新编译,因此需要一个能生成 PowerPC 目标代码的交叉编译环境。我们采用了德国 DENX 软件工程公司推出的开放源代码的嵌入式 Linux 开发套件 ELDK, ELDK 可在 <ftp://ftp.denx.de/pub/eldk> 处下载。

E1 网络驱动程序在运行过程中要调用 Linux 2.6.10 操作系统内核的通用 HDLC 驱动层模块的函数,而通用 HDLC 驱动层模块是可选模块,因此我们在编译 Linux 内核时必需选择配置该模块。选择配置的步骤是先用 make menuconfig 命令进入菜单配置内核方式,然后按以下步骤选择菜单及相关选项:

```

Device Drivers→
  Networking support→
    [ * ] Network device support
    [ * ] Network device support
  
```

```

Wan interfaces support→
< * > Generic HDLC layer
[ * ] Raw HDLC support
[ * ] Raw HDLC Ethernet device support

```

* 表示选中该驱动模块,在内核启动后加载到内核。选择 Raw HDLC support 项表示支持使用 HDLC 的点对点 IP 通信模式,选择 Raw HDLC Ethernet device support 项表示支持使用 HDLC 的以太网设备仿真模式。

在第2节中已提到的 sethdlc 是一个功能强大的实用程序,和驱动程序配合除了可以创建 hdlc 网络设备之外还可以对物理接口、时钟速率、使用 hdlc 的协议及模式等进行选择和设置。sethdlc 实用程序并不包含在嵌入式 Linux2.6.10 系统中,需要到 <http://www.kernel.org/pub/linux/utils/net/hdlc/#versions> 处下载源程序 sethdlc.c,将源程序用交叉编译命令进行编译。sethdlc 的命令格式比较复杂,不需要让驱动程序全面实现它的设置,只需实现其中一部分自己所需的模式选择及参数设置即可。在 E1 网络驱动程序中,由于使用的是 hdlc 的点对点 IP 通信模式,则采用了 sethdlc hdlc0 hdlc 这样的命令格式。命令中的 hdlc0 为 hdlc 网络设备名称,后面的 hdlc 表示使用 hdlc 协议的点对点 IP 通信模式。有关 sethdlc 的详细用法可参考上述网页中的说明。

使用交叉编译命令编译 E1 网络驱动程序、IP 网络应用程序。将编译好的 E1 网络驱动程序、sethdlc 实用程序、IP 网络应用程序加入到嵌入式 Linux 根文件系统中,在本端 E1 网络通信服务器的根文件系统的启动脚本中写入如下的启动命令:

```
ifconfig eth0 10.10.5.6 netmask 255.255.255.0    创建 eth0 以太局域网络接口
```

```
insmod eOneNetDrv    装载 E1 网络驱动程序( eOneNetDrv 为网络驱动程序名称)
```

```
./sethdlc hdlc0 hdlc    创建 hdlc0  
点对点通信网络设备
```

```
ifconfig hdlc0 10.10.6.10 netmask 255.255.255.0    创建 hdlc0  
E1 广域网络接口
```

```
./sethdlc hdlc1 hdlc    创建 hdlc1  
点对点通信网络设备
```

```
ifconfig hdlc1 10.10.7.68 netmask 255.255.255.0    创建 hdlc1  
E1 广域网络接口
```

生成根文件系统 RAMDISK 映像文件后写入到本端通信服务器目标板的闪存中。对于对端的 E1 网络通信服务器 A、B 也进行类似地处理。连接好本端通信服务器与对端服务器 A、B 通信的两条 E1 传输

线路就可进行一点对两点的 IP 网络通信的调试了。

5 结束语

开发完成的 E1 网络驱动程序在基于 MPC8270 嵌入式处理器设计的网络通信服务器上进行了运行测试。测试结果表明,基于 MPC8270 CPM 的 FCC 快速串行通信控制器的 HDLC 传输模式设计的 E1 网络驱动程序在 E1 传输电路上进行 IP 网络通信稳定可靠,传输性能达到设计要求。配备 E1 网络驱动程序的嵌入式网络通信服务器已在实际的通信系统中得到应用。

参考文献:

- [1] Freescale. MPC8280 PowerQUICC II Family Hardware Specifications[EB/OL]. <http://www.freescale.com/files/32bit/doc/data-sheet/MPC8280EC.pdf>.
- [2] JONATHAN CORBET, ALESSANDRO RUBINI & GREG KROAH - HARTMAN. LINUX 设备驱动程序[M]. 3 版. 魏永明,耿岳,钟书毅,译. 北京:中国电力出版社,2005.
- [3] HALASA K. Generic HDLC layer for Linux[EB/OL]. <http://www.kernel.org/pub/linux/utils/net/hdlc/#versions>.
- [4] Freescale. MPC8280 PowerQUICC II Family Reference Manual [EB/OL]. http://www.freescale.com/files/netcomm/doc/ref_manual/MPC8280RMAD.pdf
- [5] 吴军,周转运. 嵌入式 Linux 系统应用基础与开发范例[M]. 北京:人民邮电出版社,2007.
- [6] 孙琼. 嵌入式 Linux 应用程序开发详解[M]. 北京:人民邮电出版社,2006.
- [7] 俞永昌. Linux 设备驱动开发技术及应用[M]. 李红姬,李明吉,译. 北京:人民邮电出版社,2008.

作者简介:



高建国(1960-),男,江苏苏州人。南京邮电大学通信与信息工程学院工程师。1982年1月毕业于南京大学计算机科学系软件专业。目前主要从事计算机软件开发工作。

戴海鸿(1963-),男,浙江宁波人。南京邮电大学通信与信息工程学院高级工程师。1985年毕业于南京邮电学院电信工程系。目前主要研究方向为数字信号处理在通信中的应用。

一种支持大型多人在线游戏的覆盖网组播生成树算法

林巧民¹, 王汝传^{2,3}, 许棣华², 林萍²

1. 南京邮电大学 数字媒体研究中心, 江苏南京 210046
2. 南京邮电大学 计算机学院, 江苏南京 210046
3. 南京大学 计算机软件新技术国家重点实验室, 江苏南京 210093

摘要:提出了一种基于AOI(Area of Interest)域的可调覆盖组播生成树算法AOMST(Adjustable Overlay Multicast Spanning Tree),该算法可用于支持大型多人在线游戏MMOG(Massively Multi-player Online Games)。它的基本思想是先将MMOG按照兴趣域划分分区,在每个分区内以结点带宽及时延为可调影响因子构建组播生成树,然后再通过3种不同的结点变换操作来进一步减少组播生成树中的时延。仿真实验表明,AOMST算法是有效的。

关键词:大型多人在线游戏;兴趣域;覆盖网组播;带宽;时延

中图分类号:TP391 文献标识码:B 文章编号:1673-5439(2009)06-0096-05

The Design and Implementation of an Intelligent Game Engine Based on OGRE

LIN Qiao-min¹, WANG Ru-chuan^{2,3}, XU Di-hua², LIN Ping²

1. Digital Media Research Center, Nanjing University of Posts and Telecommunications, 210046, China
2. College of Computer, Nanjing University of Posts and Telecommunications, 210046, China
3. State Key-lab for Novel Software Technology, Nanjing University, 210093, China

Abstract: This paper presents an Adjustable Overlay Multicast Spanning Tree (AOMST) algorithm which is based on Area of Interest (AOI). The algorithm can be used to support Massively Multi-player Online Games (MMOG). Its fundamental idea is to divide the MMOG world into different areas of interest, then construct multicast spanning tree based on bandwidth and latency, whose affection factors are adjustable, within each AOI. Next, the latency of the multicast spanning tree can still be reduced via three different node swap operations. The experimental results indicate that AOMST is really effective.

Key words: massively multi-player online game; area of interest; overlay multicast; bandwidth; latency

0 引言

目前的MMOG(Massively Multi-player Online Games)大多数采用的是集群C/S模式,这种模式面临着伸缩性问题:资源、用户热区以及过多的服务器之间通信等。针对MMOG的伸缩性问题可以有两种不同的解决方法,一种是通过增加硬件资源的方法,如花费重金部署更多的服务器,这种方法的基本

思想是以牺牲硬件资源来进行系统扩容,另一种方法是通过减少通信负载来实现系统扩容,主要包括AOI(Area of Interest)技术^[1-3]、DR(Dead Reckoning)技术^[4]和ALM(Application Layer Multicast)技术^[5]等, AOI是利用游戏环境中的玩家一般只跟他周边的其他玩家打交道这一系统特征,通过只给临近的对象发送状态更新数据的方法来减少通信负载; DR的实质是以牺牲系统一致性来优化通信;

ALM 则是建立在 P2P 覆盖网模型基础上,通过合并重复信息的传输来达到减少带宽浪费和降低服务器处理负担的目的。

根据 MMOG 兴趣域的特征可以将 MMOG 的组播服务分为两种,即 AOI 域间组播和 AOI 域内组播。本文主要研究基于 AOI 的域内覆盖组播树生成及优化算法,即每一个 AOI 兴趣域对应一棵相应的组播树,因此 AOI 兴趣域的具体划分算法这里不做介绍,文献[1-3]有详细的描述。

在 AOI 域内可以采用 Scribe^[6-7]组播,文献[6-7]中将每一个兴趣域作为一个组播组,利用 Scribe 进行组播。但是 Scribe 组播在设计的时候由于没有考虑到游戏网络中所有玩家结点在计算能力和带宽上的差异性,很可能会造成低带宽结点在组播树中距离组播源较近而不能接受大量的连接(例如游戏 Quake III 中更新率为 20 次/秒,每次更新的数据量大约为 100 个字节,这样就要求每个玩家结点必须提供至少 16 kbit/s 的带宽以维持游戏的正常更新,因此,一个带宽为 128 kbit/s 的结点最多只能扩展出 8 个其它玩家结点^[8]),显然这会造成组播树的扩展性受限,进而影响整个 MMOG 的同步性和实时性。事实上,由于现有网络接入带宽的差异性较大,没有度约束的 Scribe 组播其实是不适用于 MMOG 的。因而我们提出一个综合考虑带宽和时延条件的组播树生成及优化算法,从实验结果上看,该算法能够有效满足 MMOG 对于时延和扩展性的双重要求。

1 AOMST 算法

1.1 基本概念

对于网络,其生成树中的边也带权,将生成树各边的权值总和称为生成树的权,并将权值最小的生成树称为最小生成树(Minimum Spanning Tree, MST)。MST 的生成算法主要有两种:适合于边稠密的普里姆(Prim)算法和适合于边稀疏的克鲁斯卡尔(Kruskal)算法^[9]。由于我们的工作是建立在 Prim 算法的基础上,因此,下面仅对其作简要介绍。

Prim 算法的基本思想是:

(1) 在图 $G = (V, E)$ (V 表示顶点, E 表示边) 中, 从集合 V 中任取一个顶点(例如取顶点 v_0)放入集合 U 中, 这时 $U = \{v_0\}$, 集合 $T(E)$ 为空。

(2) 从 v_0 出发寻找与 U 中顶点相邻(另一顶点在 V 中)权值最小的边的另一顶点 v_1 , 并使 v_1 加入

U 。即 $U = \{v_0, v_1\}$, 同时将该边加入集合 $T(E)$ 中。

(3) 重复(2), 直到 $U = V$ 为止。

这时 $T(E)$ 中有 $n - 1$ 条边, $T = (U, T(E))$ 就是一棵最小生成树。

定义 1 组播树 T 是图 $G = (V, E)$ 的一棵生成树, 这里我们定义它为三元组 $T(r, U, T(E))$, r 为组播源结点, $U = V - \{r\}$, 而 $T(E)$ 为 E 的子集。

下面列举算法中用到的一些符号意义:

B_i : 结点 i 的带宽

$P_i(T)$: 结点 i 的父亲

$G_i(T)$: 结点 i 的祖父

$C_i(T)$: 结点 i 的孩子数

$N_i(T)$: 结点 i 的子孙数

$L_i(T)$: 结点 i 到组播源结点 r 的总时延, 也可表示为 $L(i, r)$

$d(i, j)$: 结点 i 到结点 j 的单播时延

T_i : 结点 i 为根的组播子树

定义 2 某结点 i 的度是由其带宽决定的, 假设 MMOG 的消息更新率为 u , 则该结点的度即可用如下公式计算而得。

$$D_i(T) = \text{floor}(B_i/u) \quad (1)$$

式中, floor 为地板函数。

结点 i 的度 $D_i(T)$ 表示了其扩展的能力, 即能够派生出的子结点数。

定义 3 $L(i, j)$ 表示从结点 i 到结点 j 的组播时延, 特别地, 用 $L(T_i)$ 表示以结点 i 为根的组播子树 T_i 的时延, 因此组播树 T 的总时延可以由如下公式计算:

$$\begin{aligned} L(T) &= \sum_{\forall i \in U} L_i(T) \\ &= \sum_{\forall P_i(T)=r} (L(T_i) + d(i, r) \cdot (N_i(T) + 1)) \end{aligned} \quad (2)$$

由式(2)得到的组播树 T 的总时延可以很容易计算出各结点的平均时延。鉴于一般网络状态的频繁变动以及带宽与时延的相互制约, 单纯从时延或者带宽角度构建的最小生成树通常并不是最优的生成树, 有时甚至是不合理的, 即可能生成可扩展性很差或者实时性不好的覆盖网组播树, 因此本文设计了一种影响因子可调(见式(5))的组播生成树算法 AOMST, 由它生成的组播树虽然不一定是最优的, 但却是很有效的, 另外考虑到 AOMST 组播树存在一些可利用的剩余带宽, 我们设计了 3 种不同的局部变换操作, 通过它们可以进一步提高 AOMST 组播树的总体性能。

1.2 AOMST 组播树的生成

组播生成树的源结点一般由游戏运行商指定,当游戏运营商不提供时,也可以从 AOI 域中选择能力较强的结点代替。假设当前 AOI 组播域中的结点个数为 N , 其中组播源结点为 r 。下面详细阐述 AOMST 算法是如何生成它的组播树的。

定义 4 U_{wait} 表示等待加入组播树的结点集合, U_{attached} 表示已加入结点集, U_{attached} 再进一步细分为两部分: U_{open} 和 U_{closed} , U_{open} 表示还有能力进行扩展的结点集, U_{closed} 表示已经不能再转发数据包给其他结点的叶子结点, 即其带宽已耗尽。刚开始, $U_{\text{wait}} = U$, $U_{\text{open}} = \{r\}$, $U_{\text{closed}} = \emptyset$ 。

定义 5 U_{wait} 集中任意结点 i 到组播源结点 r 的最短时延计为 $LL_i(T)$, 假设结点 i 加入组播树的接入结点为 a , 则 $LL_i(T)$ 可以表示为

$$LL_i(T) = \min \{L_a(T) + d(i, a) | a \in U_{\text{open}}\} \quad (3)$$

定义 6 假设系统可接受的最大时延为 LL_{\max} , 当 $LL_i(T) > LL_{\max}$ 时, 结点 i 将不能加入组播树, 因此我们定义另一个结点集 $U_{\text{qualified}}$, 它仅包含 U_{wait} 集中满足 $LL_i(T) \leq LL_{\max}$ 的结点。然后, 找出 $U_{\text{qualified}}$ 集中

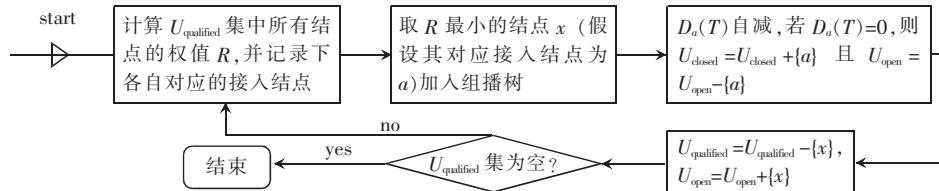


图 1 AOMST 组播树的生成

1.3 AOMST 生成树的优化

生成组播树后, 还需要根据覆盖网络的具体情况进行相应的优化, 优化操作的主要目的是减少组播树的时延。按照上述流程生成的组播树并不一定是最优的, 为此我们规定: 假定当前组播生成树为 $T_c(r, U, T(E))$, 则组播树的优化在于找出这样的生成树 $T_o(r', U', T'(E))$, 它满足如下条件:

- (1) $\{r'\} \cup U' = \{r\} \cup U$
- (2) 对于任意的结点 i , $C_i(T_o) \leq D_i(T_o)$
- (3) $L(T_o) < L(T_c)$

由于 AOMST 不是一种带宽贪婪算法, 因此靠近组播源的一些结点可能会存在可供扩展的剩余度(带宽), 这给我们进一步提高组播树的性能提供了空间。下面设计 3 种变换操作, 通过它们可以实现对当前组播树的性能优化。

1.3.1 从祖父结点扩展

如图 2 的左子图所示, b 结点的祖父结点 s 存在可扩展的度, 因此可以考虑将 T_b 子树往上迁移至如

所有结点到组播源结点 r 的最短时延序列的最小值, 并将其记为 $LL_{\min}(T)$, 则 $LL_{\min}(T)$ 可以表示如下

$$LL_{\min}(T) = \min \{LL_i(T) | i \in U_{\text{qualified}}\} \quad (4)$$

由式(1)可以计算 $U_{\text{qualified}}$ 集中所有结点的度, 取其最大值记为 $D_{\max}(T)$, 另外, 我们将 $U_{\text{qualified}}$ 集中结点 i 加入组播树的权值计为 R_i , 这里 R_i 不仅包含时延因素, 还应考虑可扩展性因素, 即结点的度, 因此, 将其设计为

$$R_i = \frac{1}{\left(\gamma * \frac{LL_{\min}(T)}{LL_i(T)} + \varphi * \frac{D_i(T)}{D_{\max}(T)}\right)}, i \in U_{\text{qualified}} \quad (5)$$

其中, γ 和 φ 为可调节因子, 它们可以让系统在时延和扩展性方面做出权衡, 当取 γ 比 φ 大时, 意味着系统更倾向于重视时延因素, 相反, 若取 γ 比 φ 小时, 则表示系统更强调可扩展性, 因此它们的取值将直接影响到组播树的生成。特别地, 当 $\gamma=1, \varphi=0$ 时, 该算法退化为以时延为唯一度量的贪婪算法, 而当 $\gamma=0, \varphi=1$ 时, 该算法又变成以度(即带宽)为度量的贪婪算法。

AOMST 组播树生成算法的流程图如图 1 所示。

右图处, 只要左子图组播树满足条件

$$d(s, b) < L(s, b) = d(s, a) + d(a, b) \quad (6)$$

即可实施此操作来减少组播树的总时延。在 MMOG 所处的互联网环境中, 条件 $d(s, b) < d(s, a) + d(a, b)$ 多数情况下是成立的, 因此这种变换在所有优化操作中是较频繁的一种。

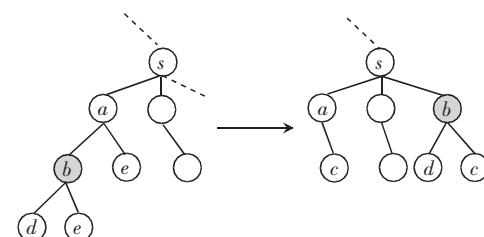


图 2 从祖父结点扩展的变换

1.3.2 从子结点扩展

这种变换通常发生在子结点的能力比其父结点强时, 比如子结点比父结点更接近祖父结点(即时延更少)且其带宽条件也不错的情况下, 子父结点进行对换操作, 如上图 3 所示, 子结点 b 向上顶替了

其父结点 a 的位置,而 a 子树(除去 T_b 分子树)则由 b 结点来扩展。具体当上图左子图的组播树满足条件

$$(d(s,a)+d(a,b)-d(s,b)) \cdot (N_b(T)+1) > \\ (N_a(T)-N_b(T)) \cdot (d(s,b)+d(b,a)-d(s,a)) \quad (7)$$

则执行此对换操作。值得指出的是,在 MMOG 所处的互联网环境中,这类操作虽然发生没有第一种操作频繁,但它却给距离组播源较远但有能力的结点提供向上迁移的途径,比如在 MMOG 的 AOI 组播域中有能力较强的玩家客户端后加入到相应的组播树中,这时该操作将被执行一次或多次,结果是该玩家客户端被对换到更能发挥其能力的组播位置,以提高系统的整体性能。

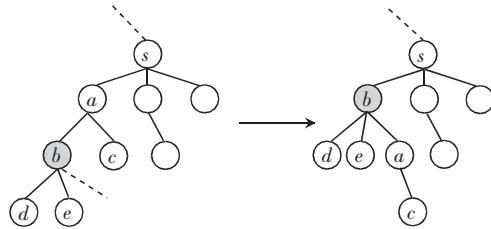


图 3 从子结点扩展的变换

1.3.3 从叔结点扩展

图 4 左子图表示,结点 b 的叔结点 u 具有可用带宽,则当该组播树满足条件

$$d(s,u)+d(u,b) < L(s,b) = d(s,a)+d(a,b) \quad (8)$$

便执行变换操作将 b 结点子树 T_b 迁移至其叔结点 u 处,如图 4 右子图所示。

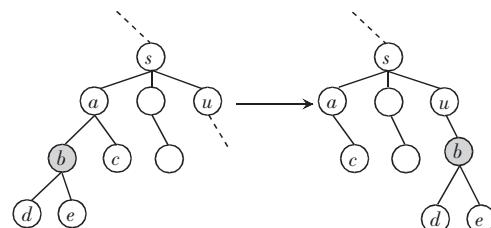


图 4 从叔结点扩展的变换

2 实验与分析

2.1 实验的分析目标

$$(1) U_{\text{qualified}} \text{ 和 } U_{\text{wait}} \text{ 集的大小比值 } \eta = \frac{|U_{\text{qualified}}|}{|U_{\text{wait}}|}.$$

之所以选择 η 来分析,是因为它能够全面反映带宽和时延条件,反过来,当时延带宽条件一样时, η 越大表明算法越好。

(2) 存在大量结点时的时延优化。

2.2 实验环境

采用乔治亚理工大学的网络拓扑生成器 GT-ITM^[10]产生基于 transit-stub 模型的网络拓扑作为我们实验的网络环境,该网络拓扑共包括 1 500 个随机产生的路由器节点,由 5 个平均拥有 6 个路由器节点的互相连通的 transit 域构成核心,平均每个 transit 域节点与 5 个 stub 域相连,每个 stub 域平均包含 10 个路由器节点。域内节点链路连接采用 Waxman 模型并假设相连的路由器之间的通信时延和它们之间的距离成正比,然后将 N (N 在实验时设定) 个虚拟玩家结点随机接入网络拓扑的路由器上。玩家结点的端到端时延分布为 [5 ms, 35 ms], 玩家结点的平均可扩展度数(即带宽)为 6, 其中上限为 30, 下限为 2, 并呈正态分布, 组播源则从可扩展度数较大的结点中随机选取一个。下面为对 AOMST 算法中 $U_{\text{qualified}}$ 和 U_{wait} 集的大小比值 η 以及玩家结点的时延优化进行仿真实验得出的结果。

2.3 实验结果和分析

- (1) case 1: $\gamma = 0.3, \varphi = 0.7$;
- case 2: $\gamma = 1, \varphi = 0$;
- case 3: $\gamma = 0, \varphi = 1$.

在 AOMST 算法中 γ 和 φ 取值为上述 3 种情况下,假设 $LL_{\max} = 300$ ms, 分别对其进行实验,结果如图 5 所示。

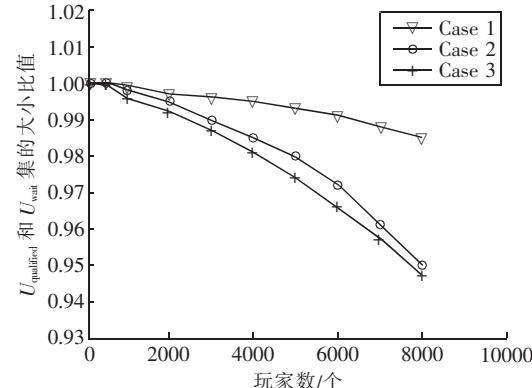


图 5 $U_{\text{qualified}}$ 和 U_{wait} 集的大小比值

从图 5 中可见,第一种情况的表现最好,这说明兼顾时延和结点度的组播树生成算法是合理的,它比只顾单方面的算法(如 Scribe 组播)有效。但同时我们也发现,在随着玩家数增多的同时,3 种情况下 $U_{\text{qualified}}$ 和 U_{wait} 集的大小比值 η 都呈下降趋势,表明系统在已有大量玩家后,新玩家的加入会越来越困难。

(2) 如下为 $N=1\,000$ 、 $N=6\,000$ 和 $N=9\,000$ 三种情况下用 AOMST 算法的 3 种变换操作对组播生

成树进行优化的结果。

从图6可见,对大量结点应用AOMST算法生成组播树后,还可以利用1.3小节的变换操作来进一步减少系统的平均时延,从而提高MMOG系统的实时性。此外,当系统的结点数较少时(如1 000个),变换操作带来的时延优化并不明显,相反,当系统结点数较多时(如9 000个),此时变换的优化效果则要明显得多。

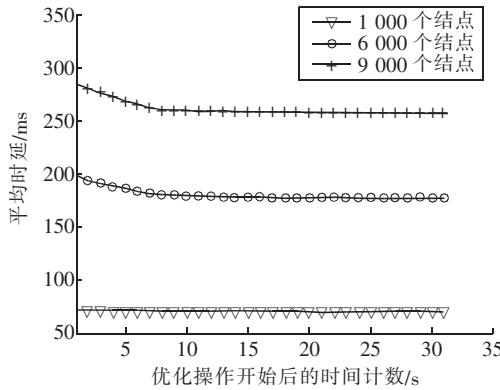


图6 平均时延的优化

3 结束语

本文提出了一种支持大型多人在线游戏的覆盖组播生成树算法AOMST,该算法比Scribe组播算法具有更好的适应性。通过调整式(5)中 γ 和 φ 的取值,系统扩展性及时延因素得以全面考虑,因此AOMST算法更可能构建出合理的组播生成树,此外,通过3种不同的变换操作,系统的平均时延还能够进一步优化。我们下一步工作还将通过仿真实验分析可调因子 γ 和 φ 的取值与不同条件MMOG系统之间的关系以及玩家结点的突然退出和动态加入对系统的影响(如时延抖动)。

参考文献:

- [1] BACKHAUS H, KRAUSE S. Voronoi-based adaptive scalable transfer revisited: gain and loss of a voronoi-based peer-to-peer approach for mmog[C]// Proceedings of the 6th ACM SIGCOMM workshop on Network and system support for games. 2007:49–54.
- [2] CASTRO M, DRUSCHEL P, KERMARREC A, et al. Scribe: a large-scale and decentralized application-level multicast infrastructure[J]. IEEE Journal on Selected Areas in Communications, 2002, 20(8): 1489 – 1499.
- [3] JIANG X, SAFAEI F, BOUSTEAD P. Enhancing the multicast performance of structured P2P overlay in supporting Massively Multiplayer Online Games[C]// Proc of 15th IEEE International Conference on Networks ICON. 2007:124 – 129.
- [4] SANDEEP K S. Effective remote modeling in large scale distributed simulation and visualization environment[D]. California: Stanford University, 1996.
- [5] CHEKURI C, CHUZHOY J, LEWIN-EYTAN L. Non-Cooperative Multicast and Facility Location Games[J]. IEEE Journal on Selected Areas in Communications, 2007, 25(6): 72 – 81.
- [6] KNUTSSON B, LU H, XU W, et al. Peer-to-Peer support for massively multiplayer games[C]// INFOCOM. 2004, 1:96 – 107.
- [7] SHI Xingbin, ZHOU Dongming. Peer-to-Peer Support for MMOG [J]. MINI-MICRO SYSTEMS, 2005, 26 (12): 2100 – 2103.
- [8] PANG Jeff. Scaling Peer-to-Peer Games in Low Bandwidth Environments[EB/OL]. <http://research.microsoft.com/en-us/um/redmond/events/iptps2007/papers/pangyedalorch.pdf>.
- [9] YAN Weimin, WU Weimin. Data Structure[M]. 2ed. Beijing: Tsinghua University Press, 2006:171 – 174.
- [10] Georgia Institute of Technology. Georgia Tech Internetwork Topology Models[EB/OL]. <http://www.cc.gatech.edu/fac/Ellen.Zegura/graphs.html>

作者简介:



林巧民(1979-),男,福建泉州人。南京邮电大学数字媒体研究中心讲师,博士研究生。主要研究方向为基于通信网络的计算机软件技术、数字媒体技术及多媒体传感器网络等。

王汝传(1943-),男,安徽合肥人。南京邮电大学计算机学院教授,博士生导师。(见本刊2009年第1期第67页)

许棣华(1975-),女,江苏南京人。南京邮电大学计算机学院讲师,博士研究生。主要研究方向为计算机网络、计算机软件在通信中的应用和信息安全。

林萍(1985-),女,福建泉州人。南京邮电大学计算机学院硕士研究生。主要研究方向为计算机软件、计算机网络和通信协议等。