

<http://www.phei.com.cn>

河南省计算机学会论文集

计算机研究 新进展(2010)



电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

河南省计算机学会 2010 年学术年会论文集

计算机研究新进展 (2010)

河南省计算机学会 组编

電子工業出版社·

Publishing House of Electronics Industry

北京·BEIJING

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

计算机研究新进展. 2010/河南省计算机学会组编. 北京：电子工业出版社，2010.9
ISBN 978-7-121-11828-9

I. ①计… II. ①庄… III. ①电子计算机—文集 IV. ①TP3-53

中国版本图书馆 CIP 数据核字（2010）第 180644 号

策划编辑：何 况
责任编辑：张贵芹 何 况 特约编辑：李云霞
印 刷：
装 订：
出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036
开 本：850×1 168 1/16 印张：23.25 字数：687 千字
印 次：2010 年 9 月第 1 次印刷
定 价：85.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888。
质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。
服务热线：(010) 88258888。

前 言

2010 河南省计算机大会暨学术年会是河南省计算机学会第四届理事会主办的第五次学术交流大会。大会由郑州轻工业学院承办。

河南省计算机学会第四届理事会自 2005 年 7 月成立以来,就以全新的姿态朝气蓬勃地开展着各项工作,努力致力于“三主一家”(即学会要作为河南省计算机学术交流的主渠道,科学普及的主战场,国内外民间学术交流的主要代表,成为会员之家)的建设,五年来发展态势良好,在河南省计算机领域的建设中发挥了积极的作用。

河南省计算机学会 2008 年创办的河南省大学生程序设计竞赛已成功举办了三届,参赛队伍逐年大幅增加,有力地推动了河南省大学生程序设计能力的锻炼与提高,必将成为学会的又一品牌项目。学会继续在组织发展、服务会员、学术交流、青少年信息学奥林匹克竞赛等方面积极开展着工作。2009 年,在全国青少年信息学奥林匹克竞赛 25 周年之际,河南省计算机学会被中国计算机学会授予优秀省组织单位。学会与电子工业出版社的合作进展顺利,已陆续出版了《计算机组成与维护》、《Linux 操作系统》、《32 位汇编语言程序设计》、《C/C++程序设计教程——面向过程分册》、《C/C++程序设计教程——面向对象分册》、《JAVA 程序设计应用教程》、《Access 数据库程序设计》、《MATLAB 基础及应用教程》、《网络安全技术及应用》等教材并得到广泛的应用(其中《计算机组成与维护》、《Access 数据库程序设计》等已推出第二版),取得了良好的社会效益。

本次计算机大会是学会在 2010 年的一项重要工作,为举办好这次大会,郑州轻工业学院进行了精心的组织,为保证会议的圆满成功付出了艰辛的劳动。学会表示衷心的感谢。同时学会相信本次会议仍将是一次老友新朋齐相聚、相互交流谋发展的促进会,是推进新技术发展的展示会,是河南省计算机科学技术水平的检阅会,是促进会员之间友好交流的联谊会。

学会真诚地希望广大计算机科技工作者及各政府机关、高等院校、企事业单位的领导关心支持河南省计算机学会,积极参与学会的各项活动,进行学术交流,为推进河南省计算机科学与技术的发展共同努力。

本论文集是学会与电子工业出版社合作的继续,感谢论文作者对学会工作的支持,感谢为论文审稿、修改的专家学者们,同时也由衷地感谢电子工业出版社的友好支持和大力帮助。

河南省计算机学会

2010 年 8 月 6 日

目 录

流媒体技术综述.....	1
A Review of Streaming Media	吕 斌 庄 雷
可重构互连网络的发展与现状.....	6
The Development and Present of Reconfigurable Interconnection Network	曾 韵 蒋烈辉 董卫宇
类比推理的计算模型研究综述.....	15
Computational Model of Analogical Reasoning Summary	周 倩 沈夏炯
恶意代码检测技术综述.....	21
Survey on Static Detection Technologies of Malware	奚 琪 王清贤 曾勇军
计算机总线的分类及发展趋势.....	28
On the Classification and Developing Trend of Computer Bus	丁彦芳 黄欢欢 王月蓉 秦风云
形式化开发非递归 Koch 曲线算法.....	32
Formal Development of Non-recursive Algorithm for Koch Curve	刘润杰 申金媛 穆维新
商空间模型下不确定本体知识推理研究.....	37
Uncertain Ontology Knowledge Reasoning Research of Based on Quotient Space Model	王晓东 孙 滨 李学威
粒子滤波多样性测度分析.....	42
Research on Diversity Measure in Particle Filter	于金霞 刘文静 汤永利
基于 CPSS 算法的 RTAI 调度器的改进.....	48
The Improvement of RTAI Real-time Scheduling Algorithm and Scheduler	李学桥 梁 爽 陈 园
改进的 UIO 序列生成算法.....	54
Improved UIO Sequences Generation Algorithm	黎中文 张来顺 肖健鹏

一个基于 IB 原理的单类算法——OCD-BA 算法.....	58
A Algorithm for One-Class Based on IB Principle——OCD-BA	
	王媛媛 叶阳东
矩阵乘法的 FPGA 并行设计与实现.....	63
The Parallel Design and Implementation Matrix Multiplication on FPGA	
	何红旗 邵 仪 蒋烈辉 赵秋霞
基于 Zigbee 技术的加权质心定位算法.....	69
Weighted Centroid Localization Based on Zigbee Technology	
	李占波 刘慧玲
基于自适应窗和 Hartley 变换的河工模型 PIV 测速.....	74
The Particle Image Velocimetry of River Model Based on Adaptive Window and Hartley Transform	
	喻 恒 赵建军
一种团队 CGA 行进中的队形维护方法.....	80
An Method for Formation Maintaining of Team CGA Advancing	
	郑延斌 李双群
基于历史时序的访问控制模型研究.....	87
Towards An Access Control Model Based on History Temporal Data	
	徐长征 王清贤 颜学雄
基于 FIDXP 的分布式入侵防御系统的设计.....	93
FIDXP-based Distributed Intrusion Detection System Prevention and Implement	
	刘 松 赵东明 周清雷
基于抽象区间域的数组边界检查技术.....	100
Array Bound Checking Technology Based on Abstract Interval Domain	
	曾勇军 王清贤 奚 琪
Kerberos 协议在单点登录中的改进及应用.....	107
Improvement and Application of Kerberos Protocol in A Single Sign-On System	
	郭甜滋 毛 楠 司志刚 陈 丽
分布式安全评估通信协议的研究与设计.....	113
Research and Design of Communication Protocol in Distributed Security Assessment	
	李金武 郑秋生
一种分布式安全评估主控中心的研究与设计.....	119
Research and Design on Control Center of Distributed Security Assessment System	
	夏 冰 夏敏捷 徐 飞 郑秋生

重要信息系统安全测评工具的研究与设计	125
Research and Design of Important Information System Security Assessment Tool	武俊芳 郑秋生
基于策略的网络安全管理研究	131
Research on Network Security Management Based on Policy	王海涛
信息安全技术在电子政务系统中的应用研究	136
The Application Research of Information Security Technology in E-government System	张 鸣
浅谈电子商务及其安全性	140
Discussion of E-commerce and Its Safety	郎士宁 秦兴桥 王月蓉
浅谈通信新技术的安全威胁	145
Discussion of the Security Threats of New Communications	杨 凯 郎士宁 黄欢欢
储粮害虫图像识别知识库研究	150
Study on Image Recognition of Stored-grain Pests Knowledge Base	王利强 张红梅
一种有效的 SAR 图像角反射器检测方法	154
An Effective Method of SAR Image Corner Reflector Detection	薛笑荣 王爱民 曾琪明
一种基于 Canny 检测算子的图像分割算法	159
An Image Segmentation Algorithm Based on Canny Edge Detection	明 生 邬长安 马 珂
G.SHDSL 技术在远距离音视频信号传输系统中的应用研究	163
G. SHDSL Technology in the Remote Audio and Video Signal Transmission System of Applied Research	黄继海 杨建国 姜鹏飞
H.264 编码技术及其应用	168
The Technology and Application of H.264 Coding	黄欢欢 王月蓉 冯少华
基于 Arnold 和 DCT 的数字水印技术研究	172
Research of Digital Watermark Based on Arnold and DCT Technology	尚 存 邬长安

基于 HLA 的仿真系统可视化数据模型研究.....	177
Study on Visual Data Models of HLA-based Simulation System	
	周龙龙 姜鹏飞 黄建廷
红外小目标背景抑制和检测方法研究.....	183
Research on IR Small Target Detection and Backgroud Suppression	
	秦兴桥 郎士宁 王 辉
基于马尔可夫链模型的软件可靠性测试研究.....	188
The Research of Software Reliability Testing Based on Markov Chain Model	
	何 焱 张来顺 石荣刚
VXI 总线在军事装备检测系统中的应用.....	194
Application of the VXI Bus in the Military Equipment Detection System	
	王书伟 杨 静 丁彦芳
基于嵌入式的车载式压实度检测系统.....	199
Design of In-vehicle Compact Degree Detection System Based on Embedded System	
	普 邑 王新勇
数据库设计中 SQL 优化策略和技巧.....	207
Some Optimization Strategies and Techniques of SQL in Designing Database System	
	秦 军
.Net 控制 Excel 自动生成表格的应用研究.....	213
Application of Automatic Table Excel controlled by .NET	
	王 辉 杨 凯 郎士宁 冯少华 王月蓉
.NET 平台下材料管理系统的设计模式研究.....	217
Study on Design Pattern of Material Management Base on .NET Platform	
	孟 军 王 辉 杨 凯 秦兴桥
基于客户满意度的第四方物流多属性指派决策机制.....	223
The 4th Party Logistics Multi-attribute Assignment and Decision Mechanism Based on Customer Satisfaction	
	周宏宇 张战峰 栗青生 葛彦强
面向用户和领域本体的 Web 信息采集系统.....	230
User and Domain Oriented Web Information Collection System	
	张素智 李宝燕 樊得强
ERP 项目中成本管理子系统的分析与设计.....	237
Analysis and Design of the Subsystem of the Cost Control System in the ERP Project	
	申 康

分布式缓存策略模式在高考网上报名系统设计中的应用	250
The Distributing Cache Mechanism by S strategy Patterns and Implementation of Higher Education Admission Information Collection System Based on Web	
杨浩杰 宋 涛 刘 刚	
无线测控通信平台中间件的设计与实现	259
A Middleware for Wireless Multi-Hop Network Based Measurement and Control System	
潘 磊 张书钦 郑秋生	
云计算及其在移动学习模式下应用初探	266
Preliminary Study of Cloud Computing and Its Application Under the Mode of Mobile Learning	
赵 萌	
网络统计及软件研究	273
Reach on Internet Statistic and Software	
贺学剑 孟光胜	
基于 CBR 的人防城市人口应急疏散预案辅助决策研究	284
Urban Population Based on CBR's Civil Air Defense Decision on Emergency Evacuation	
左 军 刘凤荣 张 涛 彭祥新	
人防城市人口应急疏散路径选择模型研究	288
Urban Air Defense Emergency Evacuation Route Choice Model	
左 军 刘凤荣 黄欢欢 王月蓉	
基于 Pastry 算法的物联网信息发现服务	292
Discovery Service Based on Pastry for Internet of Things	
李占波 刘冬冬	
看我国 SNS 社交网站现状与趋势	298
The Status and Trends of Social Networking Site in China	
牛 星	
由 AT89S52 与 TC35i 实现的短信息处理系统	304
Short Information Management System Which Realize by AT89S52 and TC35i	
王书伟 张 鸣 杨 静	
永煤集团安全监测联网系统及其应用	309
The Security Detection Networking System of Yongcheng Coal-electricity Group and Its Application	
冯少华 田 丰 王月蓉 黄欢欢	
基于三层架构的科研管理系统研究与应用	314
The Research and Application of Scientific Research Management System Base on Three-Tiers Architecture	
王 辉 孟 军 秦兴桥 丁彦芳 黄欢欢	

基于 OSG 的三维虚拟化学实验的建模技术.....	318
Three-Dimensional Virtual Chemistry Experiments Modeling Techniques Based on OSG	
谭同德 石奇波 赵新灿	
知识合作在计算机类课程教学中的应用研究.....	326
Application and Research of A Teaching Model Based on Knowledge Cooperation	
赵 妍 李玲玲	
协议分析软件在计算机网络教学中的应用.....	329
The Application of Protocol Analysis Software in the Computer Network Teaching	
张巧荣 郑娅峰	
操作系统进程同步的教学实践.....	335
Teaching Practice of Operating System Process Synchronization	
李志民 赵一丁 底 恒	
“数据库系统原理” 教学实践与改革.....	340
Practice of Teaching and Reformation of Database System Principle	
职为梅	
计算机人才培养研究.....	344
The Research of Computer Person’s Training	
曲宏山 崔清民	
高等学校教学资源共享运行机制研究.....	348
Research on the Sharing Mechanism of ,Teaching Resource in College	
郑娅峰 张巧荣	
任职教育中计算机教学方法浅析.....	353
On The Teaching Methods of Computer in Professional Education	
王月蓉 冯少华 黄欢欢 金 玉	
二叉树的四种遍历的非递归算法.....	356
Binary Tree Traversal Non-recursive Algorithm for the Four	
王正辉 姜鹏飞 张锋	

流媒体技术综述

吕 斌, 庄 雷

(郑州大学信息工程学院, 河南 郑州, 450001)

摘 要: 作为多媒体和网络领域的交叉学科, 流媒体技术的应用和研究得到了迅速发展。本文主要介绍流媒体在目前发展状况下所产生的相关技术。首先说明了流媒体的定义; 然后介绍流媒体在当前因特网的工作环境下是如何运转的, 以及发送和接收流媒体数据所使用的网络传输与控制协议; 最后综述为了提高视频服务器的服务能力, 研究人员所提出来的一些解决方法, 特别是流媒体数据调度技术。

关键词: 流媒体; 调度技术; 网络协议

中图分类号: TP393 **文献标识码:** A **文章编号:** 1006-7043 (2004) xx-xxxx-x

A Review of Streaming Media

LV Bin, ZHUANG Lei

(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, Henan China)

Abstract: As the interdisciplinary of multimedia and network, the applications of streaming media technology and research have developed rapidly. In this paper we mainly discuss a number of technical based on the development of streaming media in the current situation. Firstly, we present the definition of the streaming media and the principle of it under the Internet. After that we introduce the network transmission and control protocol of streaming media. Finally, we introduce the current data scheduling technical which is to improve the service capability of video server.

Keywords: streaming media; data scheduling technical; network protocol

引言

流媒体是宽带通信网和多媒体技术共同发展产物。近年来, 随着 Internet 的普及和对 3G 研究的深入, 视频流媒体技术得到了普遍关注, 围绕数字媒体内容生产、加工、存储、调度、传输、缓存和接收技术的研究, 已经成为近期多媒体和通信领域的重要热点。

本文主要是对流媒体在发展过程中所产生的技术, 进行简单的介绍。先粗略地说明流媒体在因特网下的工作原理, 此工作原理是流媒体系统在网络上工作的一般方式, 具有普遍性。同时也介绍了流媒体在传输过程中所要使用的协议, 通过对这些协议的分析, 更加具体地说明了流媒体系统的工作原理。最后介绍了流媒体的调度技术, 通过对这些调度技术的运用, 流媒体系统的工作效率得到了很大的提高。

流媒体 (Stream Media) 是指在网络中使用流式传输技术的连续时基媒体^[2]。

说得更具体一点, 所谓流媒体技术, 就是把连续的影像和声音信息经过压缩处理后放到流媒体网络服务器上, 通过因特网让浏览者一边下载一边观看、收听, 而不需要等到整个多媒体文件下载完成就可以即时观看的技术, 即流媒体技术实现了边传输、边下载、边播放的过程。只需经过几秒的启动延时即可在用户计算机上利用相应的播放器进行播放和观看, 甚至可以随时地进行暂停、快进、快退

基金项目: 河南省重大攻关项目 (092101210104); 河南省留学回国人员科研基金

作者简介: 吕斌 (1986—), 男, 硕士研究生;

庄雷 (1963—), 女, 教授, 博士, 博士生导师。

等操作。

流媒体与普通媒体的差别在于：对于普通媒体，在访问它之前要得到全部的内容；对于流媒体，则在完全接收到全部内容之前就可以开始访问。

1 流媒体工作原理

流传输的实现需要缓存，因为因特网是以分组传输为基础进行断续的异步传输的，对一个实时音频、视频源或存储音频、视频的文件，在传输中它们要被分解为很多包，由于网络是动态变化的，各个包选择的路由可能不尽相同，因此到达客户端的时间延迟也就不等，甚至先发的数据包还有可能后到。为此，使用缓存系统来弥补延迟和抖动的影响，并保证数据包的顺序正确，从而使媒体数据能连续输出，而不会因为网络暂时拥塞使播放出现停顿。通常高速缓存所需容量并不大，因为高速缓存使用环形链表结构来存储数据：通过丢弃已经播放的内容，就可以重新利用空出的高速缓存空间来缓存后续尚未播放的内容。

流媒体传输的一般过程是：用户选择某一流媒体服务器后，Web 浏览器与 Web 服务器之间使用 HTTP/TCP 交换控制信息，以便把需要传输的实时数据从原始信息中检索出来；然后客户机上的 Web 浏览器启动 A/V 客户端程序，使用 HTTP 从 Web 服务器检索相关参数对客户端程序初始化，如图 1 所示。这些参数可能包括目录信息、A/V 数据的编码类型或与 A/V 检索相关的服务器地址等。

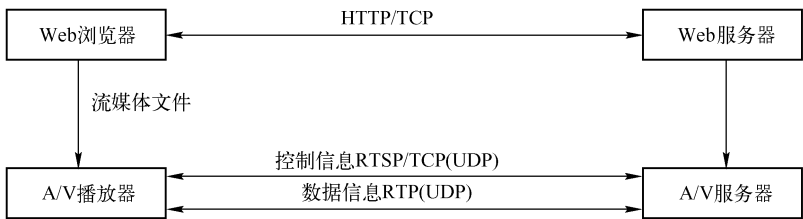


图 1 流媒体工作过程

A/V 客户端程序及 A/V 服务器运行实时流协议（RTSP），以交换 A/V 传输所需的控制信息。与 CD 播放器或 VCR 所提供的功能相似，RTSP 提供了操作播放、快进、快倒、暂停及录制等命令的方法。A/V 服务器使用 RTP/UDP 协议将 A/V 数据传输 A/V 客户程序，一旦 A/V 数据到达客户端，A/V 客户程序即可播放输出。

2 流媒体网络传输与控制协议

流媒体的实现离不开网络实时传输技术，良好的网络通信通道的设计和通信协议的选择对于视频数据传输的实时性及控制命令的准确性是至关重要的，对于协议的选择，IP 层是通用的 IP 协议，而“传输层”中应用于监控系统的主要协议有传输控制协议 TCP，用户数据协议 UDP。应用层协议有实时传输协议 RTP（Real-Time Transfer Protocol）用于数据传输，实时传输控制协议 RTCP（Real-Time Transfer Control Protocol）用于统计、管理和控制 RTP 传输。资源保留协议 RSVP（Resource reservation Protocol），实时流协议 RTSP（Real-Time Streaming Protocol）。

RTP^[1]是一个因特网标准协议，用于提供端对端传输服务的实时传输协议，以便支持实时应用。RTCP^[1]是 RTP 的同伴协议。设计 RTCP 协议是为了向 RTP 话路的参与者提供 QoS 反馈。也就是说，RTP 是一个数据传输协议，而 RTCP 是一个控制协议。在 RTP 会话期间，各参与者周期性的传输 RTCP 包。RTCP 包中含有已发送的数据包的数量、丢失的数据包的数量等统计资料。因此，服务器可以利用这些信息动态地改变传输速率，甚至改变有效载荷类型。当应用程序开始一个 RTP 会话时将

使用两个端口：一个给 RTP 进行数据流的传递，另一个给 RTCP 进行控制流的传递。一个传递有效数据，另一个帮助监视网络流量和阻塞情况，为有效数据传递提供可靠保障。RTP 和 RTCP 两者配合使用，能以有效的反馈和最小的开销使传输效率最优化，因而特别适合在 Internet 上传输实时数据。

RTSP^[1]是因特网上流媒体的话路控制协议。RTSP 的一个主要功能是支持类似 VCR 的控制操作，如停止、暂停/重新开始、快进和快退。此外，RTSP 还可以提供选择传输通道（如 UDP，组播 UDP 或 TCP）的方法，以及基于 RTP 的传输机制。RTSP 既可用于组播，也可用于单播。

RSVP^[1]是 1993 年 L. Zhang 提出的一种资源预留协议。针对 Internet 原有传输层协议不能保障 QoS 质量和不支持多点传输的缺点，RSVP 在业务流传送之前，预约一定的网络资源，建立静态或动态的传输逻辑通路，保障了每一业务流都有足够的“独享”的带宽，克服了由于网络信包过多引起的拥塞、丢失和重传，提高了网络传输的 QoS 性能。

3 流媒体的调度技术

流媒体调度技术是指在接收到客户端请求后，如何分配、管理和使用服务器及网络资源，为用户提供视频流服务，以提高系统的服务质量和处理性能。对流媒体应用而言，其业务数据量大、用户持续时间长，研究如何利用有限的系统资源（如网络带宽、服务器磁盘带宽和存储容量等）资源，提供更大的用户并发访问数量和更好的用户的服务质量具有更重要的价值，因此，调度技术是流媒体系统的关键问题。

目前，典型的流媒体调度算法分为两类：静态调度算法和动态调度算法^[8]。一般来说，静态调度算法采用服务器推模式，而动态调度算法则采用客户拉模式。服务器推模式是指，视频服务器不考虑用户动态行为而调度媒体流；客户拉模式是指，媒体流的调度首先由用户请求驱动，视频服务器根据一定调度算法响应用户请求。

3.1 静态调度策略

静态调度策略的基本原理是，将节目分成若干个段，服务器为每个段安排一个通道，在每个通道上作周期性的广播；用户的请求到达系统后，将选择加入最近的通道，经过有限等待后，开始从同步的该信道读取节目流，并根据调度算法的规定，在适当的时刻跳转到不同的通道，与缓存和平滑技术相结合以获得连续的流服务。不同的静态调度算法由于节目分段方式和数据接收策略的不同，具有不同的用户启动延迟和客户缓存空间要求。

假设节目 i 表示为 Prg_i ，根据调度算法将其划分为 n 个片段，第 j 个片段表示为 $\text{Seg}_{i,j}$ ，则有 $\text{Prg}_i = \{ \text{Seg}_{i,j} | 1 \leq j < n \}$ ，令 $\text{Length}(\text{Seg}_{i,j})$ 表示节目 i 第 j 个片段的播放时间长度， $\text{Size}(\text{Seg}_{i,j})$ 表示节目 i 第 j 个片段的存储空间大小。

最简单的静态调度算法称为等间隔广播（Equally-spaced interval periodical Broadcasting，EB），将节目 i 划分为等播放时间长度的片段，即对任意 $i \neq j$ ， $\text{Length}(\text{Seg}_{i,j}) = \text{Length}(\text{Seg}_{i,k})$ 。每个片段 $\text{Seg}_{i,j}$ 用一个组播通道 Ch_j 周期性播放；用户在接收完一个通道的数据后，转入下一个通道即可无时延的接收后续数据。在 EB 算法中，用户的平均等待时间为节目片段长度的一半，即 $\text{Length}(\text{Seg}_{i,j}/2)$ ；从理论上讲，调度算法所需的客户端缓存空间大小可以为 0。

为进一步缩小启动延迟，出现一系列改进算法。金字塔广播 PB（Pyramid Broadcasting）^[3]算法中，将节目划分为呈几何级数序列增长的片段，即 $\text{Length}(\text{Seg}_{i,j+1}) = y * \text{Length}(\text{Seg}_{i,j})$ ，其中 $y > 1$ 。因此，第 1 片段的长度可以很小，播放频率很高，从而有效降低了用户的启动延迟。但这种方法要求客户端能够同时从 2 个通道接收数据；并需要大量的空间用于缓存来自未播放通道上的数据流，当 $y = 2$ 时，该缓存空间的大小应该在节目总大小的 70% 以上。摩天大楼广播 SB（Skyscraper Broadcasting）^[4]算法，则提出了不同的数据分块和传输方法，按照指定序列将一个节目分块（如 {1, 2, 2, 5, 5,

12, 12, 25, 25, ...}), 且每个分块大小不大于宽度参数 w , 因而使得节目片段长度的增长变慢, 达到节省缓存空间的目的。

静态调度策略的优点是: 算法不受用户点播特性的影响, 并发用户数量的增加不会导致每个用户播放质量的下降, 因此能支持较多的并发用户, 每个可以得到较公平的启动延迟和服务质量; 其缺点是: (1) 用户需要一定的等待才能接受服务, 因此只能提供 NVoD 服务; (2) 单个节目占用多个通道, 适用于有大量点播请求的热门节目, 对流行度小的节目, 资源浪费严重, 系统可提供的节目数量受资源限制较大。

3.2 动态调度策略

典型的动态调度策略分为 3 类: **Batching**、**Patching** 和 **Adaptive Piggybacking**。

3.2.1 批量调度算法

批量调度 (**Batching**) 算法^[6]适用于集中式的 VoD 系统。算法的基本思想是: 对用户请求进行一定的延迟, 以等待后续对同一节目的其他请求的到达, 最终用同一个视频流为在一定时间范围内到达的对同一节目的点播请求批量提供服务。将两次对同一个节目调度之间的时间差称为调度间隔, 批量调度算法的调度间隔也就是该算法中用户启动时延的最大值。批量调度算法是在减少服务器带宽消耗和增加用户启动时延之间寻求平衡; 牺牲了部分公平性: 在一个调度间隔内, 先到达的用户需要等待后到达的用户。批量调度算法比较适合流行度大的热门节目, 因为只有热门节目才可能在一个调度间隔内有多个用户请求到达。在 **Batching** 算法中, **VCR** 操作是通过在同一个节目不同传输通道间的跳转实现的。因此, 对用户交互能力的支持非常有限。

3.2.2 补丁调度算法

为了解决批量调度 (**Batching**) 算法的用户启动时延问题, 提出了补丁调度 (**Patching**) 算法^[7]。补丁调度算法适合于分布式网络环境下的视频点播系统。其基本思想是: 允许不同时间到达的对同一个节目的请求共享一个视频流, 这种共享是通过让后到达的用户从为前面到达的用户分配的组播通道上接收数据实现的; 同时, 为后到达的用户请求分配一个专门的单播通道以传送该用户错过的部分流数据。因此, 用户在一段时间内将同时从共享的组播通道和专用的单播通道上接收数据, 利用客户端的缓存以实现无时延的用户点播服务, 使系统维持较高的效率。将该共享的组播通道上传输的数据流称为常规流, 将专用通道上的传输的数据流称为补丁流。显然, 若后一个用户请求到达的时间距最近的一个常规流的时间差越大, 补丁流的持续时间也越长, 所需的用户缓存空间也越多, 补丁算法的代价会持续增加, 直到超过指定的值, 传输补丁流的代价将超过重新开始一个常规流的代价。此时, 调度算法便会重新发起一个常规流, 开始新的调度周期, 将最大时间差 t 称做补丁门限。

在补丁调度算法的基础上, 演化出一系列优秀的算法, 构成了补丁调度算法族, 如 **Transition Patching**、**Controlled Multicast**、**Split and Merge Protocol**、**Period Patch**、**Mcache** 等, 从不同的方面对 **Patching** 算法进行了优化。

3.2.3 自适应的流搭载调度算法

自适应的流搭载调度 (**Adaptive Piggybacking**, **AP**)^[5]算法, 通过改变节目播放速率来让不同时间到达的用户请求共享同一个视频流。假设一个新的用户请求到达系统时, 系统中恰好有一个该节目的组播流, 则系统通过调整使用该组播流的用户的播放速度, 即减缓超前用户的播放速度, 提速滞后用户的播放速度, 最终使这些用户的播放达到同步, 使用同一个组播流所提供传输服务。这种方法弥补了 **Batching** 算法带来用户启动延迟并避免了补丁通道的开销。但用户播放率的改变, 对终端的处理能力提出了严峻的挑战。其次, 为了避免用户感知到服务质量的下降, 播放速度的改变率必须在 $\pm 5\%$ 之内, 降低了可合并为同一个流的用户数量, 从而限制了算法对系统性能的改善程度。

4 结论

流媒体技术作为网络宽带化发展趋势下的一个产物，未来将会得到更多的应用和发展，对流媒体技术的研究也会越来越深入。本文是对流媒体知识的简单介绍，通过这些对这些知识的说明，让大家对流媒体有一个大体上的了解。

参考文献

- [1] 钟玉琢, 向哲, 沈洪. 流媒体和视频服务器[M]. 清华大学出版社, 2003.6.
- [2] 高宗敏. 流媒体技术[J]. 有线电视技术, 2005.9:48-55.
- [3] Viswanathan S, Imielinski T. Pyramid broadcasting for video-on-demand service [A]. Proceedings of the SPIE Multimedia Computing and Networking Conference. 1995.2417:66-77.
- [4] Hua KA, Sheu S. Skyscraper broadcasting: a new broadcasting scheme for metropolitan video-on-demand systems [A]. Proceedings of SIGCOMM 97, 1997.09:89-100.
- [5] Kulkarni S. Bandwidth efficient video-on-demand algorithm (BEVA) [A]. ICT 2003.3:1335-1342.
- [6] Dan A, Sitaram D, Shahabuddin P. Scheduling policies for an on-demand video server with batching [A]. Proceedings of 2nd ACM Multimedia Conference. 1994.10:15-24.
- [7] Hua KA, Cai Y, Sheu S. Patching: a multicast technique for true video-on-demand services [A]. Proceedings of the 6th ACM International Conference on Multimedia, 1998.9:191-200.
- [8] 杨波. 流媒体系统的关键技术研究 [D]. 北京: 北京邮电大学, 2006.

可重构互连网络的发展与现状

曾 韵, 蒋烈辉, 董卫宇

(国家数字交换系统工程技术研究中心, 河南 郑州, 450002)

摘 要: 由于性能、价格、功耗等方面的优势, 可重构互连网络有望在高效能超级计算领域得到广泛的应用。本文分析了高效能计算系统对可重构互连网络的需求, 讨论了重构互连网络的几种方式, 阐述了构建可重构互连网络的三种技术及它们的优缺点, 最后分析了可重构互连网络面临的技术挑战。

关键词: 可重构互连网络; 多级互连网络; 现场可编程门阵列; 光互连

中图分类号: TP303 文献标识码: A 文章编号: 1006-7043 (2004) xx-xxxx-x

The Development and Present of Reconfigurable Interconnection Network

ZENG Yun¹, JIANG Liehui², DONG Weiyu²

(National research center of digital switch system engineering, ZhengZhou 450002, Henan China)

Abstract: Due to the advantages on performance, price and power consumption, reconfigurable interconnection network is promising to gain its wide-spread use in high efficiency computing area. This paper analyzed the need from high efficiency computer systems to reconfigurable interconnection networks, discussed several ways to reconfigure interconnection networks, summarized three related technologies, as well as their pros and cons, about building an interconnection network, and at last, give the technical challenges that the reconfigurable interconnection networks must overcome.

Keywords: reconfigurable interconnection network; multistage interconnection network; field programmable gate array; optical interconnection

1 引言

目前, 对超级计算机的评价标准已经开始从高性能 (High Performance) 向高效能 (High Productivity) 转变, 除了考虑超级计算机的峰值速度之外, 还要更多地考虑其能耗、适用范围和实测性能等方面的指标。可重构计算 (Reconfigurable Computing) 是支持计算机体系结构灵活改变的一种技术, 是实现高效能超级计算机的一种途径, 也是目前高效能计算领域的研究热点和发展趋势。

一般来说, 超级计算机体系结构的可重构包括计算节点的可重构和互连网络的可重构, 本文重点讨论后者。一个可重构的互连网络 (Reconfigurable Interconnection Network, RIN)^[1]是指在不同算法执行时或同一算法的不同执行阶段互连网络的配置可变, 即可以根据应用的不同通信需要调整互连网络的结构和计算节点间的互连关系。

可重构互连网络思想的提出可追溯到 20 世纪 80 年代, 当时随着 VLSI 技术的发展, 使得研制包含多个、甚至上千个处理器的并行处理系统成为可能, 这些并行系统多采用简单、规则的二维网格拓扑的互连网络, 为了克服体系结构固定的缺点, Miller 等设计了带可重构总线的网格网络^[2], 将处理器阵列嵌入到可重构总线系统中, 使处理器间的互连可适应不同算法的需求、在算法执行过程中动态

基金项目: 国家 863 计划新概念高性能计算机体系结构与系统研制重点项目资助, 项目编号 2009AA012200

作者简介: 曾韵 (1976—), 女, 讲师, 硕士;

蒋烈辉 (1967—), 男, 教授, 博士;

董卫宇 (1976—), 男, 讲师, 博士生。

改变。这种可重构网格网络（Reconfigurable Mesh, RMESH）可使用通用处理器进行高性能计算，因此很快流行开，为它开发了很多低复杂度的算法^[3, 4]，并且出现了很多改进的设计，如总线自动装置^[5]、多态环面网络^[6]、带可重构总线系统的处理器阵列^[7]、处理器的可重构阵列^[8]，可重构光总线阵列^[9, 10]等。从 20 世纪 90 年代中期以来，MPP、集群等超级计算机系统逐渐成为并行处理系统的主流体系结构，这些系统关注的主要是节点间的互连网络，并设计、开发了多种具有较高通信性能的网络拓扑结构，如胖树、超立方体、全互连等。但如今随着超级计算机计算规模的急剧扩大，现有的互连网络结构已经不能满足高效、低耗等指标的要求，而必须在有限的网络资源的基础上通过资源的优化配置来满足应用的通信要求。对可重构互连网络的需求主要表现在以下三个方面。

首先，从工程实现角度来讲，固定体系结构的互连网络已不能满足大规模超级计算机的需求。传统上，超级计算机的互连网络可分为高度互连和低度互连两类^[11, 12]。高度互连网络（如全互连或多级胖树拓扑网络）的性能较好，但布线复杂且可扩展性不好，随着超级计算机规模的扩展，链路数是超线性增加，网络的造价也超线性地增加，因此不适合大规模的计算机系统。低度互连网络（如 Mesh 和 3D Torus 网络）的可扩展性好，但性能不高，不能很好地满足很多具有高度并行性的科学应用的需求。因此，需要在高度互连网络和低度互连网络之间进行折中，一种较好的方案就是在低度互连网络的基础上利用节点间互连的重构来构造连接度适中的互连网络^[13]。

其次，对于不同的应用，最有效率的互连网络拓扑不尽相同。例如，快速排序和向量求和算法在树形拓扑上执行得很好，而 Jacobi 迭代算法可在环网上得到很好的性能^[14]。对并行应用进行有效率建模的最好拓扑应是与机器无关而和应用相关的，具有固定拓扑的互连网络显然无法很好地满足各种应用的需要，若能根据不同应用的算法重构互连网络，应用将能获得更好的性能。

最后，对诸如 Cactus、GTC、LBCFD 等科学应用问题的研究显示，一个应用的多个进程间具有不同的通信关系，而且进程间的通信关系通常只存在于较临近的少数进程间^[11]，因此，使用较少的互连网络资源，辅以可重构的互连网络体系结构，就有可能满足多种应用的需要。

以上分析表明，受实现复杂度、成本、应用需求等因素的影响，在超级计算机的互连网络中采用可重构技术实现通信资源的优化配置是有实际需求的。

本文的后续内容组织如下：第 2 节介绍了可重构互连的三种主要方式；第 3 节详细阐述了三种主要的可重构互连技术；最后分析了可重构互连网络面临的主要技术挑战。

2 可重构互连方式

根据 RIN 重构表现形式的不同，主要可分为以下三类^[15]。

2.1 拓扑内可重构

拓扑内可重构（Intra-Topology Reconfiguration），是指在互连网络拓扑结构不变的情况下对其进行不同的配置。例如，星型拓扑的网络可使用不同的处理器作为中心节点，网格拓扑的网络可改变每维的节点个数等。

采用拓扑内可重构技术可以提高进程间的通信效率和体系结构的容错性。如 Srinivas 等人提出的二元可重构树结构^[16]，可根据应用的通信需求，动态改变配置，采用多种不同的树结构，实现拓扑内可重构。通过可重构，可以使两个进行数据交换的节点相邻，从而降低通信延迟；还可以使出错的节点处于叶的位置，从而使系统保持连通性，在适度降级的模式中继续运行。

2.2 拓扑间可重构

拓扑间可重构（Inter-Topology Reconfiguration），是指互连网络的拓扑可变，根据算法的结构建立对应的网络结构，因此可获得更高的计算效率。

对不同的应用，最有效率的网络拓扑不尽相同。例如，快速排序算法和向量求和算法在树型拓扑上执行得很好，而 Jacobi 迭代则在环网上性能不错。程序在不合适的网络上运行，性能显然要受到影响。若能按需为应用动态提供合适的互连网络，需要构建可动态改变物理拓扑的网络。如 Wu 等人提出的 Star 系统^[17]，可以根据应用的需求将通信网络 Starnet 的星型拓扑改为二元树拓扑，并且可以根据需求将任意处理器设置为根。

2.3 分区

分区 (Partitioning)，通过分区，系统可作为很多独立的配置工作，因此可提高资源的利用率。将一个基本拓扑分成多个相同类型的互连拓扑，如 Howard J. Siegel 的 Pasm (Partitionable SIMD/MIMD System)^[18]，因此单个拓扑可作为多个专用体系结构，可有效支持特定算法的通信模式。另外，还有一些可重构互连网络的重构主要用于解决网络的容错问题，如文献[19]~[22]。

3 可重构互连技术

可重构互连网络中通常会使用开关或可编程逻辑器件，网络的配置改变是通过改变开关或可编程逻辑器件中互连资源的连接状态来实现的。

按照实现互连网络可重构主要依赖的技术，可分为以下三种：一是使用多级互连网络 (Multistage Interconnection Network，MIN)；二是利用现场可编程门阵列 (Field Programmable Gate Array，FPGA) 的互连资源；三是利用光互连技术。这几种技术各有优缺点及适用范围。

3.1 基于 MIN 的可重构互连网络

MIN^[23]通常用于在多级系统或通信网络中提供处理器之间的动态连接。 n 级 MIN($n=\log N$)通常有 N 个输入端、 N 个输出端，输入/输出之间通过 n 级交换子网络相连；每级子网由 $N/2$ 个 2×2 的开关组成，每个 2×2 开关可以置成“平行”和“交叉”两种状态，对应于输入和输出之间的两种不同连接方式；相邻各级开关之间都有固定的级间连接 (InterStage Connection，ISC)。各种多级网络的区别在于所用开关模块、ISC 模式和控制方式的不同。

动态设置开关的连接状态就可以在输入/输出之间建立所需的连接。例如，Star 系统^[17]的可重构通信网络 Starnet 是在基准网络 (一种全互连的 MIN) 的基础上，通过设置网络的交换开关使 Starnet 获得可重构性，根据应用的需求可以将通信网络 Starnet 的星型拓扑改为二元树拓扑，并且可以根据需求将任意处理器设置为根。又如，文献[24]中提出一种 m 元可重构树结构的通用体系结构设计 (其中 m 可以是大于 1 的任意整数)，使用 $\log_m N$ 级交换网络互连，每一级由 N/m 个交换单元构成，每个交换单元使用一个 $\lceil \log_2 m \rceil$ 级 MIN 实现，其中每一级又包含 $m/2$ 个 2×2 或 2×4 的交换开关，通过控制各级开关的交换状态即可在处理器节点间建立不同配置的树型网络，总共可有 $m\times 2^{\lceil \log_2 m \rceil (\log_m N - 1)}$ 种配置。这种体系结构不需要在节点中安装单独的硬件来获得可重构性，而且和自路由 MIN 网络相比，实质需要的硬件更少。由于不需要有配置节点的算法、不需要冲突解决机制，该体系结构有简单路由、低同步开销和快速可重构的优点。

另外，还可以利用 MIN 的拓扑等价特性实现另一种可重构：使一种网络可以实现其他各种网络的互连功能。互连网络拓扑等价的含义是指：对某个网络，通过某种规则或方法将其网络元件的位置名重新安排，变换成相应的逻辑名结构，使之带有逻辑名的网络元件的互连特性可以用另一个网络的数学拓扑规则描述^[25]。文献[26] 证明了简化数据变换网络、flip 网络、 Ω 网络、间接二进制 n 方体网络和规则 SW 榕树网络 ($F=S=2$) 是拓扑等价的，并根据这些网络的数学拓扑规则寻找到一套一到一的映射规则，以此来确定相互拓扑等价的网络的逻辑名结构。图 1 所示是 Ω 网络与基准网络拓扑等价的逻辑结构。

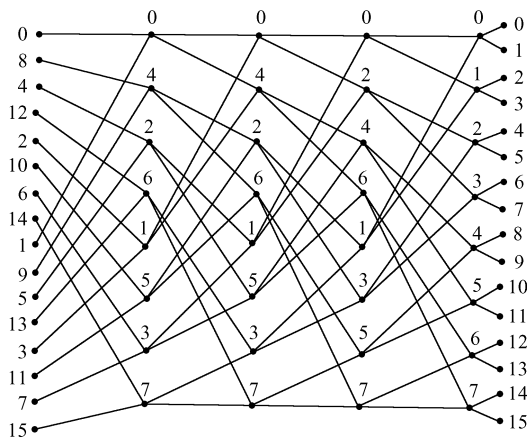


图 1 Ω 网络与基准网络拓扑等价的逻辑结构

文献[27]进一步推导出多种 MINs 之间的功能关系。用 B_N 、 B_N^{-1} 、 Ω_N 、 Ω_N^{-1} 、 C_N 、 C_N^{-1} 、 D_N 、 D_N^{-1} 、 S_N 、 S_N^{-1} 、 G_N 、 G_N^{-1} 分别代表基准网络、逆基准网络、 Ω 网络、逆 Ω 网络、间接二进制 n 方体网络、逆间接二进制 n 方体网络、简化数据变换网络、逆简化数据变换网络、STARAN 网络、逆 STARAN 网络、规则 SW 榕树网络 ($F=S=2$) 和逆规则 SW 榕树网络, $\rho(x_m x_{m-1} \cdots x_0) = x_0 x_1 \cdots x_m$, $m \geq 1$ 来表示位序颠倒置换, $\delta(x_m x_{m-1} \cdots x_3 x_2 x_1 x_0) = x_1 x_2 \cdots x_{m-2} x_{m-1} x_m x_0$, $m \geq 1$ 来表示位交换置换。利用这些定义和网络输入/输出连线逻辑名的二叉树编码法, 可导出各种多级网络的功能关系, 如表 1 所示。

表 1 各种多级网络与基准网络 B_N 的功能关系 ($P=f(\theta)$)

Network Function(P)	$f(\theta)$				
	$\theta=B_N$	$\theta=\Omega_N$	$\theta=C_N$	$\theta=D_N$	$\theta=G_N$
Baseline Network(B_N)	B_N	$\Omega_N op$	$\rho o C_N$	$D_N o \delta$	$\delta o G_N$
Reverse Baseline (B_N^{-1}) Network	B'_N	$\Omega_N op$	$\rho o C_N$	$D_N o \delta$	$\delta o G_N$
Omega Network (Ω_N)	$B_N op$	Ω_N	$\rho o C_N op$	$D_N o \delta op$	$\delta o G_N op$
Reverse Omega (Ω_N^{-1}) Network	$\rho o B_N$	$\rho o \Omega_N op$	C_N	$\rho o D_N o \delta$	$\rho o \delta o G_N$
Indirect Binary n-Cube (C_N) Network	$\rho o B_N$	$\rho o \Omega_N op$	C_N	$op D_N o \delta$	$\rho o \delta o G_N$
Reverse Indirect Binary (C_N^{-1}) n-Cube Network	$B_N op$	Ω_N	$\rho o C_N op$	$D_N o \delta op$	$\delta o G_N op$
Modified Data Manipulator (D_N)	$B_N o \delta$	$\Omega_N op o \delta$	$\rho o C_N o \delta$	D_N	$\delta o G_N o \delta$
Reverse Modified Data Manipulator (D_N^{-1})	$\delta o B_N$	$\delta o \Omega_N op$	$\delta op o C_N$	$\delta o D_N o \delta$	G_N
Flip Network (F_N)	$\rho o B_N$	$\rho o \Omega_N op$	C_N	$op D_N o \delta$	$\delta op o G_N$
Reverse Flip Network (F_N^{-1})	$B_N op$	Ω_N	$\rho o C_N op$	$D_N o \delta op$	$\delta o G_N op$

Network Function(P)	f(θ)				
	θ=B _N	θ=Ω _N	θ=C _N	θ=D _N	θ=G _N
Regular SW Banyan (G _N) Network (S=F=2)	ρoB _N	δoΩ _N oρ	δoρoC _N	δoD _N oδ	G _N
Reverse Regular SW Banyan (C _N ⁻¹) Network (S=F=2)	B _N oδ	Ω _N oρoδ	ρoC _N oδ	D _N	δoG _N oδ

从表 1 所列的功能关系可以看出，如果将互连网络的输入/输出连线重新命名，就有可能使这个网络实现其他各种网络的功能。为实现这种网络重构，需要网络的输入端与输出端有实现 ρ 和 δ 变换的功能部件。如基于 FPID（Field Programmable Interconnection Device，现场可编程互连设备）的 FDRMIN（FPID-based Dynmaic Reconfigurable MIN，动态可重构多级互连网络）^[28]，通过 FPID 实现这种由物理名到逻辑名的映射，将开关模块相同的多种多级互连网络的功能合并，根据应用算法的需求重构 MIN。但是由于 FPID 的配置时间相对较长，FDRMIN 不能实现动态实时重构。

基于 MIN 的可重构互连网络需要使用大量的交换开关，节点数为 N 的网络需要的开关数为 $O(N\log N)$ ，因此只适合构建规模不太大的计算机系统，而对节点数上万的巨大规模的超级计算机系统则不适用。

3.2 使用 FPGA 的可重构互连网络

FPGA 一般由三种可编程电路和一个用于存放编程数据的静态存储器 SRAM 组成。这三种可编程电路是：可编程逻辑块（Configurable Logic Block，CLB）、输入/输出块（Input and Output Block，IOB）和互连资源（Interconnect Resource，IR）。CLB 是实现逻辑功能的基本单元，通常包含组合逻辑部分和时序逻辑部分，组合逻辑一般包括查询表（Look-up T able）和相关的多路选择器（Multiplexer）；而时序逻辑部分包含触发器（DFF）和一些相关的多路选择器。IOB 主要完成芯片上的逻辑与外部封装引脚的接口，提供了 FPGA 内部和外部的一个接口。IR 则提供 CLB 与 IOB 之间及各 CLB 之间的通信，包括各种长度的连线线段和一些可编程连接开关。

通过重构可编程逻辑器件的互连资源，可建立不同配置的网络。经典对称式 FPGA 互连资源包括开关盒（Switch Box，SB）、连接盒（Connection Box，CB）及互连线段^[29]，CB 将 CLB 或 IOB 的引脚连接到互连线段，SB 则连接不同的互连线段，如图 2 所示。

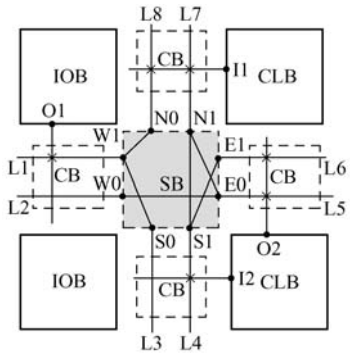


图 2 经典 FPGA 互连结构示意图

条边代表节点间一种可能的连接关系。

从功能上看，FPGA 的互连资源与互连开关十分类似，都可以实现输入和输出之间的任意连接。不同之处在于实现开关阵列的方式不同，互连开关中的交叉点由开关电路实现，由控制信号控制其开关状态，变换的速度快。FPGA 的开关状态受 SRAM 单元控制，变换速度慢，优点在于可以大规模集成，降低成本。

FPGA 可以用于构建片上网络。如 FLUX^[30]是使用 FPGA 实现的一种可重构互连网络，根据算法的实际需求，在 FPGA 的可重构区域动态建立 PEs（Processing Elements，处理单元）间不同拓扑的网络，从而获得比固定互连网络更好的性能。

由于 FPGA 构建网络使用的是不同的物理链路，因此这种网络的灵活性受到片上可用互连资源的限制，一种拓扑中可连接的 PEs 数同样受可用路由资源的限制。若想连接更多的处理器，可使用多片 FPGAs 进行多级互连。如 Christophe 等^[11]使用多片 FPGAs 构建的并行计算机系统，连接到一个 FPGA 芯片的处理器组称为一个集群，集群间通过称为“桥”的 FPGA 进行互连，系统中所有 FPGA 的配置决定了并行计算机的拓扑，其物理拓扑可根据特定应用的并行计算范例模型使用的虚拟拓扑而修改。

如今，FPGA 的动态、部分可重构特性将使网络的重构更灵活，如 DRAFT（Dynamic Reconfiguration Adapted Fat-Tree，动态可重构适应胖树）^[31]是利用 Xilinx Virtex5 的动态、部分可重构范例构建了一种灵活的互连网络，和传统胖树相比，DRAFT 需要更少的资源和更少的通信链路，具有更低的平均延迟和更高的传输与注入速率（可达 1000Mbps）。

由于 FPGA 的路由体系结构提供了一个基础的“未使用的”可重构网络，使用 FPGA 还可提供点到点的、无序连接的网络。这意味着可能不需要网络结构本身，如果有可用的互连资源（未使用的路由资源），则可以抛开任何固定的网络拓扑，按照应用的需求，使用可用的连接资源在需要通信的 PEs 间建立直接的互连。这种方法的主要问题是复杂的路由算法和时机的选择。而且由于不能预先知道每条连接的电路长度，还需要有准确的机制来保证 PEs 间正确的通信。

FPGA 也可用于多芯片间互连，但由于 I/O 引脚数的限制，片外通信带宽也将受限，通信性能不能和片上网络的性能相比。

3.3 可重构光互连网络

光互连是以光子作为信息载体、实现处理器间信息交换的连接方式^[32]。光互连技术基本上可以分为光波导（包括光纤）互连和自由空间光互连两类。

导波光互连是以折射率引导传播的，采用光纤或集成光学波导作光束传输介质，光束的传输方向由传输介质控制。光纤互连不仅具有光传输的高速率、高带宽及双向传输和多路复用等特性，而且在机械连接方面具有导线连接的方便、灵活、简单、可靠和结构简单的特性，还易于与光波导器件相结合，因此能构建高速的光交换系统。人们已开始研究使用光纤互连技术来重构互连网络，如采用 MEMS 光开关构建的基于 MIN 的混合交换体系结构 HFAST（Hybrid Flexibly Adaptable Switch Topology，混合柔性可修改交换拓扑）^[33]；使用基于 SOI（silicon-on-insulator，绝缘体上硅）微环谐振器的行-列交换阵列实现动态带宽重分配的 nD-RAPID（n dimensional-reconfigurable, all-photonic interconnect for distributed and parallel systems，用于分布式并行系统的多维可重构全光互连结构）^[34]，其中 n 可以是 1、2 或 3。

自由空间光互连技术是以空气为媒质进行信息传输的，通过光学器件转折和控制空间传输的光束进行互连的。自由空间光互连能使密集的、可重构的光信号传输和分布同时实现，灵活

地组成各种互连拓扑结构，是很有前途的一种可重构互连技术。例如，Artundo 等使用 VCSELs（Vertical Cavity Surface Emitting Lasers，垂直腔面激光发射器）、SOB（Selective Optical Broadcasting，选择光广播）部件和光接收组件设计了一种用于分布式共享内存多处理器环境中的可重构光互连网络^[35, 36]，采用自由空间光互连技术在一些需要大量通信的处理器对间临时建立可重构的光链路；Aljada 等使用光电集成电路 Opto-VLSI 处理器实现了速率为 2.5Gbps 的可重构光互连体系结构^[37]，可实现点到点或点到多点的自由空间光互连；Shen 等使用 Opto-VLSI 处理器和 4-f 映像系统实现了一种新颖的可重构光互连体系结构^[38]，可实现芯片间或板间的自由空间光互连。

光互连具有带宽高、传输速度快、抗电磁干扰能力强、互连功耗小等优点，不仅能克服电互连的带宽有限和电磁干扰瓶颈，还能提供大量的互连通道、综合和可重构性^[39~41]，是为新一代超级计算机构建可重构互连网络的理想技术。

4 总结

在超级计算机中，根据应用需求，采用某种可重构互连技术来调节互连网络的拓扑、带宽等要素，从而更好地提高资源利用率、降低处理器间通信延时、减少能耗，还能提供容错支持（在有错的情况下恢复系统的连接）。

但是任一种可重构方案都会引入两种开销：重构硬件和重构时间。重构硬件是完成重构所需要的额外的硬件，重构时间是在配置间转换时的延迟。因此设计可重构网络时需要考虑的一个主要事项是如何最小化可重构引入的这两种开销。

而要实现互连网络物理拓扑的高效动态可变，还面临着一系列的技术挑战，主要包括：（1）网络拓扑变换有一定的时延，在程序运行过程中，能否确定进行拓扑变换的合适时刻；（2）网络拓扑变换过程中，如何处理处理器间的数据交换；（3）对拓扑可变的网络，如何统一管理和进行作业分配；（4）编译器能否为每种算法确定网络拓扑、进程间的通信关系、进程的分配，如何实现；（5）对拓扑可变的网络，如何实现底层通信协议（如同步、广播、多播等）的优化。

只有解决了这些软、硬件上的挑战，才能为超级计算机系统研制出真正实用的可重构互连网络。

参考文献

[1] Arabnia H.R and Smith J.W.. A Reconfigurable Interconnection Network for Imaging Operations and Its Implementation Using a Multi-Stage Switching Box[J]. Proc. 1993 Conf. High Performance Computing: New Horizons, 1993, pp. 349-357.

[2] Miller R., Kumar V. K. P., Reis D. I., et al.. Meshes with Reconfigurable Buses[J]. Proc. 5th MIT Conference On Advanced Research in VLSI, 1988, pp. 163-178.

[3] Miller R., Kumar V. K. P., Reis D. I., et al.. Parallel computations on reconfigurable meshes[J]. IEEE Trans. Comput., 1993, vol. 42, pp.678-692.

[4] Olariu S, Schwing JL, Shen W, et al.. A simple selection algorithm for reconfigurable meshes[J]. Parallel Algorithms Appl., 1993, vol. 1, 29-41.

[5] J. Rothstein. Bus automata, brains, and mental models[J]. IEEE Trans.Syst., Man, Cybernetics, 1988, vol. 18, pp. 522-531.

[6] Li H and Maresca M. Polymorphic Torus Network[J]. IEEE Trans.Comput, 1989 vol. C-38, pp. 1345-1351.

[7] Biiing-Feng W, Gen-Huey C, Ferng L. Constant Time Sorting on a Processor Array with a Reconfigurable Bus System[J]. Information Processing Letters, 1990, 34, pp. 187-192.

[8] Kao TW, Horng SJ and Tsai HR. Designing efficient parallel algorithms on a CRAP[J]. IEEE Trans. Parallel Distrib. Systems, 1995, 6, pp. 554-559.

- [9] Pavel S and Akl SG. Matrix operations using arrays with reconfigurable optical buses[J]. *Parallel Algorithms Appl*, 1996, 8, pp. 223-242.
- [10] Rajasekaran S, Sahni S. Sorting and routing in arrays with reconfigurable optical buses[J]. In *Proc. Int'l. Conf. Par. Processing*, 1996, pp. III-90_III-94.
- [11] Kamil S, Shalf J, Oliner L, et al.. Understanding UltraScale Application Communication Requirements[J]. *IEEE International Symposium on Workload Characterization(IISWC)* Austin Texas, 2005, October 6-8.
- [12] Bodba C, Danne K. A New Approach for Reconfigurable Massively Parallel Computers[J]. In *Proceedings of the IEEE International Conference on Field-Programmable Technology*, 2003, 15-17, pp. 391-394.
- [13] Kamil S, Oliner L, Pinar A, et al.. Communication Requirements and Interconnect Optimization for High-End Scientific Applications[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2009, Vol.22, 188-202.
- [14] Brent R. P. and Luk F. T.. The solution of singular-value and symmetric eigenvalue problems on multiprocessor arrays[J]. *SIAM J. Sci. Stat. Comput.*, 1985, 6 (1) :69-84.
- [15] S. Srinivas. Dynamically reconfigurable architectures for supercomputing systems[C]. Ph.D. dissertation, Indian Institute of Science, Dec. 1988.
- [16] Biswas N N, Srinivas S. A reconfigurable tree structure with multistage interconnection network[J]. *IEEE Trans. Comput.*, 1990, vol.39, pp.1481-1485.
- [17] Wu C. L., Feng T. Y., and Lin M. C.. Star: A local network system for real-time management of imagery data[J]. *IEEE Trans. Computer*, 1982, vol. C-31, pp.923-933.
- [18] Siegel H. J., Siegel L. J., Kemmerer F. C., et al. PASM: A partitionable SIMD/MIMD system for image processing and pattern recognition[J]. *IEEE Trans. Comput.*, 1981, vol. C-30, pp. 934-947.
- [19] Negrini R., Sami M., Stefanelli R.. Fault-tolerance techniques for array structures used in supercomputing[J]. *IEEE Comput. Mag.*, 1986, pp.78-87.
- [20] Dutt S. and Hayes J. P.. On designing and reconfiguring k-tolerant tree architectures[J]. *IEEE Trans. Comput.*, 1990, vol. C-39, pp.490-503.
- [21] Singh A. D. and Youn H. Y.. A modular fault-tolerant binary tree architecture with short links[J]. *IEEE Trans. Comput.*, 1991, vol. 40, pp.882-890.
- [22] Belkhale K. P. and Banerjee P.. Reconfiguration strategies for VLSI processor arrays and trees using a modified Diogenes approach[J]. *IEEE Trans. Comput.*, 1992, vol. 41, pp.83-96.
- [23] 王鼎兴, 陈国良. 互连网络结构分析[M]. 北京: 科学出版社, 1990.
- [24] Srinivas S, Biswas N N. Design and Analysis of Generalized Architecture for Reconfigurable m-ary Tree Structures[J]. *IEEE Trans. Comput.*, 1992, vol.41, pp.1465-1478.
- [25] 艾军, 曹明翠, 李再光. 互连网络拓扑等价的图分析法[J]. *计算机研究与发展*, 1994, Vol.31, No.3, pp.29-33.
- [26] Wu C. L., Feng T. Y.. On a class of multistage interconnection networks[J]. *IEEE Transactions on Computers*, 1980, Vol. C-29, pp. 694-702.
- [27] Wu C, Feng T. The Reverse-Exchange Interconnection Network[J]. *IEEE Trans Computers*, 29(9):801-811. 1980.
- [28] 佟冬, 胡铭曾. 基于 FPID 的动态可重构的多级互连网络. *哈尔滨工业大学学报*, 33 (1) . 2001.
- [29] 代莉, 梁绍池, 王伶俐. 基于布线资源图的 FPGA 互连测试算法. *计算机工程*, Vol.35, No.14, pp. 258-260, 2009.
- [30] Vassiliadis S. and Sourdis I.. Reconfigurable fabric interconnects. In *Intl. Symposium on Soc*, pp. 41-44. Nov. 2006.
- [31] Devaux L., Sassi S. B., Pillement S., et al.. Flexible Interconnection Network for Dynamically and Partially Reconfigurable Architectures[J]. *International Journal of Reconfigurable Computing*, 2010, Vol. 2010, Article ID 390545.
- [32] 张以谟. 光互连网络技术[M]. 电子工业出版社, 2006.
- [33] Kamil S, Pinar A, Gunter D, et al.. Reconfigurable Hybrid Interconnection for Static and Dynamic Scientific Applications

tions[J]. In Proceedings of the ACM International Conference on Computing Frontiers, 2007.

- [34] Kodi AK and Louri A. Multidimensional and reconfigurable optical interconnects for high-performance computing (HPC) systems[J]. J. Lightwave Technol, 2009. 27(21), 4634-4641.
- [35] Artundo I, Desmet L, Heirman W, et al. Selective Optical Broadcast Component for Reconfigurable Multiprocessor Interconnects[J]. IEEE Journal on Selected Topics in Quantum Electronics:Special Issue on Optical Communication, 2006, 12 (4) :828-837.
- [36] Artundo I, Heirman W, Debaes C, et al. Design of a reconfigurable optical interconnect for large-scale multiprocessor networks[J]. Proc. of SPIE Photonics Europe., 2008, Vol. 6996. pp. 69961H.
- [37] Aljada M., Alameh K. E., Lee Y. T., et al. High-speed (2.5 Gbps) reconfigurable inter-chip optical interconnects using opto-VLSI processors[J], Opt. Express, 2006, 14(15), pp. 6823-6836.
- [38] M. Shen, F. Xiao, and K. Alameh. A novel reconfigurable optical interconnect architecture using an Opto-VLSI processor and a 4-f imaging system[J]. Optics Express, 2009, Vol. 17, Issue 25, pp. 22680-22688.
- [39] Lytel R., Davidson H. L., Nettleton N., et al.. Optical interconnections within modern high-performance computing systems[J]. Proc. IEEE, 2000, 88(6), 758-763.
- [40] Kash J. A.. Leveraging optical interconnects in future supercomputers and servers[J]. Proceedings of the 16th IEEE Symposium on High Performance Interconnects, 2008, 190-194.
- [41] Liboiron-Ladouceur O., Wang H., Garg A. S., et al.. Low-power, transparent optical network interface for high bandwidth off-chip interconnects[J]. Opt. Express, 2009, 17(8), 6550-6561.

类比推理的计算模型研究综述

周 倩, 沈夏炯

(河南大学计算机与信息工程学院, 河南 开封, 475004)

摘 要: 类比推理是指从两个(类)对象的相似性和一个(类)对象的已知特征推出另一个(类)对象也具有这个特征的过程。类比推理提供了一种新的问题求解机制和机器学习方法, 近年来受到了人工智能和认知科学研究者的广泛关注。本文论述了几种目前比较主要的类比推理计算模型, 并对这几种模型进行比较分析, 提出未来的发展趋势。

关键词: 类比; 类比推理; 计算模型

中图法分类号: TP181 **文献标识码:** A

Computational Model of Analogical Reasoning Summary

ZHOU Qian¹, SHEN Xiajiong²

(Henan University Computer and Information Engineering, Kaifeng 475004, Henan China)

Abstract: Analogical reasoning is generally defined as a process that individual infers the characteristics of one object based on the known characteristics of another and their similarity. Analogical reasoning provides a new mechanism for problem solving and machine learning methods, in recent years, it has received extensive attention of the artificial intelligence and cognitive science researchers. In this paper, we discussed about several current major analogical reasoning calculation models and analysed the comparison of these models, besides, we proposed the development trend of the future.

Keywords: analogy; analogical reasoning; computational models

1 引言

类比推理是人类核心的认知能力之一^[1,2]。类比推理的基础是事物、状态、关系之间的相似性。类比推理与学习的研究可以分为两大类, 一类是问题求解型类比^[3], 它的主要思想是当求解新问题时, 总是首先回忆一下以前是否解过类似问题, 若有则以此为根据, 来解决新问题。另一类是预测推定法^[4], 它又可分为两种形式, 一是传统法, 如已知 A, B 具有相似特征 a、b、c, 若 A 有特征 c, 则推出 B 也具有特征 c; 二是因果关系型, 如已知 $A \rightarrow B$, 给定 A 与 A1 相似, B 与 B1 相似, 则可能得出结论 $A1 \rightarrow B1$ 。

在 20 世纪 80 年代, 由于计算机性能的大幅度提高, 传统推理技术趋于成熟, 特别是计算机广泛应用对机器学习和知识获取提出的迫切需求, 使得人们对类比推理产生了越来越大的兴趣, 并出现了一批引人瞩目的研究成果, 如 Winston 的类比学习系统采用以对象为中心的思想, 并强调因果关系在类比推理中的重要作用^[3]。Garbonell 提出了类比转换方法, 将已知问题的解法转换成类似问题的解法^[4]。Gentner 提出了结构映射理论^[1,2,5,6]。Holyoak&Thagard 等人提出的限制满足理论^[1,7,8]。Keane 提出的累进理论^[1,7,9]。这些系统和理论各有特色, 其中以结构映射理论, 受限制满足理论, 累进理论最具有完整性。以下分别介绍和比较这三派理论, 以使我们对类比推理能有进一步的理解。

作者简介: 周倩 (1987—), 女, 河南唐河人, 硕士, 主要研究方向: 类比学习、知识发现 (zhouqian29@yahoo.com.cn);
沈夏炯 (1963—), 男, 河南开封人, 教授, 主要研究方向: 软件工程、知识发现、分布式/并行处理、分布式存储、数据挖掘。

2 类比推理理论

人们使用类比推理的历程包含从长期记忆中检索合适的类比物——检索阶段，将两个领域相似部分配对——映射阶段，将对应的结果应用到学习中——学习阶段。在整个类比过程中，映射是核心的过程。下面将详细介绍三种映射理论即结构映射理论、限制满足理论和累进理论。

2.1 以 Gentner 为主结构映射理论

Gentner 的结构映射理论 SMT (Structure Mapping Theory) 将类比论域的内容分为三类关系（高阶和低阶）、属性和对象^[1,2,5]。对象被用来表达单个的实体。属性是实体的一值论证。关系是对象之间的关系。谓词包含关系和属性。例如：我们可以将原子领域和太阳系领域的知识表征为图 1 和图 2。

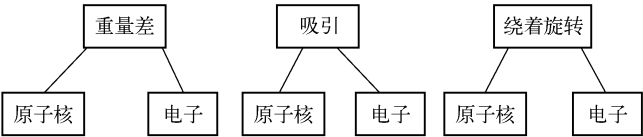


图 1 原子领域具有的知识表征

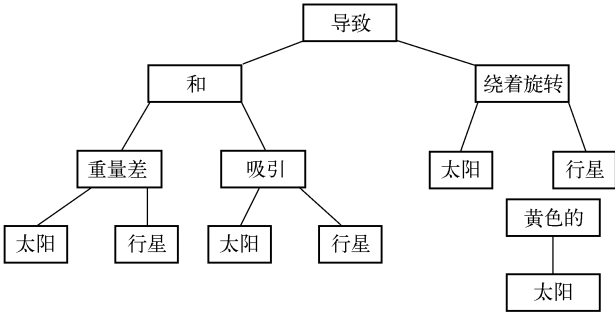


图 2 太阳系领域所具有的知识表征

在这个领域的知识表征中，太阳、行星就是对象或实体，黄色的就是属性，吸引、质量大就是低阶关系，导致就是高阶关系。在类比中源领域即类比过程中我们已知的知识领域，目标领域或靶领域是类比过程中我们要学习的知识领域。

结构映射理论中，源、目标是概念和关系的集合，其元素之间具有一定的关系。源和目标对应时要舍弃对象的属性，但这种属性与高阶关系相关时，不舍弃^[6]。关系与关系相对应，对象与对象相对应，不同关系要有高级限制关系联结，如因果关系等^[10,11]。

以 SMT 理论为基础，提出的类比对对应工具结构映射引擎 SME (Structure Mapping Engine)。SME 的类比对对应过程如下^[12,13,14,15]：

- (1) 找目标领域和源领域间所有可能的局部配对。
- (2) 合并局部配对，形成整体配对，整体配对可能是多个。
- (3) 评价整体配对，评价整体配对是根据对应关系的匹配，匹配越往上，结构评量分数 (Structural Evaluation Scores, SES) 值越大，选择 SES 最大的配对为最佳配对。

我们可以用太阳系与原子的类比来说明 SMT 工作过程。太阳系与原子之间可能存在的对象对应关系为太阳=原子核、行星=电子、太阳=电子、行星=原子核。关系的对应关系为 {重量差(太阳, 行星)} = {重量差(原子核, 电子)}，{重量差(太阳, 行星)} = {吸引(原子核, 电子)}等 (步骤一)。将这些配对合并，合并可能产生多个整体配对有：(1) 导致 {和 [重量差(电子, 原子核), 吸引(电子, 原子

核}], 绕着旋转(原子核, 电子)}。(2) 导致{和[吸引(原子核, 电子), 比…热(原子核, 电子)], 绕着旋转(电子, 原子核)}。(3) 导致{和[重量差(原子核, 电子), 吸引(原子核, 电子)], 绕着旋转(电子, 原子核)} (步骤二)。然后对整体配对进行评价, 例如, 在 (1) 中对象的对应为: 太阳=电子、行星=原子核, 它们不存在对应关系, 所以 SES 值很小, 假设为 1。在 (2) 中对象的对应为太阳=原子核、行星=电子, {重量差(太阳, 行星)}={ 比…热(原子核, 电子)}对应中, 重量差、吸引作用不相同并且缺乏系统性, SES 值就次小, 假设为 2。在 (3) 中, 太阳=原子核、行星=电子, {重量差(太阳, 行星)}={ 重量差(原子核, 电子)}等对应中, 存在更高阶的关系, 所以 SES 值更大些, 假设为 3。第 3 个整体配对的 SES 值最高, 所以它所对应的整体配对就是最优配对 (步骤三) (注释: 上文中的 “= “是映射的意思)。

显然 SME 是纯结构性的, 与内容无关。然而学术界的争论和心理证据表明, 语义性的内容又常常是重要的。因此, Gentner 在长时记忆中寻找类比物 (检索) 时加入了语义相似性的因素, 而 SME 机制并未发生改变。Gentner 的纯结构性方法加上语义相似性保持了她在这一领域中奠基人的地位。但这个理论存在很大的局限, SME 找最佳匹配时, 要对所有的匹配都进行计算, 计算量比较大。

2.2 以 Holyoak&Thagard 为主的限制满足理论

平行限制满足理论^[7]是以节点或单位来表征字母或单字, 两节点的连线代表彼此间的互动, 实线表示正向的支持—激发, 虚线表示反正—抑制。当我们知觉到字母或单字时, 节点就开始活化, 并蔓延到相关节点, 这样就构成了一个活化网络。

限制满足理论源于平行限制满足理论 (Parallel constraint satisfaction), 并加入了结构和语义相似性, 并依次开发出了对应 ACME (Analogical Constraint Mapping Engine) 映射工具^[7]。

同构性: 两者具有结构并且一一对应。

实用性因素: 例如, 如果问河南与湖南有哪些可比较的, 此时需考虑有关两地方各个方面的情況。如果问题是经济的, 则考虑的对象缩减为经济这个知识的一个子集, 考虑的知识集合都与特定的目的有关, 这就是实用性因素。

语义相似性: 对象的语义相似可以从谓词的相似性得来, 如果两个谓词之间存在相似, 则其对应的对象也是相似的。谓词的相似性可用关系的相似性, 属性共享 (即有相同的属性)。

在建立 ACME 的匹配映射网时, 使用了两条原则^[1]: 第一, 类型原则规定只有相同类型的成分可以形成匹配, 即命题与命题, N 元谓词与 N 元谓词, 对象与对象; 第二, 分部分原则当类比对象分为几个主要部分时, 只有相应部分之间的内容可以按第一条原则形成匹配。这两条原则实际表达了结构相似性的原则。

ACME 算法类似于平行限制满足理论, 将对应过程分为两步^[11,8]:

(1) 建立对应网络。建立两个领域命题间对应的节点。网络中的节点就是源、目标两个范围上对象 (谓词, 命题) 的合理匹配, 网络中除了节点外还包含两个特殊的单元, 语义单元和实用单元。以同构性、语义相似性和实用性三个限制, 评量两节点的活化程度, 建立激发或抑制的网络连接。

(2) 执行网络。首先给每一个节点一个活化的初始值, 语义单元和实用性单元值初始为 1, 其他单元设置的值介于 0 与 1 之间。然后开始执行网络, 语义约束的执行是通过谓词间的兴奋连接从语义单元到其他的映射单元。实用性约束的执行是通过谓词间的兴奋连接从实用单元到其他的映射单元。在每次循环中都会重新评量并更换节点的活化程度, 语义和实用性单元不用更新。当某节点活动超过了临界点, 那么这个节点对应的匹配就是两领域的最佳匹配。

ACME 最大的贡献就是在对应过程中加入了语义相似性, 结构性和实用性三种限制。ACME 具有两个优点: 一是它的应用范围相对来说比较广泛; 二是它对结构要求不是那么严谨和强调系统性, 可以处理非完全相同的对应关系。同时 ACME 具有两方面的缺陷: 一是可以成功地对一个命题进行匹配, 但不一定能对其所有的命题都进行成功的匹配; 二是 ACME 只能指出一个最佳匹配, 当结果在

后期验证中由于某种原因被舍弃时，需要重新修正初始映射以重新计算，求得新的结果。

2.3 以 Keane 为主的累进理论

累进理论除了结构，语义相似性，实用性限制外，加入了工作记忆和背景知识的限制。结构限制与 SMT 的结构限制一样，语义相似性、实用性两个限制与限制满足理论的两个限制一样^[7,9]。

工作记忆的限制：工作记忆的限制会使某些信息遗失从而限制类比的运作工程，而累进理论采用序列的加工方式，减少工作记忆的负载，更符合类比的认知心理运作过程。

背景知识的限制：背景知识的限制同样会影响类比的推理。如果对应部分与背景知识相符合，学习者花费的时间比较少；反之亦然。

IAM（Incremental Analogy Machine）是根据累进理论发展出来的对应工具^[1,7,9]，IAM 将类比分 6 步：

（1）选择种子群：为源领域的述次群按照实用性方面，高阶连接，谓词的数量进行排序，并以等级最高的群当做种子群。

（2）找出种子配对：在种子群中按照实用性进行排序，选等级最高的作为种子成分，种子成分与目标领域产生配对。

（3）找出种子群中其他的配对：采用实用的、语义相似性的和结构的限制来找出种子群中其他与目标领域一一同构的配对。

（4）迁移：对于源领域尚未配对的关系，我们迁移到目标领域。

（5）评估组映射：如果种子群中谓词的匹配数量超过半数，则匹配成功。如果正确，就进行步骤（6），否则就尝试另一种子配对（步骤 2），如果尝试所有配对都不是最优，则另选一个种子群。

（6）如果还需要其他的类比群对应，则重复步骤（1）～（5）。

我们可以用太阳系与原子的类比来说明 IAM 工作过程。如图 2 所示 IAM 将太阳系领域分为三群，在这三群中，第一群：导致{和[重量差(太阳，行星)，吸引(太阳，行星)]，绕着旋转(行星，太阳)}，因此群具有多物件的关系结构，所以被选为种子群（步骤 1），重量差(太阳，行星)被视为种子成分。接着，从种子成分中可以产生下列配对：重量差=重量差，太阳=原子核，行星=电子（步骤 2），除此之外，它还有具有两个同构配对：绕着旋转(行星，太阳)=绕着旋转(电子，原子核)，吸引(太阳，行星)=吸引(原子核，电子)（步骤 3）。配对之后，我们发现“导致”与“和”这两个关系尚未配对，因此将之迁移到目标领域去（步骤 4），即我们可以借由类比推论出由于两者之间的“重量差”和“吸引力”使得电子围绕原子核转，最后，评量这个对应（步骤 5），如果觉得合适，整个对应就此结束，否则就重复步骤（2），直到找到合适的对应为止。

IAM 采用分批逐次的方式，先选定某一成分，再逐次增加类比部分。它一方面考虑了工作记忆的限制，另一方面也减轻认知负荷。

3 三派学者的类比理论比较分析

由上述的对三种理论及映射工具的介绍可以看出，三派学者对类比过程的想法不尽相同，以下将从理论和对应工具两方面进行比较。

3.1 类比理论的比较

三派学者在类比过程中所受的限制不同。结构映射理论只包含结构性限制，限制满足理论包含结构性、语义相似性、实用性三个限制，累进理论包含结构性、语义相似性、实用性、工作记忆和背景知识影响五个限制。

就结构限制而言，三派理论都认同结构相似性对类比的重要性，但所提到的结构性不尽相同，结

构映射和累进理论所提出的结构性相同，它们对结构的要求比较严谨。限制满足理论要求的结构不如它们要求得那么严谨，所以可以处理非完全相同的对应。

就语义相似性而言，Gentner 认为语义相似性因素在映射过程中起到的作用很小，所以在映射过程中并未加入语义相似性的因素，只是检索过程中加入了语义相似性因素的限制。Holyoak 则认为语义相似性在类比推理的整个过程都很重要，所以不仅在检索过程中加入了语义相似性因素，在映射过程中也加入了语义相似性因素。累进理论因为没有检索过程，所以也只是在映射过程中加入了语义相似性因素。

就实用性而言，结构映射理论不考虑实用性因素，限制满足理论和累进理论所提出的实用性因素相同。

3.2 对应工具的比较

这三种对应工具有以下共同点：（1）对应过程与内容无关。（2）都是寻找整体最好的唯一对应。（3）源的谓词和对象在初始化时需要人为地给出，而不是自动地计算而来。（4）都是以结构相似性为主要的限制条件。（5）谓词和对象的相似性要人为地给出。

就映射过程的演算来看，SME 和 IAM 都是序列加工模型。它们的不同之处是：SME 穷尽所有可能的匹配，然后再来评价各个匹配，寻找最佳映射；IAM 是以渐进的方式，分批次逐次处理种子群的配对，除非另有规定，否则只负责处理种子群。ACME 也是一次性处理所有可能的配对，然后寻找最佳映射。

SME 和 IAM 的节点是对象或是关系，ACME 的节点是源领域和目标领域一组可能的配对，这些配对的内容是对象或关系。对结果的评价标准，SME 只考虑结构的相似性，ACME 和 IAM 都是考虑了结构相似性、语义相似性和实用性这三个限制，但是都以结构相似性为主。

SME 和 IAM 主要是处理因果关系型的类比，ACME 使用范围比较广，涵盖了问题类比、故事类比和因果关系型的类比等。

4 未来发展方向

类比推理研究表明，使计算机系统具有一定的类比推理能力，为我们实质上提高计算机的智能水平提供了一条途径，同时也提出了巨大的挑战。根据对类比理论现状的分析，我们提出几点未来发展的方向。（1）现在的类比推理主要用于智能翻译等，尚未能以计算机完全模拟人的思考历程，这将是未来研究的一个方向。（2）我们还可以研究将类比推理用于机器自动学习。（3）近年来类比推理研究主要与心理过程相结合，使其能更符合人类的认知过程。有一些研究者主要是把类比学习应用于某一门学科的教学过程中。这些研究偏重类比映射机制的研究和类比的应用，很少有一个完整的模型，用于帮助学习者寻找合适的类比物。（4）有一些类比推理的研究主要关注学生在老师的帮助下用类比推理来找到类比物，很少有一个自动产生类比用例的系统，来帮助不是在校学生的学习产生类比用例。（5）类比推理产生的结果不一定是正确的，结果的正确性评价一般都是人为进行的，没有一个自动评价的系统。上述这些方面都是我们有待努力的方向。

参考文献

[1] 徐冬溶，潘云鹤，张畅，王选. 类比推理综述（下）. 计算机科学.24（2）：10-14，1997.

[2] Gentner, D. Analogical reasoning, psychology of. In L. Nadel (Ed.), Encyclopedia of cognitive science (Vol. 1, pp. 106-112). London: Nature Publishing.2003.

[3] Winston. Learning and Reasoning by Analogy. Communications of the CAM, Vol.23, No.12(1980) pp.698-703.

[4] Carbonell, A Computational Model of Analogical Problem Solving (1981) pp.147-152.

- [5] Richland L E,Morrison RG,Holyoak KJ.Children's development of analogical reasoning : Insights from scene analogy problems.Journal of Experimental Child Psychology, Vol. 94, No. 3,pp.249-273,2006.
- [6] Gentner,D.Structure-Mapping:A theoretical framework for analogy. Cognitive Science, Vol. 7, No. 2,pp.155-170,1983.
- [7] 高淑芬, 邱美红. 类比的检索与对应. 科学教育学刊 6 (1) :63-80,1998.
- [8] Holyoak,K.J.&Thagard,P. a computational model of an alogical problem solving. Simila rity and analogical reasoning . 13,295-355,1989(b).
- [9] Keane.Constraints on Analogical Mapping:A Comparison of three models.Cognitive Science.pp.13-23,1994.
- [10] 张莉辛, 自强. 类比推理研究的回顾与展望. 心理研究. 2 (2) : 9-15,2009.
- [11] Samuel B.Day,DeDre Gentner.Noni ntentional analogical inference in text comprehension. Memory & Cognition.Vol.35 No. 1,pp.39-49,2007.
- [12] Gentner,D.The mechanisms of an alogical learning.In S.Vosnia dou.&A.Ortony (Eds). Simila rity and analogical reason - ing.pp.197-241.1989.
- [13] Gentner,D,Markman . Structure mapping in the comparison process. American Journal Of Psychology. Vol. 113, No. 4 , pp.501-538.2000.
- [14] Andrew Lovett,Dedre Gentnera,Kenneth Forbusa,Eyal Sagia.Using analogical mapping to simulate time-course phenom - ena in perceptual similarity.Cognitive Systems Research.Vol. 10, No. 3,pp.216-228,2009.
- [15] Matthew Klenk,Ken Forbusa, Domain transfer via cross-domain analogy. Cognitive Systems Research.pp.240-250 .2009.

恶意代码检测技术综述

奚 琪，王清贤，曾勇军

(解放军信息工程大学信息工程学院, 河南 郑州, 450002)

摘 要: 近年来, 恶意代码的广泛传播和日益泛滥, 给信息系统的发展带来不利影响。研究高效可行的检测方法, 准确地发现程序中的恶意行为是系统安全研究的热点问题。文章根据动、静态检测的分类方法对恶意代码检测技术进行了归类, 阐述了各种检测方法的基本原理、特点及优缺点, 并研究分析了形式化检测方法的设计思路, 最后讨论了恶意代码检测技术的发展方向。

关键词: 恶意代码; 静态检测; 动态检测; 模型检测; 程序验证; 代码仿真

中图分类号: TP319 **文献标识码:** A **文章编号:** 1006-7043 (2004) xx-xxxx-x

Survey on Static Detection Technologies of Malware

XI Qi, Wang Qingxian, ZENG Yongjun

(Institute of Information Engineering, PLA Information Engineering University, Zhengzhou 450002, Henan China)

Abstract: One of the biggest threats on the security of internet is the enormous explosion of malware, which prohibits the development of computer system. There are many potential hot topics in system security for researchers to work on to design an effective and efficient detection method, which could identify malicious behavior of program accurately. This paper classified mostly malware detection methods based on static and dynamic analysis techniques, summarized their principles, features, strengths and limitations, and analyzed formal detection methods. Finally, it prospected the development direction of malware detection technique.

Keywords: Malware; Static Detection; Dynamic Detection; Model Checking; Program Verification; Code Emulation

信息技术特别是因特网技术的迅猛发展, 在为人们带来方便的同时, 也为联网的计算机带来潜在的安全问题, 其中最典型的就是恶意代码的攻击和泛滥。恶意代码是经过存储介质和网络进行传播, 未经授权认证破坏计算机系统完整性的程序或代码^[1], 其类型包括病毒、蠕虫、木马、后门、间谍软件等, 攻击过程中可进行组合达到攻击的目的。目前恶意代码所使用的技术手段日趋复杂和丰富, 如广泛使用多态和变形技术提高生存能力^[2,3,4,5], 以抵抗检测工具的扫描和分析。

针对日益增长的安全威胁, 恶意代码检测技术应运而生。在与恶意代码对抗的过程中, 检测技术不断发展完善并取得了显著的效果, 成为保障信息系统安全的一道重要屏障。本文总结归纳了目前研究使用的恶意代码检测技术, 首先描述了恶意代码检测技术的分类方法, 然后从静态检测、动态检测和形式化检测等方面分析了各种检测技术的设计思路、主要特点和存在的问题, 最后给出了恶意代码检测技术面临的挑战及对未来的展望。

1 恶意代码检测技术概述

恶意代码检测器执行检测算法, 标示程序是否包含恶意行为。恶意代码检测器 D 定义为一个函

作者简介: 奚琪, 博士生, 主要研究方向为信息安全;
王清贤, 教授, 博士生导师, 主要研究方向为计算复杂性理论;
曾勇军, 讲师, 主要研究方向为信息安全。

数，其定义域为程序集合 P，值域为集合 {Y, N}：

$$\forall p \in P, D(p) = \begin{cases} Y & \text{若 } p \text{ 被感染} \\ N & \text{否则} \end{cases}$$

检测器 D 扫描程序 p，判定程序 p 是否包含恶意行为。期望的结果为：若 D 返回 Y，则检测到程序 p 包含恶意行为；否则，未检测到恶意行为。但实际的检测过程中可能存在误警（False Positive）或漏警（False Negative）现象。误警是指当一个没有包含恶意代码的对象（文件、扇区和系统内存等）触发检测器 D 产生报警，这将浪费时间和资源进行处理；漏警是一个恶意行为存在但检测器 D 没有检测到。

已经证明，任意一个程序是否包含恶意行为是不可判定的^[6]。因此无法给出通用的检测算法识别所有的恶意代码。尽管如此，由于恶意代码在感染、传播过程中存在一些特性，这些特性有别于正常程序，这为检测方法的研究提供了可能。

恶意代码检测技术可按检测时是否执行代码分为静态检测和动态检测。静态检测是在不执行任何代码时实施检测，用于主要分析文件的结构和内容；动态检测通过运行代码、观察其行为，确定代码是否包含恶意行为。

静态检测和动态检测其实是对恶意代码所有可能执行子集的不同选择。静态检测技术要考虑到恶意代码的每一种可能的执行情况，即每一次执行时恶意代码的全部可能状态，由代码内容推导出所执行的特性，因此这种检测方法是完全的。动态检测根据恶意代码一次执行或多次执行的特性，判断是否存在恶意行为，可以准确地检测出现的异常属性，但无法判定某个特定的属性是否一定存在，因此是不完全的。

2 静态检测技术

2.1 签名扫描检测技术

签名扫描检测技术是基于早期计算机病毒特征发展起来的检测技术。由于当时每一种恶意代码固定不变，通过从恶意代码中抽取不同于其他程序的字符串，称为签名，组成签名数据库。然后对目标程序进行扫描，如果在程序中发现有匹配的签名值，则判定为恶意代码。

签名扫描过程中需要用到多模式匹配算法。其中最典型的如贝尔实验室提出的 Aho-Corasick 自动机匹配算法（简称 AC 算法）^[7]。该算法的基本思想为：在进行匹配之前，先对模式串集合进行预处理，构建树型有穷状态自动机 FSA（Finite State Automata）；然后依据该 FSA，对文本串 T 扫描一次就可以找出与其匹配的所有模式串。该算法进行模式匹配的时间复杂度是 O（n），与模式串的个数和每个模式串的长度无关，若包括预处理时间在内，AC 算法总时间复杂度是 O（M+n），其中 M 为所有模式串的长度总和。此外 Veldman 算法^[8]、Wu-Manber 算法^[9]等可实现同样的功能。

签名信息提取可以手动方法和自动方法来实现^[10]。手动方法利用人工方式对二进制代码进行反汇编，分析反汇编的代码，发现非常规（正常程序中很少使用的）的代码片段，标示相应机器码作为签名值；自动方法通过构造可被感染的程序，触发恶意代码进行感染，然后分析被感染的程序，发现感染区域中的相同部分，作为候选，然后在正常程序中进行检查，选择误警率最低的一个或几个作为签名值。

签名扫描检测技术检测精度高、可识别恶意代码的名称、误警率低。但该方法也存在速度慢、不能检查未知和多态性的恶意代码，且无法对付隐蔽性（如自修改代码、自产生代码）恶意代码等缺点。

签名扫描检测技术一直是反病毒技术发展过程中的基础技术，得到广泛的应用，并对后续新的恶

意代码检测方法产生深远的影响。

2.2 启发式扫描技术

基于给定的判断规则和定义的扫描技术，检测程序中是否存在可疑的程序功能指令，并做出预警或判断的恶意代码检测方法称为启发式扫描技术^[11, 12]。启发式扫描技术能够发现已知或未知的恶意代码，不需要使用专用的签名数据库。

恶意代码和正常程序的区别可以体现在许多方面，常见的如存在垃圾代码、解密循环代码、自修改代码、调用未导出的 API（Application Programming Interface）、操纵中断向量、使用非常规指令（如编译器一般不会生成的指令）和特殊字符串等，熟练的程序员在调试状态下很容易发现这些显著的不同之处。启发式扫描技术实际上就是把这种经验和知识移植到恶意代码检测工具中的程序体现。

启发式扫描技术定义了一些采集点，通过分析采集得到的数据，并对每个采集样本赋予一定的权值，进行求和并判定。若用 F_i 表示指定的特征，而 W_i 表示对应的权值，Threshold 代表设定的门限值。若有 $\sum F_i W_i > \text{Threshold}$ ，则认为已经感染，否则未感染。

启发式扫描技术存在误警现象，它有时会将一个正常的程序识别为恶意程序，这是因为被检测程序中可能含有恶意代码所使用的可疑功能。尽管如此，启发式扫描技术仍在不断发展和完善，并已在恶意代码检测软件中得到迅速的推广和应用。

2.3 完整性检测技术

完整性检测技术是一种强有力的恶意代码检测技术，也是 Cohen 推荐采用的检测方式^[13]，既能发现已知病毒，也能发现未知病毒。该技术基于这样一个事实：在正常的计算机操作期间，大多数程序文件和引导记录的内容不会改变。这样在干净的计算机系统状态，取得每个可执行文件和引导记录的签名，将该信息存放在硬盘的数据库中。以后在文件使用、系统启动过程中，检查文件内容的签名与保存在数据库中的签名是否一致，因而可以发现文件中是否受到篡改。签名提取算法可利用常见的散列算法如 MD5、SHA1 等实现。

完整性检测技术通过检测散列值的变化作为判定恶意代码感染的依据，容易实现且保护能力较强。但技术也面临一些问题：不能识别恶意代码的类别和名称，不能清除恶意代码，对隐蔽性恶意代码无效等。此外，有些程序执行时必须修改其自身，这也使得完整性检测技会产生误报。

3 动态检测技术

3.1 行为监控检测技术

行为监控检测技术是指通过审查应用程序的执行行为来判断其是否具有恶意属性。恶意行为是在对大量恶意代码和正常程序分析的基础上归纳总结的，主要是针对程序对操作系统所产生负面影响的行为，典型的如对可执行文件进行写入操作、搜索 API 函数地址、对关键性的系统设置（如注册表启动项）进行修改等。文献[14]提出了行为监控检测系统的基本构成，一般包含数据收集模块、解释分析模块和匹配算法三个模块：数据收集模块负责捕获目标程序的执行行为；解释分析模块将收集到的行为表示为易于识别的中间语言，并进行模式提取和建模；匹配算法用于将中间语言与行为特征匹配。

行为监控检测技术归属于异常检测的范畴，其核心是如何有效地实现数据收集。一般程序对环境的操作行为是通过 API 函数来实现的，因此可采用驻留内存的监控程序监视目标程序调用的 API 及相关参数完成数据收集。典型的如微软 Detour^[15]工具可实现对所有的 Windows API 及用户自定义的函数进行监控，它使用一个无条件转移指令来替换目标函数的最初几条指令，将控制流转移到用户提供的

截获函数。由于 Detour 会修改目标程序代码，难以躲避完整性工具的扫描和分析。此外，由于 API 数量众多，很容易出现函数遗漏的现象。这种方法对于运行在内核态的恶意代码依然无能为力。

3.2 代码仿真检测技术

行为监控检测法允许代码在真实环境中运行，若代码运行失控，将会感染真实系统。此时可使用代码仿真技术让代码运行在仿真环境中，仅感染仿真环境，严格控制后不会对真实系统产生影响。

代码仿真检测法一般可分为两种类型：

(1) 动态启发式 (Dynamic heuristics)：检测原理与静态启发式类似，唯一的差别在数据收集方法，动态启发式从仿真器中收集被分析的代码。此外，动态启发式能够收集与行为监控检测法同样的特征，如 API 序列等。

(2) 通用解密 (Generic decryption)：对于多态病毒，在真实环境中确定解密环位置是非常困难的。基于仿真环境的通用解密法利用恶意代码自身的解密环，因此很容易解决这个问题。一旦解密完成，能够利用常规的扫描方法进行检测。通用解密法可利用启发式方法确定恶意代码检测位置，如恶意代码一般需要执行改写（即解密结果）的存储器位置，对该位置进行扫描可以达到检测的目的。

目前已经出现了一些利用代码仿真技术实现恶意代码检测的工具。如 PolyUnpack^[16]工具采用静态分析与动态检测相结合的方法提取程序内容，在程序运行之前首先进行静态分析，即反汇编获得程序的二进制代码；然后仿真执行被加密的代码，检查当前指令是否与静态分析的指令相匹配，以此判断是否为隐藏的代码，进而可结合签名扫描机制判断是否为恶意代码。PolyUnpack 能支持多重解密和动态链接库 DLL (Dynamic Link Library) 识别功能，但由于使用单步执行和反汇编，影响系统性能，增加分析的复杂性。Renovo^[17]是一个实现自动解密功能的通用工具，由于加密程序总是要释放被加密的代码并写存储器，因此可通过监视存储器的状态获得运行的是否为加密的代码，该工具需要逐个字节地监视存储器的访问，因此开销比较大。TTAnalyze^[18]是一个基于 QEMU^[19]仿真器的恶意代码分析工具，通过在 QEMU 仿真环境中执行加密代码，监控调用的 API 及相关参数，达到检测的目的。

代码仿真检测技术误警率较低，且代码是虚拟执行，不会影响真实系统。将代码仿真技术与传统的签名扫描技术结合，可以提高检测准确性和安全性。但是，代码仿真检测技术只能检测代码执行路径上的行为，若利用该技术进行行为检测，可能检测不完整。此外，一般代码仿真器并不是对所有指令进行仿真，且仿真的设备类型有限，很容易被探测^[20]，如何提高反探测能力也是目前面临的一个问题。

4 形式化检测方法

4.1 程序分析技术

程序分析检测法主要通过对已知恶意代码和目标程序的内部属性关系如控制流、数据流、程序依赖关系等进行分析，从中找出两者之间是否匹配，最终判断程序是否包含恶意代码。

在文献[21，22]中，Bonfante 提出以恶意代码的控制流图作为特征进行检测的方法。它主要通过检测目标程序的控制流图中是否包含基于恶意代码的控制流子图，来判断目标程序是否包含恶意代码。为了应对恶意代码可能采用的代码迷惑技术，在程序的控制流图构成过程中还结合了程序语义以提高检测的精度。检测过程通常来说分为三个阶段：首先将目标程序 P 进行反汇编得到程序控制流图 CFG_P，然后对 CFG_P 按已定义的语义模板生成标签过程控制流图 CFG_{Pn}，最后通过匹配算法在 CFG_{Pn} 中匹配已知恶意代码的控制流子图 CFG_M。Bonfante 在实现过程中也存在一些缺陷，他只考虑了三种简单的迷惑技术，对于稍微复杂的代码迷惑技术如程序结构改变等方式并未考虑。

程序分析检测法的主要缺陷是对程序代码的分析依赖于反汇编代码的精度，另外判断子图同构问

题是 NP 完全问题，因此在匹配算法上需要进一步处理。

4.2 模型检验技术

模型检测是一种自动的、基于模型的、性质验证处理的方法。最初用于硬件、协议的验证，随着技术的进步，目前也广泛应用于程序性质的验证。在模型检测中，模型 M 是一个状态迁移系统（即程序），而性质 ϕ 是时态逻辑的公式（利用公式描述恶意行为），模型检测的目的是验证系统 M 是否满足性质 ϕ 。利用模型检测技术实现恶意代码检测需要完成下述三个步骤：

- （1）用模型检测器的描述语言建立系统的模型 M 。
- （2）用模型检测器的规范语言对待验证的性质进行编码，得到时态逻辑公式 ϕ 。
- （3）以 M 和 ϕ 作为输入，运行模型检测器。

若 $M \models \phi$ ，则模型检测器输出“是”，否则输出“不是”。在后一种情况下，大多数模型检测器还会产生导致失效的系统行为轨迹，称为反例。

文献[23]描述了利用模型检测技术实现恶意代码标示的方法。该文提出了一种新的时态逻辑 CTPL（Computation Tree Predicate Logic）描述逻辑公式，该逻辑是计算树逻辑 CTL（Computation Tree Logic）的扩展，通过引入全程量词 $\forall$ 和存在量词 $\exists$ ，提供一种通用、方便的表示方法，可减少描述恶意行为所需的公式数量。在该方法中，一个模型 M 被描述成 Kripke 结构，通过将二进制代码中的每条指令映射为 Kripke 结构中的一个节点，构造待分析的程序模型。利用基于 CTPL 的模型检测算法，检测程序是否满足 CTPL 公式描述的恶意属性。

文献[23]提出的方法也存在一定的问题，如检测结果依赖于反汇编的精度、没有考虑程序状态空间爆炸的问题等。此外，对一些复杂度高的代码变形技术（如基于 3SAT 的代码变形技术），模型检测方法难以有效地发现隐藏的恶意行为^[24]。

结合文献[23]提出的方法，可进一步进行优化，如利用抽象形式表示 x86 指令的谓词，可降低 CTPL 公式的复杂性。例如，通过利用 $\text{assign}(r,0)$ 表示将寄存器设置为 0，而不管具体指令是何种形式（ $\text{mov eax},0$ 或 $\text{xor eax},\text{eax}$ ）；充分结合程序分析技术如数据流分析、程序分片、区间分析等，可提高检测的正确性。利用一些有效的实现方法如有序的二叉判定图 OBDD（Ordered Binary Decision Diagram），结合优化的数据结构提高检测的性能。

4.3 程序语义检测技术

基于语义的恶意代码检测技术是目前研究的热点方向之一。通过检测程序的语义特征，可期望获得更准确的检测结果。目前已经出现了一些基于程序语义的检测方法，如文献[25]提出了基于语义模板的检测方法，该方法使用符号变量/常量处理变量和寄存器重命名，利用控制流图处理代码重排，通过验证某个程序是否出现模板描述的行为，以此检测是否包含恶意行为。该方法能够标示、处理某些代码变形机制，提高检测的可靠性。在文献[26]中，Christodorescu 对该思想进一步扩展，提出了一种能够利用程序依赖关系自动挖掘恶意代码行为的 MINIMAL 算法，并验证了这种方法在检测恶意代码方面的有效性，但在面对一些具有复杂参数关系的函数调用时，这种方法仍存在不足。

利用程序验证机制可构造基于公理语义的恶意代码检测技术。该检测方法基于 Hoare 逻辑^[27]实现，典型的 Hoare 逻辑结构如下：

$$|\phi|P|\psi|$$

其中， ϕ 是前置断言， ψ 是后置断言， P 为程序片段， $|\phi|P|\psi|$ 为程序规范。通过定义程序规范语言和一组推理规则，可验证程序 P 在满足 ϕ 的状态下执行，其执行结果满足状态 ψ 。

根据上述程序验证机制，可构造基于程序验证的恶意代码检测方法，具体方法如下：

- 定义一套中间代码的语法和语义，提出汇编语言到中间语言转换的算法，转换过程考虑了常规的 程序结构，如顺序、跳转、分支和函数调用等，要求正确地反映待分析目标指令的行为。

- 根据中间代码语法结构，构造相应的断言语法，用于描述程序分析过程中特定状态下存储单元和寄存器的内容。利用一阶谓词语言提高断言语法的表达能力，同时给出了断言语法元素对应的语义指称，用于描述程序的行为。为自动验证断言的有效性，可利用定理证明器 *Simplify*，将断言映射成 *Simplify* 的合式公式并进行验证。
- 构建基于中间代码的形式化逻辑证明系统，同时给出一套有效的推导算法和实现机制，结合启发式方法（如利用系统调用定位代码片段等）证明代码是否满足指定的程序规范。

利用上述方法可实现机械的、基于逻辑定理证明的恶意代码检测方法。由于程序非平凡语义的不可计算性，导致定理证明器可能不会停机，因此存在代码中包含恶意行为，但无法证明的现象。另外，在逻辑证明系统中存在循环代码检测的问题，确定合适的循环不变式是实现自动定理证明的关键。上述问题可利用抽象解释技术^[28,29]，通过对程序的状态空间进行抽象，希望得到程序语义的逼近结果。

5 结论

本文从静态检测、动态检测和形式化检测等方面对目前主流的恶意代码检测技术进行了研究、比较和分析。目前恶意代码检测技术的研究仍然是计算机和信息安全领域的一个活跃分支，不断有新的思路和方法被发掘。除了本文描述的方法外，可进一步进行研究和开发以下相关技术：

- （1）由于动、静态检测技术具有各自的优缺点，在实际的检测过程可进行合理组合，可期望提高检测结果的准确性。
- （2）利用形式化方法分析程序的语法、语义，可能产生较好的检测结果。除了本文描述的方法外，程序分片技术、数据挖掘技术、指令等价性验证技术等均得到研究和关注。目前形式化方法研究得比较多，但需要进一步研究分析、优化以达到真正实用化的效果。
- （3）利用抽象解释技术实现程序的语义逼近，可通过分析获得程序在运行过程中的动态属性。目前抽象解释技术的基本理论已经比较成熟，并出现了一些实际的应用系统。研究如何使用抽象解释进行恶意代码检测是未来的一个研究方向。

参考文献

- [1] R.A.Grimes Malicious Mobile Code, Virus Protection for Windows. 1st ed. O'Reilly & Associates. 2001.
- [2] Y. Mashevsky. Watershed in malicious code evolution. <http://www.viruslist.com>. 2005.
- [3] McAfee Avert Labs. McAfee Avert Labs unveils predictions for top ten security threats in 2007 as hacking comes of age . <http://www.mcafee.com>. 2006.
- [4] Panda Research. Mal(ware)formation statistics. <http://research.pandasoftware.com>. 2007.
- [5] C. Collberg and C. Thomborson. Watermarking, tamper-proofing, and obfuscation tools for software protection. IEEE Transactions on Software Engineering, 2002.
- [6] COHEN, F. Computational aspects of computer viruses. Computer& Security. Volume 8, Issue 4, June 1989, pp.297-298.
- [7] A. V. Aho and M. J. Corasick. Efficient string matching: An aid to bibliographic search. Communications of the ACM , 18(6):333-340, 1975.
- [8] F. Veldman. Generic decryptors: Emulators of the future. IVPC Conference, 1998.
- [9] Wu S, Manber U. A fast algorithm for multi-pattern searching. Technical Report, TR 94-17, University of Arizona at Tucson, 1994.
- [10] J. O. Kephart and W. C. Arnold. Automatic extraction of computer virus signatures. In Proceedings of the 4th Virus Bulletin International Conference, pp.178-184, 1994.
- [11] Symantec. Understanding heuristics: Symantec's Bloodhound technology. Symantec White Paper Series, Volume XXXIV,

- [12] Righard Zwienenberg, “Heuristics Scanners: Artificial Intelligence?”, Virus Bulletin Conference. 1995.
- [13] Dr. Frederick B. Cohen. A Short Course on Computer Viruses. Wiley Professional Computing, Wiley 2nd Edition. 1994.
- [14] Greoigre Jacob,Herve Debar,Eric F illol, “Behavioral detection of malware: fr om a survey towards an established taxon - omy”,Springer-Verlag France 2008.
- [15] Hunt, G., Brubacher, D. :Detours: binary interception of Win32 functions.In:3rd USENIX Windows NT Symposium,1999.
- [16] P. Royal , M. Hal pin, D. Dagon, R. Edmonds, and W. Lee. PolyUnpack: Automating the Hidden-Code Extraction o f Unpack-Executing Malware. In Proceedings of 2006 Annual Computer Security Applications Conference(ACSAC), pages 289-300, Washington, DC, USA, 2006. IEEE Computer Society.
- [17] M. G. Kang, P. Poosankam, and H. Yin. Renovo: a hidden code extractor for packed executables. In Proceedings of th e 2007 ACMWorkshop on Recurring Malcode (WORM 2007), 2007.
- [18] Ulrich Bayer. TTAalyze: A Tool for Analyzing Malware 15th European Institute for Computer Antivirus Research (EICAR 2006) Annual Conference, Hamburg, Germany, April 2006.
- [19] Fabrice Bellard. QEMU, a Fast and Portable Dynamic Translator[J]. In Proceedings of the US ENIX 2005 Annual Technical Conference, pp.41-46, 2005: 41-46.
- [20] Peter Ferrie. Attacks on Virtual Machine Emulators. Security Architect. 4 October, 2007.
- [21] Guillaume Bonfante,Matthieu Kaczmarek and Jean-Yves Marion,”Control Flow Graphs as Malw are Signatures”,In WTCV, May 2007.
- [22] Guillaume Bonfante,Matthieu Kaczmarek and Jean-Yves Marion,”Control Flow to Detect Malware”, Dans Inter-Regional Workshop on Rigorous System Development and Analysis 2007 (2007).
- [23] Johannes Kinder, Stefan Katzenbeisser, Christian Schallhart, Helmut Veith. Detecting Malicious Code by Model Checking. Conference on Detection of Intrusions and Malware & Vulnerability Assessment, DIMVA 2005.
- [24] Andreas Moser, Christopher Kruegel, and Engin Kirda. Limits of Static Analysis for Malware Detection.2007.
- [25] M. Christodorescu, S. Jha, S. A. Seshia, D. Song, and R. E. Bryant. Semantics-aware malware detection. In Proceedings of the 2005 IEEE Symposium on Security and Privacy (S&P’05), pages 32 -46,Oakland, CA, USA, May 8 -11, 2005. IEEE Computer Society.
- [26] Mihai Christodorescu Somesh Jha Christopher Kruegel, “Mining Specification of Malicious Behavior” , ESEC/FSE’07, September 3-7, 2007.
- [27] C. A. R. Hoare, An axiomatic basis for computer programming, CACM 12, pp. 576-583, 1969.
- [28] Cousot P, Cousot R. Abstract interpretation: A unified Lattice model for static analysis of programs by construction o f approximation of fixpoints [J] //Proc. of the 4th POPL. Los Angeles: ACM Press, 1977. 238-252.
- [29] P. Cousot, R. Cousot. Abstract interpretation frameworks [J]// Journal of Logic and Computer, 1992,2(4):511-547.

计算机总线的分类及发展趋势

丁彦芳, 黄欢欢, 王月蓉, 秦风云

(防空兵指挥学院, 河南, 郑州, 450052)

摘 要: 计算机技术的发展过程在一定意义上来说就是总线技术的发展史, 每一次新的主流总线技术的出现都给计算机及相关系统带来了全面的革新。本文则以计算机总线的多种分类方法、作用和应用着手, 具体介绍了常见微机总线及在其基础上的外部总线类型, 还简要介绍了现场总线在相关系统中的地位与发展趋势。

关键词: 总线; 现场总线; CAN 总线; 总线协议

中图分类号: TP336 文献标识码: A 文章编号: 1006-7043 (2010) xx-xxxx-x

On the Classification and Developing Trend of Computer Bus

DING Yanfang, HUANG Huanhuan, WANG Yuerong, QIN Fengyun

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: The developing course of computer science, in a certain significance, is the developing history of BUS technology, each new main BUS technology brings full-scale renovation. This thesis begins with variety of classification, function and application of computer BUS, introduces in detail the outer BUS genre of ordinary computer BUS and the position and developing trend of Field BUS.

Keywords: BUS; field BUS; CAN; protocol

微型计算机系统及系统之间大都采用总线结构, 这种结构的目的是简化系统结构, 大大减少连线数目, 便于接口设计和软件设计, 实现系统的模块化。无论是微机系统总线, 还是在现代汽车电子工程中, 采用总线的意义已远远超出节省电线的范围, 它已成为车内各个传感器之间实施信息交互的标准接口。正是由于它的瓶颈作用, 总线逐渐得到了重视, 在将近四十年的发展历程, 呈现出百家争鸣、百花齐放的态势。因此总线的种类有很多, 各有优缺点, 有各个不同的应用场合, 分类方法也不尽相同。

1 总线的概念

总线 (Bus) 是在计算机系统中各个组件之间信息传输的一组公共通路^[1], 各个功能部件都是通过总线交换数据。从外观上来讲, 总线就是一组信号线的集合, 设备之间只要进行数据传输就可以经过总线, 但与通信电缆之类的通信技术相比, 最重要的是总线包含标准、协议和规范等内容。所谓总线标准, 可视作系统与各模块、模块与模块之间一个互连的标准界面。就像一封信从发件人送到了收件人手里, 不仅包含它是经某种运输工具到达北京完成传送, 而且要保证准确、可靠、及时。对邮局来讲, 标准或规范包含了人所共知的常识内容。比如, 信封的写法、邮局人员的分拣、纠错、封装、分发等管理制度控制下的规范化工作。在以下内容中出现的常用总线标准, 介绍的都是目前市场上较为成熟的成品或接口类型。

计算机总线不仅包含了微机总线, 随着计算机和网络技术与各个领域的应用不断深入, 在某种含

作者简介: 丁彦芳 (1979—), 女, 讲师, 硕士;
黄欢欢 (1982—), 女, 助教, 学士;
王月蓉 (1985—), 女, 助教, 学士;
秦风云 (1975—), 女, 讲师, 硕士。

义上，计算机总线还包括工业领域应用日益广泛的现场总线、仪器设备领域的测试总线等类型。

2 总线的分类与发展现状

总线的分类方法有很多，常见的有按位置和层次、通信传输方式、传输信号的性质等角度来划分总线^[2]。从在微机的位置和层次可以分为片内总线、片总线、系统总线、外部总线；按通信传输方式分为串行总线、并行总线；按传输信号的性质，可以分为地址总线、数据总线、控制总线。当然按照扩展的应用领域来分就更多了，如测试总线、网络总线、现场总线等。

片内总线和片总线是位于 CPU 内部和芯片之间的总线。而系统总线是用于微机系统各插件板之间的连接，是其中最重要的一种总线^[3]。平时只要谈到微机总线，指的就是系统总线。目前，各个领域的总线功能和特点各不相同，每种总线有自己的适用范围，在它自己的适用范围内，它是最好的，出了这个范围它就不是最好的。同时，总线是一种正在发展中的技术，也许今天存在的问题明天可能就克服了。

2.1 系统总线的发展一路小跑

系统总线发展至今，经历了 ISA、EISA、PCI、AGP、PCI-E 等标准接口，PCI 取代了 20 世纪 80 年代的 ISA、EISA 成为目前中网卡、声卡的设备接口。但 PCI 的数据传输速率对显卡来说显得力不从心，显卡渴望高速传输的总线类型，AGP 就是在这种情况下产生的。由此拉开了显卡推动总线技术一路小跑的序幕。

AGP (Advanced Graph Port) 是一种为提高视频带宽而设定的总线规范，它前后经历过了十年三代的发展，它的 AGP8X 产品传输速率高达 2Gbps 左右。虽然 AGP 8× 显卡在各种测试中对速度提升有帮助，但并没有产生质的变化。所以，AGP 8× 成为了 AGP 规格中生命周期最短、也是最后的一种标准。虽然现在仍有大部分的主板支持 AGP，但其发展已到了尽头。

显卡的总线接口呼唤新的变革！而 Intel 公司正是顺应这一潮流推出了 3GIO 标准。PCI-E (PCI Express) 于 2009 年进入市场，其传输速率可以达到 8~10Gbps，大约相当于普通 PCI 速度的 60 倍。

ISA→PCI→AGP→PCI-E 的发展演变历史，很大程度上是受显卡技术的推进。目前主板对它们都是支持的，它们在共存的同时又存在着激烈的竞争。

2.2 外部总线的竞争无处不在

硬盘总线 IDE、SATA 属于存储总线，主要完成硬盘与南桥的通信，处在主机内部。还有一些外部总线接口，大都采用标准的模块化结构设计，可以得到更多厂商的广泛支持，在这些里面，有应用普及的 USB、IEEE 1394 和 eSATA 接口。

USB (通用串行总线)，是由 COMPAQ、IBM、Inter、Microsoft、NEC 等厂商共同制定的一种通用的外部设备总线规范，是目前 PC 中应用最多最广泛的外部总线接口，支持 USB 接口的主板、产品更是充斥了整个市场。但从数据传输速率上来看，USB 并不是很高，单从理论传输速率上，现在 USB2.0 的理论传输速率是 60MB/s，能与 USB 竞争的就只有 IEEE 1394 和 eSATA。

IEEE 1394 也被称为高速串行总线，是这一领域无可争议的“速度之王”，是由苹果公司和 TI (得州仪器) 公司开发的一种高速的外部串行接口标准，也被称为 FireWire (火线)，Sony 称为 i.Link。IEEE 1394 刚推出就有很高的起点，其速率高达 100Mbps、200Mbps 和 400Mbps，高出目前的 USB 标准数十倍。

eSATA 实际上是串口硬盘 SATA 3Gb 的外置规范。eSATA 仅仅是 SATA 接口的一种扩展，主要用于连接外部而非内部 SATA 设备。用户就可以轻松地将 SATA 硬盘插到 eSATA 接口，而不用再做打开机箱的动作。

新的 USB3.0 的标准也于 2007 年下半年公布，其理论最高数据传输速率高达 4.8Gb/s (600MB/s)，相信能够缓解目前外置存储接口传输速率过慢的窘境。接口外观没有发生变化，能够实现无缝升级，因为借助现有巨大的用户群，USB3.0 依然有很大的机会继续它的统治地位。

IEEE 1394 的速度虽然很高，性能也很优越，对 USB 是一个很大的威胁，但是 IEEE 1394 为什么没有得到推广，最大的障碍在于产品，因为主板芯片组直接对 IEEE 1394 提供支持的几乎没有，要实现它必须靠外接控制芯片和接口卡，这样无疑大大提高了产品成本，这是厂家与顾客都不希望看到的 IEEE；所以 IEEE 1394 逐渐成为摄影录像、视讯应用和专业剪辑领域的专用接口，逐渐从泛用走向专用。也许若干年后，我们直接拿着带有 IEEE 1394 接口的数码相机进行拍摄，不需要连接计算机，直接轻松地进行视频处理或者和打印机直接“点对点”地传输。而在中端应用方面，它们三个的竞争将是一场旷日持久的争霸战，胜负尚难以预料。

2.3 现场总线在工控领域异军突起

现场总线被誉为自动化领域的计算机局域网^[4]。典型的现场总线基金会现场总线 FF、LonWorks 总线、PROFIBUS、CAN 总线等。

基金会现场总线（Foundation Fieldbus，FF）于 1994 年由美国 Fisher-Rosemount 和 Honeywell 为首成立。它以 ISO/OSI 开放系统互连模型为基础，取其物理层、数据链路层、应用层为 FF 通信模型的相应层次，并在应用层上增加了用户层。

Lonworks 总线是由美国 Echelon 公司推出，它采用 ISO/OSI 模型的全部 7 层通信协议，采用面向对象的设计方法，通过网络变量把网络通信设计简化为参数设置。被誉为通用控制网络。采用 Lonworks 技术和神经元芯片的产品，被广泛应用在楼宇自动化、家庭自动化、保安系统、办公设备、交通运输和工业过程控制等行业。

PROFIBUS 是德国标准（DIN 19245）和欧洲标准（EN 50170）的现场总线标准。适用于纺织、楼宇自动化、可编程控制器和低压开关等。

CAN 即控制器局域网络，又被称为网络总线，是目前为数不多拥有国际标准的现场总线。一种特别适合于组建互连网络系统或子系统。最早由德国 Bosch 公司在 20 世纪 80 年代初为了解决现代汽车中众多的控制与测试仪器之间的数据交换推出的串行总线。由于其成本低、抗干扰性、高速、高可靠性，被广泛应用在很多方面。信号传输距离达 10km 时，CAN 仍可提供高达 50Kbps 的数据传输速率。

2.4 现场总线在测试领域的更新换代

测试总线是成熟的总线技术在测试领域应用的拓展，主要用于将数据采集和控制设备与主计算机连接。测试总线从 20 世纪 70 年代至今也历经了四五代的发展，从最原始的 GPIB 总线，80 年代的 VXI 总线，到 90 年代第四代——PXI 总线，风靡一时。直到 2004 年，新一代总线标准 LXI，标志总线技术发展上了一个新台阶，其具体的设想是将成熟的以太网技术应用到自动测试系统中^[5]。

从技术发展的角度来看，虚拟仪器走的是两条技术路线：一条是向高速、高精度、大型自动测试设备（ATE）方向发展，即 GPIB（1975）→VXI（1987）→PXI（1997）的发展路线；另一条是向高性能、低成本、普及型系统方向发展，即 PC 插卡（1987）→并口式（1995）→串口 USB/FireWare 方式的技术路线。

3 结论

总线的种类特别多，不同的角度分类方法也不尽相同。本文仅对微机系统总线和现场总线进行了特征分析。目前，针对总线的发展和应用现状，我国的发展起步非常晚，而且大多数用户并不能真正

地了解和使用。虽然各大总线支撑企业都看好中国市场的巨大潜力，国外的一些总线标准在我国展开激烈竞争，国内一些厂家也意识到应该将相关的产品投入市场，但都因资金和人才不足，对市场开发的投入不足，导致产品的不成熟。所以，我们应该紧跟国际化的潮流，加大对 IEC 标准的学习、宣传力度，使更多的人了解总线发展的趋势和应用领域，积极探索途径，支持各类总线在我国的推广应用^[6]。

参考文献

[1] 艾德才，陆明等. 微型计算机总线[M]. 北京：电子工业出版社，1996.

[2] 谢静波. 计算机总线的分类与发展趋势[J]. 科技信息，pp550-551.

[3] 周明德. 微型计算机系统原理与应用[M]. 第五版. 北京：清华大学出版社，2007.

[4] 魏余，芳曾蓉. 现场总线技术与应用[J]. 兵工自动化，2002，第 21 卷第四期：13-17.

[5] 魏庆福. 现场总线技术发展与工业以太网综述[J]. 工业控制计算机，2002，15（1）:1-5.

[6] Admin. 现场总线在中国的发展趋势[EB/OL]. 2009.

形式化开发非递归 Koch 曲线算法*

刘润杰, 申金媛, 穆维新

(郑州大学信息工程学院, 河南 郑州, 450001)

摘 要: 形式化方法是构建可信软件的重要途径。Koch 曲线是典型的分形图形, 本文使用形式化方法 PAR 及循环不变式开发策略, 开发了 Koch 曲线非递归算法, 并对其进行了形式化的正确性证明。在得到求解 Koch 曲线算法的循环不变式的同时, 直接得到易读、高效且可靠的非递归算法。对使用形式化方法及循环不变式开发策略开发分形程序非递归算法做了较深入的实践和探讨。

关键词: Koch 曲线; 形式化方法; 非递归; PAR 方法; 循环不变式

中图分类号: TP311.1 文献标识码: A 文章编号: 1006-7043 (2004) xx-xxxx-x

Formal Development of Non-recursive Algorithm for Koch Curve

LIU RunJie, SHEN JinYuan, MU WeiXin

(Schools of information engineering, Zhengzhou University, Zhengzhou 450001, Henan China)

Abstract: Formal method is an important approach for construction of the trustworthy software. Koch curve is one of the typical fractals. This paper develops non-recursive algorithmic program of Koch curve employing PAR method and the strategy of developing loop invariant and verifies the program formally. This paper achieves loop invariant of Koch curve with readable, efficient and reliable non-recursive algorithm finally. The paper contributes to developing non-recursive algorithm using formal method and new strategy of developing loop invariant.

Keywords: koch curve; formal method; non-recursive; PAR method; loop invariant

分形是描述自然界和非线性系统中不光滑和不规则几何形体的有力工具^[1]。分形算法在保密通信^[3,4]、图像压缩^[6]、航天器控制^[7]等领域有着较为广泛并深入的应用。在这些应用中对算法的正确性和可靠性要求十分严格, 开发正确可靠的分形算法程序是非常重要的。

Koch 曲线是经典分形图形, 它的算法常常使用递归的方法来实现。虽然递归算法程序具有结构清晰、易用数学归纳法证明正确性许多优点。但递归程序的执行过程中存在大量参数的传递和额外空间的分配, 具有程序执行效率低、空间的耗费大的缺点。而对于某些特殊要求的程序, 如系统核心程序, 它们的效率问题至关重要, 人们更愿意使用非递归算法来解决递归问题。

PAR 方法是一种建立在程序规约和归纳断言方法基础上的程序设计方法, 有严格的数学基础。是将程序开发与程序正确证明结合的一种形式化的开发方法。本文基于形式化方法 PAR^[9], 使用循环不变式开发策略中的递归定义技术^[10], 在得到求解 Koch 曲线的循环不变式的同时, 可直接得到清晰简短、可读性好的非递归算法程序。同时基于产生的循环不变式可对算法程序进行形式化验证, 从而保证算法程序的正确性。文中所使用的技术具有通用性, 可望发展成开发一系列分形递归问题非递归算法的方法。

本文详细阐述了 Koch 曲线非递归算法程序的开发过程, 形式化证明了该算法程序的正确性, 并对其效率进行了分析。文章的最后做了总结。

基金项目: 河南省教育厅自然科学研究计划项目 (2010A510015; 2008B120010)

作者简介: 刘润杰 (1969—), 男, 河南安阳市人, 博士, 讲师; 主要研究领域为通信网络特性, 混沌分形方法;
申金媛 (1966—), 女, 山西平遥县人, 博士, 教授, 主要研究领域为神经网络, 信号处理, 非线性方法;
穆维新 (1958—), 男, 山西平定县人, 硕士, 副教授, 主要研究领域为通信网络, 交换技术。

1 开发 Koch 曲线问题非递归算法

1.1 Koch 曲线问题描述

从一条直线段开始，将线段中间的三分之一部分用一个等边三角形的两边代替，形成山丘形图形如图 1 所示。

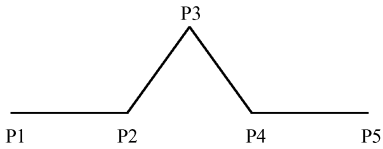


图 1 Koch 曲线

在新的图形中，又将图 1 中每一直线段中间的三分之一部分都用一个等边三角形的两条边代替，再次形成新的图形，如此迭代，形成 Koch 分形曲线。解决 Koch 分形曲线问题通常使用递归算法。

1.2 形式化方法 PAR 开发其非递归算法程序

步骤 1 用 $Koch(p_{0,1}, p_{0,2}, n)$ 表示求解原问题得到的解，即在 $p_{0,1}$ 和 $p_{0,2}$ 间画出递归 n 次的 Koch 曲线，由此可得此问题的 Radl 规约：

[X: list(list(list(real,2), 2)) ;]

AQ: $n > 0$; AR: $X = Koch(p_{0,1}, p_{0,2}, n)$

步骤 2 分划原问题。根据 Koch 曲线问题的性质，可将原问题分划为：

$Koch(p_{0,1}, p_{0,2}, n) = F(Koch(p_{1,1}, p_{1,2}, n-1), Koch(p_{1,2}, p_{1,3}, n-1), Koch(p_{1,3}, p_{1,4}, n-1), Koch(p_{1,4}, p_{1,5}, n-1))$

其中 $p_{1,1} = p_{0,1}$
 $p_{1,2} = p_{0,1} + (p_{0,2} - p_{0,1})/3$
 $p_{1,3} = p_{0,1} + (p_{0,2} - p_{0,1})/3 + ((p_{0,2} - p_{0,1})/3)'$
 $p_{1,4} = p_{0,1} + 2(p_{0,2} - p_{0,1})/3$
 $p_{1,5} = p_{0,2}$

$(p_{0,2} - p_{0,1})/3$ 表示 $p_{0,1}, p_{0,2}$ 间线段的三分之一， $2(p_{0,2} - p_{0,1})/3$ 表示 $p_{0,1}, p_{0,2}$ 间线段的三分之二， $((p_{0,2} - p_{0,1})/3)'$ 表示 $p_{0,1}, p_{0,2}$ 间线段的三分之一左旋 60 度。

步骤 3 根据分划寻找递推关系 F，同时求得循环不变式。基于上述分划，可得：

$Koch(p_{0,1}, p_{0,2}, n) = Koch(p_{1,1}, p_{1,2}, n-1) \uparrow Koch(p_{1,2}, p_{1,3}, n-1) \uparrow Koch(p_{1,3}, p_{1,4}, n-1) \uparrow Koch(p_{1,4}, p_{1,5}, n-1)$
根据此递推关系，容易得到解 Koch 曲线问题的递归算法。为了得到一个非递归的算法程序，进行下列推导：

$Koch(p_{0,1}, p_{0,2}, n) = Koch(p_{1,1}, p_{1,2}, n-1) \uparrow Koch(p_{1,2}, p_{1,3}, n-1) \uparrow Koch(p_{1,3}, p_{1,4}, n-1) \uparrow Koch(p_{1,4}, p_{1,5}, n-1)$
 $= Koch(p_{2,1}, p_{2,2}, n-1) \uparrow Koch(p_{2,2}, p_{2,3}, n-1) \uparrow Koch(p_{2,3}, p_{2,4}, n-1) \uparrow Koch(p_{2,4}, p_{2,5}, n-1) \uparrow$
 $Koch(p_{2,5}, p_{2,6}, n-1) \uparrow Koch(p_{2,6}, p_{2,7}, n-1) \uparrow Koch(p_{2,7}, p_{2,8}, n-1) \uparrow Koch(p_{2,8}, p_{2,9}, n-1) \uparrow$
 $Koch(p_{2,9}, p_{2,10}, n-1) \uparrow Koch(p_{2,10}, p_{2,11}, n-1) \uparrow Koch(p_{2,11}, p_{2,12}, n-1) \uparrow Koch(p_{2,12}, p_{2,13}, n-1) \uparrow$
 $Koch(p_{2,13}, p_{2,14}, n-1) \uparrow Koch(p_{2,14}, p_{2,15}, n-1) \uparrow Koch(p_{2,15}, p_{2,16}, n-1) \uparrow Koch(p_{2,16}, p_{2,17}, n-1)$
 $= \dots$

可以看出为求解原问题，产生出很多和原问题结构相同的子问题，其中递归深度为 0 的子问题可以通过画两点间的直线直接求解，如 $Koch(p_{0,1}, p_{0,2}, 0)$ 的解为 $line(p_{0,1}, p_{0,2})$ ，即在点 $p_{0,1}$ 和 $p_{0,2}$ 间画直线：

$X = (<xp_{0,1}, yp_{0,1}>, <xp_{0,2}, yp_{0,2}>) = line(p_{0,1}, p_{0,2}) = Koch(p_{0,1}, p_{0,2}, 0)$

由此可以得到用非递归算法解 Koch 曲线问题的总策略：将每个子问题表示成三元实数序列的形

式[起始点坐标, 终止点坐标, 递归深度]; 引进三个序列变量 X、q、S, 其中序列变量 X 用于存放部分问题的解, 也就是已得到的 Koch 曲线坐标序列, 每个坐标表示成二元序列的形式 $[p_a, p_b]$, 对应的动作是在点 p_a 和 p_b 间画直线, 循环终止时 $X = \text{Koch}(p_{0,1}, p_{0,2}, n)$ 。q 为三元序列, 用于存放正准备解决的子问题, $q[1]$ 为该子问题中起始点坐标, $q[2]$ 为终止点坐标, $q[3]$ 为递归深度, 如 $q = \text{Koch}(p_a, p_b, n)$ 表示在 p_a 和 p_b 间画出递归 n 次的 Koch 曲线; S 是一起堆栈作用的序列变量, 用于存放尚待解决的子问题, 如 $\text{Koch}(p_{1,1}, p_{1,2}, n-1) \uparrow \text{Koch}(p_{1,2}, p_{1,3}, n-1) \uparrow \text{Koch}(p_{1,3}, p_{1,4}, n-1) \uparrow \text{Koch}(p_{1,4}, p_{1,5}, n-1)$; S 的内容由函数 F 给出, F 的定义为: (1) $F([]) = []$; (2) $F(q \uparrow S) = \text{Koch}(q) \uparrow F(S)$ 。

根据总策略, 得到 X、q、S 满足如下等式, 构成所需的循环不变式:

p: $X \uparrow \text{Koch}(q) \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n)$

其中的 $\text{Koch}(q)$ 表示子问题 $\text{Koch}(p_a, p_b, n)$ 。

步骤 4 基于递推关系和循环不变式, 可以导出下列非递归 Apla 抽象算法程序:

```

program Koch;
{PQ: n>0; PR: X= Koch(p0,1,p0,2,n)}
var X: list ( list (list (real,2) ,2) );
q: list (list (list (real,2) ,2); interger );
S: list (list (list (real,2) ,2); interger );
begin
X,S,q: =[], [], [p0,1,p0,2,n];
do q≠[] ∧ q[3]≠0
→q,S: = Koch(p1,1,p1,2,n-1) ,Koch(p1,2,p1,3,n-1) ↑Koch(p1,3,p1,4,n-1) ↑Koch(p1,4,p1,5,n-1) ↑ S;
[] q≠[] ∧ q[3]=0
→X,q: =X ↑ [pa,pb] , [];
[] q=[] ∧ S≠[]
→q,S: =S[h] ,S[h+1..t];
od;
end.

```

该程序首先将 q 初始化为原问题, 而 X、S 初始化为空。do 语句的第一个分支表示子问题 q 要画的 Koch 曲线递归深度大于 0, 不能直接求解, 将子问题根据分划原则继续分划成 4 个子问题, 并将分划出来的另 3 个子问题存入序列 S; 第二个分支表示子问题 q 要画的 Koch 曲线递归深度等于 0, 可直接画出, 在点 p_a 和 p_b 间画直线, 该结果表示成二元坐标序列 $[p_a, p_b]$ 存入结果序列 X 中, 并置 q 为空; 第三个分支表示 q 为空而 S 非空, 即 q 中的子问题已经求解, 需取序列 S 的头元素 (即一个未解决的子问题) 赋给 q, 以便下次求解, 同时将该子问题 (S 头元素) 从尚未解决的序列中删除。

另外, 若只需画出 Koch 曲线而不需保存起来, 则此程序中的序列变量 X 可去掉, 同时将第二个 do 分支改成 “ $[] q \neq [] \wedge q[3] = 0 \rightarrow \text{line}(p_a, p_b); q := [];$ ” 即可。

2 形式化证明

基于所得到的循环不变式, 可用 Dijkstra 的最弱前置谓词法^[11]形式化证明上述程序正确:

(1) 证明 ρ 在循环执行前为真:

$$\begin{aligned} & \text{wp}("X,S,q: =[], [], [\text{Koch}(p_{0,1},p_{0,2},n)];", \rho) \\ & \equiv [] \uparrow \text{Koch}(p_{0,1},p_{0,2},n) \uparrow F([]) = \text{Koch}(p_{0,1},p_{0,2},n) \\ & \equiv \text{true} \end{aligned}$$

(2) 根据循环体, 证明 ρ 确实是循环不变式, 分三步:

$$(2.1) \rho \wedge q \neq [] \wedge q[3] \neq 0 \Rightarrow \text{wp}("q,S: =(p_{1,1},p_{1,2},n-1),(p_{1,2},p_{1,3},n-1) \uparrow (p_{1,3},p_{1,4},n-1) \uparrow (p_{1,4},p_{1,5},n-1) \uparrow S", \rho)$$

$$\begin{aligned}
&\equiv \rho \wedge q \neq [] \wedge q[3] \neq 0 \implies X \uparrow \text{Koch}(p_{1,1}, p_{1,2}, n-1) \uparrow F((p_{1,2}, p_{1,3}, n-1) \uparrow (p_{1,3}, p_{1,4}, n-1) \uparrow (p_{1,4}, p_{1,5}, n-1) \uparrow S) \\
&= \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \rho \wedge q \neq [] \wedge q[3] \neq 0 \implies X \uparrow \text{Koch}(p_{1,1}, p_{1,2}, n-1) \uparrow \text{Koch}(p_{1,2}, p_{1,3}, n-1) \uparrow F((p_{1,3}, p_{1,4}, n-1) \uparrow \\
&(p_{1,4}, p_{1,5}, n-1) \uparrow S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \rho \wedge q \neq [] \wedge q[3] \neq 0 \implies X \uparrow \text{Koch}(p_{1,1}, p_{1,2}, n-1) \uparrow \text{Koch}(p_{1,2}, p_{1,3}, n-1) \uparrow \text{Koch}(p_{1,3}, p_{1,4}, n-1) \uparrow \\
&F((p_{1,4}, p_{1,5}, n-1) \uparrow S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \rho \wedge q \neq [] \wedge q[3] \neq 0 \implies X \uparrow \text{Koch}(p_{1,1}, p_{1,2}, n-1) \uparrow \text{Koch}(p_{1,2}, p_{1,3}, n-1) \uparrow \text{Koch}(p_{1,3}, p_{1,4}, n-1) \uparrow \\
&\text{Koch}(p_{1,4}, p_{1,5}, n-1) \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \rho \wedge q \neq [] \wedge q[3] \neq 0 \implies X \uparrow \text{Koch}(q[1], q[2], q[3]) \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \rho \wedge q \neq [] \wedge q[3] \neq 0 \implies X \uparrow \text{Koch}(q[1], q[2], q[3]) \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv X \uparrow \text{mov}(q) \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \wedge q \neq [] \wedge q[3] \neq 0 \implies X \uparrow \text{Koch}(q) \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \text{true} \\
(2.2) \quad &\rho \wedge q \neq [] \wedge q[3] = 0 \implies \text{wp}("X, q := X \uparrow [p_{1,1}, p_{1,2}], []", \rho) \\
&\equiv \rho \wedge q \neq [] \wedge q[3] = 0 \implies X \uparrow [p_{1,1}, p_{1,2}] \uparrow \text{Koch}([]) \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \rho \wedge q \neq [] \wedge q[3] = 0 \implies X \uparrow [p_{1,1}, p_{1,2}] \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv X \uparrow \text{Koch}(q) \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \wedge q \neq [] \wedge q[3] = 0 \implies X \uparrow [p_{1,1}, p_{1,2}] \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv X \uparrow [q[1], q[2]] \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \implies X \uparrow [p_{1,1}, p_{1,2}] \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \text{true} \\
(2.3) \quad &\rho \wedge q = [] \wedge S \neq [] \implies \text{wp}("q, S := S[h], S[h+1..t];", \rho) \\
&\equiv \rho \wedge q = [] \wedge S \neq [] \implies X \uparrow \text{Koch}(S[h]) \uparrow F(S[h+1..t]) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \rho \wedge q = [] \wedge S \neq [] \implies X \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv X \uparrow \text{Koch}([]) \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \implies X \uparrow F(S) = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \text{true}
\end{aligned}$$

(3) 证明后置断言 PR 在循环终止时为真：

$$\begin{aligned}
&\rho \wedge \neg B \implies \text{PR} \\
&\equiv \rho \wedge \neg ((q \neq [] \wedge q[3] \neq 0) \vee (q \neq [] \wedge q[3] = 0) \vee (q = [] \wedge S \neq [])) \implies \text{PR} \\
&\equiv \rho \wedge \neg (q \neq [] \vee (q = [] \wedge S \neq [])) \implies \text{PR} \\
&\equiv \rho \wedge \neg (q \neq [] \vee S \neq []) \implies \text{PR} \\
&\equiv \rho \wedge q = [] \wedge S = [] \implies X = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv X \uparrow \text{Koch}([]) \uparrow F([]) = \text{Koch}(p_{0,1}, p_{0,2}, n) \implies X = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv X = \text{Koch}(p_{0,1}, p_{0,2}, n) \implies X = \text{Koch}(p_{0,1}, p_{0,2}, n) \\
&\equiv \text{true}
\end{aligned}$$

(4) 循环的终止性显然成立。

至此，完成了该程序的正确性证明。

3 Koch 曲线问题非递归算法实现

给 Apla 抽象程序加上输入/输出语句后，基于支持 Apla 抽象数据类型的可重用部件库及自动程序转换系统，可将其转换成某一可执行语言程序。递归五次的 Koch 曲线如图 2 所示。

算法程序中空间开销较大的是序列 S，由于 S 仅存放尚待求解的子问题，整个算法的空间复杂性远远小于定义栈或数组来存放结果的非递归算法。该算法不需要使用堆栈。而递归算法堆栈需要较大的内存空间。由此可看出该非递归算法在空间上占有绝对优势。针对此算法，利用类型转换函数，可将 S 定义成字符型序列来进一步优化其空间效率。

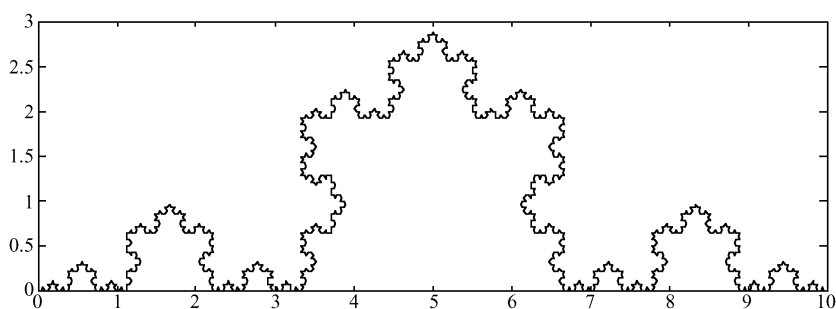


图 2 非递归程序产生的 Koch 曲线

4 总结

本文形式化开发了 Koch 曲线的非递归算法程序，并对其进行了正确性证明，所得算法程序可在相应平台及部件库的支撑下进一步得到可执行程序。

Koch 曲线的非递归算法在其他文献中也有叙述，如高军等^[12]使用 DELPHI 语言以递归方式实现了 Koch 曲线，孙继军等^[13]使用 C++ 语言以递归方式实现了 Koch 曲线。总体来说，前述工作编程思路较复杂，且所使用的技术难以掌握，不具有通用性。本文产生的算法抽象、简洁、易读，可靠性高；所使用的技术简单易用，具有通用性，可望发展成求解一系列类似递归问题的方法。另外，现有的 Koch 曲线算法程序的研究没有涉及其循环不变式的开发，无法开展对算法程序的形式化证明，难以保证算法程序的正确性。本文直接面向非递归算法程序的开发，使用循环不变式开发新策略，简捷地得到了 Koch 曲线算法程序的循环不变式的简单表达形式，从而可对产生的算法程序开展形式化证明，保证其正确性。

参考文献

- [1] Heinz-Otto Peitgen, Hartmut Jurgens, Dietmar Saupe. Chaos and Fractals: New Frontiers of Science[M]. Springer-Verlag New York, Incorporated. October 1992.
- [2] 张济忠. 分形[M]. 北京：清华大学出版社，1995.
- [3] Ljupco Kocarev, Chaos-based cryptography: a brief overview[J], IEEE Circuits and Systems Magazine, 2001, 1(3):6-21.
- [4] Yang T, A survey of chaotic secure communication systems[J], Int J Comp Cognit, 2004, 2(2):81-130.
- [5] 赵耀，王红星，袁保宗. 分形图像编码研究的进展[J]. 电子学报 2000, 28(4):95-101, 106.
- [6] LIU MQ, ZHAO Y, et al. A fast fractal image coding algorithm based on FGSE[A]. Signal Processing[C]. Beijing: IEEE Press, 2006. 1153-1156.
- [7] Alexey Bobtsov, Nikolay Nikolaev, Olga Slita. Control of Chaotic Oscillations of a Satellite[J]. Applied Mathematics and Mechanics(English Edition). 2007, 28(7):893-900
- [8] 屈文建，薛锦云. PAR 方法和循环不变式的范畴语义[J]. 计算机工程与应用 2009，45(8):50-54.
- [9] XUE JY. A unified approach for developing efficient algorithmic programs[J]. Journal of Computer Science and Technology, 1997, 12(4): 314- 329.
- [10] XUE JY. Two new strategies for developing loop invariants and their applications[J]. Journal of Computer Science and Technology, 1993, 8(2): 147- 154.
- [11] Dijkstra E W. A discipline of programming[M]. Englewood Cliffs: Prentice Hall, 1976.
- [12] 高军，孙博玲. KOCH 曲线的 DELPHI 程序设计[J]. 电脑学习 2000, (2): 35- 36.
- [13] 孙继军，卢玉蓉. Von Koch 曲线的 Visual C++ 程序实现[J]. 攀枝花学院学报 2004，21(1): 90-92.

商空间模型下不确定本体知识推理研究

王晓东, 孙 滨, 李学威

(河南师范大学 计算机与信息技术学院, 河南 新乡, 453007)

摘 要: 本文通过对商空间理论进行分析, 在其本体形式化知识表示的基础上, 对商空间模型下不确定本体知识进行推理研究。推理结果显示, 基于商空间理论的本体不确定知识的推理能很好地满足不确定知识推理的投影、合成特性, 并且对这些特性进行存在性验证。

关键字: 商空间; 本体; 不确定知识; 推理

中图法分类号: TP301 **文献标识码:** A

Uncertain Ontology Knowledge Reasoning Research of Based on Quotient Space Model

WANG Xiaodong, SUN Bin, LI Xuewei

(Institute of Computer and Information Technology Henan Normal University, Xinxiang 453007, Henan China)

Abstract: Analyzing the quotient space theory, this paper researches uncertain ontology knowledge of quotient space model on the basis of the ontology formalization knowledge representation. The reasoning results show that, uncertain knowledge reasoning of ontology based on quotient space theory can be well positioned to meet the projection property, the synthesis property, and carry on the existence confirmation to these properties.

Keywords: quotient space; ontology; uncertain knowledge; reasoning

1 引言

在语义网中, 本体具有非常重要的地位, 是解决语义层次上 Web 信息共享和重用的基础^[1]。为了对客观世界进行形式化的描述和推理, 以使计算机更具有人类的智能, 首先需要一种合理的逻辑语言对本体进行合理的形式化知识表示, 从而为知识表示提供公理和推理规则, 为智能推理提供理论基础。在文献[2]中探讨了本体和云理论下对不确定知识的研究中并没有对不确定知识的推理模型做出详细的解释, 在对不确定知识的推理等方面存在不足, 缺乏为不确定知识提供语义支持。为了更容易实现在本体构建的过程中层次之间的跳转及对不确定本体知识的推理, 为人类全局分析能力建立智能模型。本文在基于商空间理论的本体形式化知识表示基础上^[3], 重点探讨不确定本体知识的推理模型的建立及不确定推理模型的特性, 并对此推理投影、合成特性进行存在性检验。

2 粒度商空间模型

张钹、张铃教授提出的商空间理论^[4,5], 建立了基于商空间理论的粒度计算模型, 该模型用一个三元组 (X, F, T) 来描述一个问题, X 表示问题的论域, $F(.)$ 是一个映射, 表示论域的属性函数, 用 $F: X \rightarrow Y$ 表示, Y 是 N 维空间也可以是一般的空间, T 是论域的结构, 指论域 X 中各元素的相互关

基金项目: 河南省科技攻关计划项目 (No:102300410198, No: 082102210007, No: 072102210063)

作者简介: 王晓东 (1963—), 男, 教授、博士, 研究方向: 语义 Web 和 Ontology, 知识工程;

孙滨 (1983—), 男, 硕士研究生;

李学威 (1983—), 男, 硕士研究生。

系。该模型中，当 X 很复杂时，就用比较粗的粒度来考察问题，也就是在论域 X 上给出一个等价关系 R ，得到一个对应于 R 的商集 $[X]$ ，将对应的三元组变为 $([X], [F], [T])$ ，称为对应于 R 的商空间，从而将问题 (X, F, T) 转化为新层次的问题 $([X], [F], [T])$ 。逐步细化，从而将问题表示成不同的粒度世界，达到简化问题解决问题的目的^[6]。

2.1 商空间下本体形式化模型的建立

2.1.1 形式化模型的定义

定义1 对于给定术语构造符集 S ，定义三元组，记作 $O=\langle X, F, T \rangle$ ，简称本体。其中，论域 X 表示本体论域集合； $F(.)$ 表示论域集(元素)上的属性， $F: X \rightarrow Y$ ， Y 可以是 $\rightarrow$ 维空间，也可以是一般的集合，属性包括类属性和数值属性，类属性表示类间的关系，而数值属性表示类的属性； T 是论域集的结构，表示论域集中各元素之间的关系。

2.1.2 形式化模型的解释

对于给定术语构造符集 S ，定义本体 $O=\langle X, F, T \rangle$ ，就是对论域集 X 及其论域集相关的关系和论域集的属性进行分析和研究。

定义2 假设给定的本体 $O=\langle X, F, T \rangle$ 及对论域集的一种等价关系集 R ，则根据此等价关系集 R 就可得到和关系集 R 相对应的一个商集 $[X]$ ，当商集 $[X]$ 确定之后，则和商集 $[X]$ 对应的属性的商集 $[F]$ 和关系的商集 $[T]$ 也就随即确定。

根据此定义可以得到形式化本体 O 的一个商空间，记作 $O/R=[O]=\langle [X], [F], [T] \rangle$ 。

设 R 表示论域集 X 上一切关系组成的集合，可以定义如下关系，即为本体粒度的“粗”和“细”。

定义3 设 $R_1, R_2 \in R$ ，若对任意的 $x, y \in X$ ，都有 $xR_1y \Rightarrow xR_2y$ ，则称 R_2 比 R_1 细，记作 $R_1 < R_2$ 。这样一个 n 层的本体结构树对应的 n 个相应的关系就有如下的序关系：

$$R_0 < R_1 < R_2 < \cdots < R_n$$

设 R_i 对应的商集为 $[X]_i$ ($i=0, \cdots, n$)，则不同层次的粒度论域集有如下的序关系：

$$[X]_0 < [X]_1 < [X]_2 < \cdots < [X]_n$$

其中， $R_0 \in R$ 为第一层对应的等价关系，记为本体顶层论域集 $[X]_0$ （原始概念集）， $R_n \in R$ 为第 n 层对应的等价关系，记为实例集 $[X]_n$ 。

则对于 R_i 对应的本体的商空间 $O/R_i=[O]_i=\langle [X]_i, [F]_i, [T]_i \rangle$ 。其中属性商集 $[F]_i$ 解释为对应 R_i 本体商空间 $[O]_i$ 的属性集，关系商集 $[T]_i$ 解释为对应 R_i 本体商空间 $[O]_i$ 的关系集。

对于给定术语构造符集 S ，当等价关系集 R 确定之后，则商空间下的本体知识的层次关系也就随之确定。下面重点探讨不确定本体知识推理模型的建立及对商空间下不确定本体知识推理的投影、合成特性，并且对特性进行存在性验证。

3 不确定本体知识推理模型的建立

现实世界中，人们都是在知识不完全情况下进行的推理、决策并采取行动的，在这种环境下，人们对知识的认识不是绝对的，而只是一种信念。随着新知识的出现，这种信念可能加强了，也可能动摇了，于是人们原来的知识需要加以修改、补充、甚至放弃。而一阶谓词逻辑是不能解决这类问题的。传统的形式化本体都是基于描述逻辑的，而描述逻辑是一阶谓词的一个可判定子集，这就决定了对于不确定的知识，传统的本体知识的推理存在严重的不足。本节就是利用基于商空间理论的本体形式化知识模型，建立一种新的不确定的本体知识推理模型，以反映不确定本体知识推理的投影、合成

特性，并验证了合成特性的存在性^[4,7]。

首先给出不确定性的假设：对于某个不确定知识 A ，存在商空间下本体知识的一个层次 $[O]_i (i=0, \dots, n)$ ，使得 A 在 $[O]_i (i=0, \dots, n)$ 上可表示为确定的知识。反之，一个在商空间下本体知识的某一个层次 $[O]_j (j=0, \dots, n)$ 上为确定的知识 A ，在比 $[O]_j (j=0, \dots, n)$ 商空间下本体知识更小的层次 $[O]_k (k=0, \dots, n)$ 上，可能表现出某种程度的不确定性。

3.1 商空间下本体的不确定知识推理模型的建立

假设给定的商空间模型下的本体形式化知识模型 $O=\langle X, F, T \rangle$ 及论域集 X 上的等价关系集 R ，给出不确定本体知识推理模型^[6]。

设 $A \subset X$ ，以及一个定义在 A 上的函数 $f:A \rightarrow [0,1]$ 。

设 D 是本体形式化层次图中边的集合，定义推理函数 $g:D \rightarrow [0,1]$ 。

推理规则：若 b 是 a 的直接后继者， $e=(a,b)$ ， $e \in D$ 定义推理规则 $f(b)=F(f(a),g(e))$ ，其中 F 是自变量的组合函数。例如可取 $f(b)=f(a)f(e)$ 。

给定目标 $p \in X$ ，利用上述模型，若推得 $f(p)=d$ ，则称目标以可信度 d 成立。特别是，当 $d=0$ 时，称 p 不成立；当 $d=1$ 时，称 p 成立。

对于上面的推理模型，现在需要进一步讨论的是，在给定本体知识模型 $O=\langle X, F, T \rangle$ 及对论域集的一种等价关系集 R 后，应该如何在论域商空间 $[X]$ 上建立对应的推理模型（所谓投影问题），以及在论域商空间 $[X]$ 推理模型上推出的结论，可为论域集 X 的推理提供多少有用的信息。

其次，对于给定本体知识模型 $O=\langle X, F, T \rangle$ 的两个本体商空间 $[O]_1, [O]_2$ 上的推理模型（根据等价关系 R 得到的从不同层次对本体 O 的理解），如何根据这两者得出一个对本体 O 的更全面的理解。

3.2 商空间下不确定本体知识推理模型的投影

根据上面对不确定本体知识推理模型的定义，将其模型改写成

$$((X,D),(f,g),T,F,(A,p))$$

其中， X 是本体的论域集； D 是本体层次关系图中边的集合； A 是前提子集； p 是目标； f 是定义在 A 上的函数 $f:A \rightarrow [0,1]$ ， $g:D \rightarrow [0,1]$ 是推理函数； T 是商空间下不确定本体知识的半序结构； F 是定义在 $A \times D \rightarrow [0,1]$ 的函数，称为推理规则。

整个表示式的含义是：在给定的结构和推理规则下，求在前提 A 下， p 成立的可信度有多大。

投影问题：已知关于本体知识 O 的推理结构，现给定 O 的一个本体商空间 $[O]_1$ ，求 $[O]_1$ 的推理结构。

由前面定义 1 中本体形式化知识模型易知， $[O]_1$ 是 O 的本体商空间知识模型， $[X]_1$ 是论域 X 的商空间。从而可得 D 对应的商空间 $[D]_1$ ，定义 $\forall a,b \in [X]_1$ ，有 $(a,b)=\{e|e=(x,y),x \in a,y \in b,(x,y \in D)\}$ 所有这样的等价类 (a,b) 构成的集合记为 D_1' ，它还不是 D 的划分，将 D 中不属于 D_1' 中的任何等价类的边归成一个类记为 E_0 。将 E_0 加到 D_1' 中，得到 $[D]_1$ 是边集 D 的商空间。

前提子集 A 对应商空间为 $[A]_1$ ，目标 p 对应的等价类 $[p]_1$ ， $[p]_1 \in [X]_1$ 。

若 $[X]_1$ 对应的等价关系为 R_1 ，则可在 $[X]_1$ 上诱导一个商空间的半序，记为 $[T]_1$ 。

再给出 f_1, g_1 提取方法，如用组合原则，定义： $f_1(a)=G_1(f(x),x \in a \cap A), \forall a \in [A]_1$

$$g_1(E)=G_2(g(e),e \in E), \forall E \in [D]_1$$

其中 G_1, G_2 是其自变量的某一组函数，并要求定义出 g_1 的值域在 $[0,1]$ 中。

一般推理规则仍取 F ，这样就得到关于 $[X]_1$ 上的推理模型，既有

$$((([X]_1,[D]_1),(f_1,g_1),[T]_1,[F]_1,([A]_1,[p]_1)))$$

下面给出这个投影问题满足同态原则^[6]，即在粗本体商空间上，若 $[p]_1$ 成立的可信度小于 d ，则在

任一细的本体商空间知识模型上 p 成立的可信度也一定小于 d 。

命题 1 设 $F(x,y)=xy$, $f_1(a)=\max_{x \in a \cap A}^{f(x)}$, $\forall a \in [A]_1$, $g_1(E)=\max_{e \in E}^{g(e)}$ 成立, 并且已知 $((X,D),(f,g),T,F,(A,p))$ 且通过商空间下本体的不确定知识推理得 $f(p)=d$, 令 $(([X]_1,[D]_1),(f_1,g_1),[T]_1,[F]_1,([A]_1,[p]_1))$ 是本体商空间知识模型 $[O]_1$ 上对应的不确定知识的推理模型, 则在 $[O]_1$ 的本体商空间不确定知识推理模型中可得

$$f_1([p]_1) \geq f(p) = d$$

证明: 设本体 O , 论域集 X 中由 $x_1 \in A$ 经 $x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_m = p$, 推得 p , 且 $f(p)=d$ 。
现按推理步骤进行数学归纳法, 当 $n=1$ 时, 结论显然成立。现设对 $n < m$ 成立, 以下证明, 对 $n=m$ 也成立即可。

设 $x_m \in a_i \in [X]_1$, 由归纳假定有 $f_1(a_i) \geq f(x_{m-1})$ 。
若 $x_{m-1} \in a_i$, 因 $g(x_{m-1}, x_m) \leq 1$, 于是有 $f(x_m) = f(x_{m-1})g(x_{m-1}, x_m) \leq f(x_{m-1}) \leq f_1(a_i)$
因 $p = x_m \in a_i$, 得 $a_i = [p]_1$, 即 $f_1([p]_1) \geq f(x_m) = f(p) = d$
若 $x_m \notin a_i$, 设 $x_m \in a_{i+1} = [p]_1$, 因 $e = (x_{m-1}, x_m) \in D$, 故 $(a_i, a_{i+1}) \in [D]_1$, 按定义有 $g((a_i, a_{i+1})) = \max_{e \in (a_i, a_{i+1})}^{g(e)} \geq g((x_{m-1}, x_m))$ 再由 $f_1([p]_1) = f_1(a_{i+1}) = f_1(a_i)g((a_i, a_{i+1})) \geq f_1(a_i)g((x_{m-1}, x_m)) = f(x_m) = f(p) = d$ 证毕。

简单地说, 在粗粒度商空间对应的本体层次关系上, 如果无解 (即其成立的可信度小于要求的可信度), 则在任一个细粒度商空间上对应的本体层次关系也一定无解。

3.3 商空间下不确定本体知识推理的合成

根据多粒度的本体商空间模型, 从不同角度和不同侧面对同一本体的不同层次的知识进行获取, 相当于从不同粒度本体层次上观察问题, 并获得相应的知识, 这些知识具有各自的不确定性。如何根据这些知识取得所观察本体的更全面的知识, 这就是不同本体商空间的合成问题。

设 $[O]_1$, $[O]_2$ 是 O 的两个不同的本体商空间, 已知关于 $[O]_1$, $[O]_2$ 的推理模型 $(([X]_1,[D]_1),(f_1,g_1),[T]_1,[F]_1,([A]_1,[p]_1))$ 及 $(([X]_2,[D]_2),(f_2,g_2),[T]_2,[F]_2,([A]_2,[p]_2))$, 求其合成本体商空间 $[O]_3$ 上的推理模型 $(([X]_3,[D]_3),(f_3,g_3),[T]_3,[F]_3,([A]_3,[p]_3))$ 。

按照商空间模型的合成方法可得出^[4]:

- (1) $[X]_3$ 是 $[X]_1$ 和 $[X]_2$ 的最小上界空间;
- (2) $[D]_3$ 是 $[D]_1$ 和 $[D]_2$ 的最小上界空间;
- (3) $[T]_3$ 是 $[T]_1$ 和 $[T]_2$ 的最小上界半序。

其实, 当 $[X]_3$, $[T]_3$ 确定之后, $[D]_3$ 也就唯一确定了。故在具体合成时, 先合成 $[X]_3$, 再合成 $[T]_3$, 最后由 $[T]_3$ 得出 $[D]_3$ 。

当本体 O 上的属性函数的投影运算确定后, 按照商空间的属性函数的合成方法, 可求得 f_3 和 g_3 。令

$$[A]_3 = \{x \mid x \in [A]_1 \cup [A]_2, x \in [X]_3\}$$

$$[p]_3 = \{x \mid x \in [p]_1 \cap [p]_2, x \in [X]_3\}$$

推理规则 F 不变。这样, 就得到在合成的本体商空间 $[O]_3$ 上的推理模型:

$$(([X]_3,[D]_3),(f_3,g_3),[T]_3,[F]_3,([A]_3,[p]_3))$$

3.4 不确定本体知识推理特性存在性验证

在 2.1 节中建立不确定知识推理模型, 并且讨论的本体商空间的投影结构和两个不同的本体粒度商空间合成一个本体粒度商空间结构, 并证明对于特定的推理规则, 在所定义的投影下, 满足商结构的同态原则^[4]。

在本体商空间模型和推理模型建立之后, 对本体 O 中论域 X 上的投影和合成特性进行存在性验证。

下面重点验证不确定本体知识的合成特性的存在性。

已知本体 $O=\langle X, F, T \rangle$ 及对论域集的一种等价关系集 R , 其中 $R_0 \in R$ 是最粗的本体商空间 $[O]_0$ 对应的等价关系。 $R_n \in R$ 是最细的本体商空间 $[O]_n$ 对应的等价关系。

已知 X 是本体的论域集, 则 $[X]_0$ 是最粗的论域商空间。

现设 N 是 X 上的一个二元运算, 即

$$N: X \times X \rightarrow X$$

任给 X 的商空间 $[X]_i (i=1, \dots, n)$, 显然有 $[X]_0 < [X]_i < X$, 而 N 在 $[X]_0$ 中显然有对应的商空间 $[N]_0$, 只要定义 $[N]_0(a, a)=a, a \in [X]_0$ 即可。

若 $[X]_i (i=1, \dots, n)$ 上不存在商推理 $[N]_i$, 那么是否存在有最细的本体商空间 (实例商空间) $[X]_n$, 使得:

(1) $[X]_n < [X]_i$;

(2) 在 $[X]_n$ 上有商推理运算 $[N]_n$ 。

若存在, 称 $[X]_n$ 为 $[X]_i$ 关于 N 的合成推理运算下界本体商空间, $[N]_n$ 为 N 对应于 $[X]_i$ 的下合成本体商推理。

4 结束语

在利用基于商空间理论下对本体形式化模型及检验的研究成果的基础上, 本文重点探讨了商空间模型下不确定本体知识推理模型的建立, 并证明了不确定本体知识推理模型的投影、合成特性, 以及对本体商空间上不确定知识推理存在性的验证。下一步的工作将是根据本体形式化知识模型和建立的不确定本体知识推理模型, 构建一个完备商空间理论下本体的不确定知识表示系统, 实现本体知识的开发, 知识的共享, 不确定知识的推理, 从而为本体描述语言能够处理不确定知识提供语义支持。

参考文献

[1] 王洪伟, 吴家春, 蒋馥. 本体的形式化模型及在语义查询中的应用[A]. Advances of search engine and web mining in China (搜索引擎与 Web 挖掘进展) [C]. 北京: 高等教育出版社, 2003: 205-213.

[2] 林培光, 徐如志, 余正涛. 基于本体和云理论的不确定知识表示[J]. 计算机工程与应用, 2008, 44(5):51-54.

[3] 王晓东, 孙滨. 商空间模型下的 Ontology 形式化及其检验[J]. 计算机工程与应用, 2010(9).

[4] 张钺, 张铃. 问题求解理论及应用 (第二版) [M]. 北京: 清华大学出版社, 2007.

[5] Zhang Ling, Zhang Bo. The quotient space theory of problem solving. Fundamental Informaticse, 2004, 59 (2,3):287-298).

[6] 邓志鸿, 唐世渭, 张铭, 杨冬青, 陈捷. Ontology 研究综述[J]. 北京大学学报 (自然科学版), 2002, 38(5):730-738.

[7] Zhang Ling, Zhang Bo. Fuzzy reasoning model under quotient space structure(Invited Lecture). International Conference on Fuzzy Information Processing-Theories and Applications, Beijing China, March 1-4, 2003.

粒子滤波多样性测度分析

于金霞^{1,2}, 刘文静¹, 汤永利^{1,3}

(1. 河南理工大学 计算机科学与技术学院, 河南 焦作 454003;

2. 南京邮电大学 江苏省图像处理与图像通信重点实验室, 江苏 南京 210003;

3. 清华大学 计算机科学与技术系, 北京, 100084)

摘要: 有效的多样性测度对于执行重采样步骤的粒子滤波算法是十分重要的。在分析粒子滤波重采样算法固有的缺陷下, 介绍了三个多样性测度, 包括有效样本大小、种群多样性因子、粒子群多样性测度, 最后采用 Matlab7.0 设计了从固定视觉观测点进行单机目标跟踪的仿真程序, 对这三种多样性测度在粒子滤波重采样执行过程中进行了粒子多样性的评估。

关键字: 粒子滤波; 多样性测度; 有效样本大小; 种群多样性因子; 粒子群多样性测度

中图分类号: TP24 **文献标识码:** A

Research on Diversity Measure in Particle Filter

YU Jinxia^{1,2}, LIU Wenjing¹, TANG Yongli^{1,3}

(1. College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003 Henan China;

2. Jiangsu Provincial Key Lab of Image Processing and Image Communication, Nanjing University of Posts and Communication, Nanjing 210003, Jiangsu China;

3. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: It is very important to find diversity measure when to perform a resampling step in particle filter. By analyzing the inherent deficiency in resampling algorithm of particle filter, some diversity measures including effective sample size, population diversity factor and particle swarm diversity measure are introduced. Then combined with the simulation program using matlab 7.0 to track a single target motion from a fixed visual observation points, the performance of diversity measures in resampling algorithm of PF are evaluated.

Keywords: particle filter; resampling; diversity measure; effective sample size; population diversity factor; particle swarm diversity measure

粒子滤波 (Particle Filter, PF), 也称序列蒙特卡罗 (Sequential Monte Carlo, SMC), 源于 Gordon 等人^[1]提出的一种基于序列重要性采样 (Sequential Importance Sampling, SIS) 的自举 (Bootstrap) 非线性滤波方法。PF 的核心思想是用随机加权的样本 (也称粒子) 集合来近似表征后验概率密度函数。原则上, PF 可以用于任意非线性非高斯动态系统的状态递推估计或概率推理研究, 近年在移动机器人领域如定位^[2]、目标跟踪^[3]、并发定位与地图创建^[4]及故障诊断^[5]得到成功应用。

粒子滤波可以有效地解决递归贝叶斯估计问题, 但是, 基于 SIS 的 PF 算法潜在的问题是样本退化^[6] (Degeneracy), 这是无法避免。强硬的解决方法是用一个非常大的样本数目, 这通常是不切实际的。为了解决样本退化问题, 采样重要性重采样 (Sampling-Importance Resampling, SIR) 被引入 PF 中, 在 PF 的两次重要性采样之间增加重采样。但是 SIR 粒子滤波是通过大量的复制高权粒子来人为地掩盖样本退化, 并且这样导致粒子之间的高度的相关性。

在重采样的过程中 那些健壮的粒子也可能被删掉, 导致粒子退化, 因此解决办法就是减少重采样的次数。另一个解决办法是通过监控每一步骤粒子的有效性, 从而限制重采样步骤的次数, 提高

作者简介: 刘文静 (1982—), 河南焦作人, 河南理工大学硕士研究生, 研究方向为人工智能。

PF 的鲁棒性。但是过少重采样又会导致粒子滤波的发散，因此，找一个标准进行判断何时重新采样是非常重要的。这也是粒子滤波多样性测度的一个问题。针对这一问题，多种有效的方法用来改善重采样过程，目前主要有效样本大小^[8]，种群多样性因子^[9,10]，粒子群多样性测度^[11,12]用来作为粒子滤波的多样性测度。

本文的第二部分，介绍基本 PF 算法，分析重采样的固有缺陷。第三部分，针对粒子滤波重采样算法的固有缺陷并结合最新研究与应用，总结了一些多样性的测度。第四部分，用 MATLAB7.0 设计了从固定视觉观测点进行单机动目标跟踪的仿真程序，对这三种多样性测度在粒子滤波重采样执行过程中进行了评估。最后，给出结论。

1 粒子滤波算

1.1 粒子滤波的基本算法

不失一般性，假定系统的状态方程和测量方程描述如下：

$$\begin{cases} x_t = f(x_{t-1}, u_{t-1}) + d_t \\ z_t = h(x_t) + v_t \end{cases} \quad (1)$$

$$z_t = h(x_t) + v_t \quad (2)$$

其中， x_t 表示 t 时刻的系统状态； u_{t-1} 表示 $t-1$ 时刻的系统输入； z_t 表示 t 时刻的系统测量； d_t 和 v_t 分别表示 t 时刻的过程噪声和测量噪声（它们独立同分布）。状态方程式（1）刻画了系统的状态转移概率 $p(x_t | x_{t-1}, u_{t-1})$ ，测量方程式（2）刻画了似然概率 $p(z_t | x_t)$ 。

从贝叶斯滤波角度，在给定系统输入 $u_{1:t-1} = u_1, \dots, u_{t-1}$ 和测量数据 $z_{1:t} = z_1, \dots, z_t$ ，问题求解的核心是估计后验分布 $p(x_t | z_{1:t}, u_{0:t-1})$ 。假设 x_t 服从一阶 Markov 过程，且初始状态 x_0 的概率分布 $p(x_0 | z_0) \equiv p(x_0)$ ，则后验分布 $p(x_t | z_{1:t}, u_{0:t-1})$ 的估计分为预测/更新递归执行。

预测：依据前一时刻状态的后验信度 $Bel(x_{t-1})$ ，即概率分布 $p(x_{t-1} | z_{1:t-1}, u_{0:t-2})$ ，结合状态方程式（1）来预测当前 t 时刻状态 x_t 的先验信度 $Bel^-(x_t)$ 。

$$\begin{aligned} Bel^-(x_t) &= p(x_t | z_{1:t-1}, u_{0:t-1}) \\ &\stackrel{\text{Bayes规则}}{=} \int p(x_t | x_{t-1}, z_{1:t-1}, u_{0:t-1}) p(x_{t-1} | z_{1:t-1}, u_{0:t-2}) dx_{t-1} \\ &\stackrel{\text{Markov假设}}{=} \int p(x_t | x_{t-1}, u_{t-1}) p(x_{t-1} | z_{1:t-1}, u_{0:t-2}) dx_{t-1} \\ &= \int \underbrace{p(x_t | x_{t-1}, u_{t-1})}_{\text{状态方程}} \underbrace{Bel(x_{t-1})}_{\text{后验信度}} dx_{t-1} \end{aligned} \quad (3)$$

更新：利用测量方程式（2），结合当前的感知测量信息 z_t 来更新当前 t 时刻状态 x_t 的后验信度 $Bel(x_t)$ 。

$$\begin{aligned} Bel(x_t) &= p(x_t | z_{1:t}, u_{0:t-1}) \\ &\stackrel{\text{Bayes规则}}{=} \frac{p(z_t | x_t, z_{1:t-1}, u_{0:t-1}) p(x_t | z_{1:t-1}, u_{0:t-1})}{p(z_t | z_{1:t-1}, u_{0:t-1})} \\ &\stackrel{\text{Markov假设}}{=} \frac{p(z_t | x_t) p(x_t | z_{1:t-1}, u_{0:t-1})}{p(z_t | z_{1:t-1})} \\ &= \underbrace{\eta}_{\text{标准化因子}} \underbrace{p(z_t | x_t)}_{\text{测量方程}} \underbrace{Bel^-(x_t)}_{\text{先验信度}} \end{aligned} \quad (4)$$

其中，标准化因子 $\eta^{-1} = p(z_t | z_{1:t-1}) = \int p(z_t | x_t) p(x_t | z_{1:t-1}, u_{0:t-1}) dx_t$

粒子滤波是贝叶斯滤波的变种，它采用 N_s 个随机加权的样本集合 $\{x_t^i, w_t^i\}_{i=1, \dots, N_s}$ 来近似表征后验概

率密度函数 $p(x_t | z_{1:t}) = \sum_{i=1}^{N_s} w_t^i \delta(x_t - x_t^i)$ 。因而，式（3）中求解 $Bel^-(x_t)$ 的积分运算可以转化为样本的求和运算，即

$$\begin{aligned} Bel^-(x_t) &= \sum_{i=1}^{N_s} p(x_t | x_{t-1}^i, u_{t-1}) Bel(x_{t-1}^i) \\ &= \sum_{i=1}^{N_s} w_t^i \delta(x_t - x_t^i) Bel(x_{t-1}^i) \end{aligned} \quad (5)$$

其中， $\delta(\cdot)$ 表示 Dirac delta 函数，权重满足归一化条件 $\sum_{i=1}^{N_s} w_t^i = 1$ 。当 $N_s \rightarrow \infty$ 时，利用式（4）、式（5）的样本可以近似达到真实后验分布 $p(x_t | z_{1:t}, u_{0:t-1})$ 。

PF 算法开始时从初始信度 $p(x_0)$ 采样，接着对每个时间步执行预测—更新—重采样递归循环过程。式（6）形式化 PF 的执行步骤，它源于应用 Bayes 规则到后验概率中，接着使用 Markov 假设，该公式从右到左执行。

$$\underbrace{p(x_t | z_{1:t}, u_{0:t-1})}_{\text{当前的PDF}} = \underbrace{\eta}_{\text{标准化因子}} \underbrace{\overbrace{p(z_t | x_t)}^{\text{测量方程}}}_{\text{更新}} \underbrace{\int \underbrace{p(x_t | x_{t-1}, u_{t-1})}_{\text{状态方程}} \underbrace{p(x_{t-1} | z_{1:t-1}, u_{0:t-2})}_{\text{先前的PDF}} dx_{t-1}}_{\text{预测}} \quad (6)$$

1.2 存在问题

粒子滤波 PF 基于 SIS 算法进行采样，会存在样本退化问题。文献[6]指出：重要性权重的方差随着时间增长而增大，因而样本退化是不可避免的。这样，大量的计算时间浪费在更新权重较小的样本上。针对样本退化问题，可以增加样本大小 N_p 至无穷来解决，但很低效。文献[7]提议的解决方案为：选择好的建议分布及进行重采样。

重采样在基于 SIS 粒子滤波算法起着关键作用，因为：第一如果权重的分布不均，通过动态系统传播权值小的样本是一种计算能力的浪费；第二当重要性权重倾斜，重采样可以提供选择“重要”样本和恢复样本以便将来使用。重采样操作在缓解权值退化问题的同时又增加了降低样本多样、增大运算量，以及引入额外的重取样方差等问题。

在重采样的过程中 那些健壮的粒子也可能被删掉，导致粒子退化，因此解决办法就是减少重采样的次数。另一个解决办法是通过监控每一步骤粒子的有效性，从而限制重采样步骤的次数，提高 PF 的鲁棒性。但是过少重采样又会导致粒子滤波的发散，因此，找一个标准进行判断何时重新采样是非常重要的。这也是粒子滤波多样性测度的一个问题。

2 粒子滤波多样性测度

2.1 有效样本大小

Liu^[8]提出的有效样本大小（Effective Sample Size, ESS），被用于应用于自适应重采样，可以估计当前的粒子集表示后验概率的准确度，有效样本大小 N_{eff} 定义为：

$$N_{\text{eff}} = \frac{N_p}{1 + \text{Var}(w_t^{*(i)})} \quad (7)$$

式中， $w_t^{*(i)} = p(x_t^i | z^t, u^{t-1}) / q(x_t^i | x_{t-1}^i, u_{t-1}, z_t)$ 是第 i 个粒子的真实权重； N_p 为粒子集中粒子数量 $\{x_t^i, w_t^i\}_{i=1, \dots, N_p}$ 。在实践中由于真实的 N_{eff} 难以计算，所以经常用到的是估计式 $\hat{N}_{\text{eff}}$ 。

$$\hat{N}_{\text{eff}} = \frac{1}{\sum_{i=1}^{N_p} (w_t^{(i)})^2} \quad (8)$$

式中， $w_t^{(i)}$ 为第 i 个粒子的归一化权重。

2.2 种群多样性因子

遗传算法中的种群多样性因子，被用于度量粒子滤波的粒子集^[9,10]多样性，种群多样性因子如下：

$$S = k(w_{\max} - w_{\text{av}}) / w_{\text{av}} \quad (9)$$

其中， $w_{\max}$ 是所有粒子权重的最大值； w_{av} 权值较高的前 50% 的平均值； k 是标度因子通常在 3~6 之间。 k 的选择是十分重要的，如果较小，不利于适应度高的个体繁殖，并且收敛速度也会减慢，相反，多样性就会降低。

2.3 粒子群多样性测度

粒子群优化算法中的多样性测度也可以用于粒子滤波的粒子集^[11,12]，粒子群多样性测度公式如下：

$$S = \frac{1}{|S| \times |L|} \sum_{i=1}^{|S|} \sqrt{\sum_{j=1}^D (x_{ij} - \bar{x}_j)^2} \quad (10)$$

式中， $|S|$ 为种群的大小； $|L|$ 是搜索空间最长的对角度线长度； D 是问题的维数； x_{ij} 是第 i 个粒子的第 j 维的值； $\bar{x}_j$ 是所有粒子的第 j 维的均值。

3 实验

为了对粒子滤波多样性测度进行评估，采用 MATLAB 设计了从固定视觉观测点进行单机动目标跟踪的仿真程序。目标跟踪模型采用 CV 模型，其状态向量 $X(t) = [x(t) \quad v_x(t) \quad y(t) \quad v_y(t)]^T$ 中，各参数分别表示 t 时刻目标在二维平面中的 x 坐标、 x 方向的速度、 y 坐标和 y 方向的速度； T 为采样时间间隔； $m(t) = [\alpha_x(t) \quad \alpha_y(t)]^T$ 是目标随机加速度看做是随机噪声作用的结果，这里为服从高斯分布的白噪声。目标跟踪系统的状态方程为：

$$X(t) = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} X(t-1) + \begin{bmatrix} 0.5T^2 & 0 \\ T & 0 \\ 0 & 0.5T^2 \\ 0 & T \end{bmatrix} m(t-1)$$

目标跟踪系统从固定视觉观测点获得的测量方程为 $Z(t) = \arctan(x(t)/y(t)) + n(t)$ 。其中， $Z(t)$ 为观测到的目标在极坐标上的航向角； $n(t)$ 表示服从高斯分布的白噪声。粒子滤波中的相对参数描述如下：粒子数目为 1000，最大的时间步骤为 100，重采样算法选择残差重采样，重采样的门限值设定为 N 和 $0.5N$ ，在下面的实验中粒子滤波的多样性测度包括有效样本大小，群体多样性因子，粒子群优化多样性测度。

实验结果显示，如果不执行重采样算法有效样本大小应该由一个高的值转变到另一个，要是执行重采样算法，有效样本大小应该保持在一个有效的间隔。因此，重采样步骤可以缓解权值退化的问题。由于重采样导致了增加额外的蒙特卡罗方差和计算复制制度，所以 Doucet^[6]提出的如何决定是否执行重采样的思想应该被应用。有效样本大小自适应值如图 1 所示，值分别为 627 和 569 在重采样的门限值为 N 和 $0.5N$ 时。也就是说，在门限值接近 N 时执行粒子滤波算法比在门限值等于 $0.5N$ 时执行要好。计算的代价则是相反的，如图 2 和图 3 所示。在执行种群多样性因子和粒子群多样性测度是也是同样的增加，和执行有效样本大小不同，后两者应该有低值向高值改变，如果不执行重采样算法，如

果执行，应该保持在一个有效的间隔。

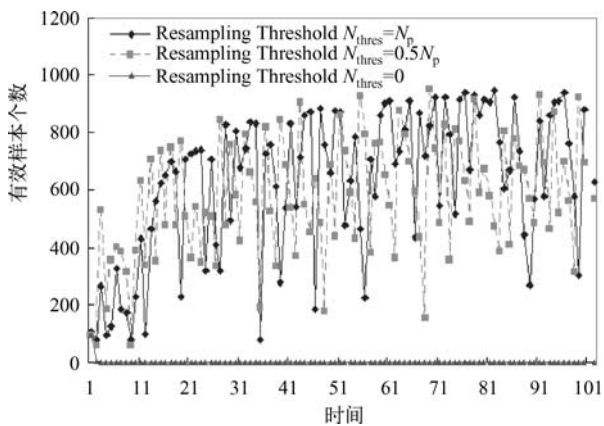


图1 有效样本大小执行分析

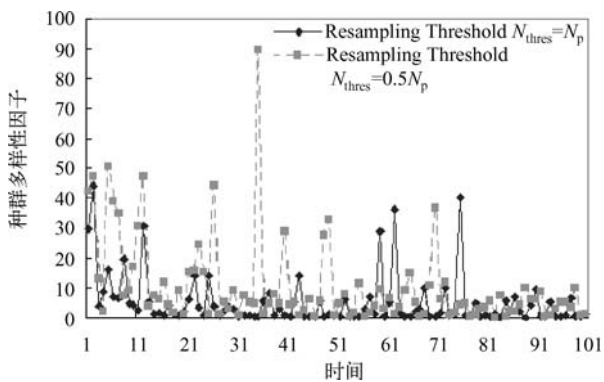


图2 种群多样性执行分析 (k=3)

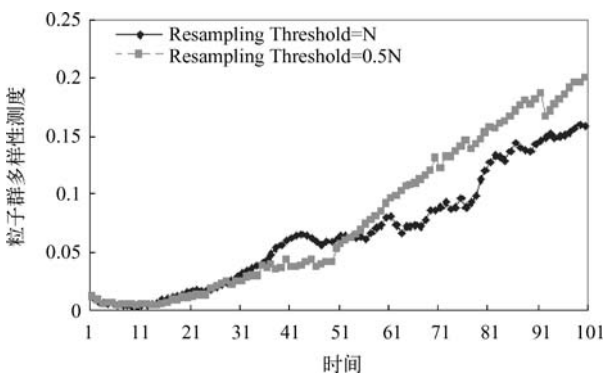


图3 粒子种群多样性测度实验

4 总结

有效的多样性测度对与粒子滤波重采样算法是十分重要的。在分析粒子滤波重采样算法固有的缺陷下，介绍了几个多样性测度，包括有效样本大小，群体多样性因子，粒子群优化多样性测度，最后采用 MATLAB7.0 设计从固定视觉观测点进行单机动目标跟踪的仿真程序，对这三种多样性测度在粒子滤波重采样执行过程中进行了评估。

参考文献

[1] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear /non-Gaussian Bayesian state estimation,[J] IEE Proceedings on Radar and Signal Processing, 1993,140(2), 107-113.

[2] F. Dellaert, W. Burgard, D. Fox, et al Monte Carlo localization for mobile robots Proceedings of the IEEE International Conference on Robotics and Automation (ICRA1999), Detroit, USA: IEEE Press,.1322-1328, 1999.

[3] D. Schulz, W. Burgard, D. Fox, et al, “Tracking multiple moving targets with a mobile robot using particle filters and statistical data association,” Proceedings of the 2001 IEEE International Conference on Robotics and Automation (ICRA2001) , Seoul, Korea: IEEE Press, 2001,1665-1670.

[4] M. Montemerlo, S. Thrun, D. Koller, et al, FastSLAM: A factored solution to simultaneous localization and mapping problem,” Proceedings of the National Conference on Artificial Intelligence, Edmonton, Canada: AAAI Press, 2002, 593-598.

[5] V. Verma, G. Gordon, R. Simmons, et al, Particle filters for rover fault diagnosis, IEEE Robotics & Automation Magazine special issue on Human Centered Robotics and Dependability,. 2004.11(2).1-4.

[6] A. Doucet, S. J. Godsill, and C. Andrieu On sequential Monte Carlo sampling methods for Bayesian filtering[J Statistical and Computing, 2000,10(3), 197-208.

[7] M. S. Arulampalam, S. Maskell, N. Gordon, et al A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking,[J], IEEE Transactions on Signal Processing, 2002,.50(20), 174-188.

[8] J. S. Liu, Metropolized independent sampling with comparisons to rejection sampling and importance sampling[J],Statistical and Computing,. 1996,6(1), 113-119.

[9] 杨振强, 王常虹, 庄显义. 自适应复制交叉和突变的遗传算法 [J] 电子科学学刊. 2000, 22(1), 112-117.

[10] 段琢华. 基于自适应粒子滤波器的移动机器人故障诊断理论与方法研究[D]. 模式识别与智能系统, 中南大学, 2007.

[11] J. Riget, and J. S. Vesterstroem . “A diversity-guided particle swarm optimizer-the AR PSO,” Aarhus, Denmark: Department of Computer Science, University of Aarhus, 2002.

[12] W. Jiao, and G. B. Liu. “Particle swarm optimization algorithm based on diversity feedback,” Computer Engineering , 2009.35(22) 202-204.

基于 CPSS 算法的 RTAI 调度器的改进

李学桥，梁 爽，陈 园

(郑州轻工业学院，计算机与通讯工程学院，河南 郑州，450002)¹⁰

摘要：在实时系统中，任务调度策略是内核设计的关键部分。如何进行实时的任务调度，使任务能在特定的周期内完成是实时操作系统领域研究的一个热点问题。本文将一种基于 RM 算法的改进算法 CPSS 算法引入 RTAI 调度器中，针对 RTAI 调度器在系统过载情况下出现调度性能下降等缺点，对 RTAI 调度器进行优化和改进。对改进后的调度器在调度时延方面和调度算法仿真方面进行了测试，实验证明了改进后的调度器能够提高 Linux 系统的实时性。

关键字：实时系统；任务调度；RTAI；调度器；CPSS 算法

中图分类号：TP316.2 **文献标识码：**A

The Improvement of RTAI Real-time Scheduling Algorithm and Scheduler

LI Xue-qiao, LIANG Shuang, CHEN Yuan

(College of Comp. and Com. Eng., Zhengzhou Univ. of Light Ind., Zhengzhou 450002, Henan China)

Abstract: In real-time system, task scheduling policy is the key part of kernel design. How to design task scheduling to ensure that all the tasks will be completed before its deadline is an important problem in the research on real-time operating system. Because the scheduler of RTAI has a bad performance when the system is in heavy load or overload, this article adds a Comprehensive Priority Static Schedule (CPSS) algorithm to RTAI scheduler, and improves and optimizes the RTAI scheduler. Finally, a simulator of the scheduling algorithms and a test of the scheduler are presented. The experiment has proved that improved scheduler can increase the real time Linux system.

Keywords: real-time system; task scheduling; RTAI; scheduler; CPSS algorithm

1 引言

随着计算机技术和电子技术的快速发展，嵌入式产业也随之迅速崛起，成为当今社会人们最关注的产业之一。而在嵌入式产业中，Linux 操作系统以其源代码开放，内核的可移植性，高效性，健壮性，高效的网络通信支持能力和丰富的开放应用软件等，已经成为了最主流的操作系统之一，上述优点也使得 Linux 系统在嵌入式领域得到了快速发展。但 Linux 系统是典型的分时操作系统，所以必须对 Linux 系统进行改进以更好的使用嵌入式实施应用的需要。目前有两种对 Linux 系统进行实时化改造的方案：对内核直接修改内核和双内核^[1]。而修改内核只能达到软实时的目的，因此本文使用双内核方案对 Linux 系统进行实时化改造。

2 RTAI 方案及其实时调度策略

RTAI (Real-Time Application Interface) 是对 Linux 内核的硬实时扩展，它遵循自由软件规范；它可以提供工业级的 RTOS 功能，而且其所有的功能都可无缝地通过 GNU/Linux 环境访问。RTAI 项目是由意大利米兰理工学院航天工程系 (DIAPM) 开发的遵循 GPL 的开源项目。目前 RTAI 已经支持 x86、PowerPC、ARM、MIPS、CRIS 等处理器，是目前支持处理器最多的 Linux 硬实时解决方案之

作者简介：李学桥（1951—），男，教授，硕士生导师，CCF 会员（E200008277S），主要研究方向：嵌入式系统与智能控制、计算机网络。
• 48 •

一。已经有很多工业控制系统采用 RTAI 作为运行支撑环境，它也表现出可以与商业实时操作系统相媲美的性能^[2]。

RTAI 还并不能算是一个完整的操作系统，而只是一个具备操作系统核心的实时内核。在双内核系统中，RTAI 代替 Linux 接管了所有的硬件资源。RTAI 调度器把 Linux 系统内核作为一个最低优先级的任务，只有当实时任务都运行完之后才能运行。

RTAI 调度器主要采用的是 RM 调度算法和 EDF 调度算法^[3~5]。RM 调度算法是一种静态的调度算法，算法简单，实现容易，但 CPU 利用率不高。EDF 算法是一种动态的调度算法，CPU 利用率高，但经常检查不出，即将错过截止期的任务，导致大量任务错过其截止期，引起系统性能下降。如今实时任务在嵌入式领域应用的越来越广泛，因此对实时任务调度器提出了很高的要求^[6]。

3 基于 RM 算法的改进调度算法——CPSS 算法

RM 调度算法是一种非常经典的静态周期任务调度算法。该算法简单，容易实现但并不容易实现最紧急的任务。本文介绍了一种基于综合优先级的静态调度算法（Comprehensive-Priority Static Schedule，CPSS），CPSS 算法是基于“综合优先级”这个参数，这里的综合优先级是结合了任务的运行时间 E_i 和重要程度 I_i 两个特征参数的综合优先级。

记任务 τ_i 的重要程度为 I_i ，再记任务 τ_i 的运行时间为 E_i ，定义任务 τ_i 的综合优先级为 P_i ，则：

$$P_i = P(\tau_i, I_i, E_i) = W_i I_i + W_e E_i$$

式中， W_i ， W_e 分别为重要程度和运行时间的权值，且 $W_i > W_e$ ，其中 $W_i + W_e = 1$ ， P_i 越大，优先级越高。

综上所述，系统的综合优先级是通过结合算法中的两个参数（重要程度 I_i 和周期 E_i ）计算得出的值，而不仅是取决于算法的周期。

算法中，将队列按综合优先级的排序分为两个队列，关键任务集 S 和非关键任务集 N，优先对关键任务集 S 的任务进行调度。首先判断 S 队列中是否还有任务，如果有，则按照 FIFO 算法继续进行下一轮的调度，如果没有，则将 N 队列中的综合优先级最高的任务放入 S 队列中继续执行。执行直至完成。本算法优先对 S 队列中的任务进行调度。只有当 S 队列中没有可以运行的任务时才对 H 队列进行调度，据此可以保证系统优先调度关键任务集 S 中的任务。

改进的 CPSS 调度算法，即延续了 RM 调度算法实现简单等优点，又改进了 RM 算法经常忽略重要任务的缺点。使一些周期长但重要程度很高的任务能尽快得到调度，改进了 RM 算法。

4 针对 CPSS 改进算法的 RTAI 调度器的改进

针对 CPSS 算法对 RM 算法的改进，RTAI 调度器需要从数据结构，任务队列管理和定时器设置三部分进行改进。

4.1 对数据结构的改进

任务控制块是 RTAI 中最重要的一个数据结构体。RTAI 与所有的进程管理程序一样，对于每一个实时任务都有一个任务控制块。任务控制块就像进程的“户口”一样，记录着进程的各种属性和进程占用的各项资源^[6]。在 `rtai_sched.c` 中以结构体 `rt_task_struct` 表示任务的控制块。

改进前的 RTAI 调度器任务控制块的主要数据结构有：`base_priority` 表示任务的优先级；`period` 表示任务的周期，`policy` 表示的是调度策略，其中大于 0 表示使用 RM 算法，小于 0 表示使用 EDF 算法，改进的 CPSS 算法对 `policy` 不好再扩展，因此需要再引进一个数据结构体表示是否使用 CPSS 算法。具体程序段如下：


```

typedef struct rt_task_struct {
    ...

    /*Append for CPSS algorithm*/
    int CPSS_policy;           //程序是否使用 CPSS 算法
    int CPSS_importance;       //任务的重要程度；值越大任务越重要
    int CPSS_compre_priority;   //任务的综合优先级
    int Wi, We;               //分别为重要程度和周期的权值
    RTIME CPSS_sched_deadline; //实时任务调度的相对截止期
    RTIME CPSS_exec_time;      //执行完当前任务需要的时间
    RTIME CPSS_current_exec_time; //任务的当前执行时间，随任务执行而变化
    /* Append for CPSS algorithm end*/
    ...
}

```

4.2 对任务队列管理的改进

RTAI 内核中本身存在着三种任务调度队列，分别为任务队列、任务延时队列和任务就绪队列。在 RTAI 内核中本身存在着一些函数，它们分别为：

```

enq_ready_task();           //将任务按照 RM 算法插入就绪队列
enq_ready_edf_task();       //将任务按照 EDF 算法插入就绪队列

```

本文引入 CPSS 算法，如果在原有的任务队列管理中新加入 CPSS 算法插入就绪队列，那样会更改很多原有操作，本文将 CPSS 算法与 RM 算法相结合，将 RM 算法本身比较任务的优先级语句改为比较 CPSS 算法中通过计算出来的综合优先级，以此来确定是否进行调度。通过在 enq_ready_task() 中增加比较综合优先级的语句来进行改进。具体扩展后的代码如下：

```

static inline void enq_ready_task(RT_TASK *ready_task)
{
    RT_TASK *task;

    ...

    /*modify for CPSS algorithm*/
    CPSS_compre_priority= Wi • CPSS_importance+ We • CPSS_exec_time; // 计算出任务的综合优先级
    if(ready_task->CPSS_policy=0) // 确定是否使用 CPSS 算法，当 CPSS_policy 为 0 时，则确定使用 CPSS
                                   算法进行调度
    {
        while (ready_task->CPSS_compre_priority >= task-> CPSS_compre_priority)
        {
            if ((task = task->rnext)->priority < 0) break;
        }
        task->rprev = (ready_task->rprev = task->rprev)->rnext = ready_task;
        ready_task->rnext = task;
    } // 比较任务的综合优先级，从链表中进行插入工作
    /*modify for CPSS algorithm end*/
    ...
}

```

4.3 对定时器部分的改进

RTAI 中定时器有两种工作模式，分别为一次性模式(ONE_SHOT_MODE)和周期模式(PERIODIC_MODE)。

由于此调度器采用的是 `one_shot` 模式的定时器中断机制，因此需要设置定时器的下一次中断时刻。由于 `RM` 算法是经典的静态调度算法，因此需要调度在任务不会抢占当前正在调度的任务，而 `CPSS` 算法虽解决了 `RM` 算法经常错失周期长的任务的缺点，但也是根据其综合优先级的大小排序来进行调度。因此和 `RM` 算法一样的是，在当前运行任务的 `possible_runtime` 这个时间片中，是不允许另外任务抢占的。因此定时器的中断时刻为 `possible_runtime` 时间片的末端时刻，如图 1 所示。

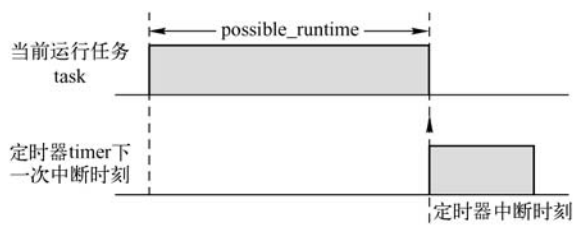


图 1 定时器下一次中断时刻图

5 R TAI 调度算法验证及其调度器的仿真与测试

5.1 改进后的算法验证

5.1.1 算法对少数实时任务的验证

验证目标：测试 `CPSS` 算法是否针对 `RM` 算法更能使一些周期长但重要性高的任务尽快得到调度。

验证方法：通过给出任务的运行时间 `Runtime` 和重要程度 `Importance` 来计算任务的综合优先级 `Compre_Priority`。对任务进行排序调度，与原先的 `RM` 算法的调度后的 `DMR` 进行比较。由于 `RM` 算法在轻载情况下能达到较好的效果，则系统直接在重负载情况下进行测试。

验证指标 `DMR` (`Deadline Missing Radio`):

截止期错失率 (`DMR`) 是衡量一个调度算法调度好坏性能的重要参数，其中错过其截止期的作业数为 n ，正常完成的作业数为 m ，则 $DMR = n/n+m^{[7]}$ 。

首先创建 4 个实时周期线程，具体线程参数如表 1 所示。

表 1 实时线程参数

TASK Period/ms		Runtime/ms	Importance	Compre_Priority
P ₁ 40		20	1	6.7
P ₂ 40		15	3	6.6
P ₃ 60		15	2	5.9
P ₄	60 20		4	8.8

任务的执行性能如表 2 所示。

表 2 实时线程的平均执行性能

TASK	RM algorithm	CPSS algorithm	DMR
P ₁ 100%		100%	RM algorithm:
P ₂ 100%		100%	75.3%
P ₃ 45%		83.3%	CPSS algorithm:
P ₄ 0		100%	34.2%

验证结果：通过在重载情况下对 4 个实时线程的调度来进行对比，发现 P₃、P₄ 两个进程因为周期长因此优先级较低，所以执行 RM 算法时执行性能比较差，而在 CPSS 算法中，通过计算综合优先级，P₄ 这个原本周期长的任务因为重要程度高而综合优先级高，因此率先得到调度，表 2 各项数据显示，CPSS 算法改进了 RM 算法对周期长的任务的调度差的缺点，从而有效地改进了 RM 算法。

5.1.2 算法对大量实时任务的验证

因为 RM 算法和 EDF 算法在轻载时 DMR 都能达到较好的效果，因此取系统负载 P 从 0.6~1.6。比较新的算法在 RTAI 中响应 10000 个任务的截止期错失率，具体仿真结果如图 2 所示。

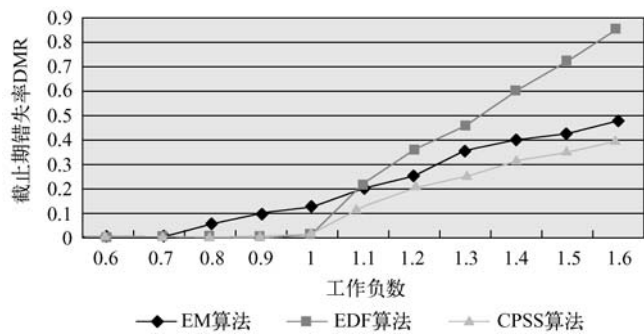


图 2 截止期错误率 DMR 比较图

仿真结果分析：图 2 中任务错失率方面，静态调度算法 RM 算法在负载为 0.8 时第一个出现了错误，而另外的两个算法在负载小于 1 时始终保持错失率为 0，而当负载大于 1 之后，EDF 算法率先出现错失率陡升的情况，而且 DMR 最高达到 85%，而 RM 算法虽然并未出现陡然上升的局面，错失率却不断上升，而 CPSS 算法无论在负载小于 1 或者大于 1 的情况下，错失率均在可承受的范围之内。由此可见，CPSS 算法相对于 RM 算法和 EDF 算法是有改进的。

5.2 改进后的调度器仿真

实时系统的实时性能指标中最主要的指标是任务响应时间。而测试调度的最主要指标是调度器时延。所谓调度器时延是指求出的期望的调度点与实际调度点之间的时间差值^[6]，图 3 中显示了调度器时延的定义，delay 即为调度器时延。

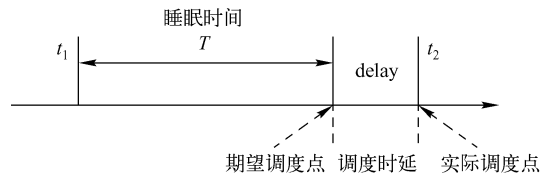


图 3 调度时延图

测试结果：
测试时选取不同的 T 值，分别测试了 T=100μs，T=200μs 时的调度时延，测试结果与分析如图 4、图 5 所示。

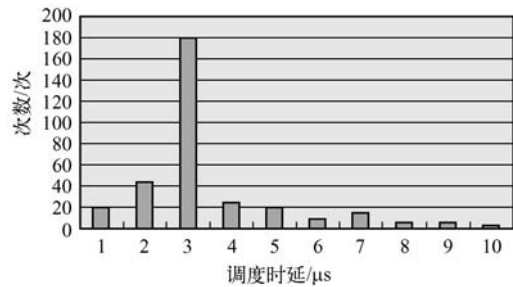


图 4 T=100μs 调度时延图

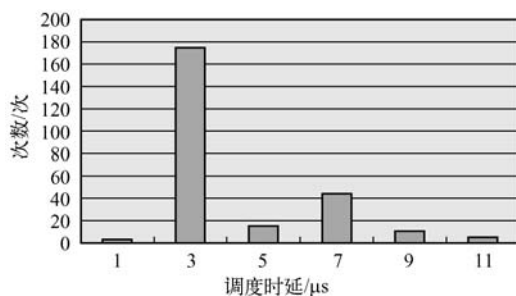


图 5 $T=200\mu s$ 调度时延图

实验显示，改进后的 RTAI 调度器的调度时延在 $2\sim 7\mu s$ 之间，大部分的调度时延集中在 $3\mu s$ 时刻，由此可以看出，改进后的调度器调度时延相对于任务的周期值是非常微不足道的，完全能够满足硬实时任务的要求。

综上所述，RTOS 本身的 RM 算法本身在任务错失率方面较之于动态算法 EDF 仍然具有较大差距，在负载小于 1 时即出现错失，而且容易错失任务周期长且重要度高的任务。而 EDF 算法虽然是典型的动态调度算法，但其缺点也明显，即在强负载情况下错失率大大增加。改进的 CPSS 静态调度算法较之 RM 算法增加了综合优先级的计算，即使周期长且重要的任务更易得到调度。且通过比较 DMR 和调度时延两个参数，发现 CPSS 算法在调度算法和调度器方面均优于典型的 RM 算法。

6 结语

本文首先分析了 Linux 操作系统的实施性改造方案——双内核 RTAI 的工作原理，针对 RTAI 中的静态调度算法 RM 算法提出了一个改进算法 CPSS 算法，并重点针对算法设计了 CPSS 调度器及其利用程序将其实现。通过仿真，对 CPSS 调度器进行了仿真和调度时延的测试。实验证明，CPSS 调度器可以在系统负载较重时仍然具有良好的实时调度性能，适合硬实时调度的要求。

参考文献

- [1] 汤子赢，哲凤屏，汤小丹. 计算机操作系统[M]. 西安：西安电子科技大学出版社，2001.
- [2] Raj Rajkumar, Kanaka Juvva, Anastasio Mola-no, et al. Resource kernel: A resource-centric approach to real-time and multimedia systems [A]. Proceedings of the SPIE/ACM Conference on Multimedia Computing and Networking, 1998
- [3] C.L.Liu, J.W.Layland. Scheduling algorithms for multi-programming in a hard real-time environment. Journal of the ACM, 1973, 20(1): 46-61.
- [4] Lehoczky J, Sha L, Ding Y. The rate monotonic scheduling algorithm: exact Characterization and average case behavior[C]. Santa Monica, CA: Proceedings 10th IEEE Real-Time Systems Symposium, 1989: 166-171.
- [5] Mok A K. Fundamental design problems of distributed systems for the hard-real-time environment [D]. Computer Science, MIT, 1983.
- [6] 王创社，周树杰. RTAI 实时调度器的优化与实现[J], 北京石油化工学院学报, 2007, 3(15): 51-55.
- [7] 王济勇，赵海，林涛等. 定时器驱动的 RM 调度机制建模及其性能优化[J]. 计算机学报, 2005, 28(2): 161-169.

改进的 UIO 序列生成算法

黎中文¹, 张来顺¹, 肖健鹏²

(1.解放军信息工程大学电子技术学院, 河南 郑州 450004; 2.中国人民解放军 65012 部队, 辽宁 沈阳 110101)

摘要: 有限自动机 (FSM) 被广泛用对软硬件进行建模。唯一输入/输出序列 (UIO Sequence) 用于构造测试序列来验证 FSM 是否到达某一状态。UIO 序列寻找过程就是构造一棵 UIO 树, 但是构造过程中使用的分解节点方法会产生许多不需要的节点。在结合已有的广度优先搜索算法和搜寻过程中构造 UIO 树的方法、本文提出将 UIO 树的节点标记为“需要”和“不需要”状态的方法, 以降低搜索空间, 加快 UIO 序列生成过程。该算法与以往算法相比, 在时间复杂度和空间复杂度两方面有较大改进。

关键词: UIO 序列; 剪枝; 标记状态; 有限自动机

中图法分类号: TP391 **文献标识码:** A

Improved UIO Sequences Generation Algorithm

LI Zhongwen¹, ZHANG Laishun¹, XIAO Jiangpeng²

(Institute of Electronic Technology of the PLA Information Engineering University, Zhengzhou 450004, Henan China; China PLA TROOP 65012, Shenyang 110101, Liaoning China)

Abstract: Finite-state-machines (FSMs) model a large of hardware and software systems. Unique input/output (UIO) sequences are used in generation of test sequences to verify that a machine is at a certain state. UIO sequence search process is to construct a UIO tree, but the structure used in the partitioning node method will produce many unwanted nodes. With the existing breadth-first search algorithm and the search method in constructing UIO tree, the paper proposed a method by classifying nodes as “necessary” and “unnecessary” state to reduce the search space, speed up the UIO sequence generation process. Comparing with previous algorithms, both of time complexity and space complexity have a greater improvement.

Keywords: UIO sequences; pruning; marked state; finite-state-machine

1 引言

有限自动机 (FSM) 模型被广泛应用于软、硬件领域, 如数字控制系统中的序列电路、微处理芯片、模式匹配和协议一致性测试。对 FSMs 进行测试用于保证系统符合规范是非常必要的。文献 [1, 2] 中对 FSMs 的测试方法进行了详细介绍, 这些测试方法主要使用区分序列 (Distinguishing Sequence) 和唯一输入/输出序列 (UIO Sequence) 来验证 FSMs 的初始状态和终止状态。UIO 序列可将一个状态与其他状态区分开。有区分序列的 FSM 的所有状态都有 UIO 序列; 反之则不同, 因此 UIO 序列比区分序列的应用范围更广。

基于 FSM 的 UIO 序列寻找过程就是构造一棵 UIO 树, 通过不断分解树的节点直至找到仅包含一个状态的节点, 那么寻找某个状态的 UIO 序列就是在 UIO 树中寻找仅含该状态的节点, 通过宽度优先搜索策略, 可以找到其最短路径。但随着 FSM 的状态数、输入向量位数等参数的增长, 必然引起 UIO 树的规模增大, 求解 UIO 序列的复杂程度增加。本文通过将 UIO 树中的节点标记为“需要”和“不需要”状态, 从而加速 UIO 序列输出, 并减小 UIO 树的规模。

作者简介: 黎中文 (1985—), 男, 硕士;
张来顺 (1963—), 男, 教授, 硕士;
肖健鹏 (1979—), 男, 硕士。

2 基础知识

定义 1 一个有限自动机 FSM 是一个五元组 $M=(I, O, S, \delta, \lambda)$, 其中 $I=\{i_1,i_2,\cdots,i_p\}$ 是有限非空输入集合, $O=\{o_1,o_2,\cdots,o_p\}$ 是有限非空输出集合, $S=(s_1,s_2,\cdots,s_n)$ 是 FSM 的状态集合, $\delta:I^*S\rightarrow S$ 是状态转换函数, $\lambda:I^*S\rightarrow O$ 是输出函数。

定义 2 一个 FSM 的状态节点 $N=(s_1s_2 \cdots s_n, s_1's_2'\cdots s_n',x/y)$, 其中 $s_i, s_i'\in S$, 对于任意 $i(1\leq i\leq n)$, $\delta(s_i, x)=s_i', \lambda(s_i,x)=y$ 。

定义 3 对于任意 I/O 序列 $a/b, N=(s_1s_2 \cdots s_n, s_1's_2'\cdots s_n',x/y)$, 则 $N'=\lambda(N, a)=(s_1s_2 \cdots s_n, s_1''s_2''\cdots s_n'',x_a/y_b)$, 其中 $s_i'=\delta(s_i, x), s_i''=\delta(s_i', a), 1\leq i\leq n$ 。

定义 4 给定一个 FSM 的状态节点 $N=(s_1s_2 \cdots s_n, s_1's_2'\cdots s_n',x,y)$, 若 $n=1$, 即 N 仅包含一个元素, 则称 N 是单元素节点。

定义 5 给定一个 FSM 定义一个 n (n 为状态的个数) 行 m 列 (用 I/O 标记列) 转换表格 T , 如果有 $\lambda(s_i,x)=y$ 和 $\delta(s_i,x)=s_j$, 则标记第 i 行标记为 x/y 列 $T(I,x/y)=s_j, 1\leq i,j\leq n, x\in I, y\in O$, 不存在转换函数的表格位置填 0。

定义 6 给定一个 FSM 的状态节点 $N=(s_1s_2 \cdots s_n, s_1's_2'\cdots s_n',x/y)$, 如果 $s_i'=s_j', 1\leq i,j\leq n$, 则称 s_i, s_j 在节点 N 中是会聚状态。

3 算法描述

生成 UIO 序列的过程可简单描述如下: (1) 对于每个状态先寻找其长度 $L=1$ 的序列, 其中 L 为 I/O 序列的长度; (2) 检查每个状态是否有 UIO 序列; (3) 如果没有, 则继续查找长度为 $L=2$ 的 I/O 序列, 检查其是否有 UIO; (4) 重复 $L=L+1$ 直到找到所有状态的 UIO 序列。

输入: 有限自动机 (FSM)
输出: 每个状态的 UIO 序列
步骤:

```
根据 M 得到一个 n 行 m 列的转换表格 T;  
UIOi=Φ(1≤i≤n); // UIOi 为状态 Si 的 UIO 序列  
VS={ (s1,s2,⋯,sn) }; //VS 为未处理状态节点的队列  
While(找到所有 UIOi 或者 VS 为空) {  
    将 NS 中的第一个元素删除, 并将被删除元素赋给状态节点 N;  
    for(1 ≤i≤n); //取转换表格 T 中每一个 I/O 序列对 N 进行运算。  
    {  
        N'=λ(N,xi);  
        If(N' 是单元素节点) {  
            计算出节点 N'的起始状态的 UIO。  
            将该节点标记为“不需要”, 丢弃。  
            对节点中的状态进行剪枝, 去掉 NS 中所有元素的 Si 状态。  
        }  
        Else {  
            删除节点 N'中的会聚状态;  
        }  
        If(N' 非空) {  
            将该节点标记为“需要”,  
            NS=NS+ N';  
        }  
    }  
}
```

}
}
}

4 算例分析

所给定的有限状态机如图 1 所示。

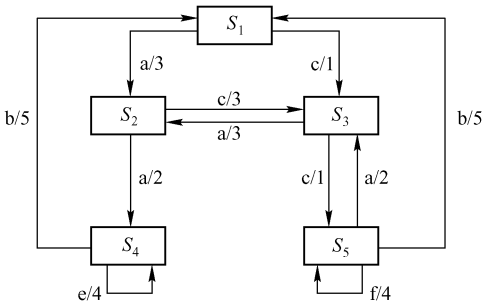


图 1 给定的有限状态机

首先根据图 1 生成一个 5 行 7 列的转换表格 T，如表 1 所示。

表 1 生成的一个 5 行 7 列的转换表格

	a/3 a/2 c/3	b/1 e/4				f/4	b/5
S ₁	S ₂ 0		0	S ₃ 0		0	0
S ₂ 0		S ₄	S ₃	0 0 0 0			
S ₃	S ₂ 0		0	S ₅ 0		0	0
S ₄	0 0 0 0				S ₄ 0		S ₁
S ₅	S ₃	0 0 0 0				S ₅	S ₁

根据上述算法可得到寻找 UIO 序列生成过程如下：

初始状态下 $NS=\{(S_1, S_2, S_3, S_4, S_5)\}$, $UIO_i=\Phi$ ，根节点为 $NS(12345,12345,null)$ 。

将 NS 中第一个元素赋给 N 后删除，此时 $NS=\Phi$ ， $N=(S_1, S_2, S_3, S_4, S_5)$ 。依据表格中每一列对 N 进行 $\lambda(N,x_i)$ 运算，得到 7 个新的节点 $(135,223,a/3)$ ， $(2,4,a/2)$ ， $(2,3,c/3)$ ， $(13,35,c/1)$ ， $(4,4,e/4)$ ， $(5,5,f/4)$ ， $(45,11,b/5)$ ，如图 3 所示。其中 $(2,4,a/2)$ ， $(2,3,c/3)$ ， $(4,4,e/4)$ ， $(5,5,f/4)$ 为单元素节点，则可得到 3 个 UIO 序列， $UIO_2=a$ ， $UIO_4=e$ ， $UIO_5=f$ ，并同时将这些节点标记为“不需要”。对所有节点进行剪枝，去掉已求出 UIO 序列的状态 S_2 、 S_4 、 S_5 和会聚状态，得到精简的节点 $(5,3,a/3)$ ， $(13,35,c/1)$ ，将它们标记为“需要”。重复以上步骤，得到新的节点 $(5,2,aa/33)$ ， $(5,5,ac/31)$ ， $(13,23,ca/13)$ ， $(1,5,cc/11)$ 等，如图 2 所示，其中 $(5,2,aa/33)$ ， $(5,5,ac/31)$ ， $(1,5,cc/11)$ 为单元素节点，则可得到剩余 2 个状态的 UIO 序列， $UIO_5=aa$ ， $UIO_1=cc$ 。

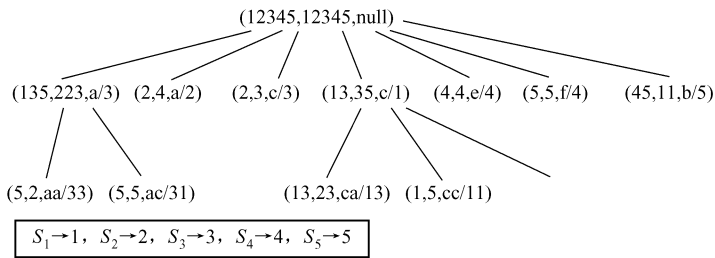


图 2 寻找 UIO 序列生成过程

5 结论

本文提出将 UIO 树中的节点标记为“需要”和“不需要”状态，减少了 UIO 树中节点和状态的数量，从而加速 UIO 序列输出，并减小 UIO 树的规模，对提高算法的效率有一定意义。

参考文献

- [1] Sidhu, D.P., and Leung, T. Formal methods for protocol testing: a detailed study. IEEE Trans. Softw. Eng. 15, (4), 413 - 426, 1989.
- [2] Lee D. and Yannakakis, M. Principles and methods of testing finite state machine —a survey. Proc. IEEE, 84, (8), 1089 - 1123, 1996.
- [3] Sun, H., Gao, M., and Liang, A. Study on UIO sequence generation for sequential machine's functional test. Proc. 4th Int. Conf. on ASIC, Shanghai, China, 22-25, 628-632, October 2001.
- [4] I. Ahmad, F.M. Ali and A.S. Das. LANG-algorithm for constructing unique input/output sequences in finite-state machine. IEE Proc.-Comput. Digit. Tech., Vol. 151, No. 2, March 2004.

一个基于 IB 原理的单类算法——OCD-B 算法

王媛媛，叶阳东

(郑州大学信息工程学院，河南 郑州，450052)

摘 要：单类问题是实际应用中经常遇到的分类问题，本文首先对单类问题的概念和研究现状做了简要的说明，然后深入阐述了如何将 IB 原理的方法论应用于单类问题的求解过程，最后从理论和算法等方面详细介绍了一个基于 IB 原理的单类算法——OCD-B 算法，并将之与传统单类算法 OC-Convex 算法进行对比和分析，分析结果表明：OCD-B 算法避免了 OC-Convex 算法中远离中心位置的部分实例被忽略的情况的出现，且 OCD-B 算法的性能优于 OC-Convex 算法。这些理论分析为进一步研究单类问题、延展 IB 原理的应用领域奠定了基础。

关键词：IB 原理；sIB 算法；单类；OC-Convex 算法；OCD-B 算法

中图分类号：TP391 **文献标识码：**A

A Algorithm for One-Class Based on IB Principle——OCD-B

WANG Yuanyuan, YE Yangdong

(School of Information Engineering, Zhengzhou University, Zhengzhou 450052, Henan China)

Abstract: One-Class Problem is common and very important in classification fields. This paper briefly describes the conception and current research of One-class problem, then clearly illustrates the process how to solve one-class problem using IB principle, at last introduces OCD-B algorithm based on IB principle from aspects of theory and algorithms. Compared with a traditional One-Class algorithm, OC-Convex algorithm, OCD-B can avoid the problem that in OC-Convex many instances far from the center are ignored, moreover, OCD-B performs better than OC-Convex. The above-mentioned theoretical analysis provides theoretical support for further study of One-Class problem and extends the application areas of IB principle.

Keywords: IB principle; sIB algorithm; One-Class; OC-Convex algorithm; OCD-B algorithm

1 引言

声音识别系统在许多领域被广泛应用，该系统主要用于辨别被测声音是否来自于某一目标声源，或从大量被测声源中识别出某一目标声源。由于目标声源的特征是从单一数据源训练得到的，因此该应用问题不同于已有的分类问题^[1]，故将该种类型的问题——训练数据来源于单一数据源，称为单类问题（One Class or Uniclass Problem）^[2]。实际应用中也经常遇到不同形式的单类问题，如用户在检索信息时，不仅需要从大量信息中找到某一类信息，而且通常更加关注检索结果中相关性最高的那些信息；又如基于内容的垃圾邮件过滤系统，可以从大量邮件中过滤出垃圾邮件，则检索“相关性最高的那些信息”和过滤“垃圾邮件”即为单类问题。

目前解决单类问题的方法有以下几种：（1）使用一个凸函数作为损失函数，被划分在单类内的实例的函数值为一个常量，单类外实例的函数值线性变化^[3]；（2）基于支持向量机（SVM）理论，用一个超平面对实例集进行划分^[4]；（3）基于信息瓶颈（Information Bottleneck，IB）原理，将单类问题描述为一个最优化问题，通过寻找一个合适的压缩代表来保留更多互信息^[5,6]。

本文介绍并分析了基于 IB 原理的单类算法：OCD-B 算法^[6]。该算法采用率失真理论的观点，

作者简介：王媛媛（1985—），女，在读，硕士研究生；

叶阳东（1962—），男，教授，博士生导师。

将单类问题描述成一个对实例进行编码的过程。也就是说，当实例属于单类时，用 0 对该实例进行编码，且用实例与质心间的距离作为对该实例进行编码后的失真，当实例不属于单类时，用其自身作为编码，同时将其失真记为零。此编码过程涉及一个求最优解的问题，解决该问题借鉴了 sIB 算法的思想，本文将深入阐述如何将 sIB 思想及 IB 原理的方法论应用于对单类问题的求解过程。

2 IB 原理

2.1 IB 原理

IB 原理是由 Tishby、Pereira 和 Bialek 于 1999 年提出的一种数据挖掘方法^[7,9]。该方法源于香农的率失真理论，率失真理论在给定的码率下寻求源变量到压缩变量的编码方案，使这种编码所产生的期望失真小于指定值，该过程中失真值是失真度量函数的数学期望，这种失真函数的定义缺乏客观性。IB 原理在对源变量进行压缩时，通过定义源变量 X 的相关变量 Y ，推导出一个合理的失真度量函数，从而有效地解决率失真理论存在的失真函数难以确定的问题。

IB 理论的思想可以概括如下：设联合分布 $(X,Y)\sim p(x,y)$ ，然后定义变量 X 的压缩代表 T ，使得在压缩 X 到 T 的过程中可以尽可能多的保持 X 关于 Y 的相关信息，即试图找到当互信息 $I(T,Y)$ 不小于某一失真值时互信息 $I(X,T)$ 的最小值，可将其形式化描述为下式^[7]， β 为权衡参数：

$$F_{\min}(p(t|x)=I(T,X)-\beta I(T,Y) \tag{1}$$

2.2 sIB 算法

sIB 算法由 Slonim 于 2002 年提出^[8]，该算法执行一个基于划分的聚类过程。相对于其他 IB 算法，sIB 算法具有较低时间和空间复杂度且可以确保得到问题的局部最优解，这些优点使 sIB 算法有广泛的应用空间。

算法初始时，将源变量 X 随机地划分成若干个聚类 $t, t\in T$ 。算法的循环过程为：将一个数据元素 $x\in X$ 从当前所属的聚类 t 中抽取出来后，将之合并进 $t_{\text{new}}=\operatorname{argmin}_{t\in T}d(\{x\},t)$ ，从而得到一个新的划分 T_{new} ，若 $t\neq t_{\text{new}}$ ，则互信息 $I(T,Y)$ 将会增加。当没有新的划分能够使互信息 $I(T,Y)$ 继续增加时，循环过程停止。

3 OCRD-BA 算法

Koby Crammer 等人于 2008 年提出了 OCRD-BA 算法^[6]。该算法基于 IB 原理将单类问题描述成一个对实例进行编码的过程，当实例属于单类时，用 0 对该实例进行编码，且用实例与质心间的距离作为对该实例进行编码后的失真，当实例不属于单类时，用其自身作为编码，同时将其失真记为零。此编码过程涉及一个求最优解的问题，在对最优化问题的求解过程中，该算法借鉴了 sIB 算法求解最优化的思想，经过迭代最终得到局部最优解。

3.1 算法思想

OCRD-BA 算法的度量函数表示如下：

$$\min_{\{q(0|x),W\}} I(T,X)+\beta D(W,q(0|x)) \tag{2}$$

整理可以得到

$$\min_{\{q(0|x),W\}} H[p(x)]+\sum_x p(x)\left[q(0|x)\log\left(\frac{p(x)q(0|x)}{q(0)}\right)\right]+\beta\sum_x p(x)q(0|x)d_x \tag{3}$$

由式 (3) 可知, 欲求解最优化问题, 须首先求出 $q(0)$, $q(0|x)$, W 。因此对式 (2) 做变换并将其写成拉格朗日乘式形式后化简, 得到以下三个等式:

$$q(0) = \sum_x p(x)q(0|x) \quad (4)$$

$$q(0|x) = \min\{q(0)\frac{e^{-\beta D(V_x||W)}}{p(x)}, 1\} \quad (5)$$

$$W = \frac{\sum_x p(x)q(0|x)V_x}{\sum_x p(x)q(0|x)} = \sum_x q(x|0)V_x \quad (6)$$

为简化求解, 假设质心 W 已知, 即 $d_x = D(V_x||W)$ 已知, 则对等式 (4)、(5) 的求解是一个循环求解过程。为了打破这个循环, 引入集合 $C = \{x; q(0|x) = 1\}$, 表示集合 C 是满足这样一个条件的集合——该集合中的实例 x 均以概率 1 属于单类。进而可根据实例 x 是否属于集合 C 及不等式 $0 \leq q(0) \leq 1$, 将式 (6) 重写:

$$\sum_{x \notin C} e^{-\beta d_x} \leq 1 - \sum_{x \in C} p(x) \quad (7)$$

通过以上分析可知, 如果能从实例集中找到集合 C , 使该集合满足不等式 (9), 则可由此求出 $q(0)$ 、 $q(0|x)$ 。为了找到集合 C , 引入如下定理:

Lemma1 Let $x_1, \dots, x_m$ be a permutation of $[1, m]$

Such that $0 < \beta d_{x_1} + \log p(x_1) \leq \dots \leq \beta d_{x_m} + \log p(x_m)$

Then $C = \{x_i; 1 \leq i \leq k\}$ for some $k \in [0, m]$

结合该定理, OCRD-IB 的算法思想可表述为: 首先对数据集内所有实例进行排序, 使排序结果满足条件 $0 < \beta d_{x_1} + \log p(x_1) \leq \dots \leq \beta d_{x_m} + \log p(x_m)$; 然后令 $k=m$ (m 为数据集内实例的总数), 并验证集合 $C = \{x_i; 1 \leq i \leq k\}$ 是否满足不等式 (9), 若不满足, 则令 $k=m-1$ 继续验证, 若满足, 则此时的集合 C 即为所求集合; 最后根据集合 C 求出 $q(0)$ 、 $q(0|x)$, 进而得到度量函数式 (4) 的最优解。

3.2 算法描述

OCRD-BA 算法^[6]:

每个实例到质心的失真 d_x , 先验概率 $p(x)$, 参数 $\beta \geq 0$ 。

对每个实例按以下条件排序:

$$0 < \beta d_{x_1} + \log p(x_1) \leq \dots \leq \beta d_{x_m} + \log p(x_m)$$

令 $k=m$, $a_k=1$, $p_k=1$, $J_k = \beta \sum_x p(x)d_x$, $C_{\text{best}}=k$, $J_{\text{best}}=J_k$

while $k>0$

(1) 计算 $a_{k-1} = a_k - e^{-\beta d_{x_{k-1}}}$ $p_{k-1} = p_k - p(x_{k-1})$

(2) 如果 $p_{k-1} \leq a_{k-1}$

● 计算 J_{k-1}

$$\begin{aligned} J_{k-1} &= J_k - p(x_{k-1})[\beta d_{x_{k-1}} + \log p(x_{k-1})] \\ &\quad + [p_k \log(p_k) - p_k \log(a_k)] \\ &\quad - [p_{k-1} \log(p_{k-1}) - p_{k-1} \log(a_{k-1})] \end{aligned}$$

● 如果 $J_{k-1} < J_{\text{best}}$, 则令 $k_{\text{best}}=k-1$ $J_{\text{best}} = J_{k-1}$

(3) 令 $k=k-1$

$$q(0|x)$$

4 算法对比和分析

以往解决单类问题时使用一个凸函数作为损失函数，以下分析中将这一方法称为 OC-Convex 算法，该算法中被划分在单类内的实例的函数值为一个常量，单类外实例的函数值线性变化。当平均召回率较低时，该方法的结果趋近于位于中心位置的实例，从而导致大量其他实例被忽略。基于 IB 原理的单类问题求解方法引入一个预先设定的阈值 R ，由于它的存在，使得基于 IB 原理的单类问题求解方法的结果趋向于实例分布较密集的区域，从而避免了类似 OC-Convex 算法中大量实例被忽略的情况的出现。

以 Reuters-21578 数据集的两个子集 money-tx 和 crude 作为实验数据集^[6]，采用 OC-Convex 算法和 OCRD-BA 算法分别进行实验，并用平均正确率（Precision）和平均召回率（Recall）作为评价标准，生成如图 1（a）、（b）所示的 precision-recall 关系图。

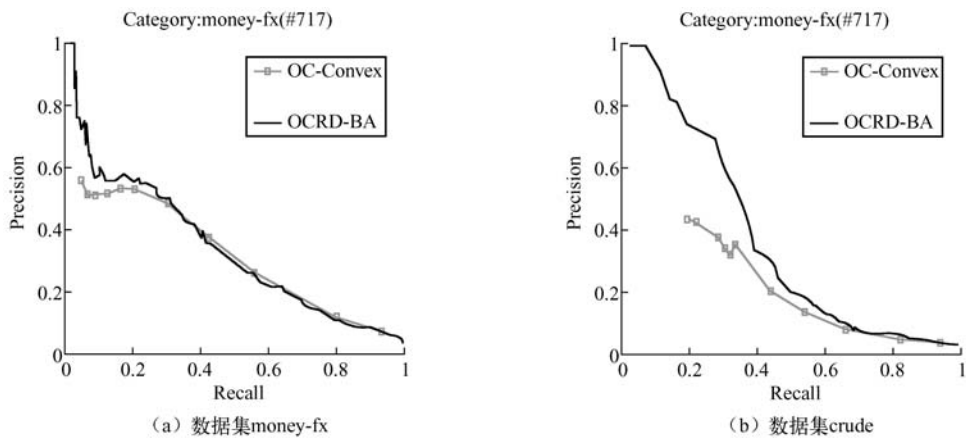


图 1 OC-Convex 与 OCRD-BA 的 precision-recall 关系图

从图 1 中可以直观地看出，当平均召回率较低时，OCRD-BA 算法的结果明显优于 OC-Convex 算法，这也从实验的角度解释了 OC-Convex 算法的结果趋近于位于中心位置的实例，从而导致大量其他实例被忽略的问题。随着平均召回率逐渐增大，OCRD-BA 算法的优势变得不那么明显，但就整体而言，在整个平均召回率区间上 OCRD-BA 算法的表现仍然是优于 OC-Convex 算法的。

5 结论

本文在对基本 IB 原理和 sIB 算法做出概括的基础之上，着重介绍了 Koby Crammer 等人提出的基于 IB 的解决单类问题的算法：OCRD-BA 算法，并对该算法做了详细的分析。今后的工作中，将对 OCRD-BA 算法做出一些改进，以期望得到更优秀的解决单类问题的算法，并可对算法的外延做进一步的扩展，拓展了 IB 原理的应用领域与范围。

参考文献

[1] J. Han, M. Kamber. Data Mining: concepts and techniques[M].2006.
[2] D.Tax, and R.Duin. Data domain description using support vectors[C]. Proceedings of the European Symposium on Artificial Neural Networks ,1999,251-256.
[3] B.Scholkopf, C.Burge, and V.Vapnik. Extracting support data for a given task[C]. First International Conference on Know-

- [4] K.Crammer, Y. Singer. Learning algorithms for enclosing points in bregmanian spheres[C]. Proceedings of the Sixteenth Annual Conference on Computational Learning theory.2003.
- [5] K.Crammer, G.Chechik. A Needle in a Haystack: Local One-Class Optimization[C]. Proceedings of the 21st International Conference on Machine Learning,Banff,Canada,2004.
- [6] K. Crammer, P. P. Talukdar, F.Pereira. A Rate-Distortion One-class Model and its Application to Clustering[C]. Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008.
- [7] N.Tishby, F. Pereira, and W. Bialek. The information bottleneck method[C]. The 37th Allerton Conference on Communication, Control, and Computing,Allerton House,Illinois.1999.
- [8] N. Slonim, N. Friedman, and N.Tishby. Unsupervised document classification using sequential information maximization[C]. Proceedings of 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 2002, 129-136.
- [9] N. Slonim. The information bottleneck: theory and applications[D].Doctoral dissertation, The Hebrew University.2002.
- [10] K. Crammer, N. Slonim. Bregman information bottleneck[C]. Presentation at the Workshop on Information Bottleneck and Information Distortion, Neural Information Processing Systems, Vancouver, Canada. 2003.

矩阵乘法的 FPGA 并行设计与实现

何红旗, 邵仪, 蒋烈辉, 赵秋霞

(国家数字交换系统工程技术研究中心 河南 郑州, 450002)

摘 要: 矩阵乘法是最基本的矩阵运算之一, 由于其计算密集的特点, 适合于在 FPGA 上实现。本文在对矩阵乘法的执行周期进行分析的基础上, 提出了一种在 FPGA 上实现矩阵乘法的并行算法, 通过 VHDL 编程实现并在 FPGA 芯片上实际测试。结果表明在频率为 154.53MHz 的情况下, FPGA 的浮点运算能力可达 19.76Gflops。

关键词: 矩阵乘法; 并行; 处理单元; FPGA

中图分类号: TP309 **文献标识码:** A **文章编号:** 1006-7043 (2004) xx-xxxx-x

The Parallel Design and Implementation Matrix Multiplication on FPGA

HE Hongqi, SHAO Yi, JIANG Leihui, ZHAO QiuXia

(National Digital Switching System Engineering & Technological R&D Center Zhengzhou 450002, Henan China)

Abstract: The matrix multiplication is one of the basic operation. Due to its compute-intensive character, it is suit to implement on FPGA. Based on the analysis of the latency of the matrix multiplication, this paper proposes a parallel design of implementing the matrix multiplication on FPGA, programs with VHDL and tests it. Th e result reflects the speedup of our designs reaches to 19.76Gflops at 154.53 MHz clock frequency.

Keywords: matrix multiplication; parallel; processing element; FPGA

1 引言

矩阵算法是许多工程问题中基本的数学工具, 其在信号处理和无线通信系统等领域有着广泛的应用, 其中矩阵乘法为最基本的矩阵运算之一。FPGA 是 Filed Programmable Gate Array 的缩写, 即现场可编程逻辑阵列, 它是在 CPLD 的基础上发展起来的新型高性能可编程逻辑器件, 一般采用 SRAM 工艺, 集成度很高, 其器件密度从数万系统门到数千万系统门不等, 可以完成极其复杂数的时序和组合逻辑电路功能, 适用于高速、高密度的高端数字逻辑电路设计。随着 FPGA 技术的发展, 在其上高效能地实现浮点算法已经成为了可能。因此, 在 FPGA 上实现矩阵乘法的运算也得到了越来越多的关注。文献[1], [2]提出了在 FPGA 上实现矩阵乘法的线性阵列结构模型。本文在对矩阵乘法的执行周期进行分析并借鉴线性阵列结构的基础上, 提出了一种在 FPGA 上实现矩阵乘法的并行计算结构模型, 使用 VHDL 硬件编程语言实现, 并在单片 FPGA 上进行了试验, 结果表明, 在频率为 154.53MHz 的情况下, FPGA 的浮点运算能力可达 19.76Gflops。

基金项目: 国家 863 计划重点课题“新概念高效能计算机体系结构及系统研究开发”(编号: 2009AA012201)

作者简介: 何红旗 (1972—), 男, 讲师, 国家数字交换系统工程技术研究中心, 硕士研究生, 主要研究方向为高性能计算

邵 仪 (1985—), 男, 硕士生, 国家数字交换系统工程技术研究中心;

蒋烈辉 (1967—), 男, 教授, 国家数字交换系统工程技术研究中心;

赵秋霞 (1964—), 女, 副教授, 国家数字交换系统工程技术研究中心。

2 相关工作

2.1 矩阵乘法

矩阵的乘法运算包括标量与矩阵的乘法、Hadamard 积、Kronecher 积和矩阵与矩阵的乘法，下面分别介绍这几种乘法。

标量与矩阵的乘法简称数乘，是将标量与矩阵各个元素相乘，得到的新矩阵与原矩阵的阶数相同，因此标量与矩阵的乘法属于矩阵的线性运算。Hadamard 积^[3]是按矩阵的对应分量进行相乘的矩阵乘法。Kronecher 积又称为直积^[3]，它是信号处理中的随机向量和随机向量过程分析，以及随机静态分析的一种基本分析的工具。以上三种矩阵的乘法运算都只涉及乘法运算，不涉及乘累加运算，重点在结果矩阵的排列上，用硬件实现比较容易，本文不做重点介绍，本文重点介绍矩阵与矩阵的乘法。

矩阵与矩阵的乘法定义如下^[4]：设 $\mathbf{A}=(a_{ij})$ 是一个 $m \times s$ 矩阵， $\mathbf{B}=(b_{ij})$ 是一个 $s \times n$ 矩阵，那么规定矩阵 $\mathbf{A}$ 与矩阵 $\mathbf{B}$ 的乘积是一个 $m \times n$ 矩阵 $\mathbf{C}=(c_{ij})$ ，其中

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{is}b_{sj} = \sum_{k=1}^s a_{ik}b_{kj} \quad (i=1,2,\cdots,m; j=1,2,\cdots,n) \quad (1)$$

把此乘积记作 $\mathbf{C}=\mathbf{AB}$ 。

只有当矩阵 $\mathbf{A}$ 的列数等于矩阵 $\mathbf{B}$ 的行数时， $\mathbf{AB}$ 才有意义， $\mathbf{AB}$ 是 $m \times n$ 矩阵， $\mathbf{AB}$ 中第 i 行、第 j 列的元素是矩阵 $\mathbf{A}$ 第 i 行各元素分别与矩阵 $\mathbf{B}$ 第 j 列对应元素的乘积之和。

2.2 现有工作分析

Ju-wook Jang 等人^[1, 2]首先提出了利用由处理单元（Processing Element，PE）组成的线型阵列结构（如图 1 所示）在 FPGA 上实现矩阵乘法运算的方法，其中，处理单元由浮点乘法和加法运算单元组成，每一个处理单元对矩阵 $\mathbf{A}$ 的每一行元素和矩阵 $\mathbf{B}$ 的一列元素进行乘加运算，得到矩阵 $\mathbf{C}$ 的一列元素，该结构需要用 $n^2 + 2n$ 个周期来完成矩阵乘法运算。文献[5]从硬件资源利用率方面就线型阵列结构中处理单元的个数和矩阵的规模之间的关系进行了分析，文献[6]从 FPGA 的能耗方面对线型阵列结构进行了改进。但是，目前线型阵列结构仍然只有相邻的处理单元才能通信，其与外部进行通信的数据通路只有一条，这使得矩阵元素不能同时送入线型阵列结构中进行运算。因此，本文对矩阵乘法的执行周期进行分析的基础上，提出了一种新的基于 FPGA 的矩阵乘法并行计算结构。

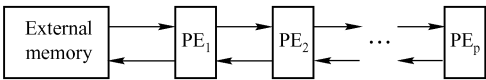


图 1 线型阵列结构示意图

3 矩阵乘法的执行周期分析

对矩阵乘法而言，其算法执行周期 L 是由计算周期 L_1 和 I/O 周期 L_2 来决定的。对于一个 $n \times n$ 矩阵，如果一个周期中 p 个处理单元同时进行计算，则 $L_1 = O\left(\frac{n^3}{p}\right)$ 。 $L_2 = \text{I/O 操作的总次数} / \text{每一个周期 I/O 操作的执行次数}$ ，I/O 操作执行的总次数即为算法的 I/O 复杂度，文献[7]中证明了所有矩阵乘法的 I/O 复杂度都等于 $O\left(\frac{n^3}{\sqrt{M}}\right)$ ，其中 M 是算法可用的存储空间大小。该等式在 $O(1) \leq M \leq O(n^2)$ 的情况

下成立。当 $M \geq O(n^2)$ 时，矩阵乘法的 I/O 复杂度仍等于 $O(n^2)$ 。

如果用 R 来表示每一个周期 I/O 操作的执行次数，那么矩阵乘法执行周期的理论下限值是

$$L \geq \max(L_1, L_2) \geq \max\left(\frac{n^3}{p}, \frac{n^3/\sqrt{M}}{R}\right), M \leq n^2 \tag{2}$$

图 2 所示的是 M 、 p 和 L 三者之间的关系，其中 $l(n^2)$ 表示参数 L 的原点为 n^2 ，参数 p 和 M 的原点为 1。当 $p = O(1)$ 时，无论 M 是多少， L 都为 $O(n^3)$ ，此时矩阵乘法的执行周期 L_1 是算法的瓶颈。通用处理器就是这种情况，假设处理器拥有足够大的 cache 可以存储所有的矩阵元素，并且每一个周期都可以从 cache 中读取数据。在这种情况下， M 为 $O(n^2)$ ，矩阵乘法的 I/O 周期 L_2 也就等于 $O(n^2)$ 。然而，由于通用处理器串行执行程序的特性，其只能算做一个处理单元，因此 L_1 为 $O(n^3)$ ，那么 L 也为 $O(n^3)$ 。由此可以看出，无论 cache 多大，单个处理器中矩阵乘法的执行周期都是 $O(n^3)$ 。另外，如果 $M = O(1)$ ，无论每一个周期进行计算的处理单元个数是多少， L 都无法得到改善，仍然是 $O(n^3)$ 。在这种情况下，矩阵乘法的瓶颈是 I/O 带宽，而不再是处理单元个数。矩阵乘法执行周期的理论下限值是在 $p = O(n)$ 、 $M = O(n^2)$ 的情况下（即图中的点 (n, n^2, n^2) ）达到的，此时 L_1 和 L_2 都是 $O(n^2)$ ， L 会达到最小值，为 $O(n^2)$ 。

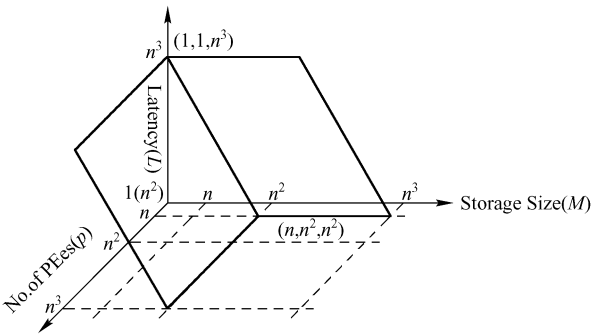


图 2 M, p, L 参数关系

4 矩阵乘法的 FPGA 固化结构设计

通过前面对相关设计参数的分析可知，当处理单元的个数 $p = n$ 、算法可用的存储空间 $M = n^2$ 时，矩阵乘法的执行周期数最少，为 $O(n^2)$ 。线型阵列结构在上述两个条件满足的情况下的计算周期为 $n^2 + 2n$ 。在线型阵列结构中，为了保证矩阵乘法计算的正确性，矩阵 A 的元素必须等待 n 个周期之后，才能送入线型阵列结构中进行运算。此外，在计算过程的前 $2n$ 个周期之内，由于矩阵 A 的元素没有完全送入线型阵列结构中，使得全部或部分的处理单元处于空闲状态，降低了硬件资源的利用率，同时也增加了算法的执行周期。

图 3 所示的是经过改进的线型阵列结构。其中，处理单元的个数 p 是一个可以变化的参数， $p = n/r$ ($1 \leq r \leq n$)。理论上， p 可以取 $1 \sim n$ 之间的任意整数，并且 p 值越大，固化执行的并行程度越高，算法的执行周期也越短。但是，在实际情况下，考虑到 FPGA 器件硬件资源的限制和编程的复杂程度， p 应当取硬件资源允许下的矩阵阶数 n 的最大约数。矩阵的划分采用基于列的循环带状划分方法^[8]。图 4 所示的是改进的线型阵列结构中处理单元的具体结构。其中，MB 和 MC 分别表示对矩阵 B 和结果矩阵 C 进行存储的本地存储单元，其存储容量都是 n^2/p ；Count 为计数单元，对算法中的循环部分进行计数，以保证算法执行的正确性。在处理单元个数改变时，其内部结构中只有本地存储器的容量发生变化，而总的存储空间和每个处理单元内部浮点运算单元的个数都不发生变化。

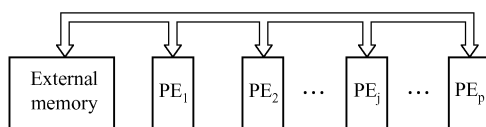


图3 矩阵乘法线型阵列固化结构的改进

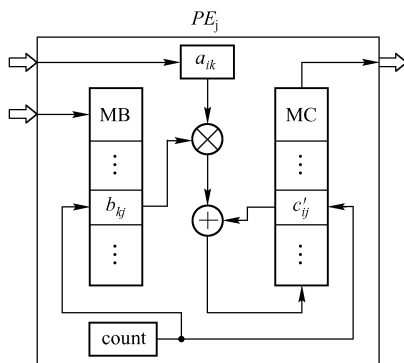


图4 改进的线型阵列结构的处理单元架构

当处理单元个数 $p = n$ 时，每一个处理单元存储矩阵 B 的一列元素，与送入该处理单元的矩阵 A 的元素进行乘加运算，得到的矩阵 C 的一列元素。以处理单元 PE_j ($1 \leq j \leq p$) 对矩阵 C 的第 j 列元素的计算为例进行说明，在初始状态下，矩阵 B 的第 j 列元素 b_{kj} ($1 \leq k \leq n$) 存储在 PE_j 的本地存储器中，矩阵 A 以列优先的顺序 ($a_{11}, a_{21}, a_{31}, \dots, a_{n1}, a_{21}, a_{22}, \dots$) 进入 PE_j ，矩阵 A 的第一列元素 a_{i1} ($1 \leq i \leq n$) 与矩阵 B 第 j 列的第一个元素 b_{1j} 进行乘法操作，得到 c'_{ij} ；矩阵 A 的第二列元素 a_{i2} 与矩阵 B 第 j 列的第二个元素 b_{2j} 进行 $c'_{ij} = c'_{ij} + a_{i2} \times b_{2j}$ 操作；依此类推对 c'_{ij} 进行更新，直至矩阵 A 的最后一列元素 a_{in} 与矩阵 B 第 j 列的最后一个元素 b_{nj} 进行 $c'_{ij} = c'_{ij} + a_{in} \times b_{nj}$ 操作，并最终得到矩阵 C 的第 j 列元素 c_{ij} 。此时，处理单元 PE_j 完成了对矩阵 C 的第 j 列元素的计算。由于 n 个处理单元是并行进行计算的，则矩阵 C 的其他列元素也都计算完成。

在改进的线性阵列结构中，由于矩阵 B 的元素在乘法计算开始之前就已经存储到处理单元中，因此矩阵 A 的元素在送入处理单元之前不需要再等待 n 个周期。此外，由于所有处理单元之间可以相互通信，矩阵 A 的元素可以同时进入每一个处理单元，又可以省去 $n-1$ 个执行周期。通过处理单元所运行程序的伪代码也可以看出，当 $p = n$ 时，改进的线型阵列结构的执行周期数为 $n^2 + q$ ，其中 q 为浮点乘加运算单元的计算周期数。对于定点数据， $q = 1$ ；而对于浮点数据，当 n 值较大时， $q \ll n^2$ 。因此，改进的线型阵列结构可以达到理论最优执行周期数，也比现有的线型阵列结构减少了 $2n$ 个执行周期，但是改进的线型阵列结构增加了处理单元的本地存储空间。

5 硬件实现分析

本文利用 VHDL 语言对上述模型进行了编程实现，利用 Quartus II 7.2 开发平台对源代码进行编译。由于矩阵的 LU 分解是在实数范围内进行的，因此在编程时需要用浮点数来表示矩阵的元素。Quartus II 软件提供了符合 IEEE 754 标准的浮点数据。本文使用单精度和双精度两种格式，其中，单精度格式用 32 位二进制数表示一个浮点数，其中最高位为符号位，用 8 位移码表示指数，用 23 位原码表示尾数；单精度格式用 32 位二进制数表示一个浮点数，其中最高位为符号位，用 8 位移码表示指数，用 23 位原码表示尾数。

本文所有处理单元中的浮点运算单元都是独立的模块，即处理单元的结构和所使用的浮点运算单元是不相关的。当用新的浮点运算单元来替换当前浮点运算单元时，矩阵算法的固化结构不需要做改动。特别的，当有面积更小或速度更快的浮点运算单元可用时，可以很容易地应用到本文的设计中。设计的模块化同时还简化了精度调整时所付出的代价。例如，设计 64 位矩阵乘法的处理单元时，只需将 64 位的浮点运算单元取代 32 位的浮点运算单元，并改变处理单元内部和处理单元之间的数据带宽即可。

6 测试及结论

本文利用 VHDL 语言对上述模型进行了编程实现，利用 Quartus II 7.2 开发平台对源代码进行编译，并在 Modelsim AE6.1g 软件上进行了仿真，选用 Altera 公司的 Stratix II GX 系列中的 EP2SGX90FF1508C3 型号 FPGA 开发板进行实际测试。在实际测试过程中，由于 FPGA 芯片的片上 RAM 资源有限（4.3Mb），且 Quartus II 提供的 RAM IP 核规定其 RAM 的大小应当为 2 的方幂。因此本文对两个 64×64 的矩阵进行了乘法运算，并对得到的结果矩阵分别进行了验证。表 1 是在处理单元个数取不同值时矩阵乘法的运行时间，其中矩阵元素采用单精度浮点数据，浮点运算能力的定义如下：

$$\text{FLOPS} = \frac{\text{总的浮点运算次数}}{\text{算法的执行时间}}$$

(3)

在矩阵乘法中，每一次循环需要执行一次乘法和一次加法操作，总共需要 n^3 次循环，因此总的浮点操作为 $2n^3$ ，浮点计算能力为 $2n^3/\text{算法的执行时间}$ 。

表 1 64 阶矩阵的乘法固化测试结果

PE 个数 p	ALUTs	$f_{\max}$ (MHz)	执行时间 (μs)	浮点运算能力 (GFLOPS)
4 2187		190.42	344.18	1.52
8 3929		190.22	172.29	3.04
16 7764		188.36	87.01	6.02
32 15106		171.15	47.88	10.95
64 29698		154.53	26.53	19.76

表 2 表示的是在矩阵乘法固化实现上，本文设计与现有设计的比较，其中矩阵元素为双精度浮点数据。通过比较可知，本文改进的矩阵乘法固化结构在时钟频率较低的情况下，得到了较好的浮点运算能力。

表 2 矩阵乘法固化设计比较

	本文	[2] [5] [6]	[8] [9]			
时钟频率 (MHz)	120 100 172	180 200 373				
浮点运算能力 (GFLOPS)	15.35 5.0		8.3	9.36 15.6	29.8	

当矩阵的阶数为 64、频率为 3.0GHz 时，通用处理器的浮点运算能力仅为 0.22Gflops，而 FPGA 的浮点运算能力在 154.53MHz 频率下可以达到 19.76Gflops，在频率不足 GPP 的十分之一情况下，FPGA 的浮点运算能力是通用处理器的 90 倍，由此可知，FPGA 对矩阵乘法的加速效果明显。然而 FPGA 的片上资源有限，要想运行大规模矩阵的 LU 分解算法，就要对矩阵乘法进行分块运算，这是下一步需要进行研究的问题。

参考文献

[1] Ju-wook Jang, Seonil Choi and Viktor K. Prasanna. Area and Time Efficient Implementations of Matrix Multiplication on FPGAs[A]. The First IEEE International Conference on Field Programmable Technology[C]. Hong Kong, 2002: 93-100.

[2] Ju-wook Jang, Seonil Choi and Viktor K. Prasanna. Energy- and Time-Efficient Matrix Multiplication on FPGAs[J]. IEEE Transactions on Electronic Computers, 2005,13(11): 1305-1319.

[3] 陈祖明, 周家胜. 矩阵论引论[M]. 北京: 北京航空航天大学出版社, 1998: 350.

[4] 同济大学数学教研室. 工程数学线性代数 (第三版) [M]. 北京: 高等教育出版社, 1998: 43-55.

[5] L. Zhuo and V.K. Prasanna, Scalable and modular algorithms for floating-point matrix multiplication on FPGAs[A]. 18th International Parallel and Distributed Processing Symposium[C]. Santa Fe, 2004: 92-101.

[6] Xizhen Xu and Sotirios G. Ziavras. H-SIMD machine: Configurable parallel computing for matrix multiplication[A] International Conference on Computer Design[C]. San Jose, 2005: 671-676.

[7] J. Hong, H. Kung. I/O Complexity : The Red Blue Pebble Game[A]. Annual ACM Symposium on Theory of Computing Proceedings of the thirteenth annual ACM symposium on Theory of computing[C]. New York, 1981: 326-333.

[8] Yong Dou, et. al. 64-bit floating-point FPGA matrix multiplication[A]. Proceedings of the 2005 ACM/SIGDA 13th International Symposium on Field programmable gate arrays [C]. New York, 2005: 86-95.

[9] Vinay BY. Kumar, et. al. FPGA based High Performance Double-precision Matrix Multiplication[J]. International Journal of Parallel Programming. 2010, 3(2): 135-140.

基于 Zigbee 技术的加权质心定位算法

李占波¹，刘慧玲²

(郑州大学信息工程学院 河南 郑州，450002)

摘 要：无线传感网已经在各领域受到广泛关注，其中节点定位技术是无线传感网络大多数应用的基础。对于无线传感节点的定位问题，目前的 CC2431 片上系统只能得到二维坐标，本文首先使用 KVP 帧实现了简单的三维定位，得到定位节点的 z 层值。实现 z 层定位后，把三维空间定位转化为二维平面坐标定位，然后，计算二维坐标的过程中，在三角形质心定位算法的基础上，为每个质心点加上权重，即用三角形质心加权算法代替原来的三角形质心算法。通过仿真实验证明，由于受各种因素的影响。定位结果还存在一定误差，但是三角形质心加权算法比原来的三角形质心算法的精确度有明显提高。

关键词：Zigbee；无线传感网络；定位；质心；加权

Weighted Centroid Localization Based on Zigbee Technology

LI Zhanbo¹,LIU Huiling²

(School of Information Engineering, Zhengzhou University, Zhengzhou 450002, Henan China)

Abstract: Wireless sensor networks (WSNs) are more and more widely used in many different scenarios. The localization information is an important criterion for the capability of WSNs. At present, as for localization in wireless sensor networks, the CC2431 system can only produce the 2-D coordinates. In this paper, we use KVP frame to achieve a simple 3-D localization and get the z floor. After finishing the z floor localization, we can change the 3-D position calculating to the 2-D localization. Then, to calculate the 2-D position, we add the weight for each centroid nodes which produced in the triangle centroid algorithm, which means replacing the triangle centroid algorithm by the weighted centroid localization algorithm. The result of simulation shows that, due to many factors, it still exists errors in the result. However, the result of the weighted centroid localization algorithm is much more precise than that of the triangle centroid algorithm.

Keywords: ZigBee; Wireless Sensor Networks (WSNs); Localization; Centroid; Weight

GPS，即全球卫星定位系统（Global Positioning System，GPS）是 20 世纪 70 年代由美国陆海空三军联合研制的新一代空间卫星导航定位系统。目前，GPS 是最著名，也是应用最广泛的定位系统，它被用来对户外移动的物体进行定位。但是，对于室内物体的定位，GPS 的定位精度达不到要求的标准。相比之下，ZigBee 是一种新兴的短距离、低速率无线网络技术，它最显著的特点是低功耗和低成本。利用 ZigBee 技术实现定位具有低成本、低功耗的优点，且信号传输不受视距的影响。

目前，已经有很多无线传感节点的定位算法，但是，无论是三维定位算法还是二维定位算法，都存在一定的问题。对于三维定位算法，存在的问题是 CC2431 片上系统现在只能实现二维坐标定位。所以这些三维算法只能停留在理论水平上。但是，在实际应用中，节点的分布往往在三维空间，地形比较复杂，简单地用二维平面中的算法在应用中也会有很多问题。最显著的问题就是，一般来说，两点间的距离是指三维的空间距离，而一般的二维算法直接用二维平面坐标计算两点间的距离，在地形比较复杂的情况下，这样会存在很大误差。所以，本文用软件实现 z 层的定位。之后，把三维定位降到二维，可以用二维定位算法实现 x 和 y 的定位。既实现了在三维空间的定位，又解决了 CC2431 只能在二维平面定位的问题。

另外，本文主要讨论单跳定位系统，在此系统中，未知节点能够与锚节点直接通信。

1 相关定位算法简介

在无线传感器网络中，按节点位置估测机制，根据定位过程中是否测量节点间的实际距离或角度，可分为基于测距（Range-based）的定位算法和距离无关（Range-free）的定位算法。前者需要测量节点间的实际距离；后者是利用节点间的估计距离来计算未知节点的位置。

基于测距定位常用的测距方式有：RSSI、TOA、TDOA、和 AOA。典型的距离无关技术的定位算法有：Bounding Box 算法^[1]、DV-Hop 算法^[2, 5]、质心定位算法^[4]。其中，RSSI 测距技术因能量消耗低，成本低廉而且易于实现而著称，已经被广泛应用到节点定位中。由于 RSSI 测距受到环境影响而产生测距误差，进而影响节点定位的精度。所以单纯的 RSSI 定位方法只能用来定位粗粒度区域。

三角形质心算法则是结合 RSSI 测距和三角形及圆的相关知识来减小误差。而本文的三角形质心加权算法是在原来三角形质心算法的基础上，利用各三角形质心点的距离特征，为每个质心点加上相应的权重。进一步提高了定位的精确度。

2 算法描述

2.1 定位 z 方法

在本文的三维定位中，z 值即水平层先用一个字节来表示，可以区分 256 个不同的层。若楼层高 m 米，则每层之间的距离为 $m/256$ 。在需要三维坐标的情况下，盲节点发送 z 请求，通过参考节点的回复和一定的 z 值选择方法即可取得待测节点的 z 层值，完成 z 层定位。

为了减少网络开销，z 坐标单独发送，因为有的情况下只需要定位水平坐标，如在平地上。若此时也对盲节点三维定位，则会浪费网络资源。

由于水平层只需要用一个字节来表示，所以 z 层定位主要通过发送 KVP 帧实现。KVP 帧是专用的比较规范的信息格式，采用键值对的形式，按一种规定的格式进行数据传输。通过一种规定来标准化数据传输格式和内容，通常用于传输一个简单的属性变量值。KVP 帧格式如表 1 所示。

表 1 KVP 帧格式

Bits:4 4		16	0/8	Variable
Command type identifier	Attribute data type	Attribute identifier	Error code	Attribute data

KVP 命令类型有：数据请求确认（Get with acknowledgement）、Get response（回复数据请求）、Set/set with acknowledgement（设置）、Set response（回复设置）、Event/event with acknowledgement（变更/需要确认的变更）、Event response（变更响应）。

定位 z 的过程中，主要用 set 命令帧设置参考节点的 z 层值，数据请确认命令帧（Get with acknowledgement）广播 z 请求，参考节点用 Get response 发送确认回复帧，即可完成 z 层的定位。

最后，求出各参考节点层数的平均值作为未知节点的层值。在距离公式 $(x_i - x_0)^2 + (y_i - y_0)^2 + (z_i - z_0)^2 = r^2$ ^[7] 中， $(z_i - z_0)^2$ 可换算成层距，记为 d^2 ，则有 $(x_i - x_0)^2 + (y_i - y_0)^2 = r^2 - d^2$ ，即把三维定位降到二维定位。

2.2 三角形质心加权定位算法

如图 1 所示，参考节点 $A_1, A_2, A_3, \dots, A_n$ ，未知节点 B ，根据 RSSI 模型计算出的节点 A_1 和 B 的距离为 r_1 ，节点 A_2 和 B 的距离为 r_2 ，节点 A_3 和 B 的距离为 r_3 ，分别以 A_1, A_2, A_3 为圆心， r_1, r_2, r_3 为半径画

圆，可得交叠区域。连接三圆交叠区域的三个交点。得到图 1 中蓝色线段围成的三角形 $C_1C_2C_3$ 。这里的三角形质心定位算法的基本思想^[3]是：计算三圆交叠区域的 3 个特征点的坐标，以这三个点为三角形的顶点，未知点即为三角形质心，点 C_1 点的计算方法为

$$\begin{cases} \sqrt{(x_{A_1} - x_{C_1})^2 + (y_{A_1} - y_{C_1})^2} \leq r_1 \\ \sqrt{(x_{A_2} - x_{C_1})^2 + (y_{A_2} - y_{C_1})^2} = r_2 \\ \sqrt{(x_{A_3} - x_{C_1})^2 + (y_{A_3} - y_{C_1})^2} = r_3 \end{cases} \tag{1}$$

同理，可计算出 C_2, C_3 ，此时质心点的坐标为 $D_1 \left(\frac{x_{C_1} + x_{C_2} + x_{C_3}}{3}, \frac{y_{C_1} + y_{C_2} + y_{C_3}}{3} \right)$,

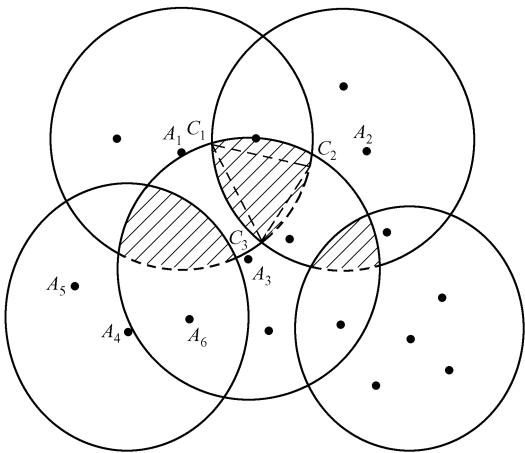


图 1 三角形质心算法

2.3 三角形加权质心算法

设参考节点 $A_1, A_2, A_3, \dots, A_n$ 的坐标分别为 $\{(x_{A_1}, y_{A_1}), (x_{A_2}, y_{A_2}), \dots, (x_{A_n}, y_{A_n})\}$ ， $A_1, A_2, A_3, \dots, A_n$ 到未知节点 B 的距离分别为 $(d_1, d_2, \dots, d_n)$ ，任意三个锚圆有交点的参考节点组合，得到以下参考节点的三元组的集合。Triple_set= $\{(A_1, A_2, A_3), (A_1, A_2, A_4), \dots, (A_{n-2}, A_{n-1}, A_n)\}$

对于每个三元组。都可以用三角形质心算法得到相应的质心点坐标。总共可得到 C_n^3 个质心点 $D_1, D_2, \dots, D_m$ 。设 $D_1, D_2, \dots, D_m$ 的坐标分别为 $\{(X_1, Y_1), (X_2, Y_2), (X_m, Y_m)\}$ 。

$$\begin{aligned} \text{令 } S_k &= \sum_{i=1}^{k-1} (d_i - d_k)^2 + \sum_{i=k+1}^m (d_i - d_k)^2 \\ &= \sum_{i=1}^{k-1} [(X_i - X_k)^2 + (Y_i - Y_k)^2] + \sum_{i=k+1}^m [(X_i - X_k)^2 + (Y_i - Y_k)^2] \quad k=1, 2, 3, \dots, n \end{aligned}$$

式中， S_k 表示第 k 个质心点到其他质心点的距离平方和。

随后，计算未知节点平面坐标。由于距离平方和越小，未知节点越近，所以，它的权重应该越大。设节点 D_k 的权重 $w_k = \frac{1}{S_k}$ ，最终 D 的平面坐标为

$$D = \left(\sum_{k=1}^m \frac{w_k}{\sum_{i=1}^m w_i} x_k, \sum_{k=1}^m \frac{w_k}{\sum_{i=1}^m w_i} y_k \right)$$

3 仿真实验及误差分析

3.1 仿真

为了验证三角形质心加权算法的性能，本文用 MATLAB 对该算法进行了仿真试验。仿真环境为：假设网络区域为 $100\text{m} \times 100\text{m}$ 的正方形，随机在区域内部署 100 个节点，从这 100 个节点里选取 30 个参考节点，让参考节点均匀分布在在在网络区域内。为每个定位节点配置 8 个参考节点，则 8 个参考节点会存在 $C_8^3 = 56$ 个质心点。现假设外界干扰服从均值为 0、标准差为 7 的高斯分布。

仿真结果如图 2 所示， x 轴和 y 轴分别表示节点的横坐标和纵坐标。蓝色点表示由 8 个参考节点产生的三角形质心点，绿色的“+”为用三角形质心加权算法得到的定位结果。黄色的“*”号是三角形质心算法。红色的圈符号是未知节点的实际坐标。从图中可以看出，用三角形质心加权算法得到的未知节点的坐标更靠近它的实际坐标。

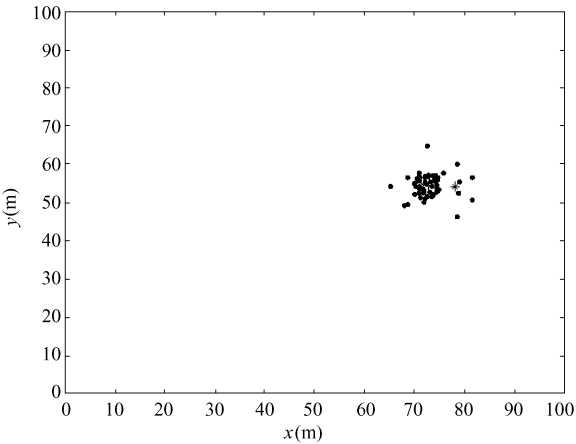


图 2 三角形质心加权算法定位结果

3.2 误差分析

如图 3 所示， x 轴和 y 轴分别表示节点的横坐标和纵坐标。红色点代表未知节点（34.62，45.82），蓝色点表示对未知节点的 50 次定位。从图中可以看出，实验结果存在一定的误差，但是，误差尚在允许的范围之内。出现定位误差的原因主要有：硬件本身存在小于 3m 的误差、障碍物对 RSSI 信号的影响、另外，三角形质心加权算法刚开始就假设存在期望为 0，标准差为 7 的高斯噪声。所以，存在误差是正常现象，误差只能尽量减小，而不可能彻底消除。

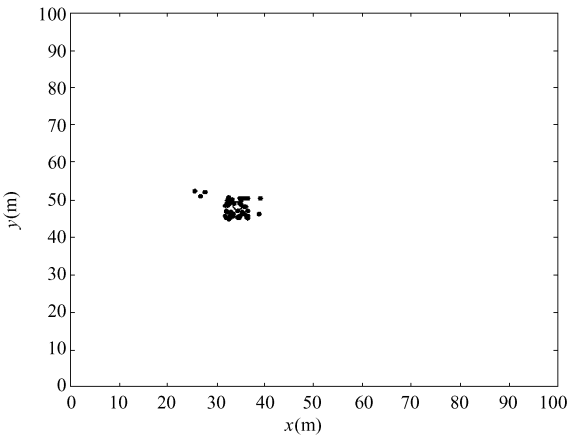


图 3 对节点（34.62，45.82）的 50 次定位

4 结束语

本文首先用一定的方法把三维空间定位降到二维，之后，在三角形质心算法的基础上，为每个质心点加上相应的权重来提高算法精度的，通过实验可以看到，三角形质心加权算法的精度确实比三角质心定位高。但是，通过后来对同一个未知节点的 50 次定位，说明误差仍然存在。另外，本文的研究重点主要在定位精度上，没有考虑时延。所以，下一步应该考虑如何在提高精度的同时减少时延。

参考文献

- [1] Kannan J . Implications of peer to peer networks on worm attacks and defenses[R] . CS29424 Project . California ,2003.
- [2] Andreas Savvides,Chih-Chieh Han,Mani B Srivastava.Dynamic Fine-Grained Localization in Ad-Hoc Networks of Sensors[C].Proceedings of Mobile Computing and Networking(MOBICOM'01),Rome.Italy:ACM Press,2001.
- [3] Binwei Deng , Guangming Huang Lei Zhang ,Improved Centroid Localization Algorithm in WSNs,Intelligent System and Knowledge Engineering ISKE 2008.
- [4] N.Bulusu,J.Heidemann,D.Estrin.GPS-Less Low Cost Outdoor Localization for Very Small Devices[R].Technical Report 00-729,Computer Science Department,University of Southern California,April,2000.
- [5] 孔庆茹，杨新宇，闫超，杨文静. 一种基于接收信号强度指示的改进型定位算法[J].西安交通大学学报, 2008 年 2 月第 2 期: 23-25.
- [6] Jan Blumenthal,Ralf Grossmann,Frank Golatowski,Weighted Centroid Localization in Zigbee-based Sensor Networks , Intelligent Signal Processing,WISP 2007.
- [7] Chih-Chun Lin ,She-Shang Xue, Leehter Yao, Position Calculating and Path Tracking of Three Dimensional Location System based on Different Wave Velocities,Dependable,Autonomic and Secure Computing,DASC 2009.
- [8] Huanjia Yang, Shuang,Hua Yang, Connectionless Indoor Inventory Tracking in Zigbee RFID Sensor Network,Industrial Electronics,IECON 2009.

基于自适应窗和 Hartley 变换的河工模型 PIV 测速

喻恒, 赵建军

(河南大学计算机与信息工程学院, 河南 开封, 475100)

摘要: 本文根据河工模型及示踪粒子的分布特点, 提出了一种自适应分析窗口的选择方法, 取代传统 PIV 技术中固定窗口大小的方法, 并采用基于 Hartley 变换的互相关计算代替传统的 Fourier 变换, 减小了测速过程中可能出现的误差, 提高了算法效率。经实验分析, 该方法达到了对河工模型表面流场的实时测量的要求, 具有一定的实际意义。

关键词: 表面流场; PIV; 自适应窗选择; 互相关计算; Hartley 变换

The Particle Image Velocimetry of River Model Based on Adaptive Window and Hartley Transform

YU Heng , ZHAO Jian jun

(Computer and Information Engineering college, Henan University, Kaifeng 475100, Henan China)

Abstract: according to the distribution of the tracer particle of the river model, we proposed the selection method of adaptive analysis window instead of the traditional fixed window method of the traditional PIV, and use the Hartley transform to achieve cross-correlation calculation, but not the traditional Fourier transform method, which has low error and high computing efficiency. Experimental results show that the improved method could meet the demands of real-time measurement of the surface flow field of river model and has some practical significance.

Keywords: surface flow field; PIV; adaptive window selection; cross-correlation calculation Hartley transform

1 引言

近几年, 在气候环境的影响下, 一些大型河流的河道河势变化逐渐剧烈, 岸线形状日益复杂, 建立河工模型, 进行相关实验从而有效分析河工模型的河势和流速变化情况, 获得大量的水文资料, 对保护河流意义重大。

在河工模型研究实验中, 模型的流场测量是一项重要的内容。河工模型的流场测量与其他的流动类别模型试验中的测量有许多不同之处: 河工模型尺寸一般很大, 有的超过数千平方米, 观测的范围大; 对于黄河模型这类动床模型或非定常流动模型, 受河岸的影响难以布置非接触式的光学测量方法的光通路; 同时, 实验时间要求在很短的范围内, 并通过一系列连续测量取得足够多的数据, 才能完全描述流体随时间变化的非定常流动。因此, 将基于数字图像处理的全场测速技术应用到河工模型流场的测量中有很大的科研价值和经济价值。

全场测速技术主要包括激光诱导磷光 (LIP)、激光诱导荧光 (LIF)、相干反斯托克斯喇曼散射 (CARS)、粒子跟踪测速 (Particle Tracking Velocimetry) [1] 和粒子图像测速 (Particle Image Velocimetry) [2] 等。其中粒子图像测速, 以及数字化粒子图像测速技术 (Digital PIV) 因为其特有的方便快捷的特点, 在流体力学研究中得到了广泛应用。

粒子图像测速技术 (Particle Image Velocimetry, PIV) 是利用单次或多次曝光的底片或 CCD 像机记录的序列图像, 经过傅里叶变换等处理, 可实现复杂环境下全流场的无接触、无扰动、高准确度测量和显示, 特别适用于非定常复杂流场的测量, 是研究复杂形态瞬态流动的有力手段。

但是 PIV 技术是一种通用的流场测速技术, 并不是针对河工模型表面流场测速的算法。本文根据河工模型及示踪粒子的分布特点, 提出了一种自适应分析窗口的选择方法, 取代传统

PIV 技术中固定窗口大小的方法, 并采用基于 Hartley 变换的互相关计算代替传统的 Fourier 变换, 减小了测速过程中可能出现的误差, 提高了算法效率, 使改进的 PIV 方法更适用于河工模型的表面流场测速, 具有一定的实际意义。

2 传统 PIV 技术原理

PIV 技术(粒子图像测速法)是 20 世纪 70 年代末发展起来的一种瞬态、多点、无接触式的流体力学测速方法, 是一种基于流场图像互相关分析的非接触式二维流场测量技术, 能够无扰动、精确有效地测量二维流速分布。

目前 PIV 测速方法有多种分类, 无论何种形式的 PIV, 其速度测量都需在流场中散播比重适当且跟随性好的示踪粒子, 由示踪粒子的跟随性运动来反映流场表面水质点的运动。如图 1 所示为 PIV 实验流程图。

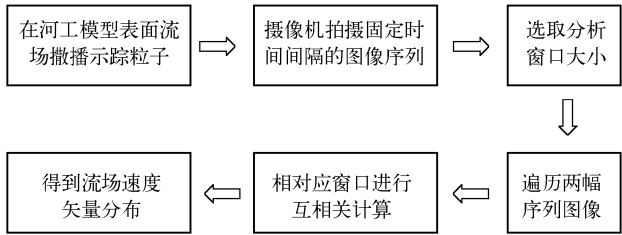


图 1 PIV 实验流程图

设图像函数 $I_1(x, y)$ 和 $I_2(x, y)$ 分别表示 t_1 时刻和 $t_1 + \Delta t$ 时刻的图像中相对应的分析窗口(如 64×64 像素), 如果该分析窗口内颗粒在 x 和 y 方向上相应时间间隔内的位移分别为 Δx 和 Δy , 根据粒子位移均匀性的假设, 第二次曝光的图像 $I_2(x, y)$ 可视为第一幅图像 $I_1(x, y)$ 经过平移后得到的, 即

$$I_2(x, y) = I_1(x + \Delta x, y + \Delta y) \quad (1)$$

I_1 和 I_2 的互相关函数如下:

$$R(\Delta x, \Delta y) = \int_{-\infty}^{\infty} I_1(x, y) I_1(x + \Delta x, y + \Delta y) dx dy \quad (2)$$

转换为离散形式为

$$R(n, m) = \sum_{X=0}^{M-1} \sum_{Y=0}^{N-1} I_1(X, Y) I_1(X + n, Y + m) \quad (3)$$

式中, $n=1, 2, 3, \dots, N-1$, $m=1, 2, 3, \dots, M-1$ 。

为了提高运算效率, 作为 PIV 技术核心的流场图像分析法目前主要采用二维快速傅里叶变换。令 $I_1(X, Y)$ 、 $I_2(X, Y)$ 的傅里叶变换分别为 $IF_1(U, V)$ 、 $IF_2(U, V)$, 则对应的 FFT 表达式为:

$$R_f(U, V) = IF_1^*(U, V) IF_2(U, V) \quad (4)$$

$IF_1^*(U, V)$ 为 $IF_1(U, V)$ 的复共轭变换, $R_f(U, V)$ 为 $R(n, m)$ 的傅里叶变换的结果, 对 $R_f(U, V)$ 进行一次傅里叶逆变换即可得到互相关函数。

实现互相关运算时, 对同一位置处的两个分析窗口进行 FFT 运算, 窗口相似程度越大时, R 的值就越大, 当 R 值达到最大峰值时离开查询窗口中心的位置, 即为窗口的平均位移, 可得到 $I_1(X, Y)$ 经过 Δt 时间后的相对位移, 即水流质点在 Δt 时刻的位移, 进而计算得到 t_1 时刻的速度。

3 河工模型表面流场的 PIV 测速

3.1 自适应窗口选择

PIV 技术的核心之一是信息分析窗口的选取, 从 PIV 应用角度考虑, 分析窗口越小越能够体现流

为了提高相关分析的准确度和可靠性,本文提出一种自适应分析窗口选择方法,提高了互相关计算的准确性,克服了因示踪粒子较大而造成的图像运动后粒子进出窗口对互相关计算精度的影响。

自适应分析窗口选择的原则是:(1)由于河工模型示踪粒子个体较大,撒播密度较小,分析窗口内示踪粒子的成像密度应最大,保证互相关计算的准确性。根据 Adrian1991 年对粒子图像测速技术的像密度 N 的定义^[2]:

$$N = C\Delta Z_0 \frac{hd_1^2}{4M^2} \quad (5)$$

式中, C 为粒子浓度; ΔZ_0 为片光源厚度; M 为照相机的放大率; d_1 为粒子直径。由于对于同一 PIV 成像系统,影响像密度的主要因素是 Cd_1^2 , 因此,我们可以将分析窗口内粒子的能量值等效为成像密度。(2)为了最大程度地避免运动造成的窗口内粒子进出造成的影响,要尽可能使分析窗口边缘不存在示踪粒子。

自适应分析窗口选择的具体步骤如下:

(1)对图像 $I(i, j)$ 每个坐标点 (i, j) 的像素值进行判断,当 $I(i, j) \geq T$ 时,记为 $P_k(i, j)$,由于示踪粒子浓度小,个体大,为保证窗口具有一定的能量,对于 200 万像素的 PIV 图像,初始窗口一般选择为 128×128 大小,以 P_k 为中心取 128×128 的矩形区域 S_k ,按照式 (6) 计算 S_k 区域的能量 E_k :

$$E_k = \sum_{(i,j) \in S_k} I(i, j) \quad (6)$$

经实验分析,针对河势表面流场灰度图,这里 T 通常取 $90 \sim 120$ 。

(2)取出 E 中的最大能量值 $E_{\max}$,得到与 $E_{\max}$ 对应的最大能量区域 $S_{\max}$ 中心的坐标 $P_{\max}(i_{\max}, j_{\max})$,以 $(i_{\max}, j_{\max})$ 为中心取 128×128 大小的窗口 $W_{\max}$,就可以称 $W_{\max}(128, 128)$ 为固定窗口最大能量区域。

(3)步骤 (1) ~ (2) 求得的窗口区域 $W_{\max}(128, 128)$ 的边缘有可能存在一个或多个示踪粒子,当粒子产生运动时,在下一帧图像中取对应的窗口区域通常会存在粒子的进出窗口现象,造成相关计算的不准确。因此,在窗口 $W_{\max}$ 向内取边缘宽度为 L 的边缘区域,记为 $A_{\max}$,包含 A_a , A_b , A_c , A_d 四个区域,依次求其每个区域的能量,即区域内像素灰度值的和:

当 E_a , E_b , E_c , E_d 超过阈值 E_T 时,则认为边缘区域存在示踪粒子,按照式 (7) 移动窗口中心坐标,同时将窗口尺寸扩大 L ,中心坐标也同时做相应的移动得到 $W_{\max}(128 + L, 128 + L)$ 。

$$(i_{\max}, j_{\max}) = \begin{cases} (i_{\max} + L/2, j_{\max}), E_a > E_T, E_a = \sum_{(i,j) \in A_a} I(i, j) \\ (i_{\max} - L/2, j_{\max}), E_b > E_T, E_b = \sum_{(i,j) \in A_b} I(i, j) \\ (i_{\max}, j_{\max} - L/2), E_c > E_T, E_c = \sum_{(i,j) \in A_c} I(i, j) \\ (i_{\max}, j_{\max} + L/2), E_d > E_T, E_d = \sum_{(i,j) \in A_d} I(i, j) \end{cases} \quad (7)$$

实验分析,通常取 $E_T = \sum_{(i,j) \in A_a} I(i, j)/6$ 时效果较好。为了防止示踪粒子运动出窗口,一般将序列图像间最大运动速度的估计值作为 L 值大小。

(4)重复步骤 (3) 直到 A_a, A_b, A_c, A_d 区域的能量均小于 E_T 。得到分析窗口 $W_{\max}(128 + nL, 128 + nL)$,

其中 n 为自适应窗口调整的次数，为防止 n 过大造成的计算效率低下，这里规定 n 值不大于 10。

(5) 在第 2 帧图像中以 $(i_{\max}, j_{\max})$ 为中心取窗口 $W'_{\max}(128 + nL, 128 + nL)$ ，通过 $W_{\max}$ 和 $W'_{\max}$ 的互相关计算得到两帧图像的运动参数 $V(v_i, v_j)$ 。

3.2 基于 Hartley 变换的互相关计算

传统 PIV 算法应用 FFT 实现互相关运算，但是 FFT 法存在两个比较大的缺陷，一是它得到的是查询窗口的平均速度，没有考虑到窗口中的速度梯度。二是测量速度范围小。Hartley 变换是类似于傅里叶变换的积分变换，其正反变换的积分核相同，具有傅里叶变换的大部分特性，且实序列的 Hartley 变换仍是实序列，避免了变换过程中的冗余性，能成倍地节约内存空间。另外，Hartley 变换的快速实现 FHT 可采用 FFT 的结构形式，能进一步提高运算速度，更适合于需要实时批量处理的 PIV 图像分析。

与二维傅里叶变换不同，二维 Hartley 变换的积分核存在两种选择： $\text{cas}(ux+vy)$ 、 $\text{cas}(ux)\text{cas}(vy)$ 。为了便于快速实现，选择可分离的第二种形式，并得到如下的正逆变换表达式^[4,5]：

$$H(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \text{cas}(ux) \text{cas}(vy) dx dy \tag{8}$$

$$f(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(u, v) \text{cas}(ux) \text{cas}(vy) dx dy \tag{9}$$

转化为离散正逆变换定义如下：

$$H(m, n) = \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} f(x, y) \text{cas}(2\pi mx / N) \text{cas}(2\pi ny / N) \tag{10}$$

$$f(x, y) = \frac{1}{N^2} \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} H(m, n) \text{cas}(2\pi mx / N) \text{cas}(2\pi ny / N) \tag{11}$$

其中， $\text{cas}(x) = \cos x + \sin x$ 。

由于 Hartley 变换与傅里叶变换存在结构相似性，函数最终都可以变换成余弦、正弦的组成式，因此相互间可以进行转换。将 $H(m, n)$ 分解为奇对称分量 $H_e(m, n)$ 和偶对称分量 $H_o(m, n)$ 之和：

$$H(m, n) = H_e(m, n) + H_o(m, n) \tag{12}$$

其中：

$$H_e(m, n) = \frac{1}{2} [H(m, n) - H(-m, -n)] \tag{13}$$

$$H_o(m, n) = \frac{1}{2} [H(m, n) + H(-m, -n)] \tag{14}$$

可得二维傅里叶变换与 Hartley 变换的关系：

$$F(m, n) = H_e(m, -n) - jH_o(m, n) \tag{15}$$

因此，已知 $f(x, y)$ 的 DHT，则 DFT 可下式求得：

$$F(m, n) = \frac{1}{2} [H(m, -n) + H(-m, n)] - \frac{1}{2} j [H(m, n) - H(-m, -n)] \tag{16}$$

二维数据 $p(x, y)$ 与 $q(x, y)$ 的互相关函数的表达式为：

$$R_r(\Delta x, \Delta y) = \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} p(x, y) q(x + \Delta x, y + \Delta y) \tag{17}$$

其 Hartley 变换表达式为：

$$\begin{aligned} R_{\text{H}}(m, n) &= P_{\text{He}}(m, n) Q_{\text{He}}(m, n) - P_{\text{Ho}}(-m, n) Q_{\text{Ho}}(m, -n) + \\ &\quad P_{\text{He}}(-m, n) Q_{\text{Ho}}(m, n) - P_{\text{Ho}}(m, n) Q_{\text{He}}(m, -n) \end{aligned} \tag{18}$$

其中， $R_{\text{H}}(m, n)$ ， $P_{\text{H}}(m, n)$ ， $Q_{\text{H}}(m, n)$ 分别代表 $R_r(\Delta x, \Delta y)$ ， $p(x, y)$ ， $q(x, y)$ 的二维 Hartley 变换。

Hartley 变换不需要进行虚数变换，大大节约了计算时间。孙鹤泉^[6]等曾比较了快速傅里叶变换和 Hartley 变换，发现 Hartley 变换几乎节省了一半的时间。

4 实验与分析

图 2 (a) 所示为一组黄河模型表面流场旋涡的局部序列图像，时间间隔为 0.12s，分别采用传统 PIV 算法和本文算法进行速度矢量的计算，效果如图 2 (b) 所示。

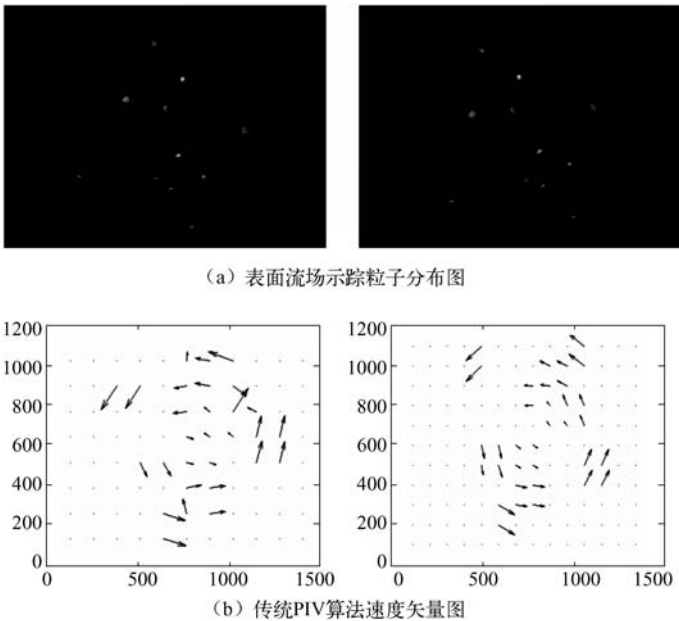


图 2 PIV 实验结果比较

可以看出，传统的 PIV 算法，存在个别错误方向的速度矢量，而本文算法则克服了这种情况。假设图 2 (a) 的运动参数为 $v_i = 0$ ， $v_j = 1, 2, 3, \dots, 10$ ，对图 2 (a) 取不同分析窗口的运动参数估计误差。

由图 3 可以看出，传统的固定窗口遍历法的运动参数互相关运动参数估计误差较大，而本文提出的自适应分析窗口法由于克服了示踪粒子进出窗口对速度测量的影响，有较小的估计误差值。

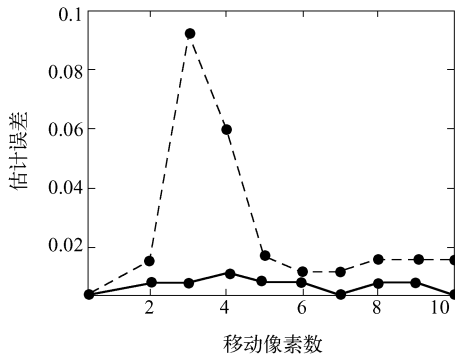


图 3 速度运动参数估计误差比较

本文算法采用自适应窗口分析法，降低了计算误差，却在一定程度增加了算法复杂度，这对需要实时批量处理图像序列是不利的。但是，由于示踪粒子浓度较小，使得图像计算目标数大大减少，同时采用了效率较高的基于 FHT 变换的互相关计算，整个测速算法效率并没有降低。本文作者在主频

为 Intel Xeon 2.13G 的 8 核 CPU，内存 2G 的主机上，同时用传统 PIV 计算方法和本文的改进方法对 20 组相同的间隔 0.12s 的表面流场图像进行速度矢量计算，所消耗的平均时间分别为 2.094s 和 1.482s，可知本文方法在效率上并没有降低反而有一定提高。

5 结论

随着计算机技术的发展，数字图像处理技术越来越多的应用于河工模型的实时测量中。本文根据河工模型的特点及表面流场示踪粒子的分布规则，提出了基于自适应分析窗口和 Hartley 变换的 PIV 改进方法，该方法能快速有效地测量出表面流场的速度矢量分布，相比较传统的 PIV 算法误差低，效率高，鲁棒性更好。符合大型河工模型表面流场的速度测量要求。从实验结果看出，应用粒子图像测速技术测量大型河工模型的表面流场速度具有广阔的前景。

参考文献

- [1] Adrian R J. Multi-point optical measurements of simultaneous vectors in unsteady flow-a review[J]. International Journal of Heat and Fluid Flow, vol. 7, pp. 127-145, June 1986.
- [2] Adrian R J. Particle-imaging techniques for experimental fluid mechanics[J]. Annual Review of Fluid Mechanics, vol. 23, pp.261-304, January 1991.
- [3] 田文栋, 魏小林, 盛宏至. DPIV 系统在河工模型试验中应用研究[J].水动力学研究与进展, 2001, 16(2): 209-215.
- [4] Bracewell R N.Discrete Hartley transform[J].J Opt Soc Am, 1983, 73(12): 182-183.
- [5] Waston A B, Allen Poirson.Separable two-dimensional discrete Hartley transform[J].J Opt Soc Am A, 1986, 3(12): 2001 - 2004.
- [6] 孙鹤泉, 沈永明, 王永学, 康海贵等. PIV 技术的几种实现方法[J], 水科学进展, 2004, 15(1): 105-108.
- [7] 王平让 .PIV 图像后处理新方法研究[D]. 大连: 大连理工大学.2004.
- [8] 陈红. “实体模型表面流场、河势数字图像测试方法及应用研究” . 南京: 河海大学, 2006.

一种团队 CGA 行进中的队形维护方法

郑延斌, 李双群

(河南师范大学计算机与信息技术学院, 河南 新乡, 453007)

摘 要: 行进是 CGA (Computer Generated Actors) 具备的基本功能, 是 CGA 完成任务的基础。团队 CGA 在行进过程中根据任务的要求, 需维持一定的队形。本文以 Leader-Following 为基础, 给出了一种基于队形分解的队形维护方法, 通过队形分解把复杂的队形维护工作分解为简单的横队和纵队的维护工作, 从而降低了队形维护工作的难度, 提高了队形维护的效率, 仿真试验表明该方法是有用的。

关键词: 团队 CGA; 队形维护; 队形分解; 避障

An Method for Formation Maintaining of Team CGA Advancing

ZHENG Yanbin LI Shuangqun

(College of Computer and Information Technology, Henan Normal University, Xinxiang 453007, Henan China)

Abstract: Advancing is basis function of CGA(Computer Generated Actors), and it is the basis for task doing. Team CGA must maintain a especial formation according to tasks. Based on Leader-Following method, presents a formation maintaining method based on the decomposed of complex formations, this method decompose the complex formations into simple horizontal formations and vertical formations, so the maintaining of complex formation becomes the maintaining of horizontal formations and vertical formations, solving the problem of formation maintaining in team CGA marching on land.

Keyword: Team CGA; formation maintaining; formation decompound; collision avoidance

1 引言

行进是 CGA (Computer Generated Actors) 所具备的基本功能, 是 CGA 完成任务的基础。在团队 CGA 的应用领域中, 要求 CGA 在执行任务或者运动过程中保持一定的队形, 尤其是在完成军事任务的过程中, 保持队形显得更为重要。目前队形问题已经成为多机器人、多 Agent 领域的一个富有挑战性的研究方向, 受到了国内外研究者的普遍重视^[10]。研究者提出的队形维护方法可以分为: Leader-Following^[1~3]、Behavior-Based Methods^[4~6]、Virtual Structure^[7~9]。针对不同的应用领域, 每种方法都有自身的缺点。本文以 Leader-Following 方法为基础, 给出一种基于队形分解的队形维护方法, 通过对复杂队形的分解, 把复杂的队形维护行为简化为简单的横队和纵队的维护行为, 针对团队中成员的角色不同, 给出了不同的行为, 简化了队形维护的难度, 提高了队形维护的效率。

2 基于队形分解的队形维护方法

2.1 基本队形

行进中的队形是比较复杂的, 研究较多的队形主要是一些几何对称的图形, 有列横队、纵队、三

河南省重点攻关项目 (102102210176, 102102210179), 河南省教育厅自然基金项目 (2010A520027)

作者简介: 郑延斌 (1964—), 男, 教授, 博士;
李双群 (1976—), 男, 讲师, 硕士。

角形、菱形和楔形（见图 1）。其他队形都可以分解为这些基本队形的某种组合形式。

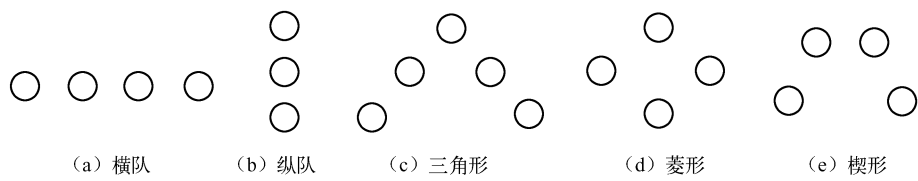


图 1 团队 CGA 编队的基本队形

2.2 队形的分解

定义（队形图）：团队 CGA 行进队形定义为一个有向图 $G = \langle V, \prec, C \rangle$ ，其中：

- (1) V 为 n 个顶点的集合， n 为团队中成员个数。每个顶点表示团队中的一个 CGA 成员；
- (2) 关系 $\prec \subseteq V \times V$ ，代表连接顶点之间的边，设 E 表示所有边的集合；
- (3) C 为 E 上的约束集合， $C = \{c_e\}_{e \in E}$ ，每个边 $e = (x_i, x_j)$ ， c_e 为一个 $\phi(e)$ 维的约束向量， $\phi(e) \in \mathfrak{R}$ 为约束条件个数， $c_e^k: \mathfrak{R} \times \mathfrak{R} \rightarrow \mathfrak{R}$ ， $k=1,2,\dots,\phi(e)$ ，定义了队形在 x_i 和 x_j 之间的所有约束，当满足所有约束时， $c_e^k(x_i, x_j) = 0$ 对所有 k 都成立。

由定义知团队成员之间的队形维护实际上是有向图边上的约束维护，若图形比较复杂，约束规则将会很多，势必增加队形维护的难度。因此为简化约束规则，提高队形维护的效率和质量，可以把复杂队形分解为简单队形的组合，图 1 给出的基本队形，横队和纵队不需要进行分解，其他的队形如三角形队形、菱形和楔形可以按照图 2 给出方法分解。图中黑圆为对象分解时增加的虚成员，并把它们作为子队形的 Leader 成员。通过这样的分解，团队中 Leader 成员形成的队形为纵队（包括虚拟 Leader），非 Leader 成员形成的队形为横队。

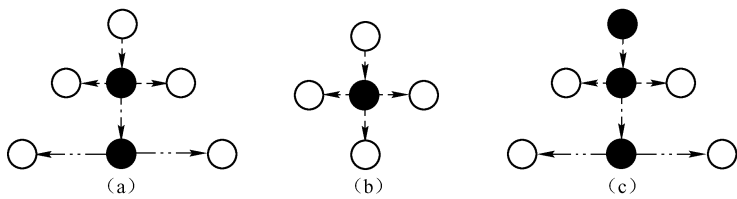


图 2 基本队形的分解

定理 1 一个队形图可以划分为若干个简单的横队队形子图和纵队队形子图。

定理 2 满足定义 1 的团队队形可以有其内的子团队通过横队队形或纵队队形来维持。

2.3 基于队形分解的队形维护方法

由队形分解可知，一个基本队形可以分解为一个纵队队形和若干个横队队形的组合，因此进行中的队形维护问题就转换为横队队形的维护和纵队队形的维护问题，下面就分别对这两种队形在行进中的维持问题进行讨论。

2.3.1 横队和纵队的维护问题

1) 纵队的维护问题

由于纵队的宽度为一个 CGA 的宽度，因此纵队的维持策略非常简单，当 Leader 的路径给定后，Leader 成员的速度和运动方向就确定了，其下属成员只需要按照 Leader 的路径前进，并且与 Leader 之间保持规定的距离，速度等于 Leader 的速度即可。

2) 横队的维护问题

横队的维护比纵队的维护相对复杂些，需要根据环境的情况来进行队形变换和压缩等处理，从横队的角度来看，环境中的障碍物分成两类：缺口障碍和独立障碍。

1) 缺口障碍

缺口障碍的形状如图 3 所示，主要指两个障碍物之间有一个通道，而该通道是团队行进的必经之地。

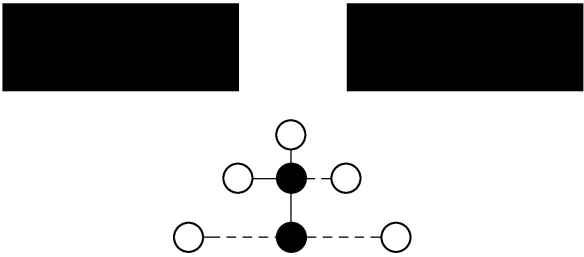


图 3 缺口障碍的形状

设团队 Leader 成员探得缺口障碍中间的缺口宽度为 $L_{\max}$ ，整个队形的宽度为 L_{Team} ，CGA 本身的宽度为 L_{CGA} （设团队中所有 CGA 具有相同的宽度）， $L_{\min}$ 为两个 CGA 之间允许的最小距离。则根据 $L_{\max}$ ， L_{Team} ， L_{CGA} ， $L_{\min}$ 之间的关系可以分为如下几种情况：

- ① 若 $L_{\text{Team}} \leq L_{\max}$ ，则整个队形能够顺利通过该障碍，不需要进行变换。
- ② 若 $L_{\text{CGA}} < L_{\max} < 2 L_{\text{CGA}} + L_{\min}$ ，表明缺口区域只能容一个 CGA 通过，不能有两个 CGA 按照横队通过，则整个队形需要变换为纵队，具体的变换方法为：按照成员的编号顺序来跟随 Leader 成员通过障碍物。
- ③ 若 $2 L_{\text{CGA}} + L_{\min} < L_{\max} < L_{\text{Team}}$ ，表明缺口区域可以允许两个 CGA 按照最小距离横队通过，此时需要对队形进行压缩处理，调整后的距离为： $(L_{\max} - 2 L_{\text{CGA}}) / 2$

3) 独立障碍

独立障碍示意图如图 4 所示，主要指开阔区域中的一个独立的障碍物，障碍物的宽度大于队形中要求的两个 CGA 之间距离，因此只能绕过该障碍物。遇到此类障碍物时，队形的维持分为如下两种情况：

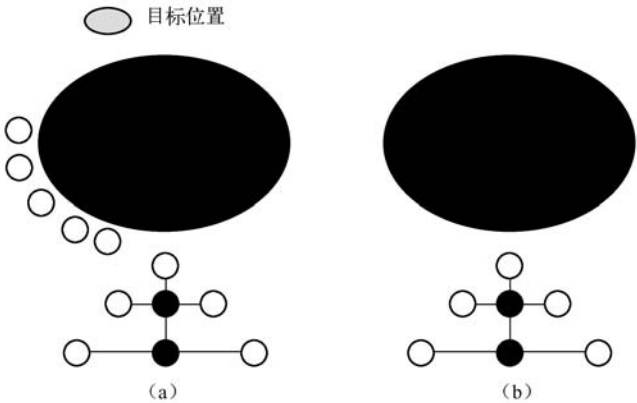


图 4 独立障碍示意图

- ① 若目标位置在障碍物的上方，在把横队队形变为纵队队形，从距离目标位置较近的一侧沿着障碍物的包围盒绕过障碍物，如图 4 (a) 所示。
- ② 若目标位置不在障碍物的上方，如果周围其他障碍对队形没有影响，则队形不变，如果周围

障碍物对队形有影响，实际上就变成上面的缺口障碍的情况，则按照缺口障碍的处理方法来处理。

2.3.2 地形的进一步处理

环境对队形的影响非常大，Leader 成员要不断地对障碍物进行检查，判断是否能够允许队形通过，虽然全局路径规划方法可以来为团队的 Leader 成员规划出一条路径，并最大限度地保证队形。但是需要在路径的每个部分进行标记，来记录该段路径是否能够允许整个队形的通过，还需要时刻通知下属成员在哪个位置需要转换队形，如何转换等。因此为了方便处理，同时也为了提高团队 Leader 处理速度，对团队 CGA 行进的地形进行如下处理。

1) 小障碍物的合并

由于团队 CGA 在进行过程中需要尽可能地保持队形，因此有些小的障碍物，它们之间的距离比较接近（小于某个阈值），不能保证整个队形顺利通过，从而导致队形频繁地变换。因此为了减少队形转换的次数，可以把距离比较接近的小障碍物合并成为一些较大的障碍物。虽然团队在行进的过程中行走的距离相对远一些，但是能够最大限度地保持队形。

2) 新障碍物包围盒

新障碍物包围盒就是重新给环境中每个障碍物做一个大的包围合，该包围合到障碍物边界的距离为 $L_{Team}/2$ （假设团队 Leader 成员的位置为队形的中心位置）。

障碍物的合并后，地形中障碍物主要包括缺口障碍和大障碍，经过包围盒处理后环境中的区域被分为三类（如图 5 所示）：障碍区域（在障碍物边界以内的点）；队形变换区域（障碍物边界外，大包围合内的点）；队形安全区域（在障碍物的大包围合外的点）。

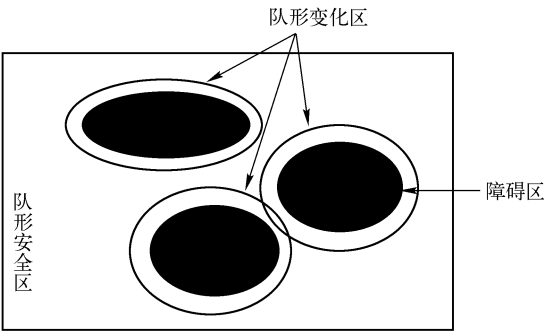


图 5 环境中的区域划分

当 CGA 行进在队形安全区域，队形可以很好地保持，当进入到队形变换区域时，有两种情况，即所遇到的障碍属于缺口障碍还是独立障碍，仍然需要 CGA 来判断。

3 团队成员的行为描述

团队的成员依据权利和义务的不同可以分为两类：即有下属的成员（包括虚拟成员，称为 Leader 成员），没有下属的成员（称为非 Leader 成员）。每类成员在队形维护的过程中发挥着不同的作用，具有不同的行为，下面分别对他们进行描述。

3.1 Leader 成员的行为

团队中的 Leader 包含整个团队的 HEADER 成员（这里把团队的最高管理者 x 定义团队的 Leader，满足： $getHeader(x)=\emptyset$ ），子团队的 HEADER 成员 x （可能是虚拟成员，满足 $getHeader(x)\neq\emptyset$ and $getMember(x)\neq\emptyset$ ）。算法 1 给出了 Leader 成员的行为。

算法 1: Leader 成员的行为(设该成员为 A)

- (1) 如果 $getHeader(A) \neq \emptyset$, 则根据其 Header 成员的位置、速度和下一目标点, 来调整自己的速度 (维持纵队队形);
- (2) 向下属成员发布自己当前状态 (位置、速度和下一目标点);
- (3) 如果 $getHeader(A) \neq \emptyset$ 则转 7;
- (4) 判断自己当前进入的区域, 如果是队形安全区转 10;
- (5) 如果遇到的是独立障碍, 则从该点向目标点做一条连线, 判断是否与该障碍物相交, 如果相交, 则按照独立障碍的第一种情况处理, 通知下属成员变为纵队队形; 转 10;
- (6) 如果是缺口障碍, 则计算缺口障碍的宽度 L_{max} , 利用缺口障碍的避障方法判断是否改变队形, 向下属成员发送队形类型及宽度; 转 10;
- (7) 接收其 Header 发送的信息;
- (8) 如果是状态信息, 则转 1;
- (9) 若是命令信息, 则把命令信息中的改变后的队形类型及宽度 (横队时有) 发送给其下属成员;
- (10) 向自己的目标行进;
- (11) 如果没有到达目标则转 1;
- (12) 如果 $getHeader(A) = \emptyset$ 则从路径列表中取出下一个关键点, 如果没取完, 则转 1; 如果取完, 则置自己的速度为零;
- (13) 如果 $getHeader(A) \neq \emptyset$ 且 Header 的速度不为零, 则转 1;
- (14) End。

3.2 非 Leader 成员的行为

团队中的非 Leader 成员在行进的过程中维持一个横队, 行进过程中队形的调整信息来自其 HEADER (即所在队形的 Leader 成员) 的命令。算法 2 给出了非 Leader 成员的行为。

算法 2: 非 Leader 成员的行为

- (1) 接收 Header 成员发送的信息;
- (2) 如果是状态信息, 则根据队形约束确定自己的速度和下一目标;
- (3) 若是命令信息, 根据需要改变的队形类型、宽度、Leader 的位置, 确定自己的位置和目标;
- (4) 向目标行进;
- (5) 如果没有达到目标则转 1;
- (6) 如果 Header 的速度不为零, 则转 1;
- (7) End。

4 仿真试验

为了验证本文算法的有效性, 我们设计了有 5 个 CGA 组成的一个团队, 按照三角形编队行进的过程, 其中编号 1 为该 CGA 团队 HEADER 成员, 它也是编队的 Leader, 编队中存在两个虚拟成员, 分别为 6、7。成员 1、6、7 构成一个纵队, 成员 1 为 Leader, 成员 2、6、3 构成一个纵队, 成员 6 为 Leader, 成员 4、7、5 构成一个纵队, 成员 7 为 Leader。图 6 给出了该团队经过缺口障碍时的运动轨迹。从图中看出, 团队经过第一个缺口障碍时, 1、6、7 构成的纵队队形不变, 2、6、3 构成的横队队形也不变, 而 4、7、5 构成的横队由于成员之间距离超过了缺口的宽度, 故把该队形调整为与 2、6、3 形成的队形一致。图 7 给出了三个 CGA 形成的团队在经过一个复杂障碍物时队形的变化情况,

其中成员 3 是团队的 Header 成员，也是编队的 Leader，增加一个虚拟成员 0，它和成员 1、2 形成一个横队。成员 0 为该队形的 Leader。由于第一个缺口障碍的宽度小于队形的宽度，因此将三角形队形变换为一个纵队队形通过该障碍。由于在行进地形中标记了队形安全点和队形变换点，因此当成员 3 到达队形变换点时，发送需要变换队形的命令给成员 0，成员 0 向成员 1、2 发布变换命令，把队形变换纵队，成员 1 和成员 2 跟随成员 3 行进。当到达队形安全点后，成员 3 发布恢复队形命令，团队成员调整自己的速度和目标，重新形成一个三角形队形，最后达到目标位置。

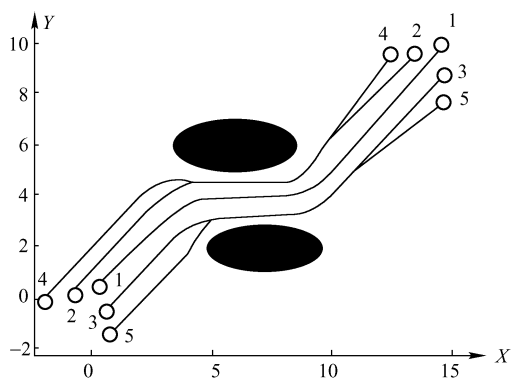


图 6 团队 CGA 通过缺口障碍时的运动轨迹

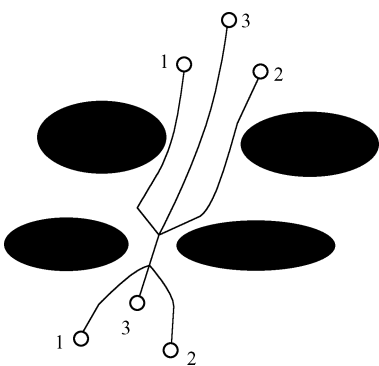


图 7 三个 CGA 构成的团队经过复杂避障时的轨迹

5 结论

本文提出了一种基于队形分解的团队 CGA 队形维护方法，将复杂的队形分解为若干个横队和纵队的组合，纵队是团队中若干 Leader 成员（包括虚拟 Leader）行进的队形，横队中成员的行为是有其 Leader 成员给定的，同时为了提供算法效率，减少 Leader 成员使用 LOD 算法的次数，对 CGA 行进的地形做了进一步的处理，从队形维持的角度将地形中的区域化分为三个区域：障碍区；队形转换区和队形安全区。只有当 Leader 成员进入队形转换区时，才需要对障碍进行判定。该算法有如下优点：（1）通过对 CGA 地形的进一步处理，减少了行进中 Leader 成员判断障碍物的次数；（2）将复杂队形分解为横队和纵队的组合，团队中 Leader 成员（包括虚拟 Leader）按照纵队行进，非 Leader 成员在其 Leader 的引导下，按照横队行进，简化了队形维护工作；（3）团队中非 Leader 成员在其 Leader 的引导下维持相互之间形成的横队，因为 Leader 了解其下属成员的位置信息，因此能够从整体上了解下属成员队形维持的效果，提高了队形维护的效果。

参考文献

[1] Wang P K C. Navigation Strategies for Multiple Autonomous Mobile Robots Moving in Formation[J]. Journal of Robotic Systems,1991,8(2):177-195.

[2] Jaydev P. Desai, Jim Ostrowski, Vijay Kumar. Controlling Formation of Mobile Robots[A]. IEEE Int Conf on Robotics and Automation[C]. Belgium, 1998:2864-2869.

[3] Hiroacki Yamaguchi, Joel W Burdick. Asymptotic Stabilization of Multiple Nonholonomic Mobile Robots Forming Group Formation[A]. IEEE Int Conf on Robotics and Automation[C]. Belgium, 1998: 3573-3580.

[4] Balch T. Ronald C A. Motor Schema-based Formation Control for Multiagent Robot Teams[C], In: Proceedings of the First International Conference on Multiagent Systems, June 12-14, 1995. 10-24.

[5] Tucher Balch, Ronald C Arkin. Behavior-based Formation Control for Multi-robot Teams[J]. IEEE Trans on Robotics and Automation, 1998, 14(6): 926-939.

- [6] Jadabaie A, Lin J, Morse A S. Coordination of Groups of Mobile Autonomous Agents using Nearest Neighbor Rules[J]. IEEE Trans on Automatic Control, 2003,18(6): 988-1001.
- [7] Anthony Lewis M, Tan K H. High Precision Formation Control of Mobile Robots using Virtual Structures autonomous [J]. Autonomous Robots, 1997,4:387-403.
- [8] Randal W Beard, Jonathan Lawton, Fred Y. Haddesh. A coordination architecture for spacecraft formation control[J]. IEEE Trans on Control Systems Technology, 2001, 9(6):777-790.
- [9] Brett J Young, Randal W Beard, Jed M Kelsey. A Control Scheme for Improving Multi-Vehicle Formation Maneuvers[A]. American Control Conf[C]. Arlington, 2001: 704-709.
- [10] 任德华, 卢桂章. 对队形控制的思考[J]. 控制与决策, 2005, 20(6): 601-606.

基于历史时序的访问控制模型研究

徐长征, 王清贤, 颜学雄

(解放军信息工程大学信息工程学院 网络工程系, 河南 郑州, 450002)

摘要: 本文提出一种以历史时序数据作为信任基础的访问控制模型。与传统的以信誉度作为信任基础的访问控制模型相比, 该模型能给资源拥有者提供细粒度且灵活的访问控制策略, 使得应用更加符合实际需求。阐明了该模型的工作思路, 提出了一种策略语言, 使用时序模态逻辑进行了描述, 并对其语法和语义进行了形式化定义, 随后给出了一种如何判定访问请求的策略可满足性判定算法。最后以电子商务应用为例说明了该模型的具体应用方式, 并与传统的模型进行了比较。

关键词: 访问控制; 信任; 时序逻辑; 格局

中图分类号: TP309 **文献标识码:** A **文章编号:**

Towards An Access Control Model Based on History Temporal Data

XU Changzheng, WANG Qingxian, YAN Xuexiong

(Department of Network Engineering, Information Engineering University, Zhengzhou 450002, Henan China)

Abstract: In this paper we propose a brand-new trust model based on history temporal data for access control. Compared with traditional trust models based on reputation which is only a numerical value, the model could provide more flexible method for user to custom specific access control policies. By using temporal logic a policy language is presented and syntax and semantics are defined formally. Later we introduce a verification algorithm to determine satisfaction of policies. E-commerce is used as an example to illustrate applications of the model and evident advantages of the fine-grained access control.

Keywords: access control; trust; temporal logic; configuration

1 引言¹

在大规模分布式网络应用中, 由于开放互连的特性, 资源拥有者不可能总是了解所有资源请求者的情况^[1]。尤其是在类似 P2P 多安全域环境中, 系统没有所谓的中心节点来充当管理者的角色, 域间主体间缺乏认知。因此, 信任关系常常被用来作为多域环境中访问控制的依据^[2, 3]。电子商务应用就是其中一个很好的例子。从顾客的角度看, 选购商品本质上是一种访问控制的过程。在这个过程中, 货款是一种资源, 顾客是这种资源的拥有者, 而商家把商品卖给顾客则是为了请求获得顾客的货款。为了尽可能取得顾客货款的拥有权, 商家会尽力满足顾客的需求, 而这种需求可以看做是顾客对其货款的一种访问控制策略。

在电子商务中, 由于顾客与商家不能面对面进行交易, 顾客一般无法看到真实商品, 因此交易在很大程度上需要依赖信任关系进行。顾客通常更愿意从他所信赖的商家处购买货物。目前, 在许多电子商务的实际应用中, 信誉度常常被用来刻画这种信任关系 (如 eBay、淘宝等), 顾客通常更愿意信任信誉度高的商家。

基金项目: 国家“863”计划基金资助项目 (编号: 2007AA01Z471); 河南省自然科学基金 (编号: 072300410260); 河南省基础与前沿技术研究计划 (082300410150)
作者简介: 徐长征 (1976—), 男, 助理研究员, 博士生, 主要研究方向为访问控制、信息安全;
王清贤 (1960—), 男, 教授, 博士生导师, 主要研究领域为网络信息安全;
颜学雄 (1975—), 男, 讲师, 博士, 主要研究领域为网络信息安全。

信誉度实际上是一个数值,关于如何计算信誉度目前已有许多研究^[4~7]。尽管具体计算方法不一样,但这些研究的思路 and 基础是一致的,即都以历史数值评价作为基础,按照一定公式进行加权累计得出。信誉度的增加比较缓慢,但如果商家因为欺骗顾客或服务态度不好或其他原因而遭到顾客的负面评价时,其信誉度将大幅降低。

以信誉度作为信任基础的访问控制存在一个很大的缺点,就是掩盖了资源请求者的历史访问细节,这使得难以进行细粒度的访问控制。而在许多实际应用中,资源拥有者常常需要了解资源请求者的历史访问细节,从而能够根据自己的特定需求来制定相应的资源访问控制策略。

针对以上问题,本文提出了一种以历史时序作为信任基础的访问控制模型,该模型可为资源拥有者提供细粒度的访问控制方法。第 2 节介绍相关概念;第 3 节使用模态逻辑语言描述访问控制的策略框架,在此基础上介绍了访问控制规程;第 4 节给出访问控制判定的验证算法,即判断资源请求者是否满足相应的控制策略;第 5 节以电子商务应用为例,说明本模型的具体应用,同时对比了以信誉度作为信任基础的访问控制模型。最后,第 6 节给出了工作小结及下一步研究内容。

2 基本概念

资源请求者访问资源的历史情况可以描述为一系列的事务记录。从抽象的角度看,资源访问事务可以描述为一个状态转换机,其中一个状态代表了在整个事务进展过程中可能出现的一个关键事件,而一次事务记录则看成是该状态转换机上从初始状态到终止状态的一条路径,该路径刻画了资源请求者在访问某个资源时所经历的整个历程。例如,在电子商务应用中,商品交易事务可抽象为如图 1 所示的状态转换机。

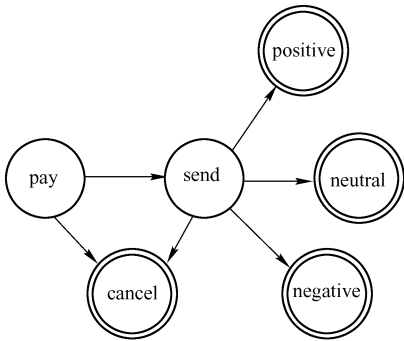


图 1 电子商务应用中商品交易事务状态转换机

图 1 中,状态 `pay` 表示顾客向商家付款, `send` 表示商家向顾客发送货物, `cancel` 表示顾客或商家取消了交易, `positive` 表示顾客对商家给出良好的评价, `negative` 表示顾客对卖家不满意,给出不良的评价, `neutral` 则表示顾客对商家评价一般。那么,该应用中的每个事务进程都可描述为其中若干事件的一个转换链。比如, `pay` → `send` → `positive` 就表示了一次正常交易事务,在该事务中,顾客向商家订购了一个商品后,向其付款,当商家收到货款后,向买家发货,最后顾客对商家的服务表示满意。

除了发展进程外,事务还有其他许多属性,如开始时间,资源请求者 id 等。因此,为了完整地表示一次事务,还应该考虑描述事务的其他属性。我们用 6 元组来表示一次事务:

$$T = \langle starttime, aid, pid, rid, accesstype, process \rangle$$

其中, `starttime` 表示事务的开始时间, `aid`, `pid` 和 `rid` 分别表示资源请求者标识号、资源所有者标识号和资源标识号, `accesstype` 表示访问类型, `process` 表示上述的关键事件状态转换链。

从本质上讲,本文介绍的访问控制模型是通过考察用户的历史资源访问记录,来查看其是否曾经产生过特定的访问事务。在这个过程中,我们并不太关心每个事务的具体进展过程,而只关心每个事

务的终止状态，即事务进程中最后的一个关键事件。据此，上述事务的 6 元组表示可简化为

$$T = \langle \text{starttime}, \text{aid}, \text{pid}, \text{rid}, \text{accesstype}, \text{finalstate} \rangle$$

其中，finalstate 是前述 process 的最后一个状态信息。我们将该 6 元组称为一个事务格局。例如，在电子商务应用中， $\langle 6, A, Q, \text{money}, \text{earn}, \text{neutral} \rangle$ 就表示一个事务格局，它表示在时间点 6，商家 A 和顾客 Q 开始了一个交易事务，从顾客角度看，商家请求访问的是顾客的货款，请求访问的方式是赚取，最后 Q 给 A 的评价一般。于是用户的历史资源访问记录就可表示为一系列按照时间先后顺序排列起来的事务格局。

3 访问控制框架

策略是访问控制的基石，任何访问控制规则都是通过策略语言来进行描述的。以下先给出所提访问控制模型的策略语言逻辑描述，然后在此基础上，介绍整个模型的访问控制过程。

3.1 策略语言的逻辑描述

任何一条策略都可分解为若干原子策略的组合，由于本文所提模型通过考察用户历史资源访问记录来进行访问控制，我们引入格局命题作为原子策略。每个格局命题是一个五元组，其表示形式如下：

$$P = \langle \text{aid}, \text{pid}, \text{rid}, \text{accesstype}, \text{finalstate} \rangle$$

其中，aid, pid, rid, accesstype 和 finalstate 的含义与前述事务格局 T 的含义一致。

对于任意一个格局命题 P 和事务格局 T，如果 $P.\text{aid} = T.\text{aid}$, $P.\text{pid} = T.\text{pid}$, $P.\text{rid} = T.\text{rid}$, $P.\text{accesstype} = T.\text{accesstype}$, $P.\text{finalstate} = T.\text{finalstate}$ ，则称 T 满足 P。

为了后面匹配验证的要求，我们需要给 P.pid 和 P.rid 预设一个特殊的取值 “*”（可以理解为程序语言的通配符）。当 $P.\text{pid} = \text{“*”}$ 或 $P.\text{rid} = \text{“*”}$ 时，它们将等于任何特定的 T.pid 和 T.rid。

上述原子策略通过与、或、非、蕴涵和等价逻辑运算符的有限次连接可以组成更为复杂的策略。

为了描述在历史事务中原子策略满足次数的特定关系，需要引入记数符 “#” 和关系符 “Rel”。对于原子策略 P，#P 表示在历史事务中 P 满足的总次数。“Rel” 被用来描述这种总次数的关系，它可以是一元关系，也可以是多元关系，例如对于原子策略 P_1 和 P_2 ， $\text{Rel}(\#P_1, \#P_2)$ 表示 # P_1 和 # P_2 之间的一种特定关系（# $P_1 > \#P_2$ 即是这样一个例子）。

此外，为了刻画历史事务的时序关系，还需要引入可描述过去时态的时序模态词，包括 P^{-1} 、 F^{-1} 、 A^{-1} 、 S^{-1} ，它们分别表示 “上一个”（Previously）， “在过去一段时间里最终”（Finally）， “过去一段时间里总是”（Always）， “自从”（Since）。

综上所述，下面采用 BNF 范式给出了策略语言的语法定义：

$$\varphi ::= P \mid \text{Rel}(\#P_1, \dots, \#P_N) \mid \neg\varphi \mid \varphi_1 \text{ op } \varphi_2 \mid P^{-1}\varphi \mid F^{-1}\varphi \mid A^{-1}\varphi \mid \varphi_1 S^{-1}\varphi_2$$

式中，N 为任意自然数；op 为二元逻辑运算符，包括 $\wedge$ 、 $\vee$ 、 $\rightarrow$ 、 $\leftrightarrow$ 。

在电子商务中顾客要选择他所信赖的商家，使用上述策略语言，可以进行如下描述：

$$\varphi : \neg F^{-1}((\# \langle \text{aid}, *, \text{money}, \text{earn}, \text{negative} \rangle) > 3)$$

上述公式意即，选取的商家 aid 需要满足：之前他在赚取顾客货款进行的所有交易中，从未得到超过 3 次的不良评价。

为了能够精确描述由上述语言表示的策略，需要对 φ 的语义进行形式化的定义。先介绍一些相关的符号说明。

设 $H = T_1 T_2 \dots T_N$ 表示用户资源历史访问记录，该记录是按时间先后顺序排列的一系列事务，令 $(H, i) (1 \leq i \leq N)$ 表示 H 的包含前 i 个事务的前缀，即 $(H, i) = T_1 T_2 \dots T_i$ ， $\text{Num}(H, i, P)$ 表示 $T_1 T_2 \dots T_i$ 中满足 P 的个数。据此，前述策略语言的语义可以形式地定义如下：

$(H, i) \models P$ 当且仅当

$T_i.aid = P.aid \wedge T_i.pid = P.pid \wedge T_i.sid = P.sid \wedge T_i.accesstype = P.accesstype \wedge T_i.finalstate = P.finalstate$

$(H, i) \models \text{Rel}(\#P_1, \dots, \#P_N)$ 当且仅当 $(\text{Num}(H, i, P_1), \dots, \text{Num}(H, i, P_N)) \in \text{Rel}$

$(H, i) \models \neg \phi$ 当且仅当 $(H, i) \not\models \phi$

$(H, i) \models \phi_1 \text{ op } \phi_2$ 当且仅当 $(H, i) \models \phi_1 \wedge / \vee / \rightarrow / \leftrightarrow (H, i) \models \phi_2$

$(H, i) \models P^{-1} \phi$ 当且仅当 $(H, i-1) \models \phi \wedge 1 < i$

$(H, i) \models F^{-1} \phi$ 当且仅当 $\exists j. (1 \leq j \leq i \wedge (H, j) \models \phi)$

$(H, i) \models A^{-1} \phi$ 当且仅当 $\forall j. (1 \leq j \leq i \wedge (H, j) \models \phi)$

$(H, i) \models \phi_1 S^{-1} \phi_2$ 当且仅当 $\exists j. (1 \leq j \leq i \wedge (H, j) \models \phi_2 \wedge \exists k. (j < k \leq i \rightarrow (H, k) \models \phi_1))$

3.2 访问控制规程

使用上述策略语言，资源拥有者可以根据自己的实际情况制定访问控制规则。实际上，资源拥有者可以为其所拥有的各个资源分别制定相应的访问策略。当一个用户向资源拥有者请求访问某个资源时，资源拥有者可根据资源 ID 号检索出相应的访问策略，然后检测该用户是否满足该策略，可按如下方式来进行检测：

(1) 查询资源请求者的历史访问记录，得到 H 。

(2) 检测 H 是否满足策略公式。

对于如何判定 $(H, N) \models \phi$ 是否满足，下一节将给出具体算法。

4 验证算法

从上一节对策略语言的语义定义可以看出，这种定义是一种归纳定义。实际上可以将这种定义改写为另一种等价的方式。

$(H, i) \models P$ 当且仅当

$T_i.aid = P.aid \wedge T_i.pid = P.pid \wedge T_i.sid = P.sid \wedge T_i.accesstype = P.accesstype \wedge T_i.finalstate = P.finalstate$

$(H, i) \models \text{Rel}(\#P_1, \dots, \#P_N)$ 当且仅当 $(\text{Num}(H, i, P_1), \dots, \text{Num}(H, i, P_N)) \in \text{Rel}$

$(H, i) \models \neg \phi$ 当且仅当 $(H, i) \not\models \phi$

$(H, i) \models \phi_1 \text{ op } \phi_2$ 当且仅当 $(H, i) \models \phi_1 \wedge / \vee / \rightarrow / \leftrightarrow (H, i) \models \phi_2$

$(H, i) \models P^{-1} \phi$ 当且仅当 $(H, i-1) \models \phi \wedge 1 < i$

$(H, i) \models F^{-1} \phi$ 当且仅当 $(H, i) \models \phi \vee (H, i-1) \models F^{-1} \phi$

$(H, i) \models A^{-1} \phi$ 当且仅当 $(H, i) \models \phi \wedge (H, i-1) \models A^{-1} \phi$

$(H, i) \models \phi_1 S^{-1} \phi_2$ 当且仅当 $(H, i) \models \phi_2 \vee ((H, i) \models \phi_1 \wedge (H, i-1) \models \phi_1 S^{-1} \phi_2)$

以上改造后的定义实际上已经显示出了一种判定 $(H, i) \models \phi$ 的方法，即采用动态规划算法来进行判断。于是针对 $(H, N) \models \phi$ ，可按如下方式来进行检测：

(1) 将 ϕ 分解为若干子公式，设分解后的子公式为 $\{\phi_0, \phi_1, \phi_2, \dots, \phi_n\}$ ，其中 $\phi_0 = \phi$ ，因为 ϕ 也是它自身的一个子公式，同时设置两个布尔数组 B_{pre} 和 B_{now} 。

(2) 对 B_{pre} 进行初始化：

如果 $\phi_i = P^{-1} \phi_j (0 \leq i \leq n, 0 \leq j \leq n, \phi_j \text{ 是 } \phi_i \text{ 的子公式})$ ，则 $B_{\text{pre}}[j] := \text{false}$ ；

如果 $\phi_i = F^{-1} \phi_j (0 \leq i \leq n, 0 \leq j \leq n, \phi_j \text{ 是 } \phi_i \text{ 的子公式})$ ，则 $B_{\text{pre}}[i] := \text{false}$ ；

如果 $\phi_i = G^{-1} \phi_j (0 \leq i \leq n, 0 \leq j \leq n, \phi_j \text{ 是 } \phi_i \text{ 的子公式})$ ，则 $B_{\text{pre}}[i] := \text{true}$ ；

如果 $\phi_i = \phi_j S^{-1} \phi_k (0 \leq i \leq n, 0 \leq j \leq n, 0 \leq k \leq n, \phi_j \text{ 和 } \phi_k \text{ 是 } \phi_i \text{ 的子公式})$, 则 $B_{\text{pre}}[i] := \text{true}$ 。

(3) 从 H 的第一个访问操作开始, 执行如下动作, 直到所有访问操作结束。假设当前考察的访问事务为 T :

如果 $\phi_i = T$, 则 $B_{\text{now}}[i] = \text{ture}$;

如果 $\phi_i = \neg \phi_j (0 \leq i \leq n, 0 \leq j \leq n, \phi_j \text{ 是 } \phi_i \text{ 的子公式})$, 则 $B_{\text{now}}[i] := \neg B_{\text{now}}[j]$;

如果 $\phi_i = \phi_j \text{ op } \phi_k (0 \leq i \leq n, 0 \leq j \leq n, \phi_j, \phi_k \text{ 是 } \phi_i \text{ 的子公式})$, 则 $B_{\text{now}}[i] := B_{\text{now}}[j] \text{ op } B_{\text{now}}[k]$;

如果 $\phi_i = P^{-1} \phi_j (0 \leq i \leq n, 0 \leq j \leq n, \phi_j \text{ 是 } \phi_i \text{ 的子公式})$, 则 $B_{\text{now}}[i] := B_{\text{pre}}[j]$;

如果 $\phi_i = F^{-1} \phi_j (0 \leq i \leq n, 0 \leq j \leq n, \phi_j \text{ 是 } \phi_i \text{ 的子公式})$, 则 $B_{\text{now}}[i] := B_{\text{now}}[j] \vee B_{\text{pre}}[j]$;

如果 $\phi_i = G^{-1} \phi_j (0 \leq i \leq n, 0 \leq j \leq n, \phi_j \text{ 是 } \phi_i \text{ 的子公式})$, 则 $B_{\text{now}}[i] := B_{\text{now}}[j] \wedge B_{\text{pre}}[j]$;

如果 $\phi_i = \phi_j S^{-1} \phi_k (0 \leq i \leq n, 0 \leq j \leq n, 0 \leq k \leq n, \phi_j \text{ 和 } \phi_k \text{ 是 } \phi_i \text{ 的子公式})$, 则 $B_{\text{now}}[i] := B_{\text{now}}[k] \vee (B_{\text{now}}[j] \wedge B_{\text{pre}}[i])$;

将 B_{now} 的值赋予 B_{pre} ;

(4) 如果 $B_{\text{now}}[0] = \text{true}$, 则 $(H, N) \models \phi$ 成立。

5 实例研究

以上介绍的逻辑语言为以历史时序作为信任基础的访问控制提供了一种灵活的策略定制方式。针对不同的具体应用, 上述策略语言不需要做较大改动即可得到应用。本节将以电子商务应用为例, 说明策略定制的灵活性。

假设顾客 G 认为只有不良评价与良好评价之比始终不超过 1:3 (限于篇幅, 这里只是举一个简单示例来说明问题, 在实际应用中该比值常常会比较小) 的商家才是其值得信赖的对象。则该标准可用如下策略语言描述。

$$\phi' : A^{-1}((\# \langle \text{aid}, *, \text{money}, \text{earn}, \text{positive} \rangle) > 3(\# \langle \text{aid}, *, \text{money}, \text{earn}, \text{negative} \rangle))$$

设商家 A 和 B 都有顾客 G 欲购买的同一种商品, 如图 2 所示, 它们的历史交易记录分别为:

H_A :	H_B :
<1, A, X, money, earn, positive>	<3, B, D, money, earn, positive>
<6, A, Q, money, earn, neutral >	<7, B, N, money, earn, positive>
<14, A, U, money, earn, positive>	<19, B, F, money, earn, positive>
<20, A, D, money, earn, positive>	<25, B, T, money, earn, positive>
<26, A, N, money, earn, negative>	<29, B, X, money, earn, positive>
<30, A, U, money, earn, neutral >	<32, B, Q, money, earn, negative>
<40, A, P, money, earn, positive>	<39, B, S, money, earn, negative>
<45, A, K, money, earn, positive>	<48, B, F, money, earn, positive>
<50, A, U, money, earn, positive>	<53, B, Y, money, earn, positive>
<53, A, Y, money, earn, negative>	<60, B, K, money, earn, neutral >

图 2 商家历史交易记录

上述记录中, 尽管商家 A 和 B 是卖商品给顾客, 但从顾客的角度看, 二者都是在请求货款这个资源, 因此 H_A 中每条记录的第 2 项都为 A , 而 H_B 中每条记录的第 2 项都为 B 。

电子商务系统根据 ϕ' 及 H_A 和 H_B , 使用上一节给出的验证算法, 将为顾客 G 推荐出商家 A 。因为从上述历史交易记录可以看出, 在任何时间点上, 商家 A 得到的不良评价与良好评价之比始终不超过 1:3; 而对于商家 B , 在时间点 39 时, 其不良评价与良好评价的比率已达到 2:5, 从而不满足顾客 G 的信赖标准。

实际上, 顾客 G 的信赖依据 ϕ' 是由其自身定制的, 他完全可以根据不同的时间或不同的环境来改变这种标准, 例如, G 在某种特定场合下可能会改变 ϕ' 中不良评价与良好评价的比率, 将其设为 1:20, 也可能会使用如下全新的策略 ϕ'' 作为其信赖依据。 ϕ'' 所表达的意思是没有得到超过 200 次不

良评价的商家或者如果商家 aid 得到 200 次不良评价之后, 其不良评价与良好评价之比不会超过 1 : 10, 都是 G 所信赖的对象。

$$\phi'' : A^{-1}(10(\#<aid, *, money, earn, posi>)<(\#<aid, *, money, earn, posi>)) S^{-1}((\#<aid, *, money, earn, negative>)=200)$$

对于传统的以信誉度作为信任基础的访问控制, 由于信誉度是一个数值且由系统统一计算出, 因此顾客只能通过比较信誉度大小来判断哪个商家更可信, 而不能根据自己的具体情况来选择所信赖的对象, 而且所有顾客得到的结果都一样。以 H_A 和 H_B 为例, 使用文献[6]中方法来分别计算 A 和 B 的信誉度, 设定良好评价记分为 1, 一般评价记分为 0, 不良评价记分为-4, 并且不考虑交易价值及交易时间久远的影响, 则可以分别得到 $Credit_A = 4$ 和 $Credit_B = 5$ 。由此, 从传统方法的观点来看, 对于所有顾客, 商家 B 更值得信赖。

在实际生活中, 每个顾客对商家通常都有自己的一个信赖标准, 即每个顾客对商家行为的看重点不一样, 因此使用统一的数值比较来确定哪个商家更值得信赖有时并不合适, 因为这并不能反映真实情况。因此在上述例子中, 对于顾客 G 来说, 选择 A 为信赖对象更加合适。

6 结束语

本文提出了一种以历史时序作为信任基础的访问控制模型, 由于考察了资源请求者的历史访问记录, 因此该模型可以为资源拥有者提供更加灵活和细粒度的访问控制方式, 各个资源拥有者可根据自身具体情况来定制相应的访问控制策略, 从而更加符合实际应用需求, 而这是传统的以信誉度作为信任基础的访问控制方式所无法提供的。虽然本文以电子商务应用为例说明了该模型的实际应用, 但该模型无须作较大改动即可应用到其他环境, 如网格计算中对移动代码请求计算资源的访问控制。本文所考察的信任关系还属于直接信任, 如何以历史时序数据为基础来考察间接信任将是本文下一步需要研究的内容。

参考文献

[1] E. Yuan and J. Tong. Attributed Based Access Control (ABAC) for Web Services. In ICWS'05: IEEE International Conference on Web Services, Orlando, p569. IEEE, July 2005.

[2] X.N. Ma, Z.Y. Feng, C. Xu, J.F. Wang. A Trust-Based Access Control with Feedback. 2008 International Symposiums on Information Processing. pp.510-514. 2008.

[3] F.J. Feng, C. Lin, D.S Peng, J.S. Li. A Trust and Context Based Access Control Model for Distributed Systems. 2008 10th IEEE International Conference on High Performance Computing and Communications. pp.629-634. 2008.

[4] C. Su, H. Zhang, F.M. Bi. A P2P-based Trust Model for E-Commerce. International Conference on e-Business Engineering (ICEBE'06). IEEE. 2006, pp.118-122.

[5] A. A. Selcuk, E. Uzun, and M. R. Pariente. A Reputation-Based Trust Management System for P2P Networks, CCGRID2004, April 2004, pp.251-258.

[6] H. Tran, M. Hitchens, V. Varadha rajan et al. A Trust based Access Control Framework for P2P File-Sharing Systems. In Proceedings of the 38th Hawaii International Conference on System Sciences. IEEE. pp.302c, 2005.

[7] T. Yu, M. Winslett, and K. E. Seamons. Supporting Structured Credentials and Sensitive Policies through Interoperable Strategies for Automated Trust Negotiation. ACM Trans. Inf. Syst. Secur., 6(1):1-42, 2003.

基于 FIDXP 的分布式入侵防御系统的设计

刘 松, 赵东明, 周清雷

(郑州大学信息工程学院, 河南 郑州, 450001)

摘 要: 面对当前严重的网络安全问题, 需要用立体防御的方法来应对。入侵检测和防火墙技术是网络安全的重要组成部分, 但它们之间相互独立, 不能联动起来保护网络安全。本文设计了一种安全联动协议, 可以结合防火墙和入侵检测各自的优点, 实现实时检测和阻断攻击, 可以用来构建分布式的入侵防御系统。实验表明该协议安全有效, 可以增强现有网络的保护能力。

关键词: 入侵检测系统; 防火墙; 联动协议

中图法分类号: TP309 **文献标识码:** A

FIDXP-based Distributed Intrusion Detection System Prevention and Implementation

LIU Song , ZHAO Dongming, ZHOU Qinglei

(School of Information Engineering,Zhengzhou University, Zhengzhou 450001, Henan China)

Abstract: Facing the serious network security problem ,we should use dynamic system to protect the internet.Intrusion detection and firewall are important components of network security ,but they are separate and can't work with each other. This paper designed a protocol FIDXP ,it can combine the virtue of firewall and ids ,it can be used to build a distribute intrusion detection system to detect and block the attacker in realtime.The experiment show that this protocol is useful and workable.It can enhance the security level of current network .

Keywords: intrusion detection system; firewall; interaction protocol

1 引言

当前网络安全状况日趋恶劣, 从互联网上下载黑客攻击工具十分容易, 所以现在发动一次攻击也非常简单。

防火墙和入侵检测系统是目前用的最广泛的两大安全产品。防火墙 (Firewall)^[1] 的主要功能是串行接入网络, 以安全策略为基础, 将内外网络隔离开, 对内网和外网的通信进行访问控制, 对不符合访问控制策略的数据包进行丢弃。入侵检测系统 (IDS)^[2] 的功能主要是旁路监听受保护网络, 以攻击特征库为基础, 检测分析网络访问是否有攻击的内容, 如果发现攻击, 就进行报警和响应。IDS 也具有一定的阻断连接功能, 但不能有效地实现对恶意攻击的连接进行实时阻断的功能, 而且无法与其他安全设备交互, 所以需要设计一种安全联动协议, 使 Firewall 和 IDS 能够协同起来^[3], 发挥 IDS 的检测优势和 Firewall 的阻断优势, 共同保护网络的安全。

本文将以开源的防火墙软件 iptable 和入侵检测软件 snort 为基础, 设计一种安全的联动协议 FIDXP (Firewall and Ids exchange protocol), 构建分布式的入侵防御立体防御体系。

863 项目资助: 基于 ASP 模式的软件服务支持技术研究 (2007AA010408)

作者简介: 刘松 (1983—), 男, 硕士, 主要研究方向: 网络安全。

赵东明 (1964—), 女, 副教授, 硕士, 主要研究方向: 算法分析与设计, 网络安全;

周清雷 (1962—), 男, 教授, 博士生导师, 博士, 主要研究方向: 模型检测, 信息安全。

2 系统分析与设计

要构建分布式入侵防御体系，当 IDS_A 检测到攻击威胁时，不但需要与 FW_A 进行联动，动态地添加 FW_A 的阻断规则。还要能够使 IDS_A 与 FW_B 和 FW_C 进行联动。构成立体防御体系。结合 IDS 的攻击检测优势和防火墙的网络连接阻断优势，可以建立分布式的入侵防御体系。系统网络拓扑图如图 1 所示。

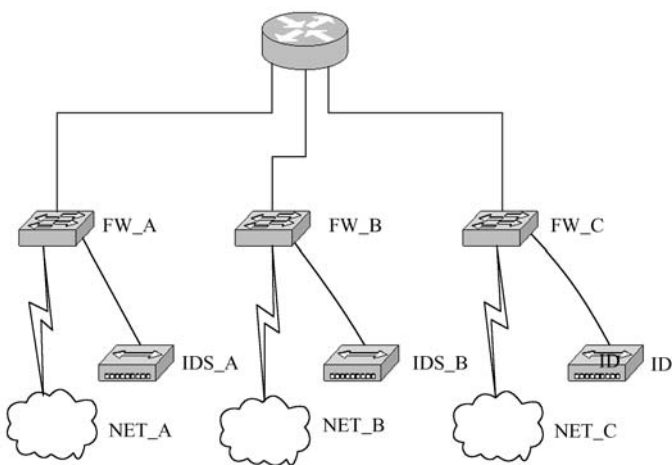


图 1 系统网络拓扑图

为实现安全设备之间的信息交互，美国国防高级研究计划署（DARPA）和互联网工程任务组（IETF）的入侵检测工作组（IDWG）发起制订了一系列建议草案^[4]，从体系结构、API、通信机制、语言格式等方面规范 IDS 的标准。用于入侵检测与响应（IDR）系统之间或与需要交互的管理系统之间的信息共享。但这样的方案需要对现有的安全软件进行升级，对于大量使用开源软件 snort 和 iptable 的用户来说代价太大。

Fwsnort^[5]是另一个开源项目，其目的是将 snort 的检测规则转换为 iptable 的阻断规则，这固然结合了二者的优点，但是没有体现出动态性，一个 snort 与多个 iptable 之间交互规则时出现困难。

本文的思路如下：根据 snort 的报警日志，分析提炼出攻击的五元组（protocol, sip, dip, sport, dport）信息，同时根据报警的优先级别，设置一个阻断时间（blocktime），将这这六元组信息发给联动的 iptable。iptables 收到联动消息后，动态添加一条带有时时间限制的阻断规则，当阻断时间到期后，再删除刚才添加的那条阻断规则。这样就能够结合 iptable 和 snort 各自的优点，构建一个分布式的入侵防御系统。

snort 检测到攻击后会报警并且记录日志。其日志格式如下：

[Classification: Misc activity] [Priority: 3]

04/13-09:55:05.849213 10.1.1.1:4190 → 10.2.2.5:7001UDP TTL:128 T OS:0x0 ID:56560 IpLen:2 0 DgmLen:60 Len: 32

其中，[Priority: 3]是报警的优先级，数值越大表明威胁越严重。其他字段都是本次通信的具体协议特征。

iptables 是 Linux 中自带的一款开源防火墙，可对经过的数据包进行 NAT 转换、DROP、LOG、ACCEPT 和 REJECT 等操作。

以上面 snort 的报警日志为例，当 iptables 收到 snort 发来的联动阻断消息之后，可将报警信息转换为如下一条 iptable 规则：

`iptables -I INPUT 1 -p udp -s 10.1.1.1 --sport 4190 -d 10.2.2.5 --dport 7001 -j DROP`

为了能实现规则的时效性，需要 `snort` 根据检测到的威胁的严重程度，将阻断的时间告诉 `iptables`。阻断时间是报警优先级的线性函数。定义如下：

$$\text{BlockTime} = \text{priority} * \text{Basetime}。$$

`Basetime` 是基准的阻断时间。报警优先级越高则阻断时间越长。到达阻断时间之后，该条规则需要被删除，否则 `iptables` 规则将无限膨胀。但是 `iptables` 本身没有动态删除规则的功能，需要安装一个 `ipset` 模块^[6]，它在指定的一段时间到期之后，会自动删除该条 `iptables` 规则，可以实现规则的时效性，这样就不会导致 `iptables` 规则的无限膨胀。举例如下：

`ipset -N list iptree --timeout 180`

它表示对 `list` 中的地址阻断 180s，时间到期之后就会自动清除该设定。为了实现联动，需要设计一个安全健壮的联动协议，下面将介绍 `FIDXP` 协议的设计和实现。

3 安全联动协议设计

`FIDXP` 协议要保证 `snort` 和 `iptables` 之间的安全通信，就要满足信息安全的机密性，完整性，保密性，可靠性等安全属性^[7]。

文中用到得术语定义：

- Ids:** 联动的 `snort` 入侵检测系统。
- Fw:** 联动的 `iptables` 防火墙。
- msg:** `Ids` 发给 `Fw` 的联动消息。
- Reply:** `Fw` 回复给 `Ids` 的确认消息。
- seq:** `msg` 包的同步序列号。
- Kind:** `reply` 包的类型。
- Content:** 联动消息体的具体内容，包括协议，地址，阻断时间。
- C_seqnum:** `Ids` 选择的同步序列号。
- S_seqnum:** `Fw` 选择的同步序列号。
- Key:** 预共享密钥。
- Auth_ct:** 校验的内容，用于身份认证。
- Auth:** 计算得到的 `HMAC` 校验值，用于完整性校验。
- Blowfish:** `Blowfish` 对称加密算法。
- HMAC:** 带密钥得摘要算法。
- Blocktime:** 阻断时间。

安全联动协议的主要交互过程是：`Ids` 发送联动消息 `msg_1` 给 `Fw`。`Fw` 校验通过之后回复一个确认响应 `reply_1`。`Ids` 收到确认响应 `reply_1` 且校验通过，发送下一条消息 `msg_2`，里面包含了对 `reply_1` 的确认，`Fw` 端收到 `msg_2` 之后要判断其中是否包含对 `reply_1` 的确认，判断通过，再发送针对 `msg_2` 的确认响应 `reply_2` 给 `Ids`。双方都要对消息进行校验和重放判断。系统运行状态如图 2 所示，其中 `si` 表示 `Ids` 的状态，`fi` 表示 `Fw` 的状态。

`Ids` 将报警信息发给防火墙，如果收不到回复，就等待一段超时时间然后重发数据包。收到回复包之后，要对数据包进行身份认证和完整性校验。其中要经过跟 `Fw` 端进行同步序列号协商。其中 `Ids` 发送的消息（`msg`）由以下内容构成。

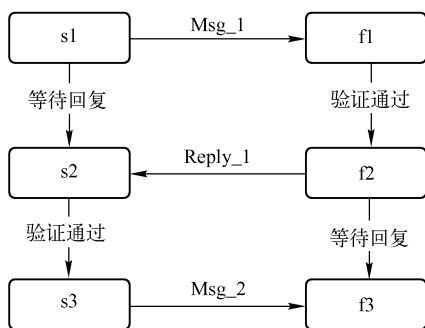


图 2 系统运行状态

数据格式定义:

$Ids.msg = Ids.auth\_ct + Ids.auth$

$Ids.auth\_ct = Blowfish(Ids.KEY, Ids.seq + Ids.c\_seqnum + Ids.s\_seqnum + Ids.content);$

$Ids.Content = (sip, dip, sport, dport, protocol, blocktime);$

$Ids.auth = HMAC(Ids.auth\_ct + Ids.key)$

Fw 端在指定端口监听, 收到数据包之后就进行完整性验证和身份验证。还要进行重放判断。验证判断完毕, 根据阻断信息, 添加一条阻断策略。

Fw 的消息 (reply) 构成内容如下:

$Fw.reply = Fw.auth\_ct + Fw.auth$

$Fw.auth\_ct = Blowfish(Fw.KEY, Fw.kind + Fw.c\_seqnum + Fw.s\_seqnum);$

$Fw.auth = HMAC(Fw.auth\_ct + Fw.key)$

联动协商和实现过程如下:

初始阶段, Ids 跟 Fw 协商一个共享密钥 key。使得 $Ids.key = Fw.key$ 。

[s1]Ids 发送联动消息给 Fw。

$Ids.Content = (sip, dip, sport, dport, protocol, blocktime);$

$Ids.auth\_ct = Blowfish(Ids.KEY, Ids.seq + Ids.c\_seqnum + Ids.s\_seqnum + Ids.content);$

$Ids.auth = HMAC(Ids.auth\_ct + Ids.key);$

$Ids.msg = Ids.auth\_ct + Ids.auth$

[s2]Ids 等待收到 Fw 发来得消息回复。

[f1]Fw 收到 Ids 的联动消息, 用 Fw.key 进行身份验证和完整性检查。

$Auth = HMAC(Ids.auth\_ct + Fw.key);$

$Ids.auth = ? = auth;$

解密 $Ids.auth\_ct$ 得到联动消息。进行同步判断和防重放判断。

$Ids.s\_seqnum = ? = Fw.s\_seqnum;$

$Fw.c\_seqnum \leq ? = Ids.c\_seqnum;$

验证通过, 更新 Fw 端得同步序列号。

$Fw.c\_seqnum = Ids.c\_seqnum;$

$Fw.s\_seqnum ++;$

构造响应消息发给 Ids。

$Fw.auth\_ct = Blowfish(Fw.KEY, Fw.kind + Fw.c\_seqnum + Fw.s\_seqnum);$

$Fw.auth = HMAC(Fw.auth\_ct + w.key);$

$Fw.reply = Fw.auth\_ct + Fw.auth$

[f2]Fw 等待收到 Ids 得回复消息。

[s3]Ids 收到 Fw 得响应包 reply,用 Ids.key 进行完整性验证。

Auth = HMAC(Fw.auth_ct + Ids.key);

Fw.auth =?= auth;

解密 Fw.auth_ct,得到 Fw 回复包的同步序列号,看是否同步,是否为消息重放。

Ids.c_seqnum =?= Fw.c_seqnum;

Ids.s_seqnum <=? Fw.s_seqnum;

验证通过,更新 Ids 端得同步序列号。

Ids.s_seqnum = Fw.s_seqnum;

Ids.c_seqnum++;

发送下一条消息给 Fw 端。

[f3]Fw 收到消息之后,先进行完整性验证, 验证通过, 解密数据包。判断消息是否是对上一个响应消息的确认。

Fw.c_seqnum = Ids.c_seqnum;

Fw.s_seqnum ++;

确认通过,更新 Fw 端同步序列号,发送对本次消息的确认。

4 安全性分析

评价一个协议的安全性, 主要包括机密性, 完整性, 可用性, 认证性等指标。下面将分析本协议的安全性。

身份认证: 本协议使用预共享密钥的方式。进行通信之前, Ids 要与 Fw 约定一个共享的 key 作为双方的共享密钥 key。其中 key=MAC(Ids.ip, Fw.ip);由通信双方的 ip 地址通过 hash 函数计算得出。

机密性: 对于传输的阻断消息内容, 使用 Blowfish^[8]加密方法加密。Ids.msg = Blowfish(Ids.key , Ids.content), 发给 Fw 之后, Fw 使用同样的密钥 key 解密得到阻断消息, 添加阻断策略。

Blowfish 是对称加密算法, 加解密速度非常快, 对于 Ids 和 Fw 这样高实时性设备来说非常重要, 不能因为计算加解密消耗大量 cpu 时间。本协议中, 每次 Ids 和 Fw 协商同步序列号时, 都会选取一个随机值, 更增加了破译难度。使用 Blowfish 加密方法, 其安全性依赖于共享 key 的安全性。

完整性: 加密过后的数据, 还要使用 md5 计算一个完整性认证尾部。auth = md5(Ids.msg, key); 即将共享密钥 key 和加密后的消息密文一起计算完整性校验值。如果数据在传输中遭遇修改或破坏, 接收方都能检测出来, 从而拒绝接收。

抗重放: 初始通信阶段, Ids 和 Fw 之间要相互通告自己随机选取的一个同步序列号, 将其作为消息的一部分被加密传输。每发送完一个消息就将同步序列号加一。如果收到序列号回退的包, 说明遭遇重放包, 将之丢弃。当序列号加到最大导致数值回绕归零时, 要重新协商同步序列号。这样每一轮的数据传输初始序列号都是不同的, 更增强了抗重放特性。

可用性: 本协议使用 UDP 传输。因为 Ids 和 Fw 都是安全设备, 要求有较高的数据处理性能。UDP 协议是无连接协议, 连接的建立、维护和终止, 其开销都较 TCP 小很多。本协议中使用重传机制保证了 UDP 协议的可靠传输。

由上述分析可知, 本协议较好地保证了信息传输的安全性。

5 实验

实验使用两台防火墙 Fw_A 和 Fw_B 及两台 snort, 即 Ids_A 和 Ids_B。Fw 使用 Redhad9.0 自带的

iptables, Ids 使用 snort2.8, 入侵测试程序使用 x-scan3.3。为了测试该联动协议的有效性, 编写了两个应用程序 Ids_client 和 Fw_server。

Fw_server 是个服务器程序, 监听 2010 端口, 通信协议使用 UDP 协议。Fw_server 通过读取配置文件 Fw.conf 得到要联动的 Ids 的地址和预共享密钥。Fw.conf 配置格式如下:

```
Ids.ip_A key_A
Ids.ip_B key_B
```

Ids_client 是客户端程序, 通过分析 snort 的报警日志给 Fw 端发送联动消息。Ids_client 也要读取配置文件 Ids.conf, 得到联动端的 Fw 地址和预共享密钥, 以及设定的基础阻断时间, 配置格式如下:

```
Fw.ip_A key_A      time_A
Fw.ip_B key_B      time_B
```

修改配置文件, 使 Ids 和 Fw 之间能相互联动。对 Ids_A 保护的网路用 x-scan 扫描, 如图 3 所示。

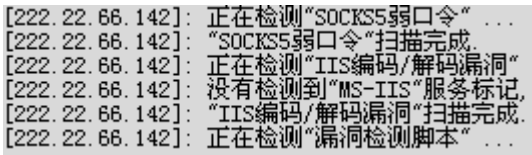


图 3 x-scan 扫描

此时在 Fw_A 端和 Fw_B 端, 可以看到添加的阻断规则, 如图 4 所示。

```
[root@localhost root]# iptables -L
Chain INPUT (policy ACCEPT)
target     prot opt source                destination
DROP       tcp  --  222.22.66.132          222.22.66.142          tcp spt:1102 dpt:http
```

图 4 iptable 添加的规则

将抓取的数据包重放给 Fw_A, 不会生成 iptable 规则, 说明 Fw 检测到这是重放数据包, 不添加规则。

对联动程序进行修改, 去掉对消息加密和计算认证尾部的动作, 实现明文传输, 以此比较使用 FIDXP 进行传输时的时间开销。

使用有重复的 1000 条报警信息, 循环发送联动消息, 实验结果如表 1 所示。

表 1 明文和密文传输时间开销实验结果

消息 (条)	明文传输 (s)	密文传输 (s)
1000 1		1
10000 1		2
20000 2		3
50000 6		8

由表 1 可知, 经过消息加密和计算认证尾部, 并没有带来很大的时间开销, 协议中使用的 Blowfish 和 md5 计算没有影响程序的实效性。

6 结论

本文使用 UDP 协议设计了一个用于 snort 和 iptable 之间传输阻断消息的安全联动协议 FIDXP, 增强了网络的立体防御能力。实验表明, 使用该协议可以构建分布式入侵防御体系, 可以增强现有网

络的安全保护能力。

参考文献

[1] 李涛. 网络安全概论[M]. 北京：电子工业出版社，2004.

[2] 唐正军，李建华编著[M]. 入侵检测技术. 北京：清华大学出版社，2004.

[3] 胡文，黄皓. 自动入侵响应技术研究[J]. 计算机工程. 2005，31(18)：143-145.

[4] <http://www.ietf.org/Ids.by.wg/idwg.html>

[5] <http://www.cipherdyneorg>

[6] <http://ipset.netfilter.org/ipset.man.html>

[7] 张淑芬，陈学斌，刘春风. RSA 公钥密码体制的安全性分析及其算法实现[J]. 计算机应用与软件. 2005，22(7)：108-110.

[8] <http://www.schneier.com/blowfish.html>

基于抽象区间域的数组边界检查技术

曾勇军¹ 王清贤³ 奚 琪²

(解放军信息工程大学信息工程学院, 河南 郑州, 450002)

摘 要: 数组访问越界是程序中常见的问题, 给程序的执行带来潜在的安全威胁。针对该问题提出了一种基于抽象解释理论的程序静态分析技术, 用于自动发现 C 程序源代码中存在的数组访问越界错误。文章介绍了抽象解释的基本理论, 描述了抽象区间域的基本概念和数组越界检查的判定条件, 给出了抽象分析器的设计思路和模拟检测过程。

关键词: Galois 连接; 抽象解释; 区间域; 不动点

中图分类号: TP309 文献标识码: A 文章编号: 1006-7043 (2004) xx-xxxx-x

Array Bound Checking Technology Based on Abstract Interval Domain

ZENG Yongjun⁴, WANG Qingxian⁵, XI Qi⁶

(Institute of Information Engineering, PLA Information Engineering University, Zhengzhou 450002, Henan China)

Abstract: Array access violation is a common bug and becomes a potential security threat to the execution of program. A program static analysis technology based on abstract interpretation is presented in this paper, in order to find the array access violations of C source code. The theory of abstract interpretation is introduced, the basic concept of abstract interval domain and the decidable condition of array bound checking are described, the design method of abstract analyzer and the array bound checking process are given.

Keywords: Galois connection; abstract interpretation; interval domain; fixed point

1 引言

数组是程序设计过程中经常使用的数据类型, 用于将相同类型的数据集合在一起, 通过下标区分不同的数据元素。一些常见的编程语言 (如 C 语言) 缺乏数组边界检查机制, 很容易产生数组访问越界的错误, 由于这样的错误可能引发安全问题 (如缓冲区溢出等), 且定位和发现比较困难, 给程序的可靠性和安全性带来很大的危害, 已经引起人们的广泛关注。目前数组越界检查技术已经出现了一些研究成果, 如可在代码中插入运行时的边界检查条件检测溢出行为^[1,2]以阻止越界错误, 但由于这种运行时检测机制增加系统性能开销, 因此通常仅用于调试, 很少集成到产品系统。针对该问题研究人员后来又提出一些方法^[3,4]消除程序中无用的边界检测, 减少无用的计算, 以提高性能。此外, 基于抽象语法树的静态检测方法^[5]得到深入研究, 相关变种也进行了分析和探讨^[6,7]。

本文描述了基于抽象解释 (Abstract interpretation)^[8,9] 理论的程序静态分析技术, 利用抽象区间域检测数组边界越界的情况, 并提出了抽象分析器的设计思路, 该分析器利用 CIL (C Intermediate Language)^[10] 分析 C 源代码, 构造中间表示形式和控制流图, 然后由抽象解释器进行分析和计算, 最后根据计算获得的语义不动点与数组存储空间的分配情况进行检查, 发现是否产生越界的错误。利用本文描述的分析技术, 可自动发现 C 程序源代码中存在的数组访问越界错误, 增加程序运行时的安全性。

作者简介: 曾勇军, 讲师, 主要研究方向为信息安全;
奚琪, 博士生, 主要研究方向为信息安全。

抽象解释是一种程序行为的逼近理论，用于在不同抽象级别逼近程序的形式语义。由于静态确定程序的非平凡动态属性是不可判定的，基于抽象解释理论的程序自动分析方法可逼近程序的语义，该分析过程收敛并具有可靠性，通过分析获得程序在运行过程中的动态属性提供了一种可行的方法。本节介绍抽象解释的基本理论。

2.1 基本概念

基于抽象解释的程序静态分析技术涉及完备格、连续函数、不动点等基本概念。

定义 1 设 $(L, \leq)$ 为偏序集，且 $L \neq \emptyset$ 。若对 L 的任意子集 S ，均存在最小上界和最大下界，则称该偏序集为完备格，可记为 $(L, \leq, \cup, \cap, \top, \perp)$ ，其中最大元为 $\top = \cup L$ ，最小元为 $\perp = \cap L$ 。

当应用格理论进行程序分析时，通常会构建格的元素序列，表示为 $(l_n)_{n \in \mathbb{N}}$ 。

定义 2 设 L 为完备格， L 的子集 Y 构成一个链，若 $\forall l_1, l_2 \in Y: l_1 \leq l_2 \vee l_2 \leq l_1$ 。 L 的元素序列 $\{l_n | n \in \mathbb{N}\}$ 为递增链，若 $n \leq m \Rightarrow l_n \leq l_m$ 。类似地可定义递减链为 $n \leq m \Rightarrow l_n \geq l_m$ 。

定义 3 完备格 L 中的递增链 $(l_n)_{n \in \mathbb{N}}$ 趋于稳定，当且仅当 $\exists n \in \mathbb{N}, \forall m > n, l_m = l_n$ 。

定义 4 设 L_1, L_2 是两个完备格，函数 $f: L_1 \rightarrow L_2$ ：

(1) $\forall l_1, l_2 \in L_1$ ，若 $l_1 \leq_{L_1} l_2 \Rightarrow f(l_1) \leq_{L_2} f(l_2)$ ，则称 f 为单调函数。

(2) $\forall Y \subseteq L_1$ ，若 $\cup_{L_2} f(Y) = f(\cup_{L_1} Y)$ ，则称 f 为连续函数。

定义 5 设 $f: L \rightarrow L$ 是完备格 L 上的函数， x 是 L 中的元素，若 $f(x) = x$ ，则称 x 为 f 的不动点；若 x 为 f 的不动点，且 $\forall y \in L$ ，若 $f(y) = y$ ，有 $x \leq y$ ，则称 x 为 f 的最小不动点，记为 $\text{lfp}(f)$ 。

根据 Knaster-Tarski 定理，完备格 L 上的连续函数 f 存在最小不动点 $\text{lfp}(f) = \cup \{f^i(\perp) | i \in \mathbb{N}\}$ 。

2.2 Galois 连接

直接在完备格上进行程序的语义计算可能开销太大，甚至是不可计算的，此时可由抽象解释逼近程序的语义，这种逼近过程可由 Galois 连接来表述。

定义 6 设 $(D, \leq)$ 和 $(A, \sqsubseteq)$ 是两个完备格， $\alpha: D \rightarrow A$ 和 $\gamma: A \rightarrow D$ 是两个映射，序偶 (D, α, γ, A) 称为一个 Galois 连接，如果满足以下条件：

$$\forall x \in D, \forall y \in A, \alpha(x) \sqsubseteq y \Leftrightarrow x \leq \gamma(y)$$

可以验证，Galois 连接具有以下基本性质：

(1) $\forall x \in D, x \leq \gamma(\alpha(x))$ 。

(2) $\forall y \in A, \alpha(\gamma(y)) \sqsubseteq y$ 。

(3) α 和 γ 为单调函数。

程序的具体语义是在具体域 $(D, \leq)$ 上计算获得的，偏序关系 $\leq$ 描述了相对精确性，即若 $a \leq b$ ，则说明 a 比 b 更精确地描述程序的属性。程序具体语义的逼近由抽象域 $(A, \sqsubseteq)$ 上的计算来实现，同样偏序关系 $\sqsubseteq$ 描述了逼近的程度。Galois 连接 (D, α, γ, A) 建立两个域之间的映射关系，其中 α 称为抽象映射， $\alpha(x)$ 是 x 的抽象表示； γ 称为具体映射， $\gamma(y)$ 是 y 的具体表示。

定义 $F: D \rightarrow D$ 为具体域上的转移函数， $F^\# : A \rightarrow A$ 为抽象域上的转移函数。语义逼近过程中利用 $F^\#$ 模拟 F 的行为获得抽象语义。由于在语义抽象的过程中可能丢失某些属性，因此需要保证逼近的可靠性，Galois 连接的可靠性条件可表述为 $\alpha \circ F \sqsubseteq F^\# \circ \alpha$ ，即 $F^\#$ 可靠地逼近 F 。对抽象域 A 上的两个函数 $F_1^\#$ 和 $F_2^\#$ ，当 $F_1^\# \sqsubseteq F_2^\#$ 时，则称 $F_1^\#$ 比 $F_2^\#$ 更精确。当可靠性条件得到加强时，可获得抽象域 A 上最精确的逼近函数为 $F^A \triangleq \alpha \circ F \circ \gamma : A \rightarrow A$ 。

2.3 Widening/Narrowing 算子

程序静态分析构造的抽象解释格可能有无限高度，应用不动点算法计算程序的语义可能不收敛。此时可考虑使用近似方法，即 Widening/Narrowing 算子在有限步骤内完成不动点计算，可以实现语义的加速逼近。

定义 7 Widening 算子 $\nabla \in A \times A \rightarrow A$ 满足以下条件：

- (1) $\forall x, y \in A: x \sqsubseteq x \nabla y, y \sqsubseteq x \nabla y$ 。
- (2) 对任意的递增链 $x^0 \sqsubseteq x^1 \sqsubseteq x^2 \cdots$ ，递增链 $y^0 = x^0, \dots, y^{i+1} = y^i \nabla x^{i+1}, \dots (i \geq 0)$ 均收敛。

利用 Widening 算子，可构造完备格 A 上的单调函数 f 的序列 $(f_{\nabla}^n)_n$ 如下：

$$f_{\nabla}^n = \begin{cases} \perp & \text{if } n = 0 \\ f_{\nabla}^{n-1} & \text{if } n > 0 \wedge f(f_{\nabla}^{n-1}) \sqsubseteq f_{\nabla}^{n-1} \\ f_{\nabla}^{n-1} \nabla f(f_{\nabla}^{n-1}) & \text{otherwise} \end{cases}$$

序列 $(f_{\nabla}^n)_n$ 将趋于稳定并可安全地逼近 $lfp(f)$ ^[11]。

Widening 算子在加速迭代过程中丢失了精度。为获得更精确的结果，可使用 Narrowing 算子进行更精确的逼近。

定义 8 Narrowing 算子 $\Delta \in A \times A \rightarrow A$ 满足以下条件：

- (1) $\forall x, y \in A$, 若 $x \sqsubseteq y$, 则 $x \sqsubseteq x \Delta y \sqsubseteq y$ 。
- (2) 对任意的递减链 $x^0 \sqsupseteq x^1 \sqsupseteq x^2 \cdots$ ，递减链 $y^0 = x^0, \dots, y^{i+1} = y^i \Delta x^{i+1}, \dots (i \geq 0)$ 均收敛。

对满足 $f(f_{\nabla}^m) \sqsubseteq f_{\nabla}^m$ 的 f_{∇}^m ，利用 Narrowing 算子可构造单调函数 f 的序列 $(f_{\nabla}^n)_n$ 如下：

$$f_{\nabla}^n = \begin{cases} f_{\nabla}^m & \text{if } n = 0 \\ f_{\nabla}^{n-1} \Delta f(f_{\nabla}^{n-1}) & \text{if } n > 0 \end{cases}$$

同样序列 $(f_{\nabla}^n)_n$ 将趋于稳定并可安全地逼近 $lfp(f)$ 。

2.4 利用抽象解释实现程序静态分析

利用抽象解释实现程序静态分析需完成以下基本工作：

- (1) 根据所关注的程序属性，定义抽象语义域，用于逼近程序的具体语义域；
- (2) 建立抽象语义和程序之间的关系，该关系可通过 Galois 连接及定义的程序抽象语义来实现；
- (3) 建立程序的具体属性和抽象属性之间的可靠性对应关系，确保抽象属性的逼近效果；
- (4) 设计抽象语义迭代过程的收敛条件，以保证尽可能地精确；
- (5) 要求抽象解释能够终止。

抽象域由所关心的程序属性的所有取值空间构成。通过为程序定义相应的抽象语义函数，利用 Galois 连接，为目标程序的属性建立一组方程，通过在程序中进行迭代求解，找出方程的最小不动点（或最大不动点）。由于在实际中使用的大多数抽象域不满足递增链条件，在这样的抽象域中计算不动点可能不会终止，此时可使用 Widening/Narrowing 算子确保语义方程的迭代求解趋于稳定并获得不动点结果。

3 抽象区间域和数组边界检查

3.1 区间域及基本操作

抽象解释的核心是构造合理的抽象域，通过执行定义在抽象域上的运算和操作获得所关注的程序属性。数组边界检查可利用整数区间上的抽象计算来实现。

设由整数的幂集组成的具体域为 $(\wp(\mathbb{Z}), \leq)$ ，定义抽象区间域为 $(Intv, \sqsubseteq)$ 。抽象区间域的元素

$$(1) \alpha: \wp(\mathbb{Z}) \rightarrow \text{Intv}$$

$$\forall S \subseteq \wp(\mathbb{Z}), \alpha(S) = \begin{cases} \perp & \text{if } S = \emptyset \\ [l, h] & \text{if } \min(S) = l \wedge \max(S) = h \\ (-\infty, h] & \text{if } \min(S) \text{不存在} \wedge \max(S) = h \\ [l, +\infty) & \text{if } \min(S) = l \wedge \max(S) \text{不存在} \\ (-\infty, +\infty) & \text{if } \min(S) \text{不存在} \wedge \max(S) \text{不存在} \end{cases}$$

$$(2) \gamma: \text{Intv} \rightarrow \wp(\mathbb{Z})$$

$$\forall \text{int} \in \text{Intv}, \gamma[\text{int}] = \begin{cases} \emptyset & \text{if } \text{int} = \perp \\ \{z \mid z \in \mathbb{Z} \wedge l \leq z \leq h\} & \text{if } \text{int} = [l, h] \\ \{z \mid z \in \mathbb{Z} \wedge z \leq h\} & \text{if } \text{int} = (-\infty, h] \\ \{z \mid z \in \mathbb{Z} \wedge l \leq z\} & \text{if } \text{int} = [l, +\infty) \\ \mathbb{Z} & \text{if } \text{int} = (-\infty, +\infty) \end{cases}$$

此外，抽象区间域上定义的一些基本运算如下：

$$(1) [l_1, h_1] \sqsubseteq [l_2, h_2] \Leftrightarrow \{(l_2 \leq l_1) \wedge (h_1 \leq h_2)\}$$

$$(2) [l, h] = \perp \Leftrightarrow l > h$$

$$(3) [l, h] = \top \Leftrightarrow l = -\infty \wedge h = +\infty$$

$$(4) [l_1, h_1] \cup [l_2, h_2] = [\min(l_1, l_2), \max(h_1, h_2)]$$

$$(5) [l_1, h_1] \cap [l_2, h_2] = [\max(l_1, l_2), \min(h_1, h_2)]$$

(6) Widening 算子 ∇ ：

$$\perp \nabla X = X$$

$$X \nabla \perp = X$$

$$[l_0, h_0] \nabla [l_1, h_1] = [\text{if } (l_1 < l_0) \text{ then } -\infty \text{ else } l_0, \\ \text{if } (h_1 > h_0) \text{ then } +\infty \text{ else } h_0]$$

(7) Narrowing 算子 Δ ：

$$\perp \Delta X = X$$

$$X \Delta \perp = X$$

$$[l_0, h_0] \Delta [l_1, h_1] = [\text{if } (l_0 = -\infty) \text{ then } l_1 \text{ else } l_0, \\ \text{if } (h_0 = +\infty) \text{ then } h_1 \text{ else } h_0]$$

Widening 算子 ∇ 保持稳定边界，而将不稳定边界扩展为无穷大。Narrowing 算子 Δ 仅更改无穷边界，以期获得更精确的逼近结果。

从直观意义上说，若用二维坐标空间中的点代表程序的语义值，则抽象区间域通过坐标空间中的矩形区域逼近程序的语义值。在语义计算过程中，由抽象区间域引入的开销并不大，对每个变量只需要存储两个边界值，所有操作可在线性时间内完成（仅与变量数目有关）。当然，在抽象计算的过程中可能会丢失某些精度，如变量之间的关系等。

3.2 Worklist 算法

基于抽象区间域的静态分析器通过分析程序的控制流图 CFG（Control Flow Graph），建立语义约

束方程，通过解方程获得语义不动点。程序约束关系的迭代求解可使用 **Worklist** 算法。设程序的控制流图 $CFG = (Block, Edge)$ ，其中 $Block = \{b_1, b_2, \dots, b_n\}$ 为基本块的集合，构成控制流图的节点， $Edge \subseteq Block \times Block$ 为有向边的集合，表示程序的控制转移关系。

定义 $F: Block \times State \rightarrow State$ 为语义转移函数，其中 $State$ 为状态集合，定义了变量与取值之间的关系。基于抽象区间域的 **Worklist** 算法如下。

算法 **W orklist** 算法， s 、 s_{new} 代表状态集。

输入：控制流图 CFG 。

输出：状态集 S 。

$S = \perp$

根据 CFG 构造 **Worklist** 列表 W

```

while (W ≠ ∅) {
    从 W 中取出基本块 b
    Snew = F(b, S)
    if (b 是循环头节点)
        Snew = S ∇ Snew    //执行 Widening 操作
    if (Snew ≠ S) {
        for ((b, b') ∈ Edge) // b' 是 b 的后继
            W = add(W, b')    //将 b' 添加到 W 中
        S = Snew
    }
}
return S

```

算法初始化时将所有的节点加入 **Worklist** 列表中，在计算过程中关注引起状态变化的基本块，并将其后继添加到 **Worklist** 列表中。尽管构成抽象区间域的格可能有无限高度，但由于在迭代计算过程中引入 **Widening** 算子，可确保算法能够快速收敛。考虑到频繁运用 **Widening** 算子将增加开销，降低精度，因此算法 1 仅对循环头节点执行 **Widening** 算子。为获得更精确的结果，可在执行完 **Widening** 算子之后执行 **Narrowing** 算子。通过综合运用 **Widening** 算子和 **Narrowing** 算子来获得最小不动点的较好近似，并使所需要的计算步数得到限制。

3.3 数组边界检查

数组边界检查是指判定程序中所有的数组访问是否在其声明的范围之内。这可通过检查访问数组元素的索引表达式，与数组分配区域的长度进行比较，以此发现是否存在越界行为。利用基于区间域的抽象解释判定数组访问越界的情况，需要对源程序进行静态分析，计算获得程序的语义不动点，通过检查程序的状态信息达到判定的目的。

定义 **Var** 程序的变量集合，**Type** 为变量的类型集合，其中数组类型 $Array \in Type$ ，**Val** 表示变量的取值区间，**Addr** 记录变量起始地址，**Size** 为数组分配空间的尺寸。

定义函数 $\sigma: Var \rightarrow Type \times Addr \times Size$ ，指定变量的类型、起始地址及分配的缓冲区长度。

定义函数 $\delta: State \rightarrow Var \times Val$ ，用于建立程序状态与变量/取值之间的对应关系。

设 **Worklist** 算法在达到程序语义不动点时，程序状态为 s 。**BC(s)**记录了所有存在越界情况的数组变量，其成员的判定条件如下：

$$BC(s) = \{v \in Var \mid (v, i) \in \delta(s) \wedge \sigma(v) = (t, \_, r) \wedge t = Array \wedge (i \cap \overline{r} \neq \perp)\}$$

其中，变量的类型可在源代码分析时获得； i 为数组下标索引的取值区间； r 为分配的数组长度区间，

若某个数组分配的存储器长度为 l ，则 $r = [0, l - 1]$ ； $\bar{r}$ 为 r 的补区间。

4 抽象分析器设计与验证

4.1 系统设计

我们设计了一个基于区间域的抽象分析器，该分析器主要由两部分组成：即 CIL 程序分析器和抽象解释器，其结构如图 1 所示。



图 1 CIL 程序分析器和抽象解释器结构

CIL 程序分析器分析输入的 C 源程序，将比较复杂的 C 程序结构转换成简单形式，这有利于程序的进一步分析和优化。CIL 输出的结果包括与 C 源程序对应的高度结构化、便于处理的中间表示形式，并构造控制流图 CFG。

中间表示形式和控制流图由抽象解释器分析、计算和检测。抽象解释器根据抽象区间域和 Worklist 算法对中间表示形式进行分析和计算，并获得程序的语义不动点，在不动点状态下根据数组访问越界判定条件检查数组访问是否存在越界错误，并产生输出结果。

4.2 实验及结果

为验证数组边界检查的有效性，我们利用抽象分析器分析了一个简单程序。我们构造的抽象分析器运行于 Linux Redhat Enterprise Server 5 操作系统下，作为 VMware Workstation 5.5.0 的客户操作系统。待测程序如图 2 所示，main 函数中的数组 buf 存在一个字节的访问越界错误。

抽象分析器分析该程序，在达到语义不动点时的输出结果如图 3 所示。由分析结果可知，CIL 对原程序进行了语法变换。程序分析结果给出了两个变量的区间值，其中 buf 为考察的目标数组，其区间值为下标索引可能的取值范围。分析器用 1073741823 表示 $+\infty$ 。由于数组分配长度为 32，取其补区间为 $[32, 1073741823]$ ，而数组访问的索引表达式的计算结果区间为 $[0, 32]$ ，两者的交集不等于 $\perp$ ，存在越界的情况。

```
void main() {
    char buf[32];
    int i;

    i=0;
    while (i<=32) {
        buf[i] = i;
        i++;
    }
}
```

图 2 存在数组访问越界错误的待测程序

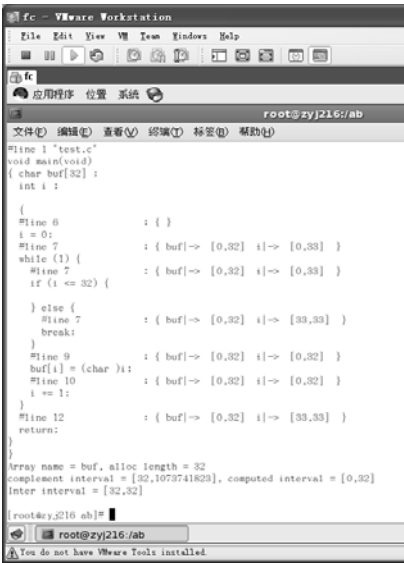


图 3 程序输出结果

5 结论

本文描述了一种基于抽象解释的程序静态分析技术，利用区间分析方法实现数组边界的检查。通过模拟仿真实验可知，该方法可用于静态检测 C 程序源代码中存在的数组访问越界错误。

当然本文描述的的方法也存在一定局限性，如仅能对 C 源代码进行分析，不能够分析处理二进制代码；CIL 仅能处理 C 程序语言的语法元素，不具备处理复杂程序结构（如 C++中的类）的能力；抽象解释器仅提供了区间域的分析方法，对于程序的其他属性没有涉及。上述问题对一个实际的静态分析系统是非常重要的，这也是我们下一步研究的重点。

参考文献

- [1] Crispan Cowan, Calton Pu, Dave Maier, Jonathan Walpole, Peat Bakke, Steve Beattie, Aaron Grier, Perry Wagle, Qian Zhang, and Heather Hinton. StackGuard: Automatic adaptive detection and prevention of buffer-overflow attacks [C] // Proc. 7th USENIX Security Conference, pp 63-78, San Antonio, Texas, Jan 1998.
- [2] Richard W. M. Jones, Paul H. J. Kelly. Backwards compatible bounds checking for arrays and pointers in C programs [J] // In Automated and Algorithmic Debugging, pp 13-26, 1997.
- [3] Rastislav Bodik, Rajiv Gupta, and Vivek Sarkar. Abcd: eliminating array bounds checks on demand [C] // Proceedings of the ACM SIGPLAN 2000 conference on Programming language design and implementation, pp 321-333, New York, NY, USA, 2000. ACM Press.
- [4] Hongwei Xi, Frank Pfenning. Eliminating array bound checking through dependent types [C] // Proceedings of the ACM SIGPLAN 1998 conference on Programming language design and implementation, pp 249-257, New York, NY, USA, 1998. ACM Press.
- [5] Xie Yichen, Chou A, Engler D. ARCHER: Using Symbolic, Pathsensitive Analysis to Detect Memory Access Errors. ESEC/FSE'03. 2003.
- [6] Dor N, Rodeh M, Sagiv S. CSSV: towards a realistic tool for statically detecting all buffer overflows in C. PLDI. 2003.
- [7] Ganapathy V, Jha S, Chandler D, et al. Buffer Overrun Detection using Linear Programming and Static Analysis. CCS'03. 2003.
- [8] Cousot P, Cousot R. Abstract interpretation: A unified Lattice model for static analysis of programs by construction or approximation of fixpoints [J] // Proc. of the 4th POPL. Los Angeles: ACM Press, 1977. 238-252.
- [9] P. Cousot, R. Cousot. Abstract interpretation frameworks [J] // Journal of Logic and Computer, 1992, 2(4):511-547.
- [10] George C. Necula, Scott McPeak, S. P. Rahul, Westley Weimer. CIL: Intermediate Language and Tools for Analysis and Transformation of C Programs [EB/OL]. <http://sourceforge.net/projects/cil>.
- [11] P. Cousot, R. Cousot. Comparing the Galois connection and widening/narrowing approaches to abstract interpretation [J] // Proc. of the PLILP'92. LNCS 631, Springer-Verlag, 1992. 269-295.

Kerberos 协议在单点登录中的改进及应用

郭甜滋¹ 毛楠¹ 司志刚² 陈丽²

(1. 华南理工大学电子与信息学院 广州 510640, 2.信息工程大学电子技术学院, 河南 郑州 450004)

摘 要: 分析了目前国内外流行的单点登录模型, 探讨了对称密钥下 Kerberos 协议的局限性, 然后结合公钥技术, 对 Kerberos 认证协议进行改进。设计并实现了基于代理的单点登录系统, 使用二次票据方法保证了单点登录系统的安全性。该方案在可实施性、运行可靠性及管理等方面都有较好的性能。

关键词: 单点登录; Kerberos 协议; 代理

中图分类号: TP309 **文献标识码:** A **文章编号:**

Improvement and Application of Kerberos Protocol in A Single Sign-On System

GUO Tianzi¹, MAO Nan¹, SI Zhigang², CHEN Li²

(1. School of Electronic and Information, South China University of Technology, Guangzhou 510640, Guangdong China

2. Institute of Electronic Technology, Information Engineering University, Zhengzhou 450004, Henan China)

Abstract: This thesis researches on existing popular Single Sign-On model and Kerberos authentication protocol, and discusses the limits of Kerberos protocol in symmetric key technology, improves Kerberos protocol with the public key technology. Then it designs and realizes the agent-based single sign-on system, and by the second ticket method, ensures the safety of the SSO system. Meanwhile, the program has better performance in implementation, operating and management.

Keywords: SSO; kerberos protocol; agent

1 引言

随着计算机和网络应用的快速发展, 计算机用户每天都要登录到许多不同的应用系统中^[1]。传统的方法需要管理员针对不同的应用系统和用户设置登录凭证、维护管理各个系统的用户信息库, 不但增加了工作量, 同时也存在安全隐患。解决上述问题的方案就是采用单点登录技术 (Single Sign-On, SSO)^[2]。当前流行的 SSO 模型主要有以下几种^[3]: 基于 Broker 的 SSO 模型、基于 Agent 的 SSO 模型和基于 Gateway 的 SSO 模型。这三种模型中, 基于 Agent 的 SSO 模型系统扩展性好, 方便系统增加新的服务。与基于认证网关的模型相比, 此种模型减小了认证服务器的负担, 经过身份认证以后, 客户机和服务器交互时不再经过认证服务器, 而是二者直接通信, 这样就没有了认证网关模型中由于认证网关而带来的瓶颈问题。配置相对灵活, 此模型中认证代理服务器和应用服务器可以分布在互联网中的不同地方, 而不必集中在一起。

2 SSO 认证协议的设计

Kerberos 协议在用户登录时, 利用 AS 验证用户身份并发放票据。用户使用该票据获得 TGS 的授权访问其他任何一个服务, 这样, 在访问各应用服务时, 用户只需进行一次身份认证, 实现了单点登

作者简介: 郭甜滋 (1988—), 女, 华南理工大学信息工程专业学生, 研究兴趣: 信息安全;
司志刚 (1965—), 男, 硕士, 副教授, 河南省计算机学会理事, 研究方向: 信息安全。

录，即 SSO。但 Kerberos 协议存在以下局限性：（1）认证票据的正确性是基于网络中所有的时钟保持同步，如果主机的时间发生错误，则原来的认证票据就是可能被替换的。（2）原有的认证服务可能被存储或替换，虽然时间戳是专门用于防止重放攻击的，但在票据的有效时间内仍然可能奏效，假设在一个 Kerberos 认证域内的全部时钟均保持同步，收到消息的时间在规定的范围内（假定规定为 3 分钟），就认为该消息是新的。（3）Kerberos 防止口令猜测攻击的能力很弱，攻击者通过长期侦听可以收集大量的票据，经过计算和密钥分析进行口令猜测。（4）随用户数增加，密钥管理较复杂。

从对 Kerberos 的局限性分析可以看出，其很多缺陷均是由于采用对称密钥技术造成的。如果能将公钥技术有机地融合到 Kerberos 中，便能提高 Kerberos 协议的安全性，图 1 是对本协议消息流程的描述。

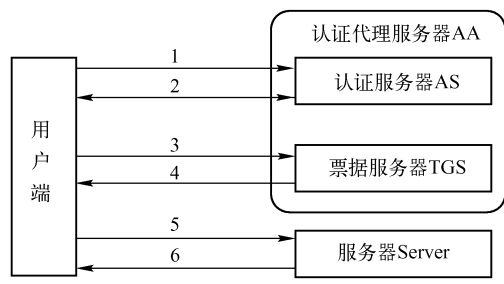


图 1 本系统协议消息流程图

改进后的协议如下：

- (1) 票据请求 (Ticket_req), C->AS:C,S,N,CertC
- (2) 票据发放 (Ticket_rep), AS->C:TGT,((Kct,N)SignAS)EncCp
TGT=((C,AS,Kct,N,lifetime)SignAS)EncTGTp
- (3) TGS 票据请求 (Ticket_req), C->TGS:TGT, (N,seq) Kct
- (4) TGS 票据发放 (Ticket_rep), TGS->C:TS, ((Kcs,N) Sign_{TGT}) EncCp
TS=((C,S,Kcs,N,lifetime) Sign_{TGT}) EncCs
- (5) 服务请求 (S_req), C->S:Ts,(N,ra,seq,opinion)Kcs
- (6) 服务器身份认证 (S_rep), S->C:(ra)Kcs (可选由 opinion 决定)

符号表示说明：

AA：认证代理服务器 (Authentication Agent)；C：用户端 (Client) 唯一标识；S：用服务器 (Server) 唯一标识；Xs：X 的私钥；Xp：X 的公钥；CertC：C 的证书序列号；(Info)SignXs：用 X 的私钥对信息 Info 签名；(Info)EncXp：用 X 的公钥对 Info 加密；(Info)K：用对称密钥 K 对 Info 加密；TGT：认证代理 AS 为用户 C 访问服务器 TGS 而颁发的票据；TS：TGS 为用户 C 颁发的访问服务器 S 的票据；Nc：用户端 C 保存的一个整数，记录用户访问次数；Ns：应用服务器 S 所保存的用户访问次数，用于防止重放；N：协议消息中使用的用于防止重放的整数，其值由 Nc 决定，在此称为“同步变量”。Lifetime：票据生存时间，(从服务器第一次接收服务请求后开始计时)；ra：随机会话密钥；seq：服务请求序列号；opinion：用户是否要求对服务器验证，1：要求，0：不要求。

在第 (2) 步和第 (4) 步中，认证代理 AA 在票据中加入了票据的生存期 lifetime，说明此票据的生存时间。在生存期内用户能够重复使用票据，而不必再向认证代理 AA 提出申请。对于票据生存时间，由认证代理 AA 在发放票据时给出。

在第 (5) 步中，用户 C 除了向应用服务器 S 提交最初协议中的内容外，还加入了 ra 和 seq，ra 为 C 生成的一个随机数。在最初的协议中，使用 Kcs 作为会话密钥，在这里使用 ra 作为二者会话密钥，因为用户 C 在票据的生存期内可以重复访问服务器 S，出于安全性考虑，每次会话都使用不同的会话密钥 ra。为了防止在票据的生存期内攻击者对服务器请求消息的重放，协议中引入了请求的序列

号 seq。seq 从 1 开始，每次加 1。

在第（6）步中，服务器将 ra 加密后返回给用户 C，由于 ra 的随机性，所以攻击者无法冒充，也无法进行重放，这样就验证了此消息的真实性和新鲜性。

3 安全性分析

对上述的协议进行安全性分析时，我们使用一种基于知识和信仰的逻辑方法——BAN 逻辑^[2~5]方法来分析协议的安全性。

初始假设如下：

- (1) $AA \models \xrightarrow{K_{CA}} CA$ （AA 相信 K_{CA} 是 CA 的公开密钥）
- (2) $AA \models \# \xrightarrow{K_C} C$ （AA 相信 K_C 是 CA 的公开密钥，并且 K_C 是新的）
- (3) $AA \models CA \models \xrightarrow{K_C} C$ （AA 相信 CA 对 C 有仲裁权，K_C 是 CA 的公开密钥）
- (4) $C \models \xrightarrow{K_{CA}} CA$ （C 相信 K_{CA} 是 CA 的公开密钥）
- (5) $C \models \# \xrightarrow{K_{AA}} AA$ （C 相信 K_{AA} 是 AA 的公开密钥，并且 K_{AA} 是新的）
- (6) $C \models CA \Rightarrow \xrightarrow{K_{AA}} AA$ （C 相信 CA 对 AA 有仲裁权，K_{AA} 是 AA 的公开密钥）
- (7) $C \models \xrightarrow{K_C} C$ （C 相信 K_C 是 C 的公开密钥）
- (8) $C \models \# N$ （C 相信 N 是新的）
- (9) $C \models AA \Rightarrow C \xleftarrow{K_{CS}} S$ （C 相信 AA 对 C, S 之间共享密钥 K_{CS} 有仲裁权）
- (10) $S \triangleleft \{ \xrightarrow{K_{AA}} AA \}_{K_{CA}^{-1}}$ （S 看到过 CA 签名的 AA 的公钥）
- (11) $S \models \xrightarrow{K_{CA}} CA$ （S 相信 K_{CA} 是 CA 的公开密钥）
- (12) $S \models \# \xrightarrow{K_{AA}} AA$ （S 相信 K_{AA} 是 AA 的公开密钥，并且 K_{AA} 是新的）
- (13) $S \models CA \Rightarrow \xrightarrow{K_{AA}} AA$ （S 相信 CA 对 AA 有仲裁权，K_{AA} 是 AA 的公开密钥）
- (14) $S \models \xrightarrow{K_S} S$ （S 相信 K_S 是 S 的公开密钥）
- (15) $S \models \# N$ （S 相信 N 是新的）
- (16) $S \models AA \Rightarrow C \xleftarrow{K_{CS}} S$ （S 相信 AA 对 C,S 之间共享密钥 K_{CS} 有仲裁权）
- (17) $S \models C \Rightarrow C \xleftarrow{ra} S$ （S 相信 C 对 C,S 之间共享 ra 有仲裁权）
- (18) $C \models \# C \xleftarrow{ra} S$ （C 相信 C,S 之间共享 ra 是新的）
- (19) $C \models C \xleftarrow{ra} S$ （C 相信 C 与 S 之间共享 ra）

由消息 1 可得：

$$AA \triangleleft \{ \xrightarrow{K_C} C \}_{K_{CA}^{-1}} \quad (1)$$

由式（1）和假设（1）得

$$AA \models CA \models \{ \xrightarrow{K_C} C \} \quad (2)$$

由式（2）和假设（2）得

$$AA \models CA \models (\xrightarrow{K_C} C) \quad (3)$$

由式（3）和假设（3）得

$$AA \models \xrightarrow{K_C} C \quad (4)$$

由以上可得

$$C \triangleleft \left(TGT, \left\{ \{ C \xleftarrow{K_{CS}} S, N \}_{K_{AA}^{-1}} \}_{K_C}, \{ \xrightarrow{K_{AA}} AA \}_{K_{CA}^{-1}} \right\} \right) \quad (5)$$

由式（5）可得

$$C \triangleleft \{ \xrightarrow{K_{AA}} AA \}_{K_{CA}^{-1}} \quad (6)$$

由式 (6) 和假设 (4) 得

$$C \models CA \mid \sim \xrightarrow{K_{AA}} AA \quad (7)$$

由式 (7) 和假设 (5) 得

$$C \models CA \models \xrightarrow{K_{AA}} AA \quad (8)$$

由式 (8) 和假设 (6) 得

$$C \models \xrightarrow{K_{AA}} AA \quad (9)$$

由式 (5) 得

$$C \triangleleft \{ \{ C \xleftarrow{K_{CS}} S, N \}_{K_{AA}^{-1}} \}_{K_C} \quad (10)$$

由式 (10) 和假设 (7) 得

$$C \triangleleft \{ C \xleftarrow{K_{CS}} S, N \}_{K_{AA}^{-1}} \quad (11)$$

由式 (9) 和式 (11) 得

$$C \models AA \mid \sim \{ C \xleftarrow{K_{CS}} S, N \} \quad (12)$$

由式假设 (8) 得

$$C \models \# \{ C \xleftarrow{K_{CS}} S, N \} \quad (13)$$

由式 (12) 和式 (13) 得

$$C \models AA \models \{ C \xleftarrow{K_{CS}} S, N \} \quad (14)$$

由式 (14) 直接可得

$$C \models AA \models C \xleftarrow{K_{CS}} S \quad (15)$$

由式 (15) 和假设 (9) 得

$$C \models C \xleftarrow{K_{CS}} S \quad (16)$$

得

$$S \triangleleft \{ TGT, \{ C, S, N, C \xleftarrow{rd} S, \text{seq} \} \}_{K_{CS}} \quad (17)$$

由式假设 (10) 及假设 (11) 得

$$S \models CA \mid \sim \xrightarrow{K_{AA}} AA \quad (18)$$

由式 (18) 和假设 (12) 得

$$S \models CA \models \xrightarrow{K_{AA}} AA \quad (19)$$

由式 (19) 和假设 (13) 得

$$S \models \xrightarrow{K_{AA}} AA \quad (20)$$

由式 (17) 得

$$S \triangleleft \{ \{ C, S, C \xleftarrow{K_{CS}} S, N, \text{lifetime} \} \}_{K_{AA}^{-1}} \}_{K_S} \quad (21)$$

由式假设 (14) 及式 (21) 得

$$S \triangleleft \{ C, S, C \xleftarrow{K_{CS}} S, N, \text{lifetime} \}_{K_{AA}^{-1}} \quad (22)$$

由式 (20) 及式 (22) 得

$$S \models AA \mid \sim \{ C, S, C \xleftarrow{K_{CS}} S, N, \text{lifetime} \} \quad (23)$$

由假设 (15) 得

$$S \models \# \{C, S, C \xleftarrow{K_{CS}} S, N, \text{lifetime}\} \quad (24)$$

由式 (23) 和式 (24) 得

$$S \models AA \models \{C, S, C \xleftarrow{K_{CS}} S, N, \text{lifetime}\} \quad (25)$$

由式 (25) 得

$$S \models AA \models C \xleftarrow{K_{CS}} S \quad (26)$$

由假设 (16) 和式 (26) 得

$$S \models C \xleftarrow{K_{CS}} S \quad (27)$$

由式 (17) 得

$$S \triangleleft \{C, S, N, C \xleftarrow{ra} S, \text{seq}\}_{K_{CS}} \quad (28)$$

由式 (27) 和式 (28) 得

$$S \models C \mid \sim \{C, S, N, C \xleftarrow{ra} S, \text{seq}\} \quad (29)$$

由假设 (15) 得

$$S \models \# \{C, S, N, C \xleftarrow{ra} S, \text{seq}\} \quad (30)$$

由式 (29) 和式 (30) 得

$$S \models C \models \{C, S, N, C \xleftarrow{ra} S, \text{seq}\} \quad (31)$$

由式 (31) 得

$$S \models C \models C \xleftarrow{ra} S \quad (d)$$

由假设 (17) 和 (d) 得

$$S \models C \xleftarrow{ra} S \quad (b)$$

得

$$C \triangleleft \{C \xleftarrow{ra} S\}_{K_{CS}} \quad (32)$$

得

$$C \models S \mid \sim C \xleftarrow{ra} S \quad (33)$$

由假设 (18) 及式 (33) 得

$$C \models S \models C \xleftarrow{ra} S \quad (c)$$

由假设 (19) 知

$$C \models C \xleftarrow{ra} S \quad (a)$$

经过以上推理, 得到了一级信仰 (a) 和 (b) 及二级信仰 (c) 和 (d), 所以本协议经 BAN 逻辑证明是正确的。

4 总结

本文深入分析了目前国内外流行的单点登录模型及 Kerberos 认证协议, 探讨了对称密钥下 Kerberos 协议的局限性, 然后结合公钥技术, 对 Kerberos 认证协议进行改进, 使用二次票据方法保证了单点登录系统的安全性。该方案在可实施性、运行可靠性及管理等方面都有较好的性能。

参考文献

[1] 孙宝林, 杨球, 吴长海. RSA 公开密钥密码算法及其在信息交换中的应用.武汉理工大学学报 (交通科学与工程版), 2000, 24(2): 169-172.

[2] 李腊元, 李春林. 计算机网络技术. 北京: 国防工业出版社, 2001.300.

[3] 肖攸安, 李腊元. 一类高效密钥协商方案的研究[J]. 武汉理工大学学报 (交通科学与工程版). 2003 年 27(6): 758-761.

[4] Burrows M, Abadi M, Needham R, A Logic of Authentication[R],Technical Report 39, Digital Syetems Reasearch Cen - ter,1989.

[5] 卿斯汉, 安全协议 20 年研究进展[J], 软件学报, vol. 14, no. 10, pp. 1740-1752, 2003.

分布式安全评估通信协议的研究与设计

李金武^{1,2} 郑秋生^{1,2}

(中原工学院 计算机学院 郑州, 450007)¹
(郑州市计算机网络安全评估技术重点实验室, 郑州, 450007)²

摘要: 根据企业评估需求, 为保证通信可控性和快速响应能力, 本文提出一种分布式安全评估通信模型, 并在此基础上研究和设计了一套性能高效、适合分布式主机安全评估的通信协议。该协议通过指令驱动, 能够实现执行流程的可控性操作和消息的及时性响应, 能够完成分布式安全评估的基本任务, 是一种应用性较强的通信协议。

关键字: 安全评估; 通信协议; 指令驱动; 可控性

中图分类号: TP391 文献标识码: A 文章编号: 1006-7043 (2004) xx-xxxx-x

Research and Design of Communication Protocol in Distributed Security Assessment

LI Jinwu^{1,2}, ZHENG Qiusheng^{1,2}

(School of Computer Science, Zhongyuan University of Technology, ZhengZhou 450007, Henan China)¹
(ZhengZhou Key Lab of Computer Network Security Assessment, ZhengZhou 450007, Henan China)²

Abstract: According to enterprise needs assessment, in order to ensure communication controllability, and rapid response capability, this paper proposes a communication model in distributed security assessment, then on the basis of the model, research and design a set of communication protocol which is efficient and suitable for the distributed host security assessment. By order-driven, the protocol can achieve controllable operation to process and timeliness response to message, can complete the basic tasks of distributed security assessment, which has strong applications.

Keywords: security assessment; communication protocol; order-driven; controllability

1 引言

网络安全风险评估一直是网络安全领域研究的重点和热点。与单机版评估系统相比, 分布式网络安全评估系统在扫描完备、扫描效率、扫描范围、扫描强度、漏洞分析、和可视化等方面具有巨大的优势。为了充分利用分布式系统的巨大处理能力, 全面多元化的收集企业全网中的原始扫描信息, 作者提出一种“主控-Sensor-数据中心”的三层框架评估体系。传统的通信协议存在传输效率低、安全性差、兼容性差等弱点, 像 AB、BB、MODICON、MOTOROLA、GE、ACTION-CONTROL 等国外大公司都致力于这种通信协议的研究, 分别开发出适合自己需要的通信协议, 其中 MODBUS、SEVBUS、DNP3.0 通信协议较为突出和流行^[1]。在研究已有分布式通信^[2~9]基础上, 结合项目背景和企业网络特殊性, 为了提高主控与 Sensor 的通信交互能力, 保证通信的可控性和快速响应能力, 突出主控的优势, 方便主控的管理, 本文提出一种可视可控的通信模型 MVCC (Model of Visual and Controllable Communication), 并在此基础上设计一套通信协议 CSCP (Center Sensor Communication Protocol)。

基金项目: 河南省科技攻关计划项目 (092102310038, 092102210029)
作者简介: 李金武 (1984—), 男, 硕士研究生;
郑秋生 (1965—), 男, 教授, 硕士。

2 问题分析与基本概念

2.1 问题分析

2.1.1 可控性分析

分布式安全评估系统包括控制中心和 Sensor 两部分。控制中心主要完成任务分发、收集数据和评估等功能，Sensor 主要完成扫描等功能。在任务执行周期内，要求控制中心控制 Sensor 运行流程，突出控制中心优势地位。

2.1.2 快速响应能力分析

分布式安全评估中，控制中心与 Sensor 的交互需要通过通信模块实现，要求双方对消息的处理要及时。另外，Sensor 执行任务时，Sensor 自调整也要求做到快速响应。

2.2 基本概念及描述

分布式安全评估中，通信环节至关重要。从分布式安全评估背景出发，为了解决上面提出的可控性和快速响应能力等问题，本文给出以下基本概念：

- (1) CC 控制中心。具有策略制定、任务调度、与 Sensor 通信、评估等功能。
- (2) S (Sensor)：负责任务的执行，完成脆弱性扫描，具有自治、社交、自适应、可推理、面向任务的特性。
- (3) DS (Designate Sensor)。指定 Sensor，充当 S 与 CC 通信间的桥梁，一个子网一个 DS。
- (4) 任务 (Task)。指在一定范围内以策略为基础，在一定的时间内并在不同级别 Sensor 上执行的基本工作单元。
- (5) 环境状态 (Environment S tatus)。指 Sensor 运行期间，计算机资源使用情况，处理任务的能力程度。
- (6) 安全状态 (Security Status)。指 Sensor 运行期间，自身的安全状况。
- (7) 运行状态 (Run Status)：指 Sensor 运行流程中表现出的一些基本的动态属性。

3 通信协议设计分析

3.1 通信模型

分布式安全评估系统通信过程包括 CC 与 DS 之间的通信和同子网内 S 的组播通信，在充分考虑现实需要和两部分通信的基础上，本文提出一种 MVCC 通信模型，如图 1 所示，图中虚方块表示一个子网，下面是我们所评估的网络资产（主机，服务器，网络设备）。

3.2 通信协议

3.2.1 协议设计背景

分布式安全评估要求任务能够快速分发与响应，并且要求任务能够可持续性执行。CC 能够通过 DS 主动获取 S 环境状态、安全状态和运行状态，并且能够可视化显示和定时更新这三种状态，显示评估报告表；S 能够完成自调整，并通过 DS 及时与 CC 进行快速响应。整个通信流程中按运行时间和激活方式，有以下几种通信：

- (1) 定时通信。一旦任务创建完成，任务具有时间属性，到达指定时间，利用设计好的任务分发指令，由 CC 通过任务调度模块对当前任务进行快速有效的分发，以求达到任务分发与执行时间最小。
- (2) 及时通信。所有以应答指令驱动的通信，它要求及时返回一些实时有用的信息，供 CC 或 S 提取。
- (3) 触发通信。满足一定条件或条件组合，需要做出回应，通告 S 运行状态的通信都属于触发通信。这类以通告指令驱动的通信，存在执行时间的先后次序，前驱指令通信构成后驱指令通信的条件。

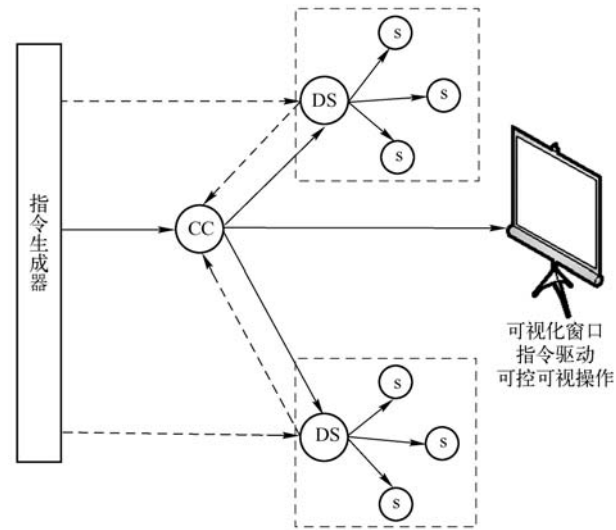


图 1 MVCC 通信模型

3.2.2 协议设计与字段描述

分布式安全评估系统中，CC 要完成 S 存活性检测、Task 分发、接受原始扫描信息和更新 S 状态信息等与通信相关的基本任务，而 S 要完成自身基本信息上传、通告自身状态信息等与通信相关的基本任务，CC 和 S 必须达成共识，构建双方都能解析的通信协议，所以安全有效的通信保障显得十分重要。因此在充分考虑 MVCC 通信模型和三种通信模式的基础上，本文参考现有协议^[10]，设计一套符合项目需要的通信协议 CSCP（Center Sensor Communication Protocol），保证通信的有效性和可控性，其通信协议设计格式如图 2 所示，协议字段描述如下：

Version (8)	Length (8)	Type (8)	Code (8)
CheckSun (16)		Reserve (16)	
DestAddr (32)			
SourceAddr (32)			
Data			

图 2 CSCP 通信协议设计格式

- (1) **Version:** 版本信息，使用 0x00 和 0x01 标示通信中是否使用 OpenSSL 技术；
- (2) **Length:** 总长度，标示传送数据的长度；
- (3) **Type:** 类型，标示指令类型，有请求、应答、通告和差错处理四种类型；

- (4) Code: 代码, 标示实现一定功能的具体指令;
- (5) CheckSum: 校验和, 验证传输数据是否被修改过, 保证数据完整性;
- (6) Reserve: 保留, 用于协议的升级;
- (7) DestAddr: 目的地址, 唯一标示通信的目的方, 填充主机 ID;
- (8) SourceAddr: 源地址, 唯一标示通信的发起方, 填充主机 ID;
- (9) Data: 通信时主要数据信息。

协议设计时充分考虑类型和代码两个字段, 这两个字段唯一标示一种具有某种功能的指令, 所有指令的运行集合最终实现一个 Task, 协议指令驱动能够很好地解决任务执行期间的可控性操作和消息的及时性响应。

3.2.3 指令单元

指令单元是指为完成某种特殊功能, 需要接收方做出反应的一组序列码。结合实际项目, 分布式安全评估系统中, CC 对 S 控制采用指令驱动, 增强 CC 主体地位。依据通信需求和协议格式, 按功能需求指令分为以下:

- (1) 请求指令。为完成本机某项作业, 而需要获取对方一些基本信息, 并需对方做出应答;
- (2) 应答指令。和请求指令相对应, 共同构成请求应答二元组;
- (3) 通告指令。通告对方自己的动态实时信息;
- (4) 差错处理指令: 能够对通信过程中的差错及时处理, 保证系统的稳定性和可用性。

在通信过程中, 对于携带数据源的指令, 定义数据格式相当重要, 携带数据主要包括单元数据 (SingleData, SD) 和多元数据 (MultiData, MD)。SD 比较简单, 本文不多做介绍, 主要介绍一下 MD。

MD 定义成结构体形式, 通信过程中便于双方解析, 后期操作便于扩展, 一次数据提取操作可以获取多方面有用信息, 降低了通信的冗余程度, 使得通信过程方便有效, 下面给出一种 MD 的表示方法。

S 存活性检测 MD 格式形式化描述如下:

SurvMD=<SensorIP,SensorMac,SensorID,SensorEnv,SensorSec,SensorRun,SensorTime>

本文充分考虑指令驱动在通信中的作用, 依据通信协议设计了 34 条指令, 涵盖了 CC 与 DS 之间的通信和同子网 S 之间的通信。在这些指令的共同驱动下, 分布式安全评估系统能够完成一个任务执行周期内所有功能。

3.3 通信交互

依据通信协议设计的指令组, 我们给出了 CC 与 DS 之间的通信交互图, 如图 3 所示。整个通信过程得以有效的保证, 能够实现分布式安全评估系统中 CC 对 S 可视化和可控性操作。

从图 3 可以看出整个通信过程以指令驱动, 指令之间存在先后顺序。指令 1、2 完成 S 存活性检测, 获取当前时间点 S 基本信息; 指令 3、4 通过最优调度算法完成任务分发; 指令 5、6、7、8、9、10 标示 S 运行状态, 供 CC 动态更新显示; 指令 11 通告 S 评估完成, 可以查询评估报告表; 指令 12、13 主要完成 S 环境、安全状态动态更新显示。为了保证通信交互的可视性、可控性、安全性和持续性, 遵循以下规则:

- Rule1: S 环境状态、安全状态每隔一段时间 (5min) 要更新一次;
- Rule2: 凡是上报给 CC 的状态信息, 都要记录这台 S 的时间点;
- Rule3: S 只负责下载扫描插件进行扫描, 没有操作数据库的权限;
- Rule4: CC 利用分析插件对扫描结果进行分析, 依据分析结果进行评估;
- Rule5: 每台 S 只拥有查看自己评估报告的权利, 而 CC 能够查看所有评估报告。

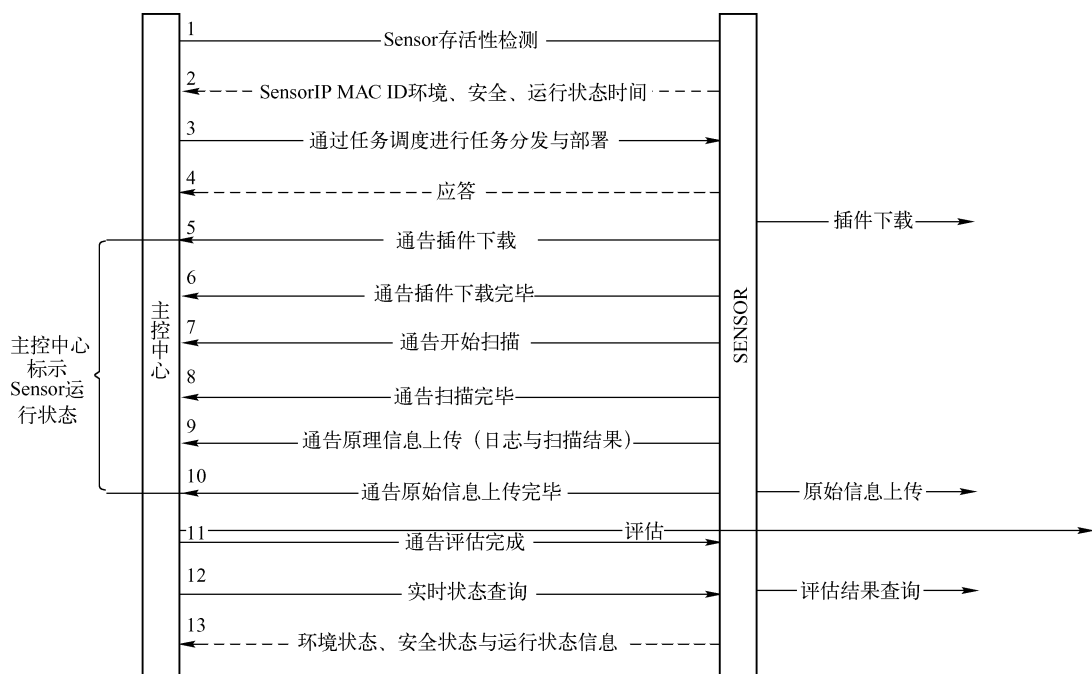


图 3 CC 与 S 通信交互图

4 实验与结果分析

以 MVCC 通信模型和三种通信模式为基础设计的通信协议能够更好地展示分布式安全评估系统的可视化和可控性操作。在“千兆骨干、百兆桌面”企业网络环境中，以高强度、中强度、低强度和自定义策略为基础的任务，针对 2 个网段内的 9 台 S，应用 11 个扫描插件进行通信协议的性能测试。

分布式安全评估中存活性探测是基础和前提，实验中 CC 的存活性探测请求与应答指令表示如图 4 所示，状态更新体现了可控性操作，如图 5 所示。

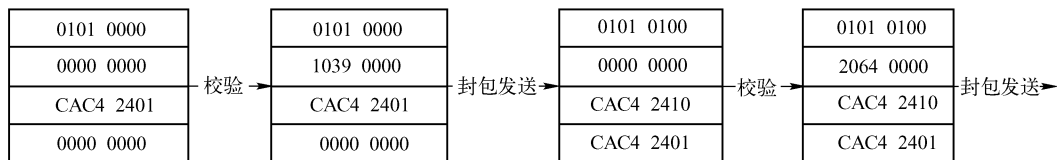


图 4 存活性探测请求与应答指令

状态	IP地址	MAC地址	sensor环境状态	sensor安全状态	sensor运行状态	备注
存活	202.196.37.146	00-E0-4C-B2...	良好	良好	评估完成	5RA3QVTE
存活	202.196.36.3	00-1D-09-9F...	良好	良好	正在运行	9RX8ZRQ5
存活	202.196.37.50	00-0B-2F-21...	良好	良好	评估完成	5RA3ZEKV
存活	202.196.37.199	00-30-18-A7...	良好	良好	扫描文件上...	9RA6OAGF

图 5 状态更新图

通过实验测试，在指令驱动下 CC 端状态信息随时间变迁进行动态刷新，能够完成我们的评估任务。可见，我们自主设计的通信协议是有效的，CC 与 S 的每次通信都是以指令驱动，指令驱动贯穿整个任务的执行周期，指令的构造、调度与发送在整个通信中占据主导地位。

5 结论

本文针对分布式安全评估特点，结合企业网实际需要，提出适合我们项目自己的通信协议。在协议的基础上设计了具有某些功能的指令单元，通过指令驱动保证了通信的可控性，同时也保证了响应的及时性。通过实验验证，在该通信协议的指导下评估任务能够顺利完成，是一种应用性较强的通信协议。下一步主要完善我们的指令，特别是差错处理指令的完善，另外还要考虑 OpenSSL 技术的应用。

参考文献

- [1] 赵宏. 分布式系统通信协议的研究与设计[D]. 北京:北京化工大学, 2003.
- [2] 魏士博, 王辉, 张春艳. 一种新的分布式通信模式的研究与实现[J]. 军事通信技术, 2006, 3(27): 27-30.
- [3] 吕锋, 唐红超, 陈德军. 基于 Agent 的分布式系统数据集成方式的研究[J]. 微计算机信息, 2006, 36(22): 239-241.
- [4] 丁忠, 刘志勤. 多服务器分布式即时通信系统模型的设计[J]. 微计算机信息, 2006, 27(22): 181-183.
- [5] 曹力, 刘晓平. 局域网中分布式仿真系统的通信模型[J]. 系统仿真学报, 2007, 13(19): 2951-2954.
- [6] 虞敏, 张为民. 分布式设备远程监控系统研究[J]. 计算机工程与应用, 2009, 45(5): 196-199.
- [7] 冯萍慧, 连一峰, 戴英侠等. 基于可靠性理论的分布式系统脆弱性模型[J]. 软件学报, 2006, 7(17): 1633-1640.
- [8] 王新志, 刘克胜. 分布式网络安全扫描任务调度模型及算法[J]. 计算机工程, 2003, 29(19): 101-103.
- [9] M.Borkar,V.Cevher,J.H.McClellan.Low Computational and Low Latency Algorithms for Distributed Sensor Network Initialization[J]. Signal,Image and Video Processing,2006,2(1):133-148.
- [10] cnpaf.RFC791-Internet Protocol[EB/OL].<http://www.cnpaf.net/Class/Rfcen?200502/1837.html>,2005-02-11.

一种分布式安全评估主控中心的研究与设计

夏 冰^{1,2}, 夏敏捷¹, 徐 飞¹, 郑秋生^{1,2}

(1.中原工学院 计算机学院, 郑州, 450007; 2.郑州市计算机网络安全评估重点实验室, 河南 郑州, 450007)

摘 要: 为满足企业网络安全管理需求, 提出一种三层框架体系的评估结构。主控中心利用 DAG、云、时间序列预测理论分别实现任务通信与调度、评估模型和网络安全态势的关键技术探索与研究, 所建立的三层框架体系结构, 在功能和性能方面, 对开发满足不同需求的网络安全管理工具具有重要指导价值。

关键词: 分布式安全评估; 主控中心; 任务通信与调度; 逆向云; 时间序列预测

中图分类号: TP391 **文献标识码:** A **文章编号:**

Research and Design on Control Center of Distributed Security Assessment System

XIA Bing^{1,2}, XIA¹ Minjie, XU¹ Fei, ZHENG Qiusheng^{1,2}

(1. School of Computer Science, Zhongyuan University of Technology,Zhengzhou 450007, Henan China;
2. Zhengzhou Key Lab of Computer Network Security Assessment Department, ,Zhengzhou 450007, Henan China)

Abstract: To meet enterprise network security management need, the paper proposed three-layer assessment architecture, and use DGA, Backward Cloud and Time Series Prediction to solve key issue of control center, such as communication and dispatch, assessment model, network security situation. The architecture can meet different needs of network security management tools and have important guiding value.

Keywords: distributed security assessment; control center; communication and dispatch; backward cloud; time series prediction

以面向网络管理而开展的网络安全风险评估近几年一直是网络安全研究的热点和重点之一^[1]。分布式网络安全评估在扫描完备、扫描效率、扫描范围、评估结果、网络态势分析、可视化多样化方面都是单机版评估系统无法比拟的。同时分布式系统固有的灵活性、可扩展性、实用性、容错性、可伸缩性, 及其面向任务的特性都使其作为一种高效的、面向网络管理与安全的工具深受网络管理者的欢迎^[2]。

为了充分利用分布式系统的巨大处理能力, 全面多元化的收集企业全网中的原始扫描信息, 郑州市计算机网络安全评估重点实验室提出一种“主控-Sensor-数据中心”三层框架的分布式评估体系, 主控中心以策略为基础, 以任务为驱动, 利用 Sensor 对企业全网进行多元扫描, 全面、快速、高效发现网络漏洞或威胁, 从而给出安全解决建议和系统加固方案的系统。

1 一种主控中心体系

主控系统负责策略制定、任务生成、任务调度、通信、评估、报告等, 是分布式安全评估系统的引擎与核心。

1.1 基本要求

从项目背景和企业网络特点出发, 分布式安全评估的主控中心具备以下基本要求:

(1) 策略的动态性: 策略制定与插件动态关联。

基金项目: 河南省科技攻关计划项目 (092102310038, 092102210029)
作者简介: 夏冰 (1981—), 男, 讲师, 硕士, 研究方向: 网络安全、网络对抗、虚拟机技术。

- (2) 可重构、面向任务的特性：以策略为核心的任务生成时，通过时间和任务名区分一个任务。
- (3) 以代价作为调度依据：在任务 T 与 Sensor 之间存在多重关系。如何快速在 T 与 S 之间进行任务分发，需要代价作为衡量依据。
- (4) 通信效率高：多 Sensor 与主控中心通信，势必造成主控中心性能低下且超负载运行。
- (5) 动态可视化特性：主控中心需要实时了解 Sensor 各种状态，并能够对未来进行环境与安全态势感知。
- (6) 多样化网络评估特性：存在单 Sensor、子网、企业全网多种评估对象，各评估对象在测评指标上存在差异势必造成评估多样化。

1.2 体系结构

依据上述设计目标，郑州市计算机网络安全评估重点实验室提出一种主控中心体系结构，如图 1 所示。

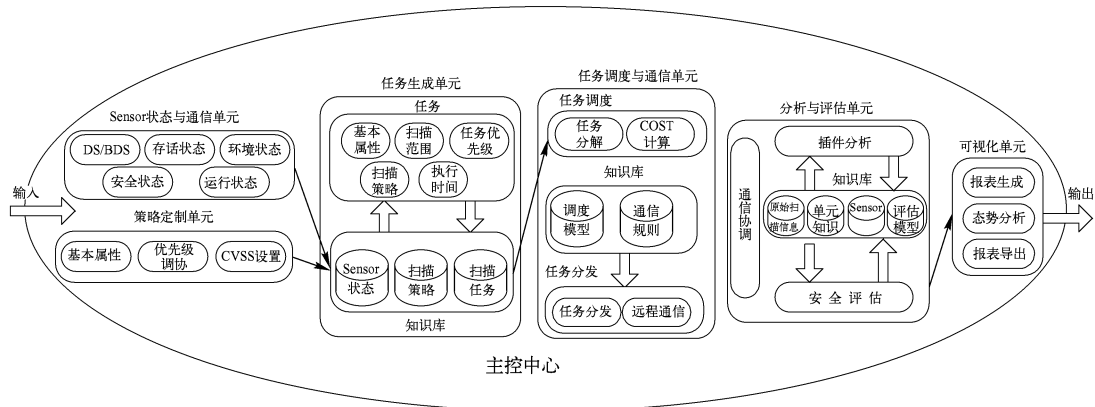


图 1 主控中心体系结构

主控中心主要有 Sensor 状态与通信单元、策略定制单元、任务生成单元、任务调度与通信单元、分析与评估单元和可视化单元组成，另外辅以升级、自身安全等功能单元。Sensor 状态与通信单元主要协调主控中心与 Sensor 的通信，表现在存活性检测、实时状态调整、与动态指定 DS 通信。该单元同时负责把收集到的信息保存在 Sensor 状态库中，以方便任务调度进行代价 Cost 计算。策略定制单元是主控中心扫描配置主要部分，基本属性、优先级和 CVSS 组成。其中 CVSS 表现在评估分值给定，以方便人性化配置。定制后的策略保存在策略库中，以方便任务生成单元提取和分析。任务生成单元和任务调度与通信单元基本内容，通过时间和任务名唯一区分一个任务。分析与评估单元从 11 方面 36 项全面对 Sensor 进行扫描。从项目组背景出发，在已有基本插件基础上采用插件分离方案，把插件分成插件扫描、插件分析和评估三部分组成。依据评估对象不同，动态调用评估模型中相关评估策略。可视化单元主要为主控中心、Sensor 提供丰富多样化的报表功能，如饼图、柱形图、曲线图。并能够提供网络安全态势分析、Sensor 环境安全态势、网络拓扑安全趋势等。

2 关键技术

由体系结构图可知，主控中心如何快速进行任务通信与调度、安全评估模型、可视化技术等是主控中心设计到的关键技术。

2.1 通信与调度^[3]

分布式通信与调度需要解决以下几个问题：（1）Sensor 过多、任务分发造成控制中心 CC（Control Center）负载过多，性能下降。（2）Sensor 存活性检测，环境安全状态查询，造成通信流量巨大。（3）可视化 Sensor 实时状态调整，造成控制中心 CC 更新慢。同时任务的调度受任务约束、链

路约束、执行环境约束和地位约束的限制，因此任务通信与调度需要冲破上述 4 种约束。

因此，设计的任务通信与调度模型在增加 Sensor 的情况下也能够使得任务达到分发和执行时间最小目标，同时提高 CC 的性能。从项目背景出发，本文提出一种采用有向加权无回路图（Directed Acyclic Graph, DAG） $G(V,E,C)$ 来搭建通信与调度模型 M3D。如图 2 所示，图中虚线部分是表示一个特定网络，简称一个“行政村”。顶点集 $V=\{CC,DS_1,DS_2,\cdots,DS_i\}$ ， $0<i<N$ （ N 表示网络个数）； E 表示 CC 与 DS（designate sensor）关系、DS 与 OS(Other Sensor) 的关系； C 表示 CC 到 DS、DS 到 OS 的代价。CC 与 DS 通信序列图如图 3 所示。

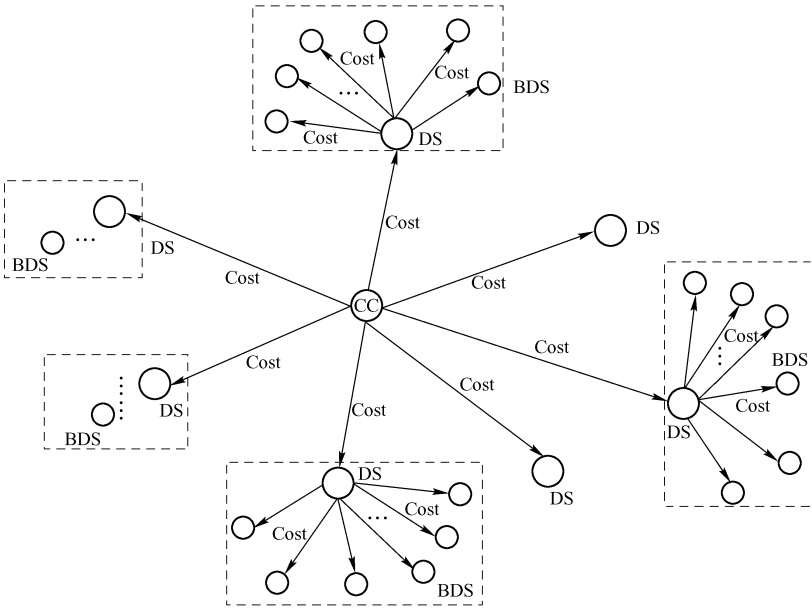


图 2 分布式调度与“行政村”

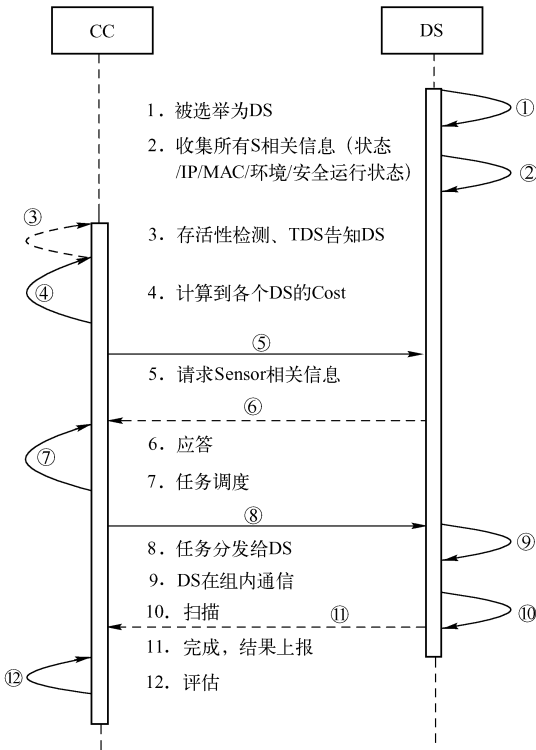


图 3 CC 与 DS 通信序列图

对于主控中心，如何给出 Cost 评定标准、如何计算 Cost 代价、主控中心 CC 如何与 DS 通信是任务通信与调度的关键。

2.2 评估模型

评估模型设计到以下几个问题：

- (1) 单 Sensor、网段、企业全网评估对象如何界定；
- (2) 单 Sensor、网段、企业全网评估指标如何给出；
- (3) 单 Sensor、网段、企业全网评估之间存在层次关系，如图 4 所示。

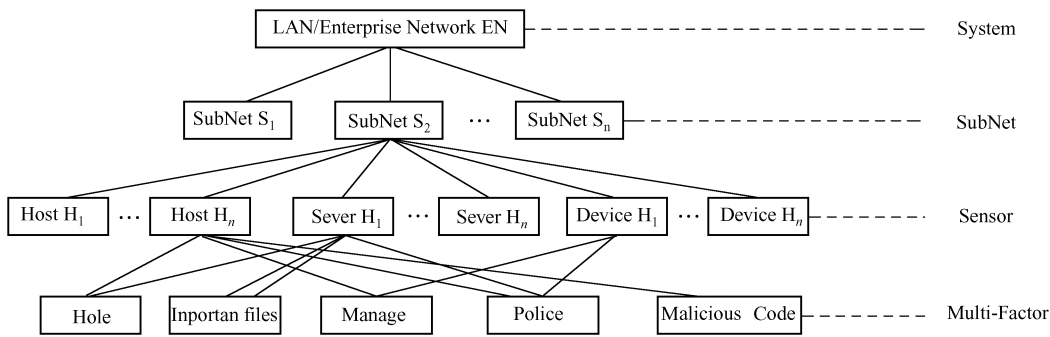


图 4 分布式、层次化安全评估模型

分布式安全评估结果既要量化也需要定性。

定义 1 定量评估。通过如下方式给出： $S = \sum_{i=1}^n P_i \times S_i$ ，其中 i 为评估指标的格式， P_i 为每一个指标的权重， S_i 为每一个指标的评估分值。

定性结果可通过云理论^[4]来解决，首先给出云的基本概念和表示。

定义 2 云。设 U 是一个用精确数值表示的定量论域 $X \subseteq U$ ， T 是 U 空间上的定性概念，若对于元素 $x (x \in X)$ 都存在一个有稳定倾向的随机数 $C_T(x) \in [0,1]$ ，称为 X 对 T 的隶属度，即： $C_T(x):U \rightarrow [0,1], \forall x \in X (X \subseteq U), x \rightarrow C_T(x)$ ，则概念 T 从论域 U 到区间 $[0,1]$ 的映射在数据区间的分布，称为云 (Cloud)。

云的数字特征反映了定性概念的定量特征，其独特之处在于仅仅用三个数值就可以勾画出由成千上万的云滴构成的云，记作 $C (Ex, En, He)$ ，其中 Ex 、 En 、 He 分别称为云的期望 (Expect value)、熵 (Entropy) 和超熵 (Hyper entropy)。

定义 3 逆向云。实现定量数值到其定性语言值的不确定性转换，它将一定数量的精确数据转换为定性语言值 $Cloud(Ex, En, He)$ 表示的概念，即从给定的云滴样本中求出正向云发生器的 3 个特征数字 Ex, En, He ，从而实现对样本数据的定性评价。

因此，本文定性评估的解决思路是以 Sensor 评估结果作为输入，利用逆向云如下：

- (1) 以 Sensor 的评估结果作为云滴 $x_i (0 \leq i \leq N, N$ 为 Sensor 总个数)；
- (2) 根据 x_i 计算这组数据的样本均值， $E_x = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ ，一阶样本中心矩 $B = \frac{1}{N} \sum_{i=1}^N |x_i - \bar{x}|$ ，样

本方差 $S^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$ ；

- (3) $E_x = \bar{x}$ ；
- (4) $E_n = (\pi/2)^2 \times B$ ；

(5) $H_e = (S^2 - E_n^2)^{\frac{1}{2}}$ 。

通过偏离靶心的程度来定性描述网络安全状况，如图 5 所示。

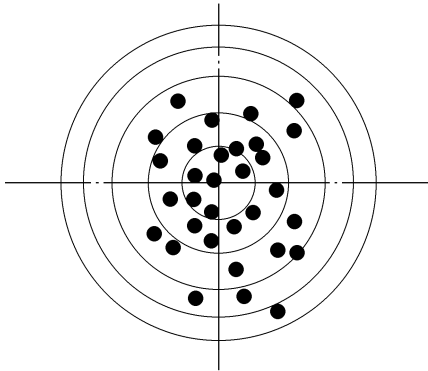


图 5 评估结果示意图

2.3 可视化技术与网络安全态势

分布式安全评估最终要以多样化评估报表结果提交给管理人员和 Sensor，因此如何提供准确、美观、丰富的报表是分布式安全评估中一个非常核心的内容^[6]。通常可采用两种方式给出可视化结果。

- (1) 对于已知数据，通过图、表、饼图、柱状图给出，实现算法较简单，在这里不讨论。
- (2) 如何从已知数据中找出数据之间的规律，这涉及数据挖掘和人工智能问题，值得研究。
- (3) 如何借助已有 n 个时刻评估数据预测 $n+1$ 时刻评估数据，这涉及网络安全评估态势技术。

在网络安全评估态势预测中，按照时间序列的统计特性来分，涉及平稳时间序列和非平稳时间序列两种情况。其基本思想是：寻求系统的当前值与其过去的运行记录数据的关系（是一种纵向关系），建立能够比较精确地反映时间序列中动态依次关系的数学模型，并借此对系统的未来行为做出预测^[7,8]。因此，主控中心将 D-S 证据理论、Bayes 网络和多元线性回归理论引入到安全预测和态势分析中。

3 结论

分布式安全评估不仅是为了评估安全，更是一个企业网络管理平台。它不仅能够对企业网中存在的安全隐患进行及时发现并给出改进建议，同时将基于 SNMP 的网络管理、基于 Sensor 的信息融合和环境安全感知等内容融入主控中心中，并以可视化方式提供给网络管理人员。其中拓扑发现与定制、Sensor 实时信息收集与上报、环境安全分析与态势感知、各网段 Sensor 的融合和态势等实现，可以较方便地对网络进行安全管理。因此这种三层体系框架对网络安全管理工具具有一定的指导作用。目前主控中心的系统开发初步实现，并在实验室环境下进行的验证，可以发现网络安全隐患并提供加固建议。下一步计划在网络安全态势、可视化、环境安全态势预测开展进一步的研究。

参考文献

[1] 夏冰, 裴斐, 郑秋生. 带策略与管理的安全评估系统研究与实现[J]. 中原工学院学报, 2009, 20(6): 29-34.
[2] 王新志, 刘克胜. 分布式网络安全扫描任务调度模型及算法[J]. 计算机工程, 2003, 29(19): 101-103.
[3] 夏冰, 李金武, 裴斐. 一种分布式安全评估通信与调度模型 [J]. 计算机工程与应用. 2010.

- [4] 李德毅, 刘常昱, 杜鹃. 不确定性人工智能[J]. 软件学报, 2004, 15(11): 1583-1594.
- [5] 李德毅, 刘常昱. 论正态云模型的普适性[J]; 中国工程科学; 2004, 08(6): 28-33.
- [6] 郑秋生, 白永红, 夏冰. 计算机网络安全评估技术的研究[A]. 河南省计算机学会论文集——2009 计算机研究新进展. 北京: 电子工业出版社, 2009.
- [7] 王慧强, 赖积保, 朱亮. 网络态势感知系统研究综述[J]. 计算机科学, 2006, 33 (10): 5-10.
- [8] 管鹏. 非平稳时间序列建模与预测[D]. 兰州: 兰州大学学位论文, 2007.

重要信息系统安全测评工具的研究与设计

武俊芳^{1,2}, 郑秋生^{1,2}

(中原工学院计算机学院, 河南 郑州, 450007)¹

(郑州市计算机网络安全评估技术重点实验室, 河南 郑州, 450007)²

摘要: 安全测评是保障整个信息系统安全的重要措施。本文首先对国内外标准进行介绍; 然后根据标准对实际测评过程中存在的问题进行了分析, 在此基础上, 设计了一个信息系统安全测评工具; 最后对该工具的关键技术做出了进一步的分析说明。

关键词: 信息安全; 安全测评; 等级保护; 测评标准

Research and Design of Important Information System Security Assessment Tool

WU Junfang^{1,2}, ZHENG Qiusheng^{1,2}

(School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, Henan China)¹

(Zhengzhou Key Lab of Computer Network Security Assessment, Zhengzhou 450007, Henan China)²

Abstract: Security assessment plays an important role for the security of information systems. This paper first introduced the existing assessment criteria both at home and abroad. Then it analyzed the problems in actual criteria-based assessment processes. On this basis, an overall framework of an information security assessment tool was designed. In the end, the paper analyzed the key techniques of the assessment tool.

Keywords: information security; security assessment; classification protection; assessment criteria

1 引言

随着计算机和网络技术的飞速发展, 人们对信息的依赖程度逐步加强。与此同时, 黑客攻击、病毒和木马的传播等严重威胁到信息系统的正常运行, 安全问题, 特别是重要信息系统的安全已经引起人们的广泛关注。我国把实施信息安全等级保护作为加强信息安全保障的一项基本制度, 而安全测评则是等级保护工作中一项承上启下的重要工作: 一方面测评机构依照国家有关法规和技术规范(目前主要是参考《测评准则》), 为重要信息系统运营、使用单位提供安全、客观、公正的信息安全标准符合程度检测服务, 被测单位通过测评结果发现安全现状与国家相关标准要求的差距, 从而进一步进行安全改建和实施; 另一方面信息安全监管部门通过测评结果了解被测系统的等级保护建设情况, 从而进一步促进其信息安全建设^[1]。然而, 目前很多单位对测评标准研究还不够系统, 不够深入, 并且缺乏针对标准的安全测评工具, 本文针对实际测评过程中存在的问题进行深入探讨, 设计出基于等级保护标准的安全测评工具。

2 测评标准研究

信息安全测评标准是信息安全测评的行动指南, 是安全测评的依据和尺度。国内外针对计算机安全的等级防护和测评制定了多个标准。

基金项目: 河南省科技攻关计划项目(092102310038 092102210029)

作者简介: 武俊芳(1979—), 女, 硕士研究生;
郑秋生(1965—), 男, 教授, 硕士。

2.1 国际标准

2.1.1 侧重于技术指标方面的标准

1985 年美国国防部公布可信计算机系统安全评估标准（TCSEC^[2]，橘皮书），是计算机系统信息安全评估的第一个正式标准。20 世纪 90 年代英、法、德、荷欧洲四国联合发布信息技术安全评估标准（ITSEC^[3,4]，欧洲白皮书），把可信计算机的概念提高到可信信息技术的高度来认识，对国际信息安全的研究实施产生了深刻的影响。1996 年由六个国家（美、英、法、加、德、荷）联合提出信息技术安全评价通用标准（CC），并逐步形成国际标准 ISO 15408。CC 标准是第一个信息技术安全评价国际标准，它的发布对信息安全具有重要的意义，是信息技术安全评价标准及信息安全技术发展的重要里程碑。

2.1.2 侧重于安全管理方面的标准

1995 年 2 月，英国标准协会（BSI）制定了世界上第一个信息安全管理标准 BS7799—1:1995，经过几年的修订完善于 1999 年形成 BS7799—1:1999 和 BS7799—2:1999 一对配套标准。BS7799 是目前国际上最知名的安全管理标准^[5]。

2.2 国内标准

我国一直高度关注信息安全标准化工作。1999 年，我国借鉴国外先进经验并且结合实际国情制定了 GB 17859—1999《计算机信息系统安全保护等级划分准则》^[6]，将计算机信息系统安全分为五级：用户自主保护级、系统审核保护级、安全标记保护级、结构化保护级和访问验证保护级。随后又在此基础上进一步细化和扩展，制定了如下相关的配套标准：

- 信息安全技术 信息系统安全等级保护定级指南 GB/T 22240—2008
- 信息安全技术 信息系统安全等级保护基本要求 GB/T 22239—2008
- 信息安全技术 信息系统安全等级保护实施指南 GB/T CCCC—CCCC
- 信息安全技术 信息系统安全等级保护测评过程指南 GB/T DDDD—DDDD
- 信息安全技术 信息系统安全等级保护测评要求 GB/T XXXX—XXXX
- 信息安全技术 信息系统通用安全技术要求（GB/T 20271—2006）
- 信息安全技术 网络基础安全技术要求（GB/T 20270—2006）
- 信息安全技术 操作系统安全技术要求（GB/T 20272—2006）
- 信息安全技术 数据库管理系统安全技术要求（GB/T 20273—2006）
- 信息安全技术 服务器技术要求 GB/T 21028—2007
- 信息安全技术 终端计算机系统安全等级技术要求（GA/T 671—2006）
- 信息安全技术 信息系统物理安全技术要求 GB/T 21052—2007
- 信息系统安全管理要求 GB/T 20269—2006
- 信息系统安全工程管理要求 GB/T 20282—2006

GB 17859—1999 是等级保护的基础性标准，技术类、管理类和 product 类一系列标准是在《划分准则》基础上制定的。《基本要求》是根据现有技术发展水平，从技术和管理两方面提出并确定了不同安全保护等级信息系统的最低保护要求。《测评要求》为等级测评机构开展等级测评活动提供了测评方法和综合评价方法。《测评过程指南》对等级测评活动提出规范性要求，以保证测评结论的准确性和可靠性。《定级指南》为信息系统定级工作提供了技术支持，《实施指南》用于指导信息系统运营使用单位了解和掌握信息安全等级保护工作的方法、主要工作内容及不同的角色在不同阶段的作用^[7]。标准间的关系如图 1 所示。

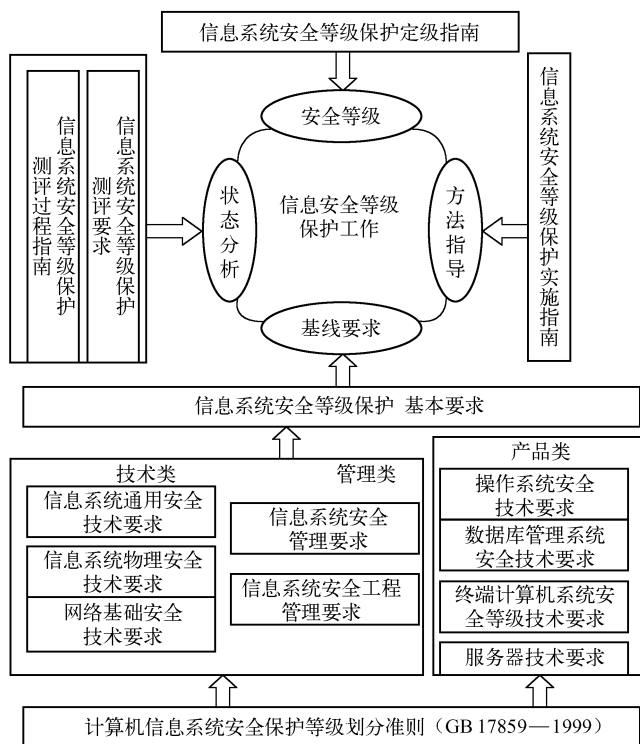


图1 等级保护相关标准间的关系

3 常用安全测评技术和工具

目前，国内外推出了很多测评技术和工具，包括漏洞扫描、渗透性测试、性能测试、入侵检测和协议分析^[8]等。比较主流的测评工具有 Nessus、Namp、X-scan、COBRA 、BDSS、CRAMM 、@RISK、启明星辰的天镜网络扫描系统、安络网络安全评估系统、金诺网安扫描器、中科网威“火眼”网络安全评估分析系统、绿盟科技极光远程安全评估系统等。

通过上述工具，用户可以很方便地大致了解到目前系统的安全状况，但有些工具仅仅对信息系统某个安全问题进行检测，不能从全局的角度对系统的整体安全状况进行测评，而且没有很好地遵循等级标准进行测评。目前，基于等级标准的安全测评工具还十分缺乏，本文依据等级标准的要求，设计出一种基于等级标准的信息安全测评工具，作为第三方认证机构进行测评的重要工具，实现对信息系统安全的有效测评。

4 信息安全测评工具的研究与设计

4.1 安全测评的内容

安全测评的内容包括安全技术和安全管理两个方面，如图2所示。

4.2 安全测评的流程

信息系统安全等级保护测评基本测评过程分为四个：测评准备过程、方案编制过程、测评实施过程、分析及报告编制过程。而测评双方之间的沟通与洽谈应贯穿整个等级测评过程^[9]。具体测评流程如图3所示。

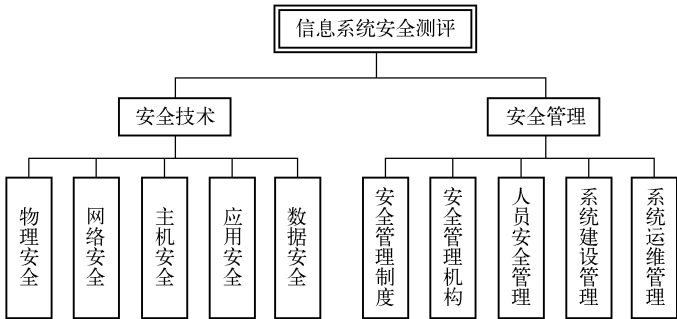


图2 安全测评的内容

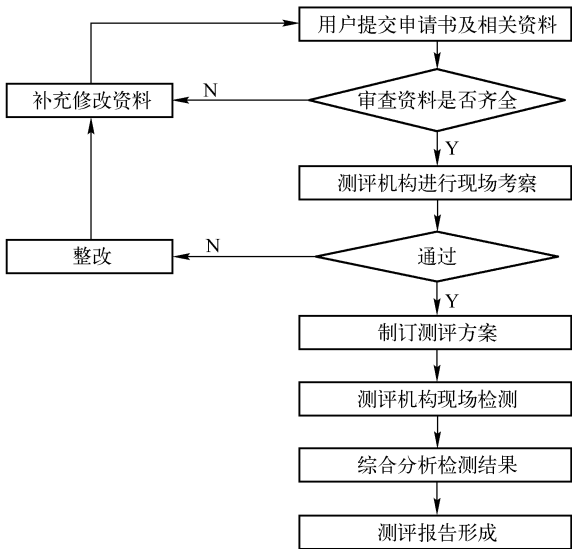


图3 等级测评基本工作流程

首先，测评委托单位按照要求提交测评所需的相关资料，测评机构对资料进行审查；如果审查通过，则整理获取的相关资料，制订测评方案，确定测评对象、测评指标及测评内容等；然后按照测评方案的总体要求进行现场测评，了解系统的真实保护情况，获取足够证据，发现系统存在的安全问题；最后根据现场测评结果和相关要求，分析整个系统的安全保护现状与相应等级的保护要求之间的差距，综合评价被测信息系统保护状况，并形成测评报告文本。

4.3 现有测评系统测评环节存在的主要问题分析

- 根据前期的调研，我们了解到目前在测评的各环节存在以下问题：
- （1）在资料申报过程中，申请单位不能完全按照要求上报资料，内容不全面、形式不够规范，为测评人员审查资料带来额外不必要的工作量；
 - （2）在资料审查过程中，目前主要依靠测评人员手工查阅相关文档，工作量大，效率低；
 - （3）在现场考察过程中，缺少对考察情况的完整记录；
 - （4）在测评方案制订过程中，存在方案制订不完备，容易遗漏检查点等问题；
 - （5）在现场测评过程中，测评组根据测评内容分为技术组和管理组。在管理组的评测过程中，大部分是对企业相关人员制度的测评，目前主要采用测评人员人工对文件与标准进行比对，工作量大，效率低；在技术组的评测过程中普遍存在着现场测评结果记录不完全，各种检测工具分散，结果不能集中展示，给现场的检测结果分析带来一定的困难。

根据以上存在的问题，重要信息安全测评工具应达到规范化、全面化、自动化、客观化、集成

化，并且界面友好、操作简单、可扩展性强等特点。

4.4 安全测评工具的设计

4.4.1 测评工具的框架结构

测评工具由测评申请软件和综合测评管理平台两部分组成，综合测评平台是整个安全测评系统的核心，主要由资料审查模块、测评方案生成模块、现场测评模块、测评结论分析生成模块四个模块组成。系统又由测评知识库、漏洞库，文档数据库三个数据库来支撑，整体结构如图 4 所示。

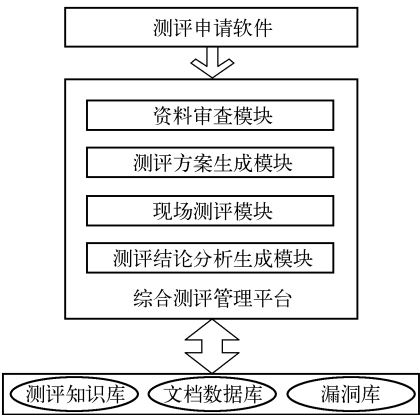


图 4 安全测评系统的整体框架结构

测评申请软件主要负责测评委托单位向测评机构提交《测评申请书》、《测评委托书》及其他资料。该软件能够以模板的形式提示用户填写或者上传资料，最后把资料以固定的格式导入到综合测评管理平台中，可以解决目前测评申请单位上报资料内容不全面、形式不规范等问题。通过将申报资料标准化，降低测评人员在资料审查过程中的手工工作量，提高工作效率。

资料审查模块主要负责将测评委托单位上报的各种资料以树形导航栏形式显示，点击名称资料显示在内容浏览区，可供测评员审阅；按关键字查找相关文档内容；采用关键字匹配技术审查资料是否合理，生成审查意见。该模块解决了测评人员制度组在对被测单位文件进行管理时，手工比对文件和标准工作量巨大，效率低的问题。

测评方案生成模块是整个测评系统的核心和灵魂。根据用户提交的资料和安全保护等级标准，确定测评对象和测评指标，结合现有辅助工具集，生成测评方案。确保所有关键点都会得到检查，并且检验检查方法的合理性及相互之间不存在冲突。由于有标准比对，测评方案中涉及的检查点比较全面，解决了方案制订不完备，容易遗漏检查点等问题。

现场测评模块主要负责根据测评方案，利用访谈、文档审查、配置检查、工具测试等方式测评被测系统的保护措施情况，并获取相关证据。该模块可以生成访谈调查表，集成合作基础好，便于整合的测评工具，并把自己开发的测评工具进行统一接口，统一界面进行集中管理，解决测评过程中检测工具分散，结果不能集中显示的问题。

测评结论分析模块主要负责汇总现场测评获得的测评结果，在汇总的基础上找出系统保护现状与等级保护基本要求之间的差距，形成等级测评结论，针对被测系统存在的安全隐患提出相应的改进建议，并编制测评报告。

4.4.2 安全测评工具实现的关键技术

(1) 测评知识库的建立与维护

测评系统关键模块的正常运行都离不开测评知识库的支持。测评知识库包括测评相关标准、具体测评点、测评指标、测评方法、以往测评的数据资料及一些文档模板，因此，测评知识库是否完整，

检查点涵盖是否全面是影响检测结果的关键。知识库为测评的有效性、权威性提供保证，随着标准的更新，测评单位情况的变更，知识库的内容需要不断地更新，因此需要对知识库进行维护和管理。

(2) 现有测评工具的集成

安全测评在技术方面主要采用不同的检测工具，工具比较分散，需要集成现有评测机构常用的测评工具，形成统一控制的界面，并且采用插件方式的集成到系统中，方便测评员更新和管理。

(3) 文档的规范化和标准化

安全测评主要包括技术、管理两大类。其中管理类多数依照被测单位的文档是否健全，是否符合标准要求，目前测评人员主要采用人工对比的方法和调查表的方式进行，工作效率不高，工作量巨大。为了便于自动评测被测单位的安全管理情况，本系统采用制定规范描述文档，可采用关键字匹配等技术帮助测评员对文档进行检查。关键字的设计和快速查找、文档的规范化设计都是项目的关键技术。

(4) 漏洞库的更新维护

漏洞库是确定测评准确性的关键因素，由于新的漏洞不断出现，该数据库需要经常更新，以便能够检测到新发现的漏洞。跟踪各漏洞发布站点、黑客论坛、搜索引擎，分析攻击手段，收集漏洞数据，及时准确地更新数据库是系统实现的关键技术也是难点。

5 结束语

本文主要从测评标准方面进行论述国内外安全测评发展的现状，并结合实际需求，设计了基于安全等级标准测评系统的框架结构，并对关键技术做了进一步的描述。下一步的工作是继续研究标准，细化各个模块，完成测评系统的开发。

参考文献

[1] 秦超, 张彬彬.重要信息系统安全测评平台的设计与实现[J].警察技术, 2008(6): 32-35.

[2] 美国国家标准《可信计算机系统评估准则》(TCESC)[S], 1983.

[3] the Department of Trade and Industry,London,Information Technology Security Evaluation Criteria,June 1991.

[4] Information Technology Security Evaluation Criteria (ITSEC),London,June,1991.

[5] 蔡昱, 张玉清, 冯登国等. 安全评估标准综述[J]. 计算机工程与应用 2004.2: 129-132.

[6] 中华人民共和国国家标准. 《计算机信息安全保护等级划分准则》[S], GB 17859—1999.

[7] 《信息安全等级保护安全建设整改工作指南》公信安[2009]1429 号.

[8] 杨磊, 郭志博.信息安全等级保护的等级测评[J]. 中国人民公安大学学报, 2007(1): 50-53.

[9] 信息系统安全等级保护测评过程指南（报批稿）.

基于策略的网络安全管理研究

王海涛

(河南理工大学计算机学院, 河南 焦作, 454000)

摘要: 随着网络的快速发展, 网络规模、结构变得日益复杂, 导致网络安全性的管理变得更困难。传统的集中式管理已不能应对复杂的大型分布式网络, 而基于策略的网络管理是一种有效解决方案; 针对该问题, 在研究和分析基于策略的管理框架后, 提出和设计了基于策略的网络安全管理系统模型, 实现对网络事件安全的自动管理, 简化网络安全的复杂性。最后, 以 Web 攻防事件为案例, 给出基于策略的响应预案, 实现其安全管理的过程。

关键词: 策略; 网络管理; 系统; 框架

中图分类号: TP393.02 **文献标识码:** A

Research on Network Security Management Based on Policy

WANG Haitao

(1. College of Computer Science & Technology, Henan Polytechnic University
Jiaozuo 454000, Henan China)

Abstract: With the rapid development of network, the scale and structure of network became more and more complex, which led network security management to become more difficult. Traditional centralized-management can no longer deal with the large complicate distributed network, whereas network management based on policy is a kind of effective solution project. To this task, this paper offered a network security management system model based on policy after researching on analyzing the management framework based on policy-based, which achieved automatic network security event management, simplified the complexity of network security management. Finally, take Web Attack event as a case, give the response preplan based on strategy and complete the security management process.

Keywords: policy; network management; system; framework

1 引言

随着网络的快速发展, 网络用户急剧增加, 网络中的应用不断丰富, 共存的技术越来越多, 网络结构变得更加复杂。在这种情况下, 网络管理变得尤为重要。传统的网络管理方式不能很好地适应这些要求, 这就需要网络管理者探求新的有效管理方案, 并要求管理方案是自适应的, 能动态地改变系统的行为。

目前, 国内外在网络安全管理方面有许多热点问题的研究, 文献[1, 2]指出了当前正在进行中的研究。其中, 基于策略的网络管理成为管理大型分布式网络的一种有效的解决方案。

使用策略的主要优势在于改善了管理系统的可扩展性和灵活性。可扩展性体现在对一系列设备和对象提供统一的策略。灵活性是通过把策略和系统中的具体实现分离获得, 策略能被动态地修改, 当一个系统行为和决策修改时, 不需要修改它的实现, 也不需要中断它的运行。因此基于策略的管理有以下优点: (1) 系统要求改变时, 只需改变或增加新的策略, 而不用重新编写程序, 就能适应变化的要求; (2) 可以根据不同的服务类型及动态信息灵活的分配资源, 使资源的利用率达到最大; (3) 可

项目名称: 通信领域网络管理数据整合软件平台; 编号: 072102212258
作者简介: 王海涛 (1974—), 男, 讲师, 硕士。

以根据不同的使用者进行策略的转换，方便用户的要求，提高系统的可扩展性和可维护性；（4）可以减少对管理员的依赖，提高系统的智能化程度。

2 基于策略的网络管理技术

2.1 策略的定义

策略是一些可以用来改变系统行为的信息，并描述了影响分布式系统中管理行为的方法，从而避免了将这些行为转变到管理代理中，实现分布式系统中的对象在不关闭整个系统的情况下得以改变，迎合变化的需要。

从内容上讲，网络策略^[3]是一组规则的集合，这些规则对网络资源的访问进行管理和控制。因此，策略具有以下特征，首先是策略的稳定性，策略相对系统的状态可以说是静态的，这意味着一次性执行一个行为的命令不是策略。其次是策略的描述特性，策略根据条件定义行为的选择（在特定的条件下预先定义好的操作或行为能被激活）而不是通过改变实际操作的功能。例如，策略指定希望实现什么样的行为，而不指定这些行为如何实现。最后是策略的来源，策略是为管理目标服务的，因此策略来源于企业决策，服务级别协议和组织间的信任关系等。

因此，基于策略的网络管理是对整个网络进行集成化管理。其中，策略是主要的管理元素，用于对系统管理、网络管理和应用管理进行协调，从而达到管理目标，提高效率，降低成本。策略可以包括 QoS 管理策略、安全管理策略、配置管理策略、性能管理策略。

2.2 基于策略的管理框架

IETF 给出了基于策略管理（PBNM）的标准体系结构，如图 1 所示。基于策略的框架最基本的优点是它允许用一个独立于机器的模式从一个控制点来管理众多的设备，实现了分布式系统管理要求解决的单一控制点技术。

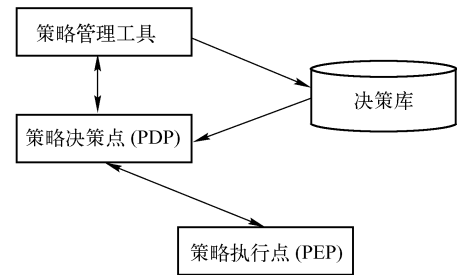


图 1 基于策略的网络管理模型

（1）策略决策点（Policy Decision Point, PDP）：负责存取存在策略库内的策略模式，并根据策略信息做出决策。策略决策点还能检测策略的变化和冲突，进而采取行动，纠正出现的问题。

（2）策略执行点（Policy Enforcement Point, PEP）：它是一些执行和实现策略的实际设备。如网络管理中的路由器、VPN 网关和防火墙。策略执行点汇总策略决策点产生的策略信息，并将这种信息存在高速缓存器内，它还可向策略决策点转发信息，以使策略决策点了解网络或设备条件的变化。

（3）策略库（Policy Repository, PR）：负责存放和检索策略。

（4）策略管理工具（Policy Management Tool, PMT）：网管员要通过策略管理工具来创建、编辑、删除、储存策略。PMT 提供了网管员对 PR 操作的接口。

PBNM 从全局的角度出发，将管理者的管理思想通过策略的方式来体现^[4]。它不像传统网络管理中进行被动检测，PBNM 从客户机/服务器方式转向服务驱动方式，从而使网管变得更加主动、灵活。通过策略将管理行为和具体实施分开，提高网络管理的自动化程度，使其朝智能化的方向迈进了一步^[5]。

2.3 基于 Ponder 的网络管理语言

Ponder 是一种面向对象的说明性语言，尤其适合描述分布式环境中的安全策略和管理策略^[6]。Ponder 可以实现基于角色的（Role-Based）访问控制，如用户的注册、登录或对重要信息的访问。也可以描述分布式系统中一般的管理策略，如事件触发的“条件—响应”规则。Ponder 定义了四种基本

策略：权限策略、应激策略、抑制策略和授权策略。

在 Ponder 中。用户可以定义策略的类型，然后根据具体应用将策略类型实例化。通过这种机制 Ponder 实现了策略定义的复用。见下例，用户首先定义一种应激策略类型 `perfIncreaseT`，描述设备在性能下降时自动保留更多带宽的策略；然后把该种策略实例化应用于核心路由器。

```
type
oblig perfIncreaseT(subject s, target t){
    on      perfdegradation(bw, source)
    do.t.bwReserve(bw)}

inst
oblig pl=perfIncreaseT(brEngineer,coreRouter)
```

在四种基本策略之上，Ponder 还定义四种复合策略。复合策略有助于在复杂的企业级分布式环境中建立策略体系，把基本策略结构化以反映企业的组织结构，并实现策略定义的重用。Ponder 复合策略有：Group，Role，Relationship，Management S tructure。除了基本策略和复合策略外，Ponder 另外设定了一种标记策略（Meta-Policy）。Meta-Policy 用 OCL 语言（Object Constraint Language ）描述，它定义了对一组策略的限制。例如，Meta Policy 可用于检查策略之间有投有语义冲突。

3 基于策略的网络安全系统实现

3.1 实现框架

本文设计了基于策略的网络管理系统模型，并以 Web 服务攻击防御为例，用策略的方法描述实现过程。在设计中引入基于域管理和事件关联分析的思想，实现对网络设备的集中、统一管理配置和设备间的联动协作^[7, 8]。系统由管理工具集、策略部署运行设施、事件关联模块、被管设备集等组成，各模块间关系如图 2 所示。

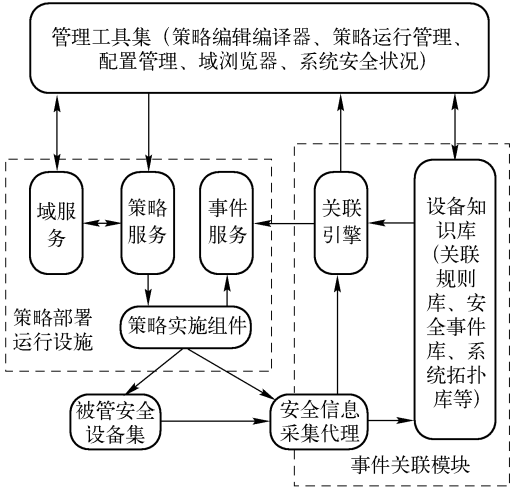


图 2 基于策略的网络安全系统各模块间的关系

- （1）管理工具集提供一个图形化的管理员界面，提供资源配置管理、策略管理、网络安全状况显示，允许管理员对安全策略进行定义、存储和运行时管理操作。网络安全状况显示界面，用于显示和查询安全信息和各自的安全状况，并给出网络安全态势图。资源配置管理实现设备的统一配置，集中管理。
- （2）策略部署、运行设施：是系统的核心部分，负责系统安全策略的统一定制、分发及自动执行，指导各用户和安全设备的行为，实现管理员的管理意图。该模块包括域服务、策略服务、事件服务及策略实施组件等部分，其核心部分是域服务，策略服务及事件服务。策略服务用作策略管理的接

口，它存储编译好的策略类、创建并分发新的策略对象，策略服务为每个策略对象创建相应的策略控制对象，对策略对象进行运行时管理。域服务用来管理域对象的层次结构并支持运行时主体和目标集的有效性评估，每个域对象包含指向其管理对象的引用及当前应用于域的策略对象的引用；域服务通过 LDAP 目录服务实现。事件服务收集系统事件和系统中被管对象发布的事件，并将它们通知给已预定该事件的策略管理组件，以触发职责策略。

（3）事件关联模块：由信息采集代理、设备知识库和关联引擎组成；信息采集代理主要负责收集各设备产品的报警信息和操作日志信息，对所有信息进行标准化、过滤融合等操作；知识库包含关联规则库、安全事件库、设备信息库及拓扑信息等，为事件关联模块和管理工具集提供及时的和基于历史的数据支持；关联引擎是该模块的核心，以信息采集代理的输出信息为输入，在知识库的支持下，利用智能关联算法对网络运行时产生的大量安全信息进行跨边界、跨设备、跨时空的关联分析，形成准确的事件报告。在事件关联模块中，为减少报警的误报率和漏报率，将不同的报警信息送入关联模块，进一步进行关联分析，发现新的异常情况和复杂的攻击模式。事件关联方法的引入实现了各设备间的信息共享，并为实现及时、准确、有效的响应提供了可能。

3.2 系统案例与应用

本文使用 Java RMI 机制实现远程通信和分布式操作为例，系统安全管理员统一定制安全策略模板，自动分发给各相应模块，当中心事件服务收到安全事件后，触发部署在相应的策略模板，策略模板指导响应模块进行攻击响应和系统恢复，并对防护模块进行新安全措施部署，同时对检测模块加强检测规则配置，如果相当一段时间内攻击现象不再发生，策略管理中心可停止策略模板的响应措施，并记录系统当前的安全状态，向系统安全管理员进行报告。以简单 Web 服务攻击防御为例来介绍系统的运行细节，当检测模块检测到一个 Web 攻击时，并向策略管理中心告警，告警信息进入安全事件关联引擎，计算告警的严重度与可能性，并重新配置信息收集器以获得更多的攻击特征，从而减少误报率。如果攻击的可能性和严重度很高，则事件服务通过事件通知接口通知响应模块触发相应的响应机制。

针对 Web 服务攻击，有效的响应机制应该包括以下几个步骤：

（1）向安全管理员报警；（2）关闭攻击源和受害对象的 TCP 连接；（3）重新配置防火墙规则，拒绝所有来自攻击源的连接请求；（4）重新配置检测模块的检测规则，挖掘报警的上下文信息，获得更多的攻击特征。下文给出一个响应预案的例子，如图 3 所示。

```
Inst oblig /SMPolices/signatureBasedAlertRe {
  on WebAttack(attackSignature, confidence);
  subject /PMA/SM;
  do getRisk(attackSignature)->
alert(attackSignature)->generateEvent(AlertRespond,
  attackSignature)->generateEvent(reConfigFW,attackSignature);
  when confidence >MIN_CONFIDENCE; }
Inst oblig /SMPolices/signatureBasedAlert {
  on WebAttack(attackSignature, confidence);
  subject /PMA/SM;
  target t=/PMA/DM;
  do t.reConfigureDM(DecRules)-> t.getContext(attackSignature);
  when confidence <MIN_CONFIDENCE;
}
Inst oblig /SMPolices/signatureBasedAlertFW
{ on reConfigFW (attackSignature);
  subject /PMA/PM;
  target t=/PMA/FW;
  do t.deny(attackSignature);
}
```

图 3 基于策略的攻防响应预案

4 结束语

针对当前网络管理方面的不足，本文设计并实现了基于策略驱动的网络管理系统，系统引入策略管理的思想和安全事件关联分析方法，以策略管理实现网络的统一配置和自动管理，提高了系统的可扩展性和灵活性；以安全事件关联分析实现信息共享，提高检测能力和准确快速的安全响应。具体来说，完成了两方面的工作：（1）将基于策略的管理思想应用于网络安全管理中，摆脱了传统的面向设备的管理模式，为网络安全管理提供了新颖、有效的方法支持。（2）设计基于策略的网络安全管理系统模型，提供管理策略的统一定义、自动分发实施，安全事件智能关联分析，能有效处理网络运行时产生的大量事件。以策略驱动的安全设备联动，能对突发安全事件做出及时有效的自动响应。最后以一个简单 Web 攻防为例，给出安全响应预案的案例。

参考文献

- [1] 马育峰, 胡修林, 张蕴玉. 网络管理热点问题研究的现状、问题与展望[J]. 计算机应用研究, 2005, (3): 10-13.
- [2] 孟洛明, 网络管理研究中的问题、现状和若干研究方向[J]. 北京邮电大学学报, 2003, (2): 1-4.
- [3] 卢世凤, 刘学敏等. 基于策略的管理综述[J]. 计算机工程与应用, 2004, (9): 85-89.
- [4] 邹一鸣, 王汝传. 一种基于策略和移动代理的网络管理体系结构[J]. 南京邮电学院学报, 2005, 25(1): 90-94.
- [5] 赵季红. 下一代网络的网络管理技术[J]. 西安邮电学院学报, 2004, 9(1): 1-5.
- [6] 李祥军. 基于策略的网络管理关键技术及其应用的研究[D]. 北京: 北京邮电大学, 2005.
- [7] 单康康, 张兴明等. ACR 网管系统中基于策略的管理方法及应用[J]. 计算机工程, 2008, 34(15): 123-125.
- [8] 张若英, 邱雪松, 孟洛明. SLA 的表示方法和应用[J]. 北京邮电大学学报, 2003, 26(10): 12-17.

信息安全技术在电子政务系统中的应用研究

张 鸣

(防空兵指挥学院, 河南 郑州, 450052)

摘 要: 电子政务信息涉及国家安全、经济利益和公民隐私, 进行电子政务信息安全保障体系研究, 对提升我国电子政务信息安全保障能力、保障政府信息化与信息安全协调发展具有重要的理论意义和实践价值。

关键词: 信息安全技术; 电子政务系统; 网络安全

The Application Research of Information Security Technology in E-government System

ZHANG Ming

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: The information of E-government relates to the national security, the economic interest and the citizen privacy , the research of the E-government information security safeguard system has theoretical significance and the practice value for the ability of the E-government information security enhance, the government informatization and information security coordinated development.

Keywords: information security technology; E-government system; network security

1 引言

20 世纪 90 年代以来, 伴随着信息技术、特别是网络技术的飞速发展, 信息化已成为各国普遍关注的一个焦点。在国家信息化体系建设中, 政务信息化是其中的关键。目前, 我国电子政务建设从整体上取得了重要进展。

所谓电子政务, 就是政府机构应用现代信息和通信技术, 将管理和服务通过网络技术进行集成, 在因特网上实现政府组织结构和 workflows 的优化重组, 超越时间和空间及部门之间的分隔限制, 向社会提供优质和全方位的、规范而透明的、符合国际水准的管理和服务。电子政务必须借助电子信息化硬件系统、数字网络技术和相关软件技术的综合服务系统。

电子政务是提升一个国家或地区特别是城市综合竞争力的重要因素之一, 电子政务系统中有众多的政府公文在流转, 其中不乏重要情报, 有的甚至涉及国家安全, 这些信息通过网络传送时不能被窃听、泄密、篡改和伪造。如果电子政务网络安全得不到保障, 电子政务的便利与效率便无从保证, 给国家利益将带来严重威胁。因此, 研究信息安全技术及其在电子政务系统中的应用具有重要的现实意义。

2 当前电子政务系统存在的信息安全问题

为了保证电子政务业务系统的信息安全, 电子政务安全系统应对进行电子政务业务的实体定义唯一的电子身份标识, 并通过该标识进行身份认证, 保证身份的真实性; 把信息资源划分成不同级别, 并把使用信息资源的用户划分成不同类型, 实现不同类型人员对不同级别信息访问的控制策略; 对于

作者简介: 张鸣 (1984—), 女, 助教, 在读硕士研究生。

传输中需要保密的信息，采用密码技术进行加解密处理，防止信息的非授权泄漏；保证收发双方数据的一致性，防止信息被非授权修改。

电子政务网在建网初期都是按照双网模式来进行规划与建设的，即电子政务内网和外网，内网作为信息内部信息和公文传输平台，开通政府系统的一些日常业务，外网通过路由会聚加防火墙过滤接入，主要向公众发布一些公共服务信息和政府互动平台。双网实现了物理隔离，建网初期往往只看重网络带来的便利与高效，但是并没有同步充分考虑安全问题，也没有对互联网平台的潜在安全威胁进行过全面综合的风险评估，将受到来自网络的安全威胁，如网络的数据窃贼、黑客的侵袭、病毒发布者等。网络安全方面的问题主要涉及以下几个方面：

一是来自内部的恶意攻击：这种攻击往往带有恶作剧的形式，虽然不会给信息安全带来什么大的危害，但对本单位正常的业务和敏感数据还是会带来一定的威胁。

二是移动存储介质的交叉使用：U 盘、移动硬盘如果在内网计算机与互联网的计算机混用，则 U 盘、移动硬盘将会被“摆渡”间谍木马植入病毒（汇编程序），将对方所需的信息打包回传给木马的制作者（窃密）。并且，这种病毒还会在内网上传播，消耗系统资源，降低平台的性能，影响政务内网各种业务的正常流转。

三是内网的身份认证和权限管理：要防止内网一般用户越权管理数据库，服务器和高权限的数据平台（破译相关密码）。

四是内网中使用的存储介质、外设终端设备的维修、销毁问题（潜在失密隐患）。以上各个环节管理不好都会造成国家重要机密的泄密。

3 安全技术 在政务系统中的具体应用

通常情况下，政务网建设将市、区县政务网连接在一起，为政府的内部办公和对外服务提供了极大的便利。但是电子政务应用中内网与专网、外网间存在着信息交换需求，然而基于内网数据保密性的考虑，不能将内网暴露在对外环境中。解决该问题的有效方式是设置安全域。通过安全域来实现内外网间信息的过渡和两个网络间的物理隔离，从而在内外网间安全地实现数据交换。

3.1 安全区域的有效划分

安全区域是指由实施共同或相似安全策略的主体和客体组成的集合。安全区域的划分主要遵从以下原则：系统功能和应用相似；信息资产价值相似；安全需求相似；环境威胁相似。安全区域既可以从物理上划分，也可以从逻辑上划分。

安全区域的物理划分是依据网络系统所处的物理位置，如地理位置、建筑大楼等。安全区域的逻辑划分则依据国家政策和管理规范，如政府专网，政府外网等。从逻辑上划分的安全区域，更易于反映出安全政策的要求。不同信息安全区域之间的信息交互主要通过访问控制、鉴别服务、数据完整性检测、数据保密和抗抵赖性等技术措施来实现，以保证信息在交换和共享过程中的保密性、完整性和可用性。

为解决好这些信息共享与保密性、完整性的关系、开放性与保护隐私的关系、互联性与局部隔离的关系，实现安全目标，在典型的电子政务内网、外网和专网架构基础上，将其划分为四个层次安全体系。

第一层是核心决策层，就是国家涉密的、最核心的、最机密的那一层。对于这一层来说，高度的认证，高度的预审和用核心密码加密，是非常必要和重要的。

第二层是政府办公业务处理层，是政府内部的电子办公环境，该层内的信息只能在内部流动。这一层一般是有防火墙、访问控制等安全措施组成。一、二层合起来就是内网（即局域网）。

第三层是信息交换层，一部分需要各部门交换的信息可以通过专网域进行交换，还可以将信息从

一个内网传送到另一个内网区域，它不与外网域有任何信息交换，这就是专网。

第四层是最外层，即公共服务层，是政府部门公共信息发布场所的外部网，三网间应根据国家保密局要求，实施物理隔离。

3.2 信息的有效控制方法

电子政务系统的数据控制主要目的是阻止攻击者利用政务系统的管理主机作为平台去攻击其他的机器，当然，针对政务内部网络任何的扫描、探测和连接管理主机是允许的，但是对从主机出去的扫描、探测、连接，网络安全系统却必须有条件的放行，如果发现出去的数据包有异常，系统管理员必须加以制止。

隔离网闸（GAP）技术是实现安全域的关键技术，它如同一个高速开关在内外网间来同切换，同一时刻内外网间没有连接，处于物理隔离状态。在此基础上，GAP 作为代理从外网的网络访问包中抽取数据然后通过反射开关转入内网。完成数据中转，在中转过程中，GAP 会对抽取的数据做应用层的协议检查、内容检测，也会对 IP 包地址实施过滤控制，由于 GAP 采用了独特的开关切换机制，因此，在进行这些检查时网络实际上处于断开状态，只有通过严格检查的数据才有可能进入内网，即使黑客强行攻击了隔离网闸，由于攻击发生时内外网始终处于物理断开状态，黑客也无法进入内网。另外，由于 GAP 仅抽取数据交换进内网，因此，内网不会受到网络层的攻击，这就在物理隔离的同时实现了数据的安全交换。以 GAP 技术为核心，通过添加 VPN 通信认证、加密、入侵检测和对数据的病毒扫描，就可构成一个在物理隔离基础上实现安全数据交换的信息安全域。

防火墙技术是指网络之间通过预定义的安全策略，对内外网通信强制实施访问控制的安全应用措施。它对两个或多个网络之间传输的数据包按照一定的安全策略来实施检查，以决定网络之间的通信是否被允许，并监视网络运行状态。由于它简单实用且透明度高，可以在不修改原有网络应用系统的情况下，达到一定的安全要求，所以被广泛使用。

3.3 系统 VPN 的合理设计

使用 VPN，可以在电子政务系统所连接不同的政府部门之间建虚拟隧道，使得两个政务网之间的相互访问就像在一个专用网络中一样。使用 VPN，可以使政务网用户在外网就像在内网一样地访问政务专用网的资源。使用 VPN，也可以实现政务网内特殊管理的需要。VPN 的建立有三种方式：第一种是 Internet 服务商（ISP）建设，对企业透明；第二种是政府部门自身建设，对 ISP 透明；第三种是 ISP 和政府部门共同建设。

在政务网的基础上建立 VPN，第二种方案比较合适，即政府部门自身建设，对 ISP 透明。因为政务网是地理范围在政务网内的计算机网络，它有运行于 Internet 的公网 IP 地址，有自己的路由设备，有自己的网络管理和维护机构，对政务网络有很强的自主管理权和技术支持。所以，在政务网基础上建立 VPN，完全可以不依赖于 ISP，政府部门自身进行建设。这样可以有更大的自主性，也可以节省经费。

4 电子政务中安全问题的对策

4.1 物理安全

保证计算机信息系统各种设备的物理安全是整个电子政务系统安全的前提。具体来说，电子政务中的物理安全主要包括以下几个方面。一是系统运行中的安全隐患，主要包括电源问题、水患与火灾、电磁干扰与泄漏及其他的环境威胁。二是物理访问风险与控制。物理安全威胁不仅来自于环境，还来自于人为操作失误及各种计算机犯罪行为。三是电子政务信息系统的场地与设施安全管理。是指

中华人民共和国国家标准 GB 50173 —1993 《电子计算机机房设计规范》、GB 2887 —1989 《计算站场地技术条件》、GB 9361—1988 《计算站场地安全要求》对应用信息系统的场地与设施进行的安全管理。

4.2 系统安全

（1）硬件系统安全的问题及对策

物理设备（CPU、硬盘、内存、网络设备）遭到损毁，整个系统就不可避免地会受到严重的影响。因此，首先考察硬件系统的安全问题。一般来说，硬件系统的漏洞主要包括设计疏忽、元件质量低劣、人为留下的“后门”等。对于这些问题，解决办法就是做好电子政务系统硬件采购时的分析工作。

（2）软件系统的安全问题及对策

软件系统是电子政务系统的灵魂，因此保证软件系统的安全运行、降低由于软件问题而导致的系统风险对电子政务系统来说至关重要。

首先是计算机病毒的防范措施，其次是软件漏洞与后门。为了解决电子政务平台软件系统中的安全问题，应该针对操作系统、应用程序、数据库等方面采用如下安全防护措施。

一是在电子政务系统的实际开发、构建过程中，应根据具体运行环境、安全级别，分别采用不同类别的操作系统。

二是尽可能采用具有自主知识产权且源代码对我国政府公开的产品；采用国产软件来构建电子政务系统成为一项切实可行的措施。

三是尽量避免在电子政务关键部门、要害环节使用外国的产品，以免受制于外国。对于核心应用系统和关键政务环节，必须确保实施方案的技术自主性。

4.3 信息安全

数据是电子政务的核心资源之一，因此其安全问题也格外重要，解决数据安全问题的主要办法就是建立完善的管理措施。一是做好物理访问控制，防止外部人员接触到存有重要数据的主机、存储设备和打印机等输出设备；二是做好数据的逻辑访问控制；三是做好数据备份和灾难恢复计划；四是标准可信时间源的获取；五是信息传递过程中的加密。

4.4 管理安全

对于电子政务系统来说，在组织管理措施方面要想做到合理分工，在实际建设和应用电子政务系统时把系统的设计者、管理者、操作者、利害关系者等角色分开，尽量避免一个人同时担任多个角色。在人力资源管理方面，一是对人员的安全教育和保护；二是对内部成员的安全防范。

参考文献

[1] 李军. 地理空间信息及技术在电子政务中的应用[M]. 北京：电子工业出版社，2005.
[2] 王琰，徐玲. 电子政务理论与实务[M]. 北京：电子工业出版社，2004.

浅谈电子商务及其安全性

郎士宁，秦兴桥，王月蓉

(防空兵指挥学院，河南 郑州，450052)

摘 要：针对电子商务普及化趋势所带来的安全威胁，提出电子商务网络安全技术策略与安全交易标准方法，给出防范电子商务安全问题的对策。

关键词：电子商务；安全威胁；安全技术；交易标准

中图分类号：TP39

文献标识码：A

文章编号：1006-7043 (2010) xx-xxxx-x

Discussion of E-commerce and Its Safety

LANG Shining, QIN Xingqiao, WANG Yuerong

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: Security threat which brings in view of the electronic commerce universalization tendency, proposes the electronic commerce network security technology strategy and the safe transaction standard method and the guard electronic commerce security problem countermeasure.

Keywords: E-commerce; safety threat; safety technology; trading standards

随着互联网技术的蓬勃发展，基于网络和多媒体技术的电子商务应运而生并迅速发展，然而由于互联网的开放性，网络安全问题日益成为制约电子商务发展的一个关键性问题。电子商务的安全运行要从多方面进行考虑。

1 电子商务概述

1.1 电子商务的定义

电子商务通常是指是在全球各地广泛的商业贸易活动中，在因特网开放的网络环境下，基于浏览器/服务器应用方式，买卖双方不谋面地进行各种商贸活动，实现消费者的网上购物、商户之间的网上交易和在线电子支付及各种商务活动、交易活动、金融活动和相关的综合服务活动的一种新型的商业运营模式。

电子商务系统是保证以电子商务为基础的网上交易实现的体系。广义上是指支持电子商务活动的电子技术手段的集合。狭义上是指在 Internet 和其他网络的基础上，以实现企业电子商务活动为目标，满足企业生产、销售、服务等生产和管理的需要，支持企业的对外业务协作，从运作、管理和决策等层次全面提高企业信息化水平，为企业提供商智能的计算机系统。

在网上进行交易，交易双方在空间上是分离的，为保证交易双方进行等价交换，必须提供相应货物配送手段和支付结算手段。货物配送仍然依赖传统物流渠道，对于支付结算既可以利用传统手段，也可以利用先进的网上支付手段。此外，为保证企业、组织和消费者能够利用数字化沟通渠道，保证交易顺利进行的配送和支付，需要由专门提供这方面服务的中间商参与，即电子商务服务商。

作者简介：郎士宁（1973—），女，讲师，硕士；

秦兴桥（1976—），男，讲师，硕士；

王月蓉（1985—），女，助教，学士。

1.2 电子商务的普及化趋势

1996 年 12 月，当时的美国总统克林顿签署了由 19 个政府机构参与起草的全球电子商务政策框架（也称实施纲领），后来经过半年的讨论，于 1997 年 7 月作为美国政府的正式文件，在这份文件中第一次使用了电子商务这个词汇。在以后的国际论坛上，美国代表团以此文件为准则，与全世界各国商讨有关全球电子商务的政策法规问题。该文件从电子商务的发展战略及法律、税收等方面谈到如何促进电子商务的发展，以实现全球一体化。

中国电子商务始于 1997 年，已经有 13 年的历史了。近年来，各电子商务企业营收呈快速增长状态，主要原因是随着中小企业逐渐认识到电子商务的好处，越来越多的中小企业开始使用第三方电子商务平台，以及网购理念的普及、网购用户的增长有着密切关联，而电子商务渗透率也随之保持持续高速增长。

国务院新闻办公室近日发布的《中国互联网状况》白皮书显示，互联网成为推动中国经济发展的重要引擎。2009 年，中国电子商务交易额超过 3.6 万亿元人民币。电子商务专业化服务体系正在形成，数字认证、电子支付、物流配送等电子商务应用支撑体系正在逐步形成。

日前，艾瑞咨询发布的 2010 年第一季度中国电子商务市场数据显示，今年第一季度中国电子商务交易额达到了 10152.7 亿元，单季交易额突破万亿规模。从投融资规模来看，第一季度互联网行业 14.3 亿元的投融资额当中，电子商务的占比高达 74.3%。从电子商务市场结构来看，企业间电子商务仍然是电子商务市场的主体。中小企业 B2B 电子商务交易规模占比最高，达到了 52.6%，成为电子商务市场发展最大的推动力。艾瑞咨询分析认为，宏观经济的复苏、中小企业利用电子商务的意识逐步提升、B2B 运营商相关扶持政策的推出等，是中小企业 B2B 电子商务交易额增长的主要原因，未来随着中小企业电子商务渗透率的稳定提升，中小企业 B2B 电子商务交易额还将稳定速度。

从 2010 年 6 月起，淘宝和软银旗下日本雅虎合作的跨国电子商务贸易平台正式启动商用。而此类的跨国电子商务，也在 eBay 和走秀网等一批电子商务推进者的努力下，逐渐推平 B2C 电子商务的全球壁垒。面对国内的激励竞争，越来越多的电子商务企业跨出国门，将海外战略作为扩张的重要途径。

1.3 电子商务面临的安全威胁

电子商务是互联网应用发展的必然趋势，也是国际金融贸易中越来越重要的经营模式，越来越多的人通过 Internet 进行商务活动。电子商务的发展前景十分诱人，而其安全问题也变得越来越突出。近年来，网络安全事件不断攀升，电子商务金融成了攻击目标，以网页篡改和垃圾邮件为主的网络安全事件正在大幅攀升。

2010 年 3 月 30 日，中国互联网络信息中心（CNNIC）和国家互联网应急中心（CNCERT）在京联合发布《2009 年中国网民网络信息安全状况调查系列报告》^[1]，调查显示：2009 年，52% 的网民曾遭遇过网络安全事件。其中，21.2% 的网民带来直接经济损失，包括即时通信、网络游戏等账号被盗造成的虚拟财产损失，网银密码、账号被盗造成的财产损失，以及因网络系统、操作系统瘫痪、数据、文件等丢失或损坏，对其找回或修复产生的费用等。

对网民用于处理网络安全事件支出的费用进行统计显示：2009 年，网民处理安全事件所支出的服务费用共计 153 亿元人民币；在实际产生费用的人群中，费用在 100 元及以下的占比 51.2%；人均费用约 588.9 元；如按国内 3.84 亿网民计算，人均处理网络安全事故费用约为 39.9 元。

调查发现，计算机病毒木马对信息网络安全的威胁越来越突出。未修补系统安全漏洞仍然是导致安全事件发生的最主要原因。根据国家计算机病毒应急处理中心病毒样本库的统计，2009 年新增病毒样本 299 万个，是 2008 年新增病毒数的 3.2 倍，其中木马程序巨量增加。截至 2009 年年底，木马样本共 330 万多个，占病毒木马样本总数的 72.9%，而 2008 年这一比例只有 54%；2009 年发现新增木马 246 万多个，是 2008 年新增木马的 5.5 倍。

赛门铁克预计 2010 年网络威胁趋势^[2]：传统的防病毒方法都不足以防范当前的威胁；社会工程成为主要的攻击媒介；社交第三方应用软件将成为诈骗目标；快速变化的僵尸网络将会增加；垃圾邮件数量将会继续波动；特定用途的恶意软件等都将使网络安全的防护更加复杂。

根据调查显示，目前电子商务主要存在两大部分的安全问题：计算机网络安全和商务交易安全（包括商品的品质、商家的诚信、货款的支付、商品的递送、买卖纠纷处理和网站售后服务等）。计算机网络安全与商务交易安全实际上是密不可分的，两者相辅相成，缺一不可。电子商务的一个重要技术特征是利用 IT 技术来传输和处理商业信息。没有计算机网络安全作为基础，商务交易安全就犹如空中楼阁，无从谈起。没有商务交易安全保障，即使计算机网络本身再安全，仍然无法达到电子商务所特有的安全要求。只有解决好以上的矛盾，电子商务才能保证又快又好的发展。

2 网络安全技术策略

电子商务的安全问题，可以归结两大类问题：一是支付安全，二是认证安全。

2.1 支付安全

由于网络天生的不安全性，特别是其网上支付领域有着各种各样的交易风险。但无论是何种风险，其根本原因都是由于登录密码或支付密码泄露造成的。

（1）密码管理问题

大部分公司和个人受到网络攻击的主要原因是密码政策管理不善。大多数用户使用的密码都是字典中可查到的普通单词姓名或者其他简单的密码。有 86%的用户在所有网站上使用的都是同一个密码或者有限的几个密码。许多攻击者还会直接使用软件强力破解一些安全性弱的密码。

因此，建议用户使用复杂的密码，降低被病毒破译密码的可能性，提高计算机系统的安全性。需要注意：一是密码不要设置为姓名、普通单词、电话号码、生日等简单密码；二是结合字母、数字、大小写共组密码；三是密码位数应尽量大于 9 位。

（2）网络病毒、木马问题

现今流行的很多木马病毒都是专门用于窃取网上银行密码而编制的。木马会监视 IE 浏览器正在访问的网页，如果发现用户正在登录个人银行，直接进行键盘记录输入的账号、密码，或者弹出伪造的登录对话框，诱骗用户输入登录密码和支付密码，然后通过邮件将窃取的信息发送出去。

因此，需要做好自身计算机的日常安全维护，注意以下几点：

一是经常给计算机系统升级；二是安装杀毒软件、防火墙，经常升级和杀毒；三在平时上网是尽量不上一些小型网站，选大型网站，知名度比较高的网站，避免网站挂有病毒、木马造成中毒；四是尽量不要在公共计算机上使用自己的有关资金的账户和密码；五是在有条件的情况下，在初装系统后确认计算机安全以后，给自己的计算机做备份，在使用资金账户前做一次系统恢复。

（3）钓鱼平台

钓鱼网站及通过网络行骗统称为“网站钓鱼”，其行为成为目前最具威胁的网络安全问题。“网络钓鱼”攻击者利用欺骗性的电子邮件和伪造的 Web 站点来进行诈骗活动，如将自己伪装成知名银行、在线零售商和信用卡公司等可信的品牌。受骗者往往会泄露自己的财务数据，如信用卡号、账户号和口令等。

因此，在登录支付资金时，应注意：一是确认该网是否是官方网站；二是仔细核对该网的域名是否正确，注意小写“l”与“L”、“0”与“O”等情况；三是保证良好的上网习惯，收藏常用的网址，减少网上链接。

（4）硬件数字认证

在电子商务体系构建的过渡时期，道高一尺，魔高一丈。各类病毒层出不穷，木马也在天天更

新，今天这种技术安全，明天就不一定安全。

因此，数字证书的引入是在线支付安全问题的最终解决方案之一。网上支付不安全，选择网下加以弥补。

以工商银行 2003 年推出并获得国家专利的客户证书 USBkey（U 盾）为例。从技术角度看，U 盾是用于网上银行电子签名和数字认证的工具，它内置微型智能卡处理器，采用 1024 位非对称密钥算法对数据进行加密、解密和数字签名。确保网上交易的保密性、真实性、完整性和不可否认性。它顺利地解决了当前网银密码泄露的问题。有了硬件数字证书的应用，即使你的密码泄露了。没有证书，黑客还是不能够使用你的账户。

动态电子密码的应用也可以确保电子银行账号的安全。现行的有两种方式：一种是在使用时查看当前的动态电子密码；另一种是临时通过绑定手机、密宝等通信工具，向账户所在银行申请临时密码。由于具有较强的时效性，从而保障账户资金的安全。

还有其他消极的防护措施。如某些网上银行交易金额限制，单次为 300 元，每日限额为 3000 元。主要是为了降低电子支付交易风险，但在一定程度上会给大额交易带来不便。这种措施其实治标不治本。

2.2 认证安全

电子商务为了保证网络上传递信息的安全，通常采用加密的方法。但这是不够的，如何确定交易双方的身份，如何获得通信对方的公钥，并且相信此公钥是由某个身份确定的人拥有的，解决方法就是找一个大家共同信任的第三方，即认证中心（Certificate Authority, CA）颁发电子证书。用户之间利用证书来保证安全性和双方身份的合法性，只有确定身份后，交易的纠纷，才得到有效的裁决。

总之，电子商务的安全是个非常复杂的问题，它的保障机制必须是有机的，多层次的，需要企业有管理方面，技术支持方面的协调来实现。它是一个系统有机的整体，不仅需要计算机网络安全保障，也需要商务交易安全上的保障，更需要管理上的进步，才能确保电子商务的安全。

同时我国在电子商务技术性较为落后，必须加强具有自主知识产权的信息安全产品的研究，注意加强信息安全人才的培养，多方共同努力建立科学的电子商务安全机制，才能为我国电子商务又快又好的发展保驾护航。

3 电子商务安全交易的有关标准和实施方法

3.1 安全交易的雏形

在电子商务实施初期，曾采用过一些简易的安全措施，这些措施包括：

- （1）部分告知（Partial Order）：在网上交易中将最关键的数据如信用卡号码及成交数额等略去，然后再用电话告之，以防泄密。
 - （2）另行确认（Order Confirmation）：当在网上传输交易信息之后，再用电子邮件对交易作确认，才认为有效。
 - （3）在线服务（Online Service）：为了保证信息传输的安全，用企业提供的内部网来提供联机服务。
- 以上所述的种种方法，均有一定的局限性，且操作麻烦，不能实现真正的安全可靠性。

3.2 安全交易标准的制定

近年来，IT 业界与金融行业一起，推出不少更有效的安全交易标准。主要有：

- （1）安全超文本传输协议（S-HTTP）：依靠密钥对的加密，保障 Web 站点间的交易信息传输的安全性。

(2) 安全套接层 (Secure Socket Layer , SSL) 协议是由网景 (Netscape) 公司推出的一种安全通信协议, 是对计算机之间整个会话进行加密的协议, 提供了加密、认证服务和报文完整性。它能够对信用卡和个人信息提供较强的保护。SSL 被用于 Netscape Communicator 和 Microsoft IE 浏览器, 用于完成需要的安全交易操作。在 SSL 中, 采用了公开密钥和私有密钥两种加密方法。

(3) 安全交易技术 (Secure Transaction Technology, STT) 协议: 由 Microsoft 公司提出, STT 将认证和解密在浏览器中分离开, 用于提高安全控制能力。Microsoft 将在 Internet Explorer 中采用这一技术。

(4) 安全电子交易 (Secure Electronic Transaction, SET) 协议: 是由 VISA 和 MasterCard 两大信用卡公司于 1997 年 5 月联合推出的规范。SET 协议主要是为了解决用户、商家和银行之间通过信用卡支付的交易而设计的, 以保证支付信息的机密、支付过程的完整、商户和持卡人的合法身份, 以及可操作性。SET 中的核心技术主要有公开密钥加密、电子数字签名、电子信封、电子安全证书等。

4 防范电子商务安全问题的对策

电子商务的安全问题涉及电子商务的各个环节和参加交易的各个方面, 解决电子商务的安全问题是一个系统工程和社会问题, 需全社会的参与。我们可以从以下几个方面对电子商务安全问题进行防范。

(1) 构建电子商务信息安全技术框架体系

在电子商务的交易中, 电子商务的安全性主要是网络安全和交易信息的安全。而网络安全是指网络操作系统对抗网络攻击、病毒, 使网络系统连续稳定的运行, 常用的保护措施有防火墙技术。交易信息的安全是指保护交易双方的不被破坏、不泄密和交易双方身份的确认, 可以用信息加密技术、数字证书和认证技术、SSL 安全协议、SET 等技术来保护。

(2) 积极推进电子商务立法

我国政府要加强对电子商务的研究, 要加快立法进程, 健全电子商务法律体系, 建立规范电子商务的灵活法律框架, 使电子商务实现公开、合理、合法化。这样不仅可保障进行电子商务各方面的利益, 而且还可保障电子商务的顺利进行。

综上所述, 由于在电子商务的交易过程中, 安全问题涉及电子商务的各个环节和参加交易的各个方面, 因此需要采取不同的对策来解决。另外, 交易过程除涉及交易双方外, 还涉及网上银行、认证中心和法律等各方面的问题, 因此电子商务安全问题的解决是一个系统工程。

参考文献

[1] CNCER T/CC. 2009 年中国网民网络信息安全状况调查系列报告.
[2] 赛迪网. 2009 网络安全热点及 2010 信息安全威胁趋势. 2010.

浅谈通信新技术的安全威胁

杨 凯，郎士宁，黄欢欢

(防空兵指挥学院，河南 郑州，450052)

摘 要：针对通信技术的发展现状，分析了无线网络技术、第三代移动通信技术和蓝牙技术的安全威胁，提醒开发者和使用者构建安全的无线通信环境。

关键词：无线网络技术安全；第三代移动通信技术安全；蓝牙技术安全

中图分类号：TP39 **文献标识码：**A **文章编号：**1006-7043 (2010) xx-xxxx-x

Discussion of the Security Threats of New Communications

YANG Kai , LANG Shining, HUANG Huanhuan

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: Aiming at the developing status of the communication, it analys the wirele ss networking, the third generation mobile communication technology and the blue tooth technology security threat, reminds the e xploiter and the user constructs the secu- rity the wireless communication environment.

Keywords: wireless network technology security; the third genera tion mobile communication technology security; bluetoot h security

进入 21 世纪以后，通信技术以人们难以想象的速度向前发展，为我们的生活带来了巨大的变化。在现代通信新技术中，无线网络技术、第三代移动通信技术及蓝牙等无线通信技术无不深入影响着我们的工作和生活，其安全性也已经成为大家关注的重要问题。

1 无线网络技术的安全¹

从 20 世纪 90 年代以来，移动通信和 Internet 是信息产业发展最快的两个领域，它们直接影响了亿万人的生活，移动通信使人们可以任何时间、任何地点和任何人进行通信，Internet 使人们可以获得丰富多彩的信息。那么如何把移动通信和 Internet 结合起来，达到可以任何人、任何地方都能联网呢？无线网络解决了这个问题。无线网络和个人通信网（PCN）^[1]代表了 21 世纪通信网络技术的发展方向。PCN 主要用于支持速率小于 56bps 的语音/数据通信，而无线网络主要用于传输速率大于 1Mbit/s 的局域网和室内数据通信，同时为未来多媒体应用（语音、数据和图像）提供了一种潜在的手段。计算机无线联网方式是有线联网方式的一种补充，它是在有线网的基础上发展起来的，使联网的计算机可以自由移动，能快速、方便地解决以有线方式不易实现的信道连接问题。然而，由于无线网络采用空间传播的电磁波作为信息的载体，因此与有线网络不同，辅以专业设备，任何人都有条件窃听或干扰信息，因此在无线网络中，网络安全是至关重要的。

1.1 无线网络技术的安全级别

无线网技术的安全性有以下 4 级定义^[2]：

作者简介：杨凯（1969—），男，讲师，学士；
郎士宁（1973—），女，讲师，硕士；
黄欢欢（1982—），女，助教，学士。

第一级，扩频、跳频无线传输技术本身使盗听者难以捕捉到有用的数据。

第二级，采取网络隔离及网络认证措施。

第三级，设置严密的用户口令及认证措施，防止非法用户入侵。

第四级，设置附加的第三方数据加密方案，即使信号被盗听也难以理解其中的内容。

无线网的站点上应使用口令控制，如 Novell NetWare 和 Microsoft NT 等网络操作系统和服务器的提供了包括口令管理在内的内建多级安全服务。口令应处于严格的控制之下并经常变更。假如用户的数据要求更高的安全性，要采用最高级别的网络整体加密技术，数据包中的数据发送到局域网之前要用软件加密或硬件的方法加密，只有那些拥有正确密钥的站点才可以恢复，读取这些数据。无线局域网还有些其他好的安全性。首先，无线接入点会过滤掉那些对相关无线站点而言毫无用处的网络数据，这就意味着大部分有线网络数据根本不会以电波的形式发射出去；其次，无线网的节点和接入点有个与环境有关的转发范围限制，这个范围一般很小。这使得窃听者必须处于节点或接入点附近。再次，无线用户具有流动性，可能在一次上网时间内由一个接入点移动至另一个接入点，与之对应，进行网络通信所使用的跳频序列也会发生变化，这使得窃听几乎无可能。无论是否有无线网段，大多数的局域网都必须要有有一定级别的安全措施。在内部好奇心、外部入侵和电线窃听面前，甚至有线网都显得很脆弱。没有人愿意冒险将局域网上的数据暴露于不速之客和恶意入侵之前。而且，如果用户的数据相当机密，比如银行网和军用网上的数据，那么，为了确保机密，必须采取特殊措施。

1.2 无线网络安全加密措施

常见的无线网络安全加密措施可以采用为以下几种。

(1) 服务区标示符 (SSID)

无线工作站必须出示正确的 SSID 才能访问 AP，因此可以认为 SSID 是一个简单的口令，从而提供一定的安全。如果配置 AP 向外广播其 SSID，那么安全程序将下降；由于一般情况下，用户自己配置客户端系统，所以很多人都知道该 SSID，很容易共享给非法用户。目前有的厂家支持“任何”SSID 方式，只要无线工作站在任何 AP 范围内，客户端都会自动连接到 AP，这将跳过 SSID 安全功能。

(2) 物理地址 (MAC) 过滤

每个无线工作站网卡都由唯一的物理地址标示，因此可以在 AP 中手工维护一组允许访问的 MAC 地址列表，实现物理地址过滤。物理地址过滤属于硬件认证，而不是用户认证。这种方式要求 AP 中的 MAC 地址列表必须随时更新，目前都是手工操作；如果用户增加，则扩展能力很差，因此只适合于小型网络规模。

(3) 连线对等保密 (WEP)

在链路层采用 RC4 对称加密技术，钥匙长 40 位，从而防止非授权用户的监听及非法用户的访问。用户的加密钥匙必须与 AP 的钥匙相同，并且一个服务区内的所有用户都共享一把钥匙。WEP 虽然通过加密提供网络的安全性，但也存在许多缺陷：一个用户丢失钥匙将使整个网络不安全；40 位加密钥匙在今天很容易破解；钥匙是静态的，并且要手工维护，扩展能力差。为了提供更高的安全性，802.11 提供了 WEP2，该技术与 WEP 类似。WEP2 采用 128 位加密钥匙，从而提供更高的安全。

(4) 虚拟专用网络 (VPN)

虚拟专用网络是指在一个公共 IP 网络平台上通过隧道，以及加密技术保证专用数据的网络安全性，目前许多企业及运营商已经采用 VPN 技术。VPN 可以替代连线对等保密解决方案及物理地址过滤解决方案。采用 VPN 技术的另一个好处是可以提供基于 Radius 的用户认证及计费。VPN 技术不属于 802.11 标准定义，因此它是一种增强性网络解决方案。

(5) 端口访问控制技术 (802.1x)

端口访问控制技术也是用于无线网络的一种增强性网络安全解决方案。当无线工作站 STA 与无线

访问点 AP 关联后，是否可以使用 AP 的服务要取决于 802.1x 的认证结果。如果认证通过，则 AP 为 STA 打开这个逻辑端口，否则不允许用户上网 802.1x。要求工作站安装 802.1x 客户端软件，无线访问点要内嵌 802.1x 认证代理，同时它作为 Radius 客户端，将用户的认证信息转发给 Radius 服务器。802.1x 除提供端口访问控制之外，还提供基于用户的认证系统及计费，特别适合于公共无线接入解决方案。

2 第三代数字蜂窝移动通信系统（3G）的安全

自 GSM 诞生以来，移动通信很快就成为当今社会不可缺少的一部分。出于质量和效益的考虑，移动通信的无线电波具有较强的穿透力，并向各个方向传播。因此，无线传输比有线传输更容易被窃听。20 世纪 80 年代的模拟移动通信系统（1G）深受其害，使用户利益受损。GSM 移动通信系统等 2G 系统在安全方面有了极大的改进，但是仍然存在一些缺陷。除语音通信的安全风险外，近年来引入的一些新技术如 GPRS 等，扩大了数据业务的范围，并允许移动用户接入公共网络资源和互联网。与语音业务相比，这些业务会遭受到更多类型的攻击。2G 系统考虑了一些安全因素，但绝大部分的安全规范是从运营商的角度设计的：防止欺骗和网络误用。这种处理方法不能提供可信的环境，不能给移动用户足够的信心开展电子商务和交换敏感信息。随着技术的成熟和移动数据业务的出现，用户比以前更加关注移动通信的安全问题。

在第三代移动通信系统（3G）中，除了传统的语音业务外，它还将提供多媒体业务、数据业务，以及电子商务、电子贸易、互联网服务等多种信息业务。因此，如何在第三代移动通信系统中保证业务信息，以及网络资源使用的安全性已成为 3G 系统中重要而迫切的问题。

2.1 3G 安全威胁

3G 系统的安全威胁大致可以分为如下几类：

（1）对敏感数据的非法获取，对系统信息的保密性进行攻击，主要包括：

- 窃听：攻击者对通信链路进行非法窃听，获取消息；
- 伪装：攻击者伪装合法身份，诱使用户或网络相信其身份合法，从而窃取系统信息；
- 业务分析：攻击者对链路中信息的时间、速率、长度、源及目的等信息进行分析，从而判断用户位置或了解正在进行的重要的商业交易；
- 浏览：攻击者对敏感信息的存储位置进行搜索；
- 泄露：攻击者利用合法接入进程获取敏感信息；
- 试探：攻击者通过向系统发送信号来观察系统反应。

（2）对敏感数据的非法操作，对信息的完整性进行攻击，主要包括对信息进行操作：攻击者故意地对信息进行篡改、插入、重放或删除。

（3）对网络服务的干扰或滥用，从而导致系统拒绝服务或导致系统服务质量的降低，主要包括：

- 干扰：攻击者通过阻塞用户业务、信令或控制数据使合法用户无法使用网络资源；
- 资源耗尽：用户或服务网络利用其特权非法获取非授权信息；
- 误用权限：用户或服务网络可以利用它们的权限来越权获得业务或信息；
- 服务滥用：攻击者通过滥用某些特定的系统服务，从而获取好处，或者导致系统崩溃。
- 拒绝：用户或网络拒绝发出响应。

（4）否认，主要指用户或网络否认曾经发生的动作。

（5）对服务的非法访问，主要包括：

- 攻击者伪造成网络和用户实体，对系统服务进行非法访问；
- 用户或网络通过滥用访问权利非法获取未授权服务。

2.2 针对 3G 的攻击方法

2.2.1 针对系统无线接口的攻击

针对 3G 系统无线接口的攻击方式主要有：

(1) 对非授权数据的非法获取，基本手段主要包括对用户业务的窃听、对信令与控制数据的窃听、伪装网络实体截取用户信息，以及对用户流量进行主动与被动分析。

(2) 对数据完整性的攻击，基本手段主要包括：对系统无线链路中传输的业务与信令、控制消息进行篡改，包括插入、修改、删除等。

(3) 拒绝服务攻击，拒绝服务攻击可分为以下三个不同层次：

- 物理层干扰：攻击者通过物理手段对系统无线链路进行干扰，从而使用户数据与信令无法传输，物理攻击的一个例子就是阻塞；
- 协议层干扰：攻击者通过诱使特定的协议失败流程干扰正常的通信；
- 伪装成网络实体拒绝服务：攻击者伪装成合法网络实体，对用户的服务请求拒绝回答。

(4) 对业务的非法访问攻击。攻击者伪装其他合法用户身份，非法访问网络，或切入用户与网络之间，进行中间攻击。

(5) 主动用户身份捕获攻击。攻击者伪装成服务网络，对目标用户发身份请求，从而捕获用户明文形式的永久身份信息。

(6) 对目标用户与攻击者之间的加密流程进行压制，使加密流程失效，基本的手段有：

- 攻击者伪装成服务网络，分别与用户和合法服务网络建立链路，转发交互信息，致使加密流程失效；
- 攻击者伪装成服务网络，通过发送适当的信令，致使加密流程失效；
- 攻击者通过篡改用户与服务网络间信令，使用户与网络的加密能力不匹配，致使加密流程失效。

2.2.2 针对系统核心网的攻击

针对系统核心网的攻击有如下几种：

(1) 对数据的非法获取。基本手段包括对用户业务、信令及控制数据的窃听，冒充网络实体截取用户业务及信令数据，对业务流量的被动分析，对系统数据存储实体的非法访问，以及在呼叫建立阶段伪装用户位置信息等。

(2) 对数据完整性的攻击。基本手段包括对用户业务与信令消息进行篡改，对下载到用户终端或 USIM 的应用程序及数据进行篡改，通过伪装成应用程序及数据的发起方篡改用户终端或 USIM 的行为，篡改系统存储实体中存储的用户数据等。

(3) 拒绝服务攻击。基本手段包括物理干扰、协议干扰、伪装成网络实体对用户请求做出拒绝回答、滥用紧急服务等。

(4) 否定。主要包括对费用的否定、对发送数据的否定、对接收数据的否定等。

(5) 对非授权业务的非法访问。基本手段包括伪装成用户、服务网络、归属网络，滥用特权非法访问非授权业务。

2.2.3 针对终端的攻击

主要是针对终端和 USIM 的攻击，包括：使用借来偷窃的终端和 USIM；对终端或 USIM 中数据进行篡改；对终端与 USIM 间的通信进行窃听；伪装身份截取终端与 USIM 间的交互信息；非法获取终端或 USIM 中存储的数据。

3 蓝牙的安全

无论是哪种无线网络，安全都是人们最关心的问题。某些设备很容易捕获到空中的无线电波，因

此在通过无线连接发送敏感信息时，需要采取适当的防范措施，以确保这些信号不会被截获。蓝牙技术同样如此：它是无线的，因此容易被窃听和远程访问，就像在网络不安全时无线上网容易受到攻击一样。只是对于蓝牙来说，它省时省力的自动连接特性同时也为不经接收者许可而向其发送数据提供了方便。

蓝牙提供了若干种安全模式，蓝牙装置采用何种安全模式由设备制造商确定。在几乎所有情况下，蓝牙用户都可以建立无须请求许可即可进行数据交换的“可信设备”。当其他设备试图与用户的手机建立连接时，用户必须决定是否允许建立连接。服务级别安全性和设备级别安全性共同保护蓝牙设备不受未授权数据传输的影响。安全手段包括授权和身份识别两个步骤，这两个步骤限制注册用户对于蓝牙服务的使用，并要求用户确定是否打开文件或者接受数据传输。只要用户的电话或其他设备中采用了这些措施，就不大可能发生未授权的访问。此外用户还可以将其蓝牙模式轻松切换至“不可发现”模式，彻底避免与其他蓝牙设备连接。如果用户使用蓝牙网络主要是为了同步家中的设备，则该功能可能是在公共场合避免安全漏洞的好方法。

虽然早期的手机病毒编写者曾利用蓝牙的自动连接功能发送感染病毒的文件，但自从大多数手机使用安全的蓝牙连接（即来自未知设备的数据在被接受之前需得到授权和鉴权）以来，感染病毒的文件一般都无法传输。病毒一旦进入用户的手机，用户必须决定是否打开和安装它。这样到目前为止，大多数手机病毒已得到控制，无法形成大的危害。但由于蓝牙自身存在的安全性问题，现在已经陆续出现了其他形式的入侵，如“蓝劫（Bluejacking）”、“蓝牙窃听”和“汽车偷听软件”。“蓝劫”指蓝牙用户向 10 米范围内的其他蓝牙用户发送名片（实际上只是文本信息）时，如果其他用户不知道信息的内容，可能会允许将该联系人添加到自己的通信簿中，这样该联系人就可以向该用户发送可以自动打开的信息，因为这些信息来源于已知的联系人。“蓝牙窃听”的危害更严重，黑客可以使用它在用户毫无察觉的情况下远程访问用户的手机，以及使用手机的功能，包括拨打电话和发送文本信息。“汽车偷听软件”是一种软件，黑客使用它可以从蓝牙汽车立体声系统中发送和接收音频信息。正如计算机的安全漏洞一样，这些脆弱性在技术创新过程中是不可避免的，针对出现的新问题，设备制造商也在不断发布固件升级。

4 结束语

无线网络技术、第三代移动通信技术及蓝牙等无线通信技术以其便利的安装、使用，高速赢得了用户的青睐，随着应用范围和新业务的逐渐扩展，对安全性的要求会不断增加，需要不断地探索新方法、新技术，以满足更高等级的安全需求。

参考文献

[1] 毛京丽等编著. 现代通信新技术. 北京：北京邮电大学出版社，2008.
[2] 邱天爽等编译. 无线网络与移动通信的资源、移动与安全管理. 北京：电子工业出版社，2008.

储粮害虫图像识别知识库研究

王利强¹, 张红梅¹

(1.河南工业大学信息学院, 河南 郑州, 450001)

摘要: 本文在知识模型及知识库相关理论和技术的基础上, 结合本体论, 构建了一个可以被各专家系统所共享的储粮害虫图像识别运行知识库。

关键词: 储粮害虫; 图像识别; 本体; 知识库

中图分类号: TP

文献标识码: A

文章编号: 1006-7043 (2004) xx-xxxx-x

Study on Image Recognition of Stored-grain Pests Knowledge Base

WANG Liqiang¹, ZHANG Hongmei¹

(1. School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, Henan China)

Abstract: Base on the correlation theories and technologies of knowledge model and knowledge base, this article which combined ontology construct image stored-grain pests' knowledge base which can be shared by the expert system.

Keywords: stored-grain pests; image recognition; ontology; knowledge base

根据资料显示, 全世界每年至少有 5% 的粮食被害虫糟蹋, 如果人力、物力和技术跟不上, 可能会达到 20%~30%。众所周知, 我国是世界上最主要的产粮、储粮大国, 更是消费大国, 搞好粮食储藏是关系到国计民生的大事。近年来, 我国的粮食生产形势十分喜人, 年产量已达 5000 亿公斤, 库存量高达年产量的一半以上。为了确保粮食的安全储藏, 每年国家用于粮食储备方面的补贴费用就有数百亿元, 但仍有不少粮食因管理决策不善等原因而遭受损失, 其中储粮害虫的危害不容忽视。目前, 国内外主要利用化学药剂防治储粮害虫, 挽回了大量的虫害损失。化学防治虽然具有防治效果好、作用快及防治费用较低等优点, 但其对环境的污染、对非防治对象的毒害及害虫产生的抗药性等副作用很难避免。因此, 在储粮害虫防治中避免使用化学药剂, 搞好害虫监测预报, 做好基础防治, 是“以防为主, 综合防治”的害虫防治方针的具体体现, 也是探索储粮害虫防治新方法的一项紧迫任务。

储粮害虫检测的目的是及时掌握储藏物中害虫发生的种类等, 为防治决策提供科学的依据。随着科技的发展, 机器视觉、数字图像处理及模式识别技术得到了前所未有的发展与应用, 利用计算机技术、机器视觉、图像处理与模式识别技术相结合实现储粮害虫检测^[1]。储粮害虫图像的识别是实现储粮害虫防治技术的重要环节, 虽然在储粮害虫图像的识别研究中已经取得了一定成果, 但在识别的过程中, 大量的可识别信息常以图像形式来记录, 如何将这些图像信息转变为计算机能够识别的信息, 并从这些信息中获取知识, 无疑是非常重要的。本文将本体的概念引入储粮害虫图像识别领域知识表达, 构建了一个可以被各专家系统所共享的储粮害虫图像识别运行知识库。

1 本体的概念

本体的概念起源于哲学领域, 用于描述事物的本质, 是对客观存在的系统的解释或说明。在近

一二十年来，本体被计算机领域所采用，用于知识表达、知识共享及重用。许多学科和研究均在使用本体这个术语，存在不同的定义。研究者们普遍接受的呈现高引用率的本体定义是 T.Gruber 于 1993 年提出的：“本体是对共享的概念化进行形式的显式规范说明^[2]。”其中：概念化是现实世界中现象的抽象模型，明确标志与现象相关的概念；显式指概念的类型及概念在使用中的约束应该明确地定义出来；形式的意思即本体应该是机器可读的；共享是反映本体中的知识是中立的、一致认可的^[3]。

Perez 等^[4]提出 5 个基本的建模元，即本体的 5 种组成元素：

(1) 概念：指任何事务，如工作描述、功能、行为、策略和推理过程，这些概念通常构成一个分类层次。从语义上讲，它表示的是对象的集合，其定义一般采用框架结构，包括概念的名称，与其他概念之间的关系的集合，以及用自然语言对概念的描述。

(2) 关系：在领域中概念之间的一类关联，形式上定义为 n 维笛卡儿积的子集： $R: C_1 \times C_2 \times \dots \times C_n$ 。在语义上关系对应于对象元组的集合。基本的关系有四种：**part-of** 表达概念之间部分与整体的关系；**kind-of** 表达概念之间的继承关系；**instance-of** 表达概念的实例与概念之间的关系；**attribute-of** 表达某个概念是另一个概念的属性。在实际建模过程中，可以根据领域的具体情况定义相应概念之间的关系。Relations:

(3) 函数：一类特殊的关系。该关系的前 $n-1$ 个元素可以唯一决定第 n 个元素。形式化的定义 $F: C_1 \times C_2 \times \dots \times C_{n-1} \rightarrow C_n$ 。如 **mother-of** 就是一个函数，**mother-of**(x,y)表示 y 是 x 的母亲。

(4) 公理：代表永真断言，如概念乙属于概念甲的范围。

(5) 实例：代表元素，从语义上讲实例表示的就是对象。

本体的目标是捕获相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些术语和词汇间相互关系的明确定义。本体是语义级的知识表示，实现了不同应用之间的翻译和互操作。利用本体表示图像识别领域中的专业知识，可以统一术语和概念，实现知识共享。同时，本体可以重用，从而避免知识表达的重复。

2 知识库

知识是人类对客观世界的认识，通常知识是先由底层数据经过分类、归纳、综合等处理过程而得到的上层信息，这种信息再经过解释、比较、推理得到我们所获取的知识。这种过程主要是在语义的层面来进行的。抽象地说，知识是由有名论域内容和有名论域内容之间关联的符号来表示的。

知识库是关于某一领域的陈述性知识、过程性知识和策略性知识的集合。在该集合中各类知识通过一定的表示方法表示，并建立相互之间的联系。知识库中不但包含了大量的简单事实，还包含了规则、过程性知识和策略性知识。从存储知识的角度来看，以描述型方法来存储和管理知识的机构称为做知识库。从使用知识的角度来看，知识库是由知识和知识处理机构组成的。

知识库是知识工程中结构化、易操作、易利用、全面有组织的知识集群，是针对某一（或某些）领域问题求解的需要，采用某种（或若干）知识表示方式在计算机存储器中存储、组织、管理和使用的互相联系的知识片集合。这些知识片包括与领域相关的理论知识、事实数据，由专家经验得到的启发式知识，如某领域内有关的定义、定理和运算法则及常识性知识等。本体知识库系统结构如图 1 所示^[5,6]。

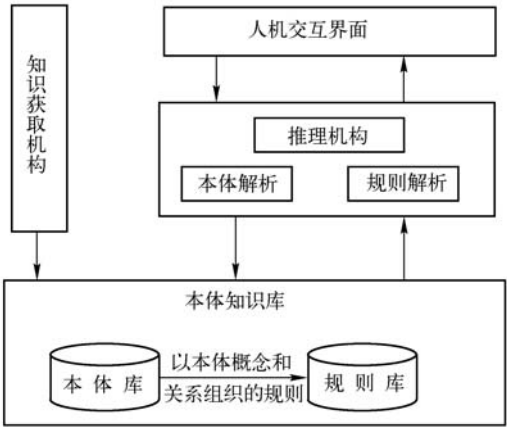


图 1 本体知识库系统结构

3 基于本体的知识库的构建

3.1 单位的书写规则

图像识别系统^[7-9]主要分为三个模块，即图像处理、图像识别和图像理解，如图 2 所示。

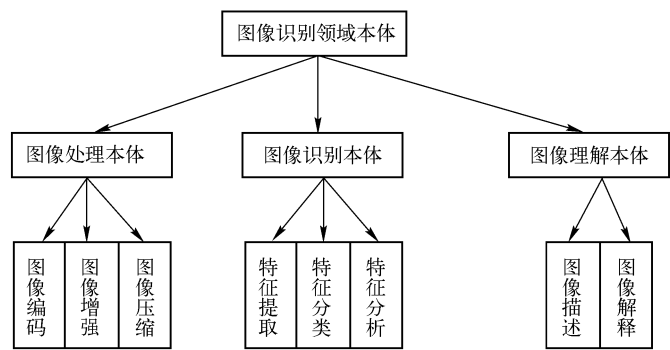


图 2 图像识别本体模型

3.2 本体构建

3.2.1 储粮害虫图像处理本体

在研究储粮害虫图像时，首先采集图像，然后对获得的图像信息进行预处理，主要工作是滤去干扰和噪声等。这样可提高信噪比，有时由于信息微弱，无法辨识，还得进行增强。增强是为了提供一个满足一定要求的图像，或对图像进行变换，以便人、机分析。为了从图像中找到需要识别的东西，还得对图像进行分割，即定位和分离。为了给观察者以清晰的图像，还要对已经退化了的图像加以重建或恢复，即进行复原处理，以便改进图像的保真度。在实际处理中，由于图像信息量非常大，在存储及传送时，还要对图像信息进行压缩。为了能让计算机处理，还要进行编码等工作。编码的作用，是用最少数量的编码位，表示单色和彩色图像，以便更有效地传输和存储。对图像处理环节来说，输入是储粮害虫图像，输出也是储粮害虫图像，也就是处理后的储粮害虫图像。

储粮害虫图像处理本体概念集包括图像编码、图像增强、图像压缩、图像复原、图像分割等概念。属性集包括储粮害虫亮度和色度信息、储粮害虫形状信息、储粮害虫纹理信息、储粮害虫尺寸信息等。

3.2.2 储粮害虫图像识别本体

图像识别是对处理后的图像进行特征提取、匹配和分类，确定类别名称。特征提取时是在分割的基础上进行的，并对某些参数进行测量，再提取这些特征；最后根据测量结果做分类。为了更好地识别图像，还要对整个图像作结构上的分析，并对图像做出一个详细的描述，以便对图像的主要信息得到一个解释和理解，并通过许多对象相互间的结构关系对图像加深理解，从而提高识别的成功率。所以图像识别是在上述分割后的每个部分中，找出它的形状及纹理等特征，即特征抽取，以便对图像进行分类，并对整个图像做结构上的分析。因而对图像识别环节来说，输入是经过预处理的图像，输出是类别和图像的结构分析，而结构分析的结果则是对图像做描述，以便对图像的重要信息得到一种理解和解释。

储粮害虫图像识别本体概念集包括储粮害虫图像特征提取、储粮害虫图像特征分类和储粮害虫图像特征分析。属性集包括参数测量结果。

3.2.3 储粮害虫图像理解本体

图像理解是一个总称。图像做描述和解释是图像处理及图像识别的最终目的，以最终理解它是什么图像。所以它是在图像处理及图像识别的基础上，再根据分类做结构句法分析，去描述图像和解释图像。对理解部分来说，输入是储粮害虫图像，输出则是储粮害虫图像的描述与解释。

储粮害虫图像理解本体概念集包括描述图像和解释图像。属性集包括图像理解结果，比如此图像是否为知识库中所包含。

4 结论

本文介绍了基于本体的储粮害虫图像识别知识库的实现，重点论述了运用本体论的思想，组织和实现本体模型。知识库是智能系统的核心部件，构建本体知识库是一个非常艰难的过程，本体模型需要不断建立、修改、测试，需要不断进行本体模型表示的知识的一致性、完备性、冗余性检验。从目前已经开发出的本体知识库来看，系统需要进一步扩展知识库规模，扩大知识服务的领域和范围，是我们不断努力的方向。

参考文献

[1] 胡丽华, 郭敏, 张景虎等. 储粮害虫检测新技术及应用现状. 农业工程学报, 2007, 23(11): 286-289.
Hu Lihua, Guo Min, and Zhang Jinghu. New detection technology and application status of stored-grain insects. Transactions of the Chinese Society of Agricultural Engineering, 2007,23(11):286-289.

[2] Gruber T R.A translation approach to portable ontology specifications [J]. Knowledge Acquisition,1993,5(2):199-220.

[3] Rudi Studer, V Richard Benjamins, Dieter Fensel. Knowledge Engineering: Principles and Methods [J].IEEE Transactions on Data and Knowledge Engineering, 1998(25):161-197.

[4] Perez AG,Benjamins VR.Overview of knowledge sharing and reuse components:ontologies and problem solving methods[D].IJCA,1999.

[5] 闫洪森, 张野, 孙娜等. 基于本体的知识库构建方法. 情报科学, 2007, 23(11): 1398-1399.
Yan Hongsen, Zhang Ye, and Sun Na. Construction Method of Knowledge Database Based on Ontology. Information Science, 2007,23(11):1398-1399.

[6] 刘成亮, 李涵. 本体知识库系统研究. 电脑知识与技术, 2008, 18(33): 1646-1648.
Liu Chengliang and Li Han. Research on Knowledge Base System Based on Ontology. Computer Knowledge and Technology, 2008,18(33):1646-1648.

[7] 蒋伟, 胡学刚. 基于对数图像处理和二阶微分的图像增强新模型. 西南大学学报(自然科学版), 2009, 31(9):142-146
Jiang Wei, Hu Xuegang. A New Image Enhancement Model Based on Logarithmic Image Processing and a Second-Order Differential. Journal of Southwest University(Natural Science Edition), 2009, 31(9):142-146.

[8] 刘飒, 邱宏. 用图像处理技术计算薄膜厚度和表面粗糙度. 信息技术, 2009, 9:117-120.
Liu Sa, Qiu Hong. Application of digital image processing in calculation of the film thickness and surface roughness. Information Technology, 2009, 9:117-120.

[9] 马晶, 雷勇, 涂国强等. 基于 DSP 的图像处理在车牌识别中的应用. 微计算机信息, 2009, 25(10):133-135.
Ma Jing, Lei Yong, Tu Guoqiang. The Application of The Image Process Based on DSP in the LPR. Microcomputer Information, 2009, 25(10):133-135.

一种有效的 SAR 图像角反射器检测方法

薛笑荣¹, 王爱民¹, 曾琪明²¹

(1. 安阳师范学院计算机与信息工程学院 河南安阳 455002;

2. 北京大学遥感与地理信息系统研究所 北京 100871)

摘 要: 在用 InSAR (干涉雷达测量) 技术进行地表变形检测时, 往往存在一些因素影响着检测精度, 解决该问题很重要的一个途径是设置人工角反射器, 或者引进所谓永久散射体的概念。而要利用角反射器, 就必须研究角反射器检测技术。本文根据 SAR 图像角反射器的特点对角反射器点目标图像进行模拟, 并结合模式匹配方法, 提出了一种有效的 SAR(合成孔径雷达)图像角反射器点检测方法, 实验结果表明该角反射器点检测方法是有效的。

关键词: SAR; InSAR; 角反射器检测

An Effective Method of SAR Image Corner Reflector Detection

XUE Xiao-rong¹, WANG Aimin¹, ZENG Qiming²

(1.The College of Computer Science and Technology, Anyang Normal University, Anyang 455002, Henan China;

2.The Institute of Remote Sensing and GIS,eking University, Beijing 100871, China.)

Abstract: InSAR(Synthetic aperture radar interferometry) technique has been used widely to examine deformation of the earth's surface, however, there are often some factors of affecting examination accuracy. One of the good ways of solving the upper problems is to put up artificial corner reflectors, or to introduce concept of permanent scatterers. However, to use corner reflectors, corner reflector detection method must be studied. In the paper, based on the characteristic of SAR image corner reflectors, the point target image of corner reflector is simulated, and pattern matching method is combined, an efficient method of corner reflector detection for SAR Image is proposed. The experiment results show that the new method of corner reflector detection is efficient.

Keywords: SAR; InSAR; corner reflector detection

滑波、泥石流等也是严重的地质灾害。用 GPS 或在地测量技术在地面站设点观测可以进行监测。D-INSAR 技术的出现引起了用卫星遥感手段监测的兴趣。但是, 这方面的实际应用尚有些困难。例如, 滑坡体的面积一般不是很大, 目前的 SAR 影像的分辨率在许多情况下还是不够的。潜在的滑坡山体往往有植被覆盖, 对于干涉对的相干性有很大的影响。解决的途径一个是有待传感器的改进, 提高空间分辨率等; 另一个是人工角反射器(如图 1、图 2 所示), 或者引进所谓永久散射体(是指在相当长的时间内仍然保持稳定反射特性的散射体)的概念。初步研究发现, 在各种地物条件下, 总存在一些强散射体在长时间间隔内保持较好的相干性, 可以等效为人工布设的角反射器。在这些点上实施差分干涉测量来检测滑坡体的微小位移, 点目标的空间分辨率可达到像元。而要利用角反射器检测微小位移, 就必须研究有效的角反射器点检测方法^[1~7]。

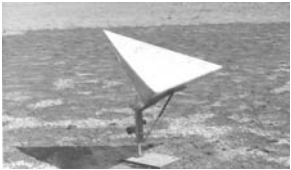


图 1 人工角反射器 1



图 2 人工角反射器 2

基金项目: 国家 863 计划项目 (2006AA12Z150)、为新基金项目 (40001015200906)。

作者简介: 薛笑荣 (1974—), 男, 博士。

1 角反射器点检测方法主要步骤

本文提出了一种有效的角反射器点检测方法，其主要步骤为：

- (1) 角反射器模拟图像生成。
- (2) 根据一些已知其经度、纬度及在 SAR 图像对应的纵横坐标来建立实际点的经度、纬度及其在 SAR 图像上对应的纵横坐标之间的近似关系式。
- (3) 角反射器点在 SAR 图像中的初步估计。
- (4) 实际角反射器点检测。

1.1 角反射器模拟图像生成

1) 角反射器强度特点

在遥感 ERS 中，雷达分辨率为 25m×25m，而角反射器的回波信号可看做为点散射或脉冲响应^[6~7]：

$$h(x,t;R)=e^{-j\frac{4\pi R}{\lambda}}\cdot\tau_p e^{-j\pi kt^2}\frac{\sin\left(\pi\frac{t}{\rho_r}\right)}{\pi\frac{t}{\rho_r}}\cdot L_s e^{j\frac{2\pi}{\lambda R}x^2}\frac{\sin\left(\pi\frac{x}{\rho_a}\right)}{\pi\frac{x}{\rho_a}} \tag{1}$$

式中， x 为方位角 (azimuth coordinate)， t 为斜距时间 (slant range time)， R 为目标和卫星轨道之间的最小距离 (minimum distance of object and satellite trace)； L 为合成孔径 (synthetic aperture)； τ_p 为脉冲长度 (impulse length)； k 为斜距率 (frequency rate)； B 为载波宽度 (bandwidth of carrier wave)； D 为天线长度 (antenna length)； ρ_r 为斜距分辨率 (slant range resolution)； ρ_a 为方位分辨率 (azimuth resolution)。

角反射器的回波信号反应包络面可表示为：

$$E_{nv}[h(x,t)]=\tau_p L_s \left| \sin c\left(\pi\frac{t}{\rho_r}\right) \sin c\left(\pi\frac{x}{\rho_a}\right) \right| \tag{2}$$

在斜距离向和方位向都呈现出 sinc 函数形状

$$E_{nv}[h(x,t)]=\tau_p L_s \left| \sin c\left(\pi\frac{t}{\rho_r}\right) \sin c\left(\pi\frac{x}{\rho_a}\right) \right|$$
$$\rho_r=\frac{1}{B}, \quad t=\frac{i}{B_{sample}}, \quad i=0,\pm1,\pm2,\cdots$$

(B_{sample} ：斜距采样率)

$$x=\frac{LineSpacing}{LS} * j, j=0,\pm1,\pm2,\cdots \rho_a=\frac{V_B}{V_{s/c}} * \frac{\lambda}{2\beta_A}=\frac{V_B}{V_{s/c}} * \frac{D}{2k} \quad (\beta_A=k\frac{\lambda}{D})$$
$$\frac{V_B}{V_{s/c}}=\frac{R_E \cos \vartheta_E}{h+R_E}=\frac{1}{1+a} \cos\left(\sin^{-1}\left(\frac{R/R_E}{1+a} \sin \theta_i\right)\right), \quad a=\frac{h}{R_E}$$

2) ASAR (Advanced Synthetic Aperture Radar) 的一些相关参数

倾斜角(Incidence Angle Center)(degree)=22.780418

行间距(Line_spacing)(m)=4.01244310

脉冲重复频率(Pulse Repetition Frequency) (PRF)=1652.443237

斜距离采样率(Range Sampling Rate)(Hz)=19208000.000000

方位向单视带宽(Bandwidth Perlook in Azimuth) (m)=1360.562744

斜距向单视带宽(Bandwidth Perlook in Range) (m)=16000000.000000

行间距离(Line_spacing)(m)=4.01244310

像元间距(pixel_spacing_range)(m) = 7.80384366

3) ASAR 角反射器图像模拟结果（见图 3）

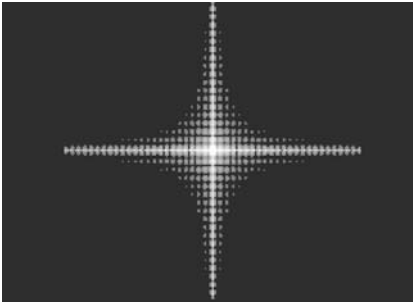


图 3 人工角反射器模拟图像

1.2 角反射器图像坐标近似计算关系式

根据已知其经度、纬度及在 SAR 图像对应的纵横坐标来建立实际点的经度、纬度及其在 SAR 图像上对应的纵横坐标之间的近似关系式。我们采用的拟合关系式如下：

$$\begin{cases} X = A_0 + A_1x + A_2y + A_3xy \\ Y = B_0 + B_1x + B_2y + B_3xy \end{cases} \tag{3}$$

式中，x 与 y 为目标点的经纬度，X 与 Y 为相应的图像坐标。

1.3 角反射器点检测采用的方法

在图 4 所示的 SAR 图像中用模拟的图像做模式图像，采用模式匹配方法以检测到较为准确的图像坐标，所采用的模拟匹配方法算法如下：

设模式 $X(x_1, x_2, \dots, x_n)$, $Y(y_1, y_2, \dots, y_n)$

$$DX = \sum_{i=1}^n x_i * x_i, \quad DY = \sum_{i=1}^n y_i * y_i, \quad DXY = \sum_{i=1}^n x_i * y_i, \quad \text{Correlation} = DXY / \sqrt{DX * DY}$$

如果 Correlation 在允许的范围，则认为模式 X、Y 相似。

2 实验

在本文中，采用了湖北新滩的相关 SAR 图像做实验进行角反射器点检测。下面的图像为在新滩地区拍摄的一幅 SAR 图像，如图 4 所示。

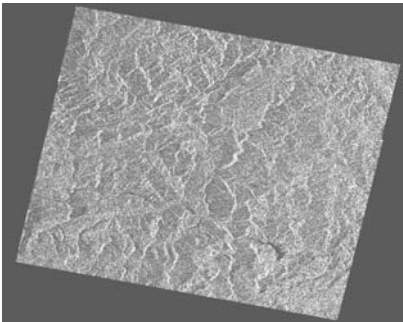


图 4 新滩角反射器的相关 SAR 图像

该图像的四个角点的经度、纬度及相应的图像坐标如表 1 所示。

表 1 新滩地区 SAR 图像一些角反射器点的经度、纬度值

点	经 度	纬 度	横 坐 标	纵 坐 标
左上角点 1	10.277886	31.667601	0	0
右上角点 1	11.459464	31.456620	0	5619
左下角点 1	10.071713	30.807246	24063	0
右下角点 1	11.241358	30.596629	24063	5619

新滩地区的一些角反射器点坐标表 2 所示。

表 2 新滩地区一些角反射器点的经度、纬度值

宜昌角反射器站点分布（宜昌岩崩所安装测量，涂国保）				
编号	地 点	北 纬	东 经	高程（m）
1	八字门滑坡（BZM） 30	° 58'20" 11	0° 45'32" 213	
2	张家湾滑坡（ZJU） 30	° 57'01" 11	0° 44'23" 176	
3	黄岩（HY） 30	° 56'45" 11	0° 47'38" 244	
4	猴子岭斜坡变形体（HZL） 30	° 56'23" 11	0° 47'39" 245	
5	新滩滑坡（XT） 30	° 56'37" 11	0° 48'20" 227	
6	上孝仁村滑坡（SX） 30	° 55'57" 11	0° 48'00" 247	
7	前山坡南滑坡（QSP） 30	° 55'46" 11	0° 49'25" 217	
8	路口子变形体（LKZ） 30	° 54'44" 11	0° 49'15" 235	
9	聚集坊大桥桥头崩塌危岩体（JJF）	30° 53'10" 11	0° 50'19" 334	
10	野猫面滑坡（YMM） 30	° 53'29" 11	0° 51'32" 237	
max		30° 59' 11	0° 52'	
min		30° 53' 11	0° 44'	

采用拟合关系式

$$\begin{cases} X = A_0 + A_1x + A_2y + A_3xy \\ Y = B_0 + B_1x + B_2y + B_3xy \end{cases}$$

同时选择 X、Y 的一些已知数据：

$$\begin{aligned} \mathbf{X} &= [0 \ 0 \ 24063 \ 24063]; \\ \mathbf{Y} &= [0 \ 5619 \ 0 \ 5619 \quad \quad \quad]; \end{aligned}$$

由已知数据可求解得

$$\begin{aligned} \mathbf{AS} &= [A_0 \ A_1 \ A_2 \ A_3] \quad ' = 1.0\text{e}+006 * [1.5479 \quad -0.0322 \quad -0.0063 \ 0.0000] \quad ' \\ \mathbf{BS} &= [B_0 \ B_1 \ B_2 \ B_3] \quad ' = 1.0\text{e}+005 * [-6.6538 \ 0.0518 \ 0.0635 \quad -0.0006] \quad ' \end{aligned}$$

将实验所采用的角反射器点的经度和纬度代入近似关系式，可得到它们相应的近似纵、横坐标：第一、二、三、四、五、六、七、八、九、十个角反射器点近似纵横坐标分别为：（16331，2972）、（17011，2909）、（16870，3162）、（17032，3171）、（16873，3218）、（17198，3205）、（17166，3317）、（17640，3324）、（18254，3435）、（18015，3522）

在上述要检测的图像中，以所要检测到角反射器的近似图像坐标为中心，取 201×201 图像，然后在这小图像中进行较为精确的坐标检测，具体来说，就是以前述 201×201 图像中的每一点为中心取 101×101 图像，将这 101×101 图像与模拟图像进行匹配，根据上述 201×201 图像中各点和模式图像匹配的情况，挑出最相近的点作为角反射器较为精确的图像坐标。所检测到各个角反射器在 SAR 图像中的坐标为：

第一、二、三、四、五、六、七、八、九、十个角反射器点检测所得到的图像坐标分别为：

(16431,3072)、(16920,2947)、(16774,3112)、(16972,3270)、(16774,3118)、(17105,3305)、(17247,3417)、(17540,3424)、(18347,3535)、(18086,3555), 检测到的角反射器如图 5 所示的小白圆圈标注。

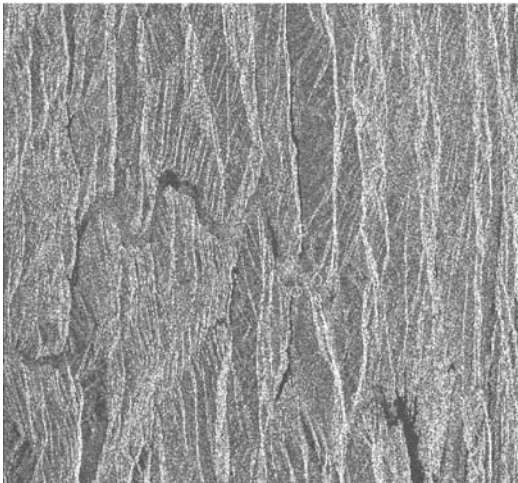


图 5 新滩地区的角反射器标注

3 结论

通过对万州地区和新滩地区的 SAR 图像用我们的算法进行角反射器点检测, 按照所检测角反射器在 SAR 图像中的坐标, 找出角反射器点在 SAR 图像中的对应位置, 并和用角反射器点经纬度所找到角反射器的对应位置相比照, 发现本文的角反射器检测算法是有效的。

参考文献

[1] Li F, Goldstein R M.S tudies of multibaseline spaceborne interferometric: synthetic aper ture radars[J].IEEE Transactions on Geoscience and Remote Sensing, 1990,28(1):88-97.

[2] Gray A L, Farris-Manning P J.Repeat-passinterferometry with an airborne synthe tic aperture radar[J].IEEE T ransactions on Geoscience and Remote Sensing, 1993,31:180-191.

[3] 廖明生, 林琚著. 雷达干涉测量—原理与信号处理基础, 测绘出版社[M], 2003.

[4] 刘智, 王超, 张红著. 星载合成孔径雷达干涉测量,科学出版社[M], 2002.

[5] [http://topex.ucsd.edu/SAR/proposals/reflector .html](http://topex.ucsd.edu/SAR/proposals/reflector.html).

[6] Ye Xia; Kaufmann, H., Xiaofang Gu o. Differential SAR interferometry using corner reflectors, IGARSS '02,2002, Vol(2): 1243-1246.

[7] Sakurai-Amano, T.; Kobayashi, S.; Fujii, N; Detection of singular corner reflectors in residential and mountainous areas from SAR images, IGARSS '99 Proceedings,1999, Vol(2): 1454-1456.

一种基于 Canny 检测算子的图像分割算法

明 生， 邬长安， 马 珂

(信阳师范学院计算机与信息技术学院， 河南 信阳， 464000)

摘 要：本文提出了一种基于 Canny 检测算子的图像分割算法。该算法首先对图像进行 Canny 算子边缘检测，然后采用粒子群优化方法寻找基于最大二维信息熵的最佳分割阈值，并进行图像分割。实验结果表明，该算法能有效抑制噪声图像边缘检测的不连续性，提高了图像分割的准确率。

关键词：图像分割；粒子群算法；Canny 算子；二维信息熵

中图分类号：TP393.08 **文献标识码：**A **文章编号：**1006-7043 (2004) xx-xxxx-x

An Image Segmentation Algorithm Based on Canny Edge Detection

MING Sheng, WU Chang'an, MA Ke

(College of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, Henan China)

Abstract: This paper presented an image segmentation algorithm based on Canny edge detection. This algorithm detects the image by Canny edge detection, and searches for the best threshold with the algorithm of maximal value of two dimensions entropy, finally carries out image segmentation. Experimental results show that the algorithm restrained the non-continuity of the noise image edge detection, and increased the accuracy of image segmentation.

Keywords: image segmentation; particle swarm optimization; canny Algorithm; two dimension entropy

1 引言

图像分割是数字图像处理研究的重要领域之一，其在目标识别、景物分析、图像理解、计算机视觉等众多领域有着重要作用，也是计算机视觉领域低层次视觉中的主要问题^[1]。

粒子群优化 (PSO) 算法是由 Kennedy 和 Eberhart 于 1995 年提出的一种基于智能的演化搜索技术^[2, 3]。该算法既考虑了当前粒子的搜索信息，又考虑了粒子群全局搜索信息，由于其具有独特的搜索性能和全局优化能力，被广泛应用于函数优化、神经网络、模糊系统控制、数据聚类等领域^[2, 3]。

边缘检测的目的是要检测图像局部特征值（如灰度）不连续或变化较为剧烈的像素点，然后将这些点连接就构成目标的边界。为了检测出边缘信息,通常是利用其周围像素灰度或颜色有阶跃性变化或屋顶变化的特性判断该像素是否为边缘点。Canny 边缘检测算子是 Canny 于 1986 年提出了基于最优化算法的边缘检测算子。Canny 算子提取的图像边缘时速度较慢，给定的两阈值不能较好地定位真假边缘。本文提出的一种改进的 Canny 边缘检测算子能够自适应地、快速地寻找两个最优阈值提取图像边缘。

2 图像最大二维信息熵

对灰度范围为 $\{0,1,\cdots,l-1\}$ 的图像直方图，以其中各像素及其 8 邻域的 8 个像素为一个区域，计算出其灰度均值图像。设 $n_{i,j}$ 为图像中点灰度 i 及其区域灰度均值为 j 的像素个数， $p_{i,j}$ 为点灰度—区

基金项目：河南省教育厅自然科学项目（2009A520022）
作者简介：明生（1981—），男，硕士，主要研究方向为数字图像处理；
邬长安（1959—），男，教授，主要研究方向为模式识别、数字图像处理；
马珂（1984—），男，硕士研究生，主要研究方向为数字图像处理。

域灰度均值对 (i, j) 发生的概率，则

$$p_{i,j} = \frac{n_{i,j}}{N \times N} \tag{1}$$

其中， $N \times N$ 为图像的大小。

假设阈值设在 (s, t) ，目标区域记为 O ，背景区域记为 B ，其他边缘区域是关于噪声信息，将其忽略不计，则目标和背景区域的概率 $p_{i,j}$ 分别为：

$$P_O = \sum_i \sum_j p_{i,j} \quad i = 1, 2, \dots, s, j = 1, 2, \dots, t \tag{2}$$

$$P_B = \sum_i \sum_j p_{i,j} \quad i = s+1, s+2, \dots, l, j = t+1, t+2, \dots, l \tag{3}$$

对于数字图像中的离散二维熵为：

$$H = - \sum_i \sum_j p_{i,j} \lg p_{i,j}$$

目标和背景区域二维熵分别为：

$$\begin{aligned} H(O) &= - \sum_i \sum_j (p_{i,j}/P_O) \lg (p_{i,j}/P_O) \\ &= -(1/P_O) \sum_i \sum_j (p_{i,j} \lg p_{i,j} - p_{i,j} \lg P_O) \\ &= (1/P_O) \lg P_O \sum_i \sum_j p_{i,j} - (1/P_O) \lg P_O \sum_i \sum_j p_{i,j} \lg p_{i,j} \\ &= \lg P_O + H_O / P_O \end{aligned} \tag{4}$$

其中

$$H_O = - \sum_i \sum_j p_{i,j} \lg p_{i,j} \quad i = 1, 2, \dots, s, j = 1, 2, \dots, t$$

同理可得：

$$H(B) = \lg P_B + H_B / P_B \tag{5}$$

其中

$$\begin{aligned} H_B &= - \sum_i \sum_j p_{i,j} \lg p_{i,j} \quad i = s+1, s+2, \dots, l, j = t+1, t+2, \dots, l \\ P_B &= 1 - P_O \quad H_B = H_i - P_O \end{aligned}$$

3 一种 canny 边缘检测的图像分割算法

基于粒子群优化的最大二维信息熵及 Canny 边缘检测图像分割算法是在经过 Canny 算子检测边缘之后，存在许多由噪声和纹理引起的假边缘^[4,5]，在这里采用基于粒子群优化的最大二维信息熵阈值分割去除假边缘。具体算法如下：

步骤 1 载入图像 $N(x, y)$ ，图像大小为 $M \times N$ ，用高斯滤波器对图像滤波，得到消除噪声后的图

像 $f(x, y)$ 。例如，方差为 1.4 的高斯函数近似模板可为：

$$\begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix}$$

但高斯滤波器的方差和模板大小应该根据图像中所含的目标情况来适当选择。

步骤 2 对滤波后图像中的每个像素，计算其梯度幅值 A 和方向角 ϕ ，采用水平模板

$$P = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \text{ 和垂直模板 } Q = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \text{ 作为计算 } x \text{ 方向和 } y \text{ 方向梯度分量的计算模板。}$$

则图像 $f(x,y)$ 上点 (i,j) 处的梯度幅值 $A(i,j) = \sqrt{P^2(i,j) + Q^2(i,j)}$ ，方向角 $\phi(i,j) = \arctan(Q(i,j)/P(i,j))$ 。

步骤 3 对梯度幅值 $A(i,j)$ 进行非极大值抑制。

步骤 4 用基于粒子群优化的最大二维信息熵阈值算法检测。

在这里采用基于粒子群优化的最大二维信息熵二阈值算法处理。将会得到两个最优阈值 τ_1 和 τ_2 ，且 $\tau_2 > \tau_1$ 。用这两个阈值处理边缘图像 $f(i,j)$ ，把梯度值小于 τ_1 的像素的灰度设为 0，得到边缘图像 T_1 ，同理，得到 T_2 。把图像 T_2 作为基础，以 T_1 为补充连接图像的边缘。

步骤 5 边缘连接。

首先在图像 T_2 中扫描，一旦遇到一个非零灰度的像素 R ，跟踪以 R 为始点的轮廓线，直到该线的终点 S 。接着在图像 T_1 中比较与图像 T_2 中 S 点位置对应的 S' 的 8-邻域。如果在 S' 点的 8-邻域有非零像素 U' 存在，则将其包括到图像 T_2 中，作为 U 。同理，重复在 T_2 中继续寻找跟踪以 U 为开始点的轮廓线。这样循环直到在图像 T_1 中和 T_2 中都无法继续为止。包含 R 的轮廓线的连接已经完成，可标记为已经访问过。然后依次可以重复寻找图像中的每一条边缘线，直到在图像 T_2 中再也找不到新的轮廓线为止。

4 实验结果分析与比较

本试验在 VC 环境中进行，种群规模设为 30，最大迭代代数设为 100，粒子群优化算法的参数设置为 $c_1=c_2=2$ ，这里设加权因子 $w_{\max}=1.0$ ， $w_{\min}=0.6$ ，进行试验，首先选取了如图 1 所示的一幅不含噪声的原始图像，经过基于粒子群优化的最大二维信息熵及 Canny 边缘检测算法分割结果。进行二阈值分割时，经过 12 代迭代就能收敛。对服从正态分布的随机噪声的噪声图像进行分割，进行二阈值分割时，经过 15 代收敛。分割效果还是较满意，分割的结果边缘比较连续，而且误分区域降低到最低限度。



图 1

5 结论

在进行图像分割时，无论是不含噪声的图像还是含有噪声的图像，利用改进的 Canny 边缘检测算法图像分割，能够自适应地寻找到最佳的二阈值，从而快速准确地提取出图像边缘。而且二维最大熵

阈值分割法充分利用图像的空间信息，抑制了噪声干扰，提高了分割图像的准确率，分割的结果较为理想。

参考文献

[1] Rafael C.Gonzalez,Richard E.Woods.Digital Image Processing Second Edition.Publishing House of Electronics Industry . BeiJing, Publishing House of Electronics Industry,2001,567-630.

[2] Kennedy J, Eberhart R, Shi Yuhui. Swarm Intelligence[M]. San Francisco: Morgan Kaufmann,2001.

[3] Chen Guo,Zuo Hong-fu.2-D maximum entropy method of image segmentation based on genetic algorithm[J].Journal of Computer-Aided Design&Computer Graphics,2002,14(6):530-534.

[4] Wachovia, Renate Smolíková. An approach to multimodal biomedical image registration utilizing[J]. IEEE Transactions On Evolutionary Computation, 2004, 8(3): 289-291.

[5] Van den Bergh F, Engelbrecht A P. A Cooperative Approach to Particle Swarm Optimization [J]. IEEE Transaction on Evolutionary Computation, 2004, 8(3): 225-239.

G.SHDSL 技术在远距离音视频信号传输系统中的应用研究¹

黄继海，杨建国，姜鹏飞

(防空兵指挥学院，河南 郑州，450052)

摘 要：论文研究基于 G.SHDSL 技术，构建一个音视频远距离传输系统，将音/视频信号从几公里以外传送到网络的节点。把音/视频信号交到网络或计算机。解决野外音/视频信号网络接入和显示的问题。

关键词：G.SHDSL 音/视频信号 远距离传输

G. SHDSL Technology in The Remote Audio and Video Signal Transmission System of Applied Research

HUANG Jihai, YANG Jiangguo, JIANG Pengfei

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Thesis based on G. SHDSL technology, audio and video to build a long-distance transmission system, the audio and video signals transmitted from a few kilometers away to the network nodes. The audio and video signals submitted to the network or computer. Address field network access audio and video signals and display problems.

Keywords: G. SHDSL; audio and video signals; long-distance transmission

1 问题的提出

在信息化建设的大背景下，对野外环境下场景的监控越来越常见，这就使得在野外环境下对场景的音视频远距离传输技术的研究变得十分重要。这种野外环境的方圆半径要求在 10km 以内，设备收放方便，便于携带，机动性好，为实现抗拉耐磨损性的要求，用线最好是军用被覆线。

目前，常见的视频传输有如下几种技术方式：

(1) 视频基带传输：是最为传统的电视监控传输方式，对 0~6MHz 视频基带信号不作任何处理，通过同轴电缆（非平衡）直接传输模拟信号。其优点是短距离传输图像信号损失小，造价低廉，系统稳定。缺点是传输距离短，距离在 300m 以上时高频分量衰减较大，无法保证图像质量。

(2) 光纤传输：常见的有模拟光端机和数字光端机，是解决几十千米甚至几百千米电视监控传输的最佳解决方式，通过把音/视频及控制信号转换为激光信号在光纤中传输。其优点是传输距离远、衰减减小，抗干扰性能好，适合远距离传输。其缺点是对于几千米内监控信号传输不够经济；光熔接及维护需专业技术人员及设备操作处理，维护技术要求高，不易升级扩容。

(3) 网络传输：是解决城域间远距离、点位极其分散的监控传输方式，采用 MPEG2/4、H.264 音/视频压缩格式传输监控信号。其优点是：采用网络视频服务器作为监控信号上传设备，只要有 Internet 的地方，安装上远程监控软件就可监看和控制。其缺点是：受网络带宽和速度的限制，目前的 ADSL 只能传输小画面、低画质的图像；每秒只能传输几帧到十几帧图像，动画效果十分明显并有延时，无

作者简介：黄继海（1955—），男，防空兵指挥学院基础部副教授，研究生导师，从事嵌入式技术教学和研究；
杨建国（1960—），男，防空兵指挥学院地方系主任；
姜鹏飞（1983—），男，防空兵指挥学院基础教研室计算机课助教。

法做到实时监控。

(4) 微波传输：是解决几千米甚至几十千米不易布线场所监控传输的解决方式之一。采用频率调制或幅度调制的办法，将图像搭载到高频载波上，转换为高频电磁波在空中传输。其优点是：综合成本低，性能更稳定，省去布线及线缆维护费用；可动态实时传输广播级图像，图像传输清晰度不错，而且完全实时；组网灵活，可扩展性好，即插即用；维护费用低。

其缺点是：由于采用微波传输，相对容易受外界电磁干扰；微波信号为直线传输，中间不能有山体、建筑物遮挡；如果有障碍物，需要加中继加以解决，Ku 波段受天气影响较为严重，尤其是雨雪天气会有比较严重的雨衰现象。不过现在也有数字微波视频传输产品，抗干扰能力和可扩展性都提高不少。

(5) 宽频共缆传输：视频采用幅度调制、伴音调频搭载、FSK 数据信号调制等技术，将数十路监控图像、伴音、控制及报警信号集成到“一根”同轴电缆中双向传输。其优点是充分利用了同轴电缆的资源空间，三十路音/视频及控制信号在同一根电缆中双向传输，频分复用技术解决远距离传输点位分散，布线困难问题；射频传输方式只衰减载波信号，图像信号衰减比较小，亮度、色度传输同步嵌套，保证图像质量达到 4 级左右；采用 75Ω 同轴非平衡方式传输使其具有很强抗干扰能力，电磁环境复杂场合仍能保证图像质量。其缺点是采用弱信号传输，系统调试技术要求高，必须使用专业仪器，如果干线线路有一台设备有问题，可能导致整个系统没图像。

(6) 双绞线传输（平衡传输）：它是视频基带传输的一种，将 75Ω 的非平衡模式转换为平衡模式来传输的，可解决监控图像 1km 内的传输。它是电磁环境相对复杂、场合比较好的解决方式，将监控图像信号处理通过平衡对称方式传输。其优点是布线简易、成本低廉、抗共模干扰性能强。其缺点是只能解决 1km 以内监控图像传输，而且一根双绞线只能传输一路图像，不适合应用在大中型监控中；双绞线质地脆弱抗老化能力差，不适于野外传输；双绞线传输高频分量衰减较大，图像颜色会受到很大损失。

从以上介绍来看，比较接近本文野外传输要求的是双绞线平衡传输。军用被覆线的传输性能接近双绞线，但双绞线远不如军用被覆线的抗拉、抗腐蚀性和收放方便。故本系统选用军用被覆线作为传输介质。

2 G.SHDSL 接口技术

2.1 G.SHDSL

2.1.1 xDSL 技术

xDSL 是以铜电话线为传输介质的传输技术组合，它按上行（用户到交换局）和下行（交换局到用户）的速率是否相同可分为速率对称型和速率非对称型两种。它包括普通 DSL、HDSL（对称 DSL）、ADSL（不对称 DSL）、VDSL（甚高比特率 DSL）、SDSL（单线制 DSL）、CDSL（ConsumerDSL）等，一般称为 xDSL。

2.1.2 ADSL 技术

对于普通需要高带宽接入的用户而言，ADSL 是一种值得考虑的技术，因其下行速率很高，适用于下行数据量很大的 Internet 业务。从电信网络提供商到用户的下行速率范围一般在 1.5~8Mbps 之间，而反向的上行速率则是在 16~640kbps 之间，所对应的最大传输距离为 5.5km，这是目前高速接入 Internet 的主要方法。

ADSL 使用一对电话线，在用户线两端各安装一个 ADSL 调制解调器，该调制解调器采用了频分复用(FDM)技术，将带宽分为三个频段部分：最低频段部分为 0~4kHz，用于普通电话业务，中间频段部分为 20~50kHz，用于速率为 16~640kbps 的上行数据信息的传递；最高频段部分为 150~550kHz 或 140~1.1MHz，用于 1.5~6.0Mbp/s 的下行数据信息的传送。

ADSL 技术能同时提供电话和高速数据业务，为此应在已有的双绞线的两端接入分离器，分离承载音频信号 4kHz 以下的低频带和 ADSL Modem 调制用的高频带。分离器实际上是由低通滤波器和高通滤波器合成的设备，为简化设计和避免馈电的麻烦，通常采用无源器件构成。

ADSL 使用调制解调器 Modem 。用户侧的 ADSL Modem 内部结构与 V.34 等模拟 Modem 几乎相同。主要由处理 D/A 变换的模拟前端（analog front end）、进行调制/解调处理的数字信号处理器（DSP），以及减小数字信号发送功率和传输误差，利用“网格编码”和“交织处理”实现差错校正的数字接口构成。交换局侧的 ADSL Modem 产品大多具有多路复用功能（DSL Access Multiplexer，DSLAM）。各条 ADSL 线路传来的信号在 DSLAM 中进行复用，通过高速接口向主干网侧的路由器等设备转发，这种配置可节省路由器的端口，布线也得到简化。目前已有将数条 ADSL 线路集束成一条 10BASE-T 的产品和将交换机架上全部数据综合成 155 Mbps ATM 端口的产品。

2.1.3 SHDSL （Symmetrical High bite Digital Subscriber Line）

SHDSL（Symmetrical High bite Digital Subscriber Line）对称高速数字用户线路。由于 ADSL 速率的不对称性，使得 ADSL 的应用存在不少局限。特别是商用宽带需求环境是一个双向的、对称的流量环境，对性能波动的容忍度比较低，ADSL 接入技术已越来越不能满足人们对频宽和流量的需求。于是，人们开始关注 SHDSL 技术。

G.SHDSL 原意是 SHDSL group（SHDSL 工作组），是 ITU-T 负责制定 SHDSL 标准的一个部门。后来以 G.SHDSL 作为这个标准的称呼。SHDSL 是由 ITU-T 定义的在单对双绞线上提供传输双向对称带宽数据业务的一种技术，符合国际电联 G.991.2 推荐标准，由于采用性能优越的电平网格编码脉冲幅度调制（TC-PAM）技术，压缩了传输频谱，提高了抗噪性能，延长了传输距离，因此与 ADSL，HDSL 技术相比有着明显的技术优势。

2.2 G .SHDSL 接口电路

G.SHDSL 接口电路由一片具有成帧功能的双通道 DSP 和一片单路模拟驱动前端组成。如美国敏迅（Mindspeed）公司开发的芯片组 M28945 为 DSL 成帧芯片，M28927 为 AFE（模拟前端）芯片，其结构框图如图 1 所示。

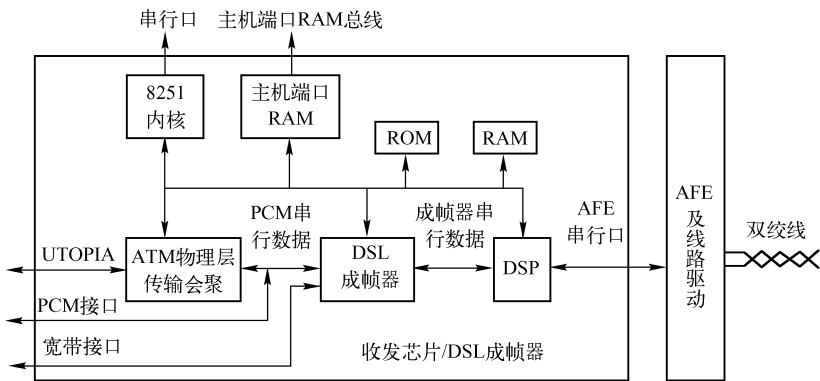


图 1 G .SHDSL 接口组成

1) DSL 成帧芯片

DSL 成帧芯片主要完成 16TC-PAM 编码、解码、回波抵消、自动均衡和成帧等功能。DSL 成帧芯片通常由三个功能模块组成。DSL 成帧器模块、采用 16TC-PAM 构 DSP 模块和微处理器内核。DSL 成帧器是一个高性能的比特流处理引擎，支持 G.SHDSL、HDSL 和 HDSL2 等 DSL 成帧模式。通过 PCM 接口将原始速率的 T1/E1 成帧信号和非成帧模式信号作经过有效载荷比特的插入和提取数据的加扰处理、比特填充等操作，输出相应的 DSL 帧数据流。

DSP 模块主要完成数据的编/解码，产生发送端码元定时和恢复提取接收端码元定时。线路均衡、回波抵消、根据对线路功率衰减的探测，调整发送功率电平等。它接收来自 AFE 的串行数据和比特泵发送的经过预编码的符号，同时将这些符号送到回波抵消器（EC），然后由回波抵消器对回波响应进行评估，从 AFE 发来的信号中减去回波响应。同时，回波处理后的信号再通过前馈均衡和判决反馈均衡，最后由格栅编码调制译码器恢复出信息比特。

芯片内嵌有微处理器内核。微处理器通过主机口或 RS-232 接口将 API 底层操作码下载至 DSL 成帧芯片内部 RAM，并通过 API 消息对系统进行配置和状态信息读出。

2) 模拟驱动前端

模拟驱动前端主要完成 D/A、A/D 变换、滤波、线路驱动、增益控制等功能，如图 2 所示。其中线路驱动包括线路的发送和接收，发送功放、接收放大、线路阻抗匹配等功能。数字接口同 DSL 成帧芯片相连，与 DSP 进行数字通信；模拟接口通过外围线路驱动反馈电阻。阻抗匹配电阻。平衡混合电路。变压器和保护电路与双绞线相连。

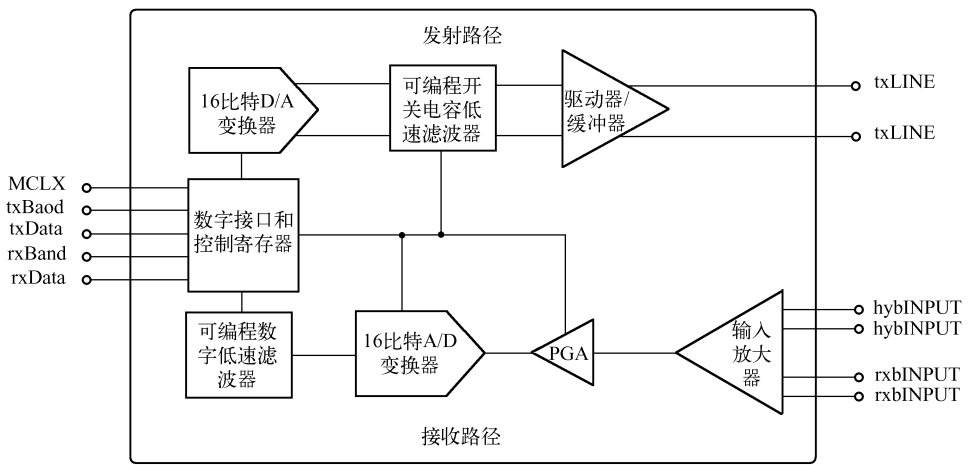


图 2 模拟前端芯片

TI 公司开发的模拟前端芯片 AFE1230 在数字信号处理器（DSP）和本地环路之间提供了一个作为线路接口的收发信机。AFE1230 能够处理速率范围在 64kbps~2.5Mbps 的上行和下行数据传输。从功能上讲,该器件分为发射和接收两部分。发射部分包括一个 16 位 $\Delta\Sigma$ 数/模变换器（DAC）、一个数字可编程 5 阶或 7 阶开关电容（SC）低通滤波器（LPF）和一个差分输出线路驱动器；接收部分包括一个输入可编程增益放大器、一个 16 位 $\Delta\Sigma$ 模/数变换器（ADC）和一个可编程抽取滤波器。

AFE1230 通过串行接口接收一个 16bit 数据字和一个 8bit 控制字节，从而简化了 D/A 变换和控制功能。后续模拟信号被传送至片上线路驱动器，该驱动器在 G.SHDSL 工作状态下可为 135 Ω 负载线路提供 14.5dBm 的功率。此外，这个片上线路驱动器还可用作输出缓冲器，和外部线路驱动器（如 OPA2677）一道，在 HDSL2 工作状态下可为 135 Ω 负载线路提供 17dBm 以上的功率。利用合适的 DSP，发射功率谱密度（PSD）既可以符合 G.SHDSL 标准，也可以符合 HDSL2 标准（以 OPA2677 作为外部驱动器）。在接收通路中，输入放大器通过对线路信号和混合路径信号的加法运算来进行一阶模拟回波抵消。合成信号经数字化处理变成 16bit 数字，并被传送至外部 DSP。

2.3 G .SHDSL 技术特点

G.SHDSL 的特点如下。

1) 支持数据速率高，传输距离远

在 26AWG（美国电线标准），电缆上传输距离为 1800m、6000m 的单线对，对应所能支持的数据速率分别为 2.312Mbps 和 192kbps；双线对操作速率可达 4.624Mbps。当距离一定时，速率比传统的

对称 DSL 高 35%~50%; 当速率一定时, 传输距离比传统的对称 DSL 提高 15%~20%。

2) 频谱兼容性

G.SHDSL 选择了 TC-PAM 调制技术, 从而确保了与其他基于 DSL 的服务 (如 ADSL) 的兼容性。

3) 传输速率自适应

G.SHDSL 是自适应的, 可以在不同的传送宽度之间做调整, 随着传送距离的不同, 实现的传送速率也不同。

3 远距离音/视频信号传输系统

3.1 系统组成框图 (见图 3)

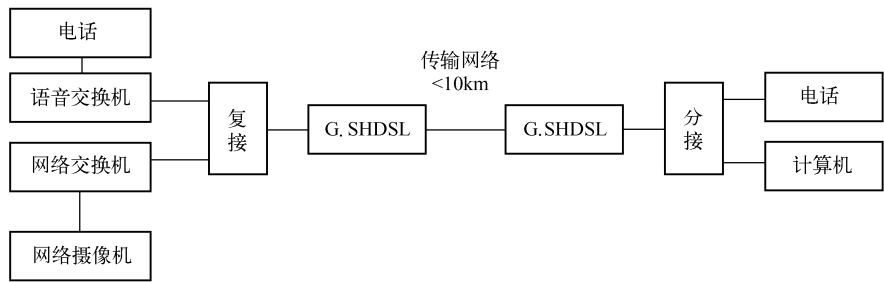


图 3 音/视频信号传输系统组成框图

3.2 系统工作过程

网络摄像机是一种结合传统摄像机与网络技术所产生的新一代摄像机, 它可以将摄像机所摄的图像转换为基于 TCP/IP 网络标准的数据包, 通过 RJ-45 以太网接口或 WiFi W-LAN 无线接口传送到网络上, 通过网络即可远端监视画面。网络摄像机内置一个嵌入式芯片, 采用嵌入式实时操作系统。摄像机传送来的视频信号数字化后由高效压缩芯片压缩, 通过网络总线传送到 Web 服务器。

监控系统的视频和语音的复接和分接处理采用同步数字复分接技术。同步数字复分接器是由数字复接器和数字分接器组成, 数字复接器是把两个或两个以上的支路同步数字信号按时分复用方式合并成为单一的同步数字信号的设备; 数字分接器则是把一个合路数字信号分解为原来的支路同步数字信号的设备。同步复分接是指如果输入支路数字信号相对于复接器的对应时钟信号是同步的, 则只需调整相位就可以实施数字复接。同步复分接有复接效率较高, 复接损伤较小等特点。

G.SHDSL 设备的作用是将音/视频信号从几千米以外传送到网络的节点。把音/视频信号交到网络或计算机。解决野外音/视频信号网络接入的问题。G.SHDSL 设备提供 RJ-11 接口和 RJ-45 接口, 以实现网络到电话线路的互换连接。

4 结论

基于 G.SHDSL 技术的远距离音/视频信号传输系统结构简单, 音/视频信号传输抗干扰性能强, 声音图像清晰, 携行方便, 展开时间短。采用的军用被覆线抗拉性好。非常适合野外条件下使用。

参考文献

[1] 4-wire Operation Using Globespanvirata SHDSL Chip Sets[J]. May,2002.
[2] 王兴亮等. 数字通信原理与技术[M]. 西安电子科技大学出版社, 2000.

H.264 编码技术及其应用

黄欢欢, 王月蓉, 冯少华

(防空兵指挥学院, 河南 郑州, 450052)

摘 要: 编码算法标准 H.264 引入了许多当前视频编码中的新技术, 使得在相同的重建图像质量下, 编码效率比 H.263 和 MPEG-4 高 50%左右。H.264 的应用场合相当广泛, 包括固定或移动的可视电话, 实时视频会议系统、视频监控系統、因特网视频传输及多媒体信息存储等。H.264 编码系统通常采用平台 DSP。该文介绍了 Blackfin533 DS P 为主 H.264 编码技术的系统组成、原理和应用。

关键词: H.264 ; 压缩; 码流; DSP

中图分类号: TP311.56 文献标识码: A 文章编号: 1006-7043 (2010) xx-xxxx-x

The Technology and Application of H.264 Coding

HUANG Huanhuan, WANG Yuerong, FENG Shaohua

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: The standard of the coding arithmetic of H.264 includes many new technologies, the technologies are the currently coding technology of video frequency. It makes the coding efficiency enhances 50% than H-263 and MPEG-4. The application of H-264 is quite abroad, including immobile and mobile visual telephone, real-time net-meeting, monitor systems, internet video transmission and multimedia, information storages and so on. The coding arithmetic of H.264 commonly adopts platform DSP. It introduces the systems form, elements and application of the coding arithmetic of H.264.

Keywords: H.264; compress; code stream; DSP

1 视频压缩与压缩算法

视频压缩通过减少和去除冗余视频数据的方式, 达到有效发送和存储数字视频文件的目的。在压缩过程中, 需要应用压缩算法对源视频进行压缩以创建压缩文件, 以便进行传输和存储。要想播放压缩文件, 则需要应用相反的解压缩算法对视频进行还原, 还原后的视频内容与原始的源视频内容几乎完全相同。压缩、发送、解压缩和显示文件所需的时间称为延时。在相同处理能力下, 压缩算法越高级, 延时就越长。

编码算法标准 H.264 是由国际电信同盟 ITU-T 的视频编码专家组 (VCEG) 和国际标准化组织国际电工委员会 ISO/IEC 的活动图像专家组 (MPEG) 两个不同的组织共同制定的^[1]。由于该标准是由两个不同的组织共同制定的, 因此有两个不同的名称: 在 ITU-T 中, 它的名字叫 H.264; 而在 ISO/IEC 中, 它被称为 MPEG-4 的第 10 部分, 即高级视频编码 (AVC)。

2 H.264 编码算法标准

(1) 将编码分为编码层 VCL (Video Coding Layer) 和传输层 NAL (Network Abstraction Layer)。将编码层和传输层分离, 有利于 H.264 的扩展。

(2) H.264 采用了空域内的帧内预测, 共两种预测模式: intra16×16 和 intra4×4。其中 intra16×16

作者简介: 黄欢欢 (1982—), 女, 助教, 学士;
王月蓉 (1985—), 女, 助教, 学士;
冯少华 (1983—), 男, 助教, 学士。

有四种预测方式，intra4×4 有九种预测方式。

- (3) 对于帧间预测，增加了预测模式，共七种预测模式。预测块从 16×16 可以最小细分为 4×4。
- (4) 增加了参考帧的数目，使预测更为准确。
- (5) 将去块效应滤波放在编码环内，提高图像的主观质量。
- (6) B 帧可以作为参考帧，同时将图像的解码顺序与显示顺序分离。
- (7) 采用整系数变换，提高变换速度。
- (8) 采用 CAVLC、CABAC 等新的熵编码方法以提高编码效果。
- (9) 提高了码流的抗误码能力，如对编码数据进行分割，一帧图像可以灵活地分为几个 slice 等。

3 H.264 编码系统构成

3.1 系统基本组成

H.264 编码算法模块构成见图 1。

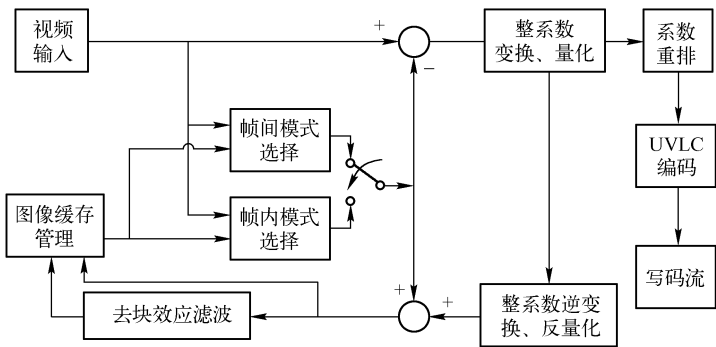


图 1 H.264 编码算法模块

目前，H.264 编码系统开发系统通常采用 DSP 平台。在 DSP 平台上进行视频产品开发有以下几方面的优势：第一，用户开发自由度大，支持多种个性化开发，可以适应市场不断提出的新要求，在第一时间提升产品性能，增强产品的竞争能力；第二，DSP 处理能力强，可以在一个 DSP 上同时实现多路音/视频信号的压缩处理；第三，开发周期短，能实现快速技术更新和产品换代，各种新出现的快速及优化算法可灵活进行升级。以 DSP 为 H.264 编码系统开发的硬件平台构成见图 2。

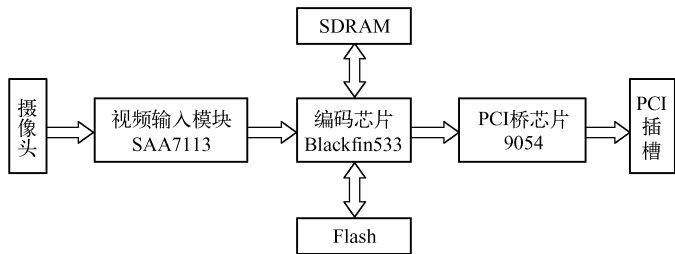


图 2 Blackfin533 平台总体框架图

H.264 编码器由视频采集、数据格式转换、H.264 编码 3 部分组成。视频采集部分负责捕获图像，并且将捕获到的图像通过 PPI 接口填充到指定的视频帧缓冲区中。数据格式转换部分完成将输入的 4：2：2 格式的图像转换成 H.264 编码器能够编码的 4：2：0 格式的数据。H.264 编码部分负责对 4：2：0 格式图像编码。在本视频编码器设计中，DSP 用于运行操作系统和协议栈，实现 H.264 的编

码算法。

3.2 Blackfin533 介绍

Blackfin533 是 ADI 公司 Blackfin 系列中的一款高性能 DSP 视频处理芯片。其主频最高能达 600MHz，每秒可处理 1200M 次乘加运算。具有大量针对视频的专用指令，可以并行处理多条指令。

从总体上看，Blackfin533 分为内核和系统接口两大部分。内核指处理器、L1 存储器、事件控制器、内核定时器等；系统接口指 SPORT 接口、PPI 接口、SPI 接口、外部存储控制器、DMA 控制器及与它们接口的外部资源等。

3.3 系统工作原理

Blackfin533 开发平台原理图如图 2 所示。SAA7113 为视频模数转换芯片。SAA7113 芯片从视频端子读入摄像头输出的模拟信号，通过并口将数字信号输出给 Blackfin533。此信号从 Blackfin533 的 PPI 接口进入 Blackfin533，压缩后的码流由 PCI 总线桥传给 PC。此系统通过 Flash 中的代码启动。编码过程中的原始图像、参考帧及其他变量存储在 SDRAM 中。

Blackfin533 通过 I²C 总线对 7113 进行配置，使其彩色视频为 YUV（Y 为亮度信号，U、V 为色差信号）输出模式、ITU656 模式及增强 ITU656 模式等。Blackfin533 与 7113 的连接电路见图 3。

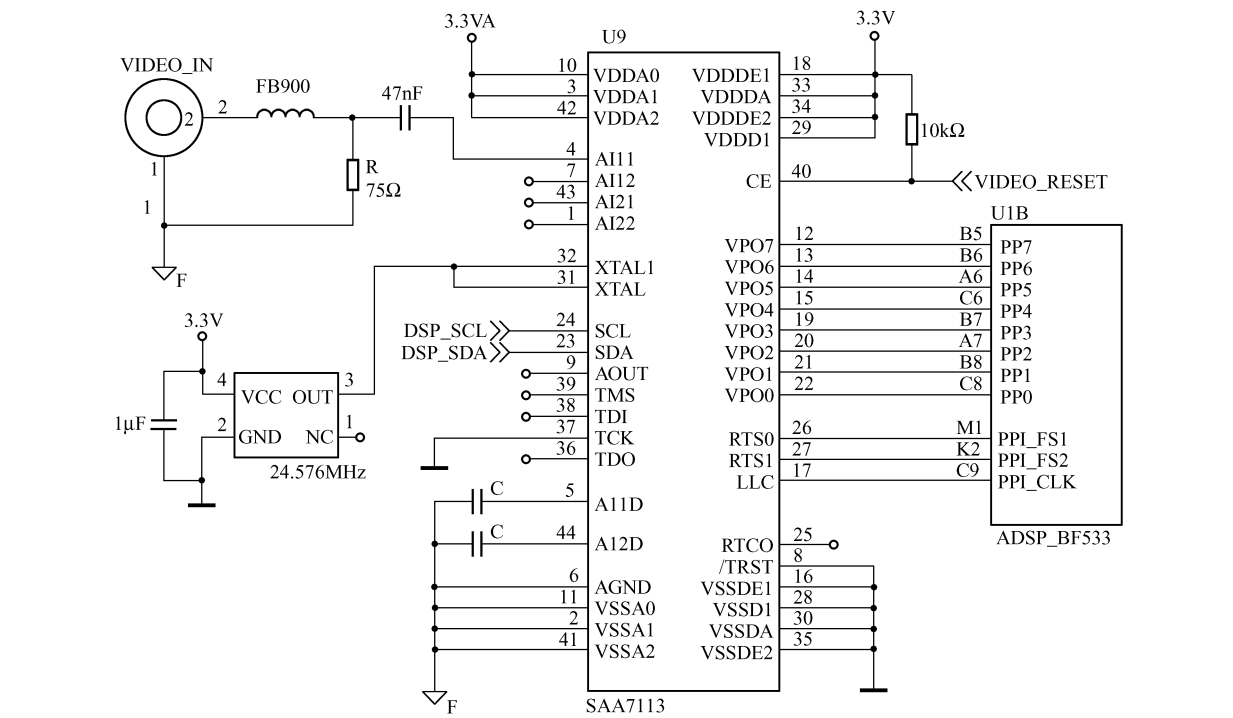


图 3 Blackfin533 与 7113 的连接电路

Blackfin 处理器用直接存储器访问（DMA）在存储器之间或存储器与外设之间传送数据。DMA 控制器可在存储器和片上外设（外设 DMA）之间进行数据传送，以及在 L1 / L2 / L3 存储器间进行数据传送（存储器 DMA 或 MDMA）。DMA 控制器是 Blackfin 处理器架构中的重要组件，完全独立于内核，不会进行周期挪用，完全无须占用处理器内核周期。在理想的应用配置中，内核只需要设置 DMA 控制器，并在数据调用过程中响应中断。

4 H.264 编码算法标准

编码算法标准 H.264 的应用场合相当广泛，包括固定或移动的可视电话，实时视频会议系统、视

频监控系统、因特网视频传输及多媒体信息存储等。新一代视频编码标准 H.264 与以往标准相比具有压缩率高，网络传输性能好，视频质量优越等优点。H.264 引入了许多当前视频编码中的新技术，使得在相同的重建图像质量下，编码效率比 H.263 和 MPEG-4 高 50%左右。在网络摄像机中，采用专用压缩芯片，实现 H.264 Main profile 压缩算法，可以使摄像机具有双码流功能。实现本地高清存储，远程流畅监控。远程监控时，还可根据带宽状况选择主码流或次码流实现高画质、低码流传输。线路带宽占用比 MPEG-4 节省 30%，使得网络摄像机在视频监控系统中得到广泛的应有。

5 H.264 编码的特点与优势

与以往的标准一样，H.264 使用运动估计和运动补偿来消除时间冗余，但是它具有以下特点：

1) 预测时所用块的大小可变

由于基于块的运动模型假设块内的所有像素都做了相同的平移，在运动比较剧烈或者运动物体的边缘处这一假设会与实际出入较大，从而导致较大的预测误差，这时减小块的大小可以使假设在小的块中依然成立。另外小的块所造成的块效应相对也小，所以一般来说小的块可以提高预测的效果。

为此，H.264 一共采用了 7 种方式对一个宏块进行分割，每种方式下块的大小和形状都不相同，这就使编码器可以根据图像的内容选择最好的预测模式。与仅使用 16×16 块进行预测相比，使用不同大小和形状的块可以使码率节省 15%以上。

2) 更精细的预测精度

在 H.264 中，Luma 分量的运动矢量（MV）使用 1/4 像素精度。Chroma 分量的 MV 由 Luma MV 导出，由于 Chroma 分辨率是 Luma 的一半，所以其 MV 精度将为 1/8，也就是说，1 个单位的 Chroma MV 所代表的位移仅为 Chroma 分量取样点间距离的 1/8。如此精细的预测精度较之整数精度可以使码率节省超过 20%。

3) 多参考帧

H.264 支持多参考帧预测（Multiple Reference Frames），即可以有多个（最多 5 个）的在当前帧之前解码的帧可以作为参考帧产生对当前帧的预测（Motion Compensated Prediction）。使用于视频序列中含有周期性运动的情况，用这一技术可以改善运动估计（ME）的性能，提高 H.264 解码器的错误恢复能力，但同时也增加了缓存的容量及编解码器的复杂性。不过，H.264 的提出是基于半导体技术的飞速发展情况，因此，这两个负担在不久的将来会变得微不足道。较之只使用一个参考帧，使用 5 个参考帧可以节省码率 5%~10%。

H.264 最大的优势是具有很高的数据压缩比率，在同等图像质量的条件下，H.264 的压缩比是 MPEFG-2 的 2 倍以上，是 MPEFG-4 的 1.5~2 倍。

参考文献

[1] H.264 视频压缩标准，AXIS communications.
[2] ITU-T H.264 TELECOMMUNICATION STANDARD OF ITU.

基于 Arnold 和 DCT 的数字水印技术研究

尚 存¹, 邬长安²

(1.信阳农业高等专科学校, 河南 信阳, 464000; 2.信阳师范学院计算机与信息技术学院,
河南 信阳, 464000)

摘 要: 数字水印技术是解决多媒体数据版权保护问题的有效手段之一。本文提出了一种基于 Arnold 和 DCT 的数字水印技术。采用 HSI 色彩空间, 将原始彩色图像的各分量实现 8×8 分块, 并对每个分块完成 DCT 变换及系数量化操作, 随后嵌入用于版权保护的鲁棒水印。为了提高水印鲁棒性, 本算法采用多重嵌入的方法, 在各个彩色分量相同位置像素块的 DC 系数上嵌入同一经过 Arnold 变换处理过的鲁棒水印。实验表明, 本算法对常规图像处理及恶意攻击具有较好的鲁棒性, 同时能够有效标识图像版权信息。

关键词: 数字水印; HSI 色彩空间; DCT 变换; 鲁棒性

中图分类号: TP391.3 文献标识码: A 文章编号: 1006-7043 (2004) xx-xxxx-x

Research of Digital Watermark Based on Arnold and DCT Technology

SHANG Cun¹, WU Chang'an²

(1. Xinyang Agricultural College, Henan, Xinyang 464000; 2. Xinyang Normal University,
Xinyang 464000, Henan China)

Abstract: Digital watermarking is one of the effective methods which can protect the copyright of multimedia data. This thesis presented a digital watermarking technology based on Arnold and DCT techniques. The watermarking technology for color image based on the HSI color space, each component of the original color image is divided into non-overlapping 8×8 block, DCT and quantization is performed to each block, then a robust watermark for copyright protection. To improve the robustness of watermark, this thesis adopted the multi-embedded strategy. The DC coefficients in the block of the same location of each color components, are embedded the same robust watermark. The results show that, the proposed watermarking algorithm is robust to normal image process and attack, and effective to identify the copyright of image.

Keywords: Digital watermarking; HSI color space; DCT transform; robustness

1 引言

随着数字技术和英特网的快速发展, 近年来, 如何保证数字图像信息的安全成为一个重要的研究课题。数字水印技术 (Digital Watermarking) 在数字产品的知识产权保护方面起到了越来越重要的作用。本文提出了一种基于 Arnold 和 DCT 技术的数字水印算法。提出的彩色图像水印算法基于 HSI 色彩空间, 对颜色的描述更符合人对颜色的视觉理解, 也更有利于彩色图像处理。将原始彩色图像由 RGB 空间转换到 HSI 空间, 选取其中 I 分量和 S、H 分量分别进行 8×8 分块, 然后对各子块进行 DCT 变换。为了增加破译难度, 采用改进后的 Arnold 变换对二值水印图像进行加密。由于 DCT 变换在水印技术中应用比较成熟, 且具有计算量小, 与国际流行的压缩和编码标准兼容等优点, 本文考虑在 DCT 域中进行水印的嵌入, 而 DCT 变换后图像的能量主要集中在 DC 系数和 AC 系数的中低频分量上, 将数字水印嵌入到这些分量, 能够获得较好的鲁棒性。本文选取在低频系数中进行水印信息的嵌入。通过实验对比, 本算法对常规图像处理及恶意攻击具有较好的鲁棒性, 且能够有效标识图像版权信息。

2 彩色空间的选取¹

对于彩色图像数字水印技术的研究, 由于原始载体图像是彩色图像, 其中最为关键的问题是如何

基金项目: 河南省教育厅自然科学项目 (2010C520014)

作者简介: 尚存 (1982—), 男, 助教, 硕士, 主要研究方向为数字图像处理;

邬长安 (1959—), 男, 教授, 主要研究方向为模式识别、数字图像处理。

选择色彩空间去嵌入水印。彩色图像的颜色大多数都为 RGB 色彩空间描述，但 RGB 及其他相似的色彩模型不能很好地适应实际中人们对颜色的敏感度。由于 HSI 色彩空间在彩色描述时更加自然和直观，更适合人类视觉的特性；且在色度和明亮度的分离上，能独立处理图像中的亮度分量^[1]。所以，本文选择 HSI 色彩空间嵌入水印。以达到既实现水印鲁棒性良好，又使彩色图像具有良好的视觉效果的目的。

3 水印图像预处理

水印系统中在水印嵌入前都要完成预处理，这不但在系统鲁棒性方面起着关键作用，而且在数字水印系统安全性方面也具有很重要的意义，它增加了攻击者在猜测攻击中的难度。预处理有很多种方法，由于 Arnold 变换简单、方便使用且拥有周期性，本文在预处理中使用了 Arnold 变换。

Arnold 变换定义如下：设有单位正方形上的点 (x,y) ，将点 (x,y) 变到另一点 (x',y') ，变换如式 (1)：

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \bmod m \tag{1}$$

其中， (x,y) 是原图像的像素点坐标， (x',y') 是变换后新图像的像素点坐标， (x,y) 及 $(x',y') \in \{0, 1 \cdots M-1\}$ ，在这里 M 表示数字图像矩阵的阶数^[2]。

将图像进行置乱变换，在水印图像空间及载体图像空间中都能够使用，通常选择对水印图像空间进行置乱。原始水印图像经过 Arnold 置乱变换后，得到杂乱无章的图像，如图 1 和图 2 所示。

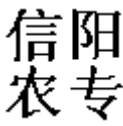


图 1 未嵌入时原水印图像

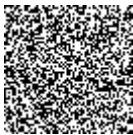


图 2 置乱后的水印图像

经过 Arnold 置乱变换后水印图像和置乱后的图像，只是空间位置发生了变化，而其像素的值和图像大小并没有发生改变。

4 水印嵌入位置的分析 and 选取

水印嵌入位置的选择对保证系统的不可见性及鲁棒性非常重要，为了充分利用人眼对频率的掩蔽效应，要将图像从空间域变换为变换域进行处理。本文选择离散余弦变换 DCT 作为变换方法。离散余弦变换 DCT 具有图像能量压缩的特点，且由于它是实数域内的变换，算法相对简单、处理速度快^[3]。利用离散余弦变换将图像分为不同频率系数，这些频率系数不仅与人类视觉的敏感程度有关，且它们的抗攻击能力也不相同。因此当选取嵌入系数时要平衡这两个指标进行选择。

4.1 离散余弦变换

进行图像处理时，二维 DCT 变换使用像素块的方式实现操作，像素块的大小一般是根据需求来定的，因为对图像整体实现二维 DCT 变换计算量比较大，本文对图像完成分块 DCT 变换。当实现 DCT 变换时，取 N 为 8，将图像分成 8×8 的分块主要因为 N 比 8 大时效率增加不多但复杂性增加很多。通过离散余弦变换后，图像左上角是低频系数，其系数值较大，能量绝大部分集中在图像的左上角，因此左上角表现出较明亮的白色，能量低的区域表现出黑色。这表明人类视觉系统对 DCT 低频系数的改变比较敏感，但对高频系数的改变不易察觉。

4.2 水印嵌入系数选取

在 DCT 域的数字水印算法中，选择不同的 DCT 系数主要是考虑了人类视觉对频率的掩蔽特性，以及变换系数自身的特性。为了使水印具有良好的不可见性和鲁棒性，通常用于嵌入水印的 DCT 系数在通过常规图像处理或攻击后仍然可以比较好的保留，此方法能够保证水印的鲁棒性；同时可在保证不可见性的基础上嵌入比较大强度的水印信息。本算法选择鲁棒性强的低频系数作为待嵌入水印系数。

5 基于 Arnold 和 DCT 技术的数字水印实现过程

5.1 水印的预处理

设原始图像大小是 $M \times N$ 的 24 位真彩色图像 K ，嵌入的版权标识水印是大小 $\frac{M}{8} \times \frac{N}{8}$ 的二值水印图像 WR ，当水印即将被嵌入时，要先将水印图像实现预处理，将它实现二值序列的转化。

(1) 将二值水印图像 WR 实现 M 次 Arnold 置乱变换，用来除去水印图像中像素空间相关性，增强水印算法的鲁棒性及安全性。其中 M 作为加密水印图像的密钥。

(2) 将 WR 转换成 $\{0,1\}$ 的序列，序列里每位元素定义为 $WR_k \left(k=1,2 \cdots \frac{M \times N}{8 \times 8} \right)$ 。

(3) 将 WR 中“0”转化成“-1”，以实现 $WR_k \in \{-1,1\}$ ，这样可以除去舍入误差的影响。

5.2 水印的嵌入算法

要增强检出水印信息的可靠性，提高水印抗剪切、JPEG 压缩等操作的鲁棒性，在嵌入水印过程中使用多重嵌入的方案，将同一个鲁棒水印信息分别嵌入 H 、 S 和 I 三个彩色分量相同位置块内经过 DCT，以及量化处理后的 DC 系数上，在操作过程中将亮度分量每个分块都进行了 DCT 和量化处理，在提取时按照多数原则确定本分块所对应的水印信息。在亮度分量 I 及两个色差分量 H 和 S 中嵌入鲁棒水印，先将亮度分量 I 中嵌入水印，其过程为：

1) 色彩空间的相互转换

先将原始彩色图像 K 从 RGB 方式变换到 HSI 方式，本文采用了柱体转换，转换公式 (1) 如下：

$$I = \frac{1}{\sqrt{3}}(R + G + B); \quad S = 1 - \frac{\sqrt{3}}{I} \min(R, G, B); \quad H = \begin{cases} \theta, & \text{当 } G \geq B \\ 2\pi - \theta, & \text{当 } G < B \end{cases};$$
$$\theta = \cos^{-1} \left[\frac{\frac{1}{2}[(R - G) + (R - B)]}{\sqrt{(R - G)^2 + (R - B)(G - B)}} \right] \tag{2}$$

2) 提取亮度分量 I 且实现分块

提取图像的亮度分量 I ，将它完成 8×8 分块，得到一个由 $\frac{M}{8} \times \frac{N}{8}$ 个分块组成的互不覆盖的分块矩阵，各分块表达成 $f_k^I(x,y) (x,y=1,2,\cdots 8, k=1,2,\cdots \frac{M \times N}{8 \times 8})$ ，每个分块 $f_k^I(x,y)$ 对应于鲁棒水印 WR 中相应水印信息 WR_k 。

3) DCT 变换及系数量化

将图像各个分块 $f_k^I(x,y)$ 实现 DCT 变换，以便获得 DCT 系数矩阵 $F_k^I(u,v)$ ，其中 $(u,v=1,2,\cdots 8, k=1,2,\cdots \frac{M \times N}{8 \times 8})$ 。对其实现量化处理，处理如式 (3) 所示，其中 $F_k^I(u,v)$ 是需要量化的系数， β 作为量化因子。

$$B_K^I(u,v)=\left\lfloor \frac{F_K^I(u,v)}{\beta^*Q(u,v)}+0.5\right\rfloor \quad (u,v=1,2,\cdots,8)$$

(3)

4) Zig-Zag 排序

采用 Zig-Zag 排序模式把每个分块中的 DCT 量化系数实现排序，从而使得每个分块的 DCT 量化系数为 $C_k^I(i)(i=1,2,\cdots,64)$ 。

5) 鲁棒水印嵌入

在每个分块中选取要嵌入鲁棒水印信息的 DC 系数 $C_k^I(1)$ ，使用加性嵌入方式嵌入水印信息，其嵌入的方法如式（4）所示，当中， $C_k^{I'}(1)$ 作为分块中嵌入了水印信息的 DC 系数， α 为嵌入强度，它的取值将会直接影响算法的有效性，通过实验校对， α 取值 0.79 比较合适，这样能同时兼顾水印的不可见性及鲁棒性。

$$C_k^{I'}(1)=C_k^I(1)+\alpha WR_k$$

(4)

6) 反量化和逆 DCT

将每个分块实现反量化处理及逆 DCT 变换，得到嵌入水印后的亮度分量 I' 。然后在原始图像的 S 和 H 两个色差分量上实现鲁棒水印的嵌入，步骤和亮度分量 I 的相同，最后得到嵌有鲁棒水印信息的色差分量 S' 及 H' ，将嵌入水印的各彩色分量 H' ， S' ， I' 反变为 RGB 色彩空间，以实现了嵌入了水印图像 K' 。

6 实验结果

本算法的测试主要在 Matlab 7.0 平台下进行。根据本文的算法，将原始图像中嵌入鲁棒水印，鲁棒水印嵌入强度 α 选定为 0.79，原始载体图像如图 3 所示，水印被嵌入后图像如图 4 所示，通过比较，发现被嵌入水印的图像在观察上并未发生显著的降质，视觉系统总体上也未察觉差异。经过计算，它的峰值信噪比（PNSR）为 32.48，表明本文所提数字水印技术具有较好的不可见性。在未被攻击的条件下，从被嵌入水印的图像里提取出鲁棒水印如图 5 所示，其提取的水印信息与原始水印的归一化相关系数 NC 值为 0.996。



图 3 原始载体图像



图 4 嵌入水印的图像

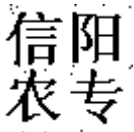


图 5 提取的版权水印

表 1 嵌入水印图像中值滤波后提取水印效果

中值滤波器	1×3 3×3 5×5		
NC	0.981 0.963 0.913		

对于经过不相同质量因子的 JPEG 压缩后提取的水印效果如表 2 所示。

表 2 不同 QF 的 JPEG 压缩后提取的水印效果

质量因子 QF	90 80 70 60			
NC	0.973 0.956 0.922 0.884			

通过实验，在对嵌入水印的图像进行适度 JPEG 压缩后，仍然能够有效提取水印，这表明该算法对 JPEG 压缩具有良好的鲁棒性。

7 结论

本文选择了 24 位 RGB 模型的彩色图像为原始图像，集合了 HSI 彩色空间更接近人对彩色的认识 and 解释，然后对基于 DCT 技术数字图像水印算法的实现效果进行测试，主要在不可见性和鲁棒性等方面对水印算法的性能完成评估和分析。通过对比试验表明，此方法具有良好的不可见性，在保证水印不影响原始彩色图像质量的前提下，具有较强的鲁棒性。

参考文献

- [1] 丁玮，齐东旭.基于 Arnold 变换的数字图像置乱技术[J].计算机辅助设计与图形学学报，2001，13:338-341.
- [2] 金聪. 数字水印理论技术[M].北京：清华大学出版社，2008，6-9.
- [3] HSUCT,WUJL.Hidden Signature in Image processing,1999,8(1):58-68.
- [4] Cox I J,Miller M L. The First 50 years of Electronic watermarking .EURASIP J of Applied Signal Proceessing, 2002 , 2 :126-132.
- [5] BenderW.,GruhlD.,MorimotoN.. Technique for data hiding.IBM Systems Journal ,1996,35(3-4):313-336.

基于 HLA 的仿真系统可视化数据模型研究

周龙龙, 姜鹏飞, 黄建廷

(防空兵指挥学院, 河南 郑州, 450052)

摘 要: 在基于 HLA 的仿真系统开发中, 设计和建立 HLA 仿真系统的各种模型是促进仿真互操作和仿真组件重用的关键过程。本文根据作战仿真系统的可视化需求, 提出了作战仿真实体可视化数据模型的分类方法和层次体系结构, 分析了可视化数据模型的建模过程和建模方法, 并提出了改进的基于 HLA 的仿真系统开发模型体系, 为各模型的快速、规范化开发打下了基础, 并能够缩短基于 HLA 的仿真系统的开发周期, 提高建模与仿真的效率。

关键词: 高层体系结构; 可视化数据模型; 模型体系结构; 概念模型

Study on Visual Data Models of HLA-based Simulation System

ZHOU Longlong, JIANG Pengfei, HUANG Jianting

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: Within the development of HLA-based simulation system, designing and establishing various models is one of the main processes for promoting the interoperability and aiding the reuse. The paper simulation system according to the visual needs, put the battle simulation entities visualization data model classification method and the hierarchy structure, analyzed data modeling process visualization and modeling method, and put forward the improvement of the simulation system is developed based on HLA model system. These lay a foundation for the rapid and standardized development of various models, can shorten the development period of HLA-based simulation and improve the efficiency of modeling and simulation.

Keyword: HLA; visualization data model; model architecture; conceptual model

引言

高层体系结构 (High level architecture) 仿真技术框架是美国国防部国防与仿真办公室 (DMSO) 于 1995 年 10 月提出的。其用意是为解决美国国防领域仿真应用之间的互操作和可重用性的问题。HLA 是用于产生计算机分布式仿真系统的通用技术框架, 其目的是希望通过提高仿真应用的互操作性和仿真资源的可重用性, 来提高建立仿真系统的效率和效费比。HLA 的核心思想是互操作和重用, 通过运行 RTI 提供通用的、相对独立的支撑服务程序, 将仿真应用同底层的支持环境分开, 即具体的仿真功能的实现、仿真运行管理、底层通信传输三者分离, 隐蔽各个细节, 从而使各部分可相对独立开发。

建模与仿真高层体系结构 (HLA) 已成为 IEEE 1516 系列标准^[1]。HLA 采用面向对象的思想来开发仿真系统, 已成为当前分布交互仿真技术的研究重点。模型问题是整个 HLA 联邦开发中的核心问题。建立被研究对象的模型是进行基于 HLA 的仿真系统研究和开发的重要步骤。

1 可视化数据模型分类与层次体系

军事模型是对军事事物进行模拟的表现形式, 包括实体模型、图形模型、数学模型和综合模型实

作者简介: 周龙龙 (1984—), 男, 助教, 本科;
姜鹏飞 (1983—), 男, 助教, 本科;
黄建廷 (1984—), 男, 助教, 本科。

现实体可视化数据模型的重用、动态选取、定制、组合及实时控制以满足不同的演习想必是日前作战仿真界亟待解决的问题之一。

1.1 可视化数据模型的分类

可视化数据模型^[2]所要描述的具体对象，可以是武器装备，也可以是部队、指挥员及指挥机构，或者是战场环境中的某个因素，是指对具备行为能力的实体进行描述的数据模型。可视化数据模型主要是确定模型的属性、行为之间或与其他相关因素间的关系，其中属性包括名称、类型、级别、性质、运动要素、当前位置、当前任务等。可视化数据模型包含行动实体可视化数据模型、指挥实体可视化数据模型、综合自然环境实体可视化数据模型。

(1) 行动实体可视化数据模型。担负各类最基本作战任务的实体称为行动实体，通常是指对具备基本作战行动能力的实体进行描述的数据模型，包含行动实体属性可视化数据模型和行动实体关系可视化数据模型。

(2) 指挥实体可视化数据模型。指挥实体指按指挥控制体制承担某级任务的指挥机构或指挥员。指挥实体可以被消灭、损伤或干扰破坏，可以移动，但不能像行动实体那样，执行所有的作战行动，具备直接攻击对方的能力。指挥实体可视化数据模型是指对指挥实体的名称、状态和属性等进行描述的可视化数据模型。

(3) 综合自然环境实体可视化数据模型。综合自然环境（SNE）建模是指对包括陆地、海洋、大气、太空在内的整个自然环境空间领域具有权威性、完整性、多态性和一致性的数据描述模型表示。

1.2 可视化数据模型的层次体系

作战仿真实体可视化数据模型包含以下 7 个层次：

(1) 全貌层。使用几何体、层次结构和属性来描述实体，包括外部引用，模型各个部分的几何模型定义、位置。

(2) 集合层。用逻辑组的形式来组织和定义模型的各个组件，基于整体模型的构建和实时渲染。

(3) 对象层。提供更好的结构细节和模型的实时渲染。其扩展功能包括文本、三维或二维显示、静态或动态的文本、定义光源的类型、位置和方向、声音的定义和附加声音文件到动态二维实体等。

(4) 表面层。为了对渲染提供更细致的控制。利用表面层模型定义和组织的属性，为色彩与纹理的相互协调选择材质，定义表面渲染、明暗模式等。

(5) 顶点层。组织和定义数据库中几何造型，提供对顶点的位置、色彩、纹理映像和光亮的绝对控制。

(6) 关系层。描述前 5 层的关系：全貌层的模型可以分解为集合层模型，集合层模型可以分解为对象层模型，对象层模型可以分解为表面层模型，表面层则可以分解为顶点层。面是由点和线组成，每个面的顶点所对应的就是顶点层。

(7) 属性层。用于描述作战仿真实体（或其部分）的属性，包括性能与参数、各项指标、所属的实体模型类别及相应关联的数据等。

2 HLA 仿真系统建模过程

在整个 HLA 联邦（仿真系统）开发过程中会涉及各种各样的模型。基于 HLA 的仿真系统的建模过程如图 1 所示，建模工作者将建立仿真模型需要的成分和现象从真实世界中提炼出来，在这一抽象的基础上，建立概念模型。这个概念模型通过分析和设计形成 HLA 对象模型，在此基础上，通过开发者实现为仿真模型。仿真模型是真实世界的近似表现；开发者通过校核过程确定仿真模型是否实现

了对象模型；对象模型对概念模型的合理设计通过验证来确认；通过确认过程判断是否正确地从真实世界抽象出概念模型。

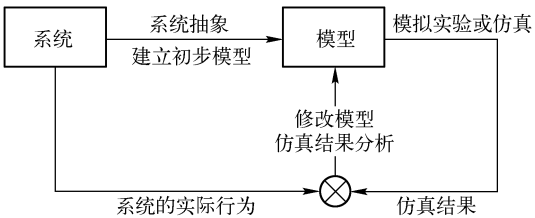


图 1 建模的一般过程

概念模型（Conceptual Model）是对真实世界的抽象，描述仿真或联邦的真实表示、限制真实表示的假设和需要满足用户需求的其它能力。

对象模型（Object Model）是对给定系统固有对象的规格描述，包括对象特征（属性）的描述和对象之间存在的静态与动态关系的描述。每一个仿真模型都是一个仿真应用程序，是联邦运行的基本元素，它是实体存在及活动的容器，负责仿真的运行管理、实体的处理，计算、仿真的表现和对外界的信息交换。

3 HLA 仿真系统模型体系

3.1 模型体系

考虑到仿真实体的重用性及联邦成员的快速开发需要，在 HLA 联邦开发与运行过程（FEDEP）的基础上，提出改进的基于 HLA 的仿真系统开发模型体系，如图 2 所示。

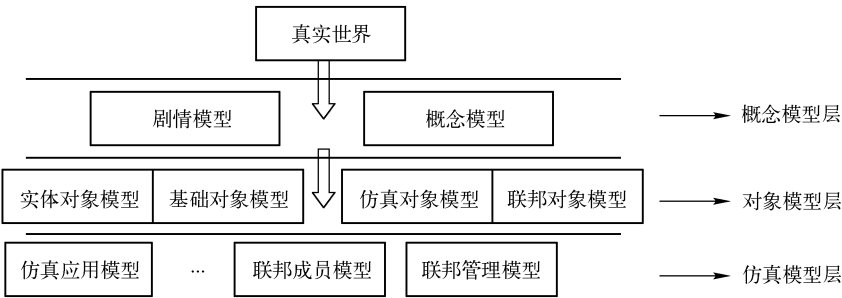


图 2 基于 HLA 的仿真系统开发的模型体系

建模工作者首先将真实世界的事物和现象根据面向对象的思想抽象为任务空间概念模型（CMMS）；在 CMMS 中，剧情模型和概念模型互为补充。剧情模型需要概念模型中描述的实体，剧情才具体，概念模型也需要剧情模型的剧情，实体才能活动。在对象模型层中，实体对象模型是从概念模型具体实现得到的组件，而基础对象模型（BOM）、仿真对象模型（SOM）和联邦对象模型（FOM）根据仿真目的从剧情模型和概念模型两方面来划分联邦及联邦成员的职责，是 HLA 仿真的对象模型，实体对象模型与 HLA 的对象模型之间没有必然的联系。仿真模型层中的联邦成员模型、仿真应用模型和联邦管理模型联合起来形成 HLA 仿真的联邦，它们每一个都是联邦的平等成员，其中联邦管理模型有效帮助联邦的集成测试和监视、控制联邦的执行。基于 HLA 的仿真系统的模型层次划分有助于仿真系统开发的分工和协作，有助于各层次模型的重用，也为各模型的快速、规范化开发打下了基础。

3.2 概念模型

概念模型是对真实世界某些关注成分或现象的一种描述，它并非最终仿真模型，但它是建模过程中的第一步。仿真首先要对仿真对象进行全面、深入的认识，建立起被研究对象的概念模型，然后在此基础上通过设计初样。HLA 对象模型，开发实现为仿真模型。开发剧情模型的目的是开发联邦剧情的功能性说明，它的主要输入是目标说明中指定的行动上、下文限制，当然已有想定数据库也可以为剧情开发提供可重用的起始点，输出就是概念分析和联邦开发所需要的剧情。想定模型中描述的实体都要被真正的可运行的仿真模型所代替。剧情模型可以实现领域专家知识向概念模型的转化，这体现出剧情模型与概念模型相互依赖关系。借助于联邦剧情可直接从 CMMS 中抽取相应的知识以形成联邦概念模型（FCM）。建立概念模型需要系统知识、工程判断力和建模工具。在建立概念模型时必须详细理解实际系统的需求和操作规则，提取实际系统的本质并滤出不必要的细节。建立概念模型的关键问题集中在合理地简化假设，确定哪些部件应包含在概念模型中，哪些互操作应包含在概念模型中。概念模型中所包含的细节及数量应根据仿真需求进行确定，那些可能引起不同决策的部件需要认真考虑，并确定仿真系统的过程流。仿真概念模型是仿真系统概念模型开发过程中十分重要的环节。仿真概念模型是指开发人员对于要仿真内容的信息描述，描述的信息包括实体、对象、算法、关系、数据、各种假设和限制。仿真概念模型是仿真开发人员把建模需求转化为详细的设计框架的纽带，它为仿真开发人员建立连续的、权威的模型与仿真表示提供了基础。根据仿真概念模型，组成仿真系统的软件，硬件，网络及系统才能够建立。仿真概念模型的开发过程一般有以下四个基本步骤：

- （1）收集有关仿真背景的权威信息；
- （2）确定仿真（分解）中要素的实体和过程；
- （3）开发仿真元素；
- （4）定义仿真元素间的互操作和关系。

3.3 对象模型

在基于 HLA 的仿真系统的设计和开发中，为了达到互操作与可重用的目标，一个关键步骤是设计和建立仿真应用的对象模型。HLA 中对象模型标准化的描述格式和内容由 HLA 对象模型模板（Object Model Template, OMT）^[3]给出，OMT 为描述 HLA 对象模型提供了标准化的方法。HLA 对象模型主要包含 FOM、SOM、MOM（管理对象模型）^[5]和 BOM^[4]四种。其中，FOM 用于描述某具体联邦中相互存在信息交换的那些联邦成员之间需交换哪些“有关对象的信息”及其具体特性，并将这些交换采用标准的格式对 FOM 进行描述；SOM 用于说明每一个联邦成员在参与联邦运行过程中能给联邦提供及需要哪些“有关对象的信息”及其具体特性，它反映联邦成员的本质能力；MOM 由 RTI 负责维护其信息，用于联邦管理、监控等功能；BOM 描述了仿真互操作某个方面的特性，是可以重用的仿真组件。HLA 对象模型的开发是基于 HLA 的仿真系统开发的重要过程，是 HLA 实现互操作和可重用的重要基础，联邦中只有确定了正确的 SOM 和 FOM，才能进一步开发实现联邦成员，进行联邦集成与测试。

FOM 和 SOM 的开发方法有多种，可以从头开发 FOM/SOM、在现有的 FOM/SOM 基础上扩展形成新的 FOM/SOM 或者使用 BOM 库创建 FOM/SOM。从无到有开发 SOM 的过程如下：

- （1）采用面向对象分析与设计的方法对系统进行分析，建立系统的对象模型。详细分析系统需求，为开发对象模型做好充分准备。
- （2）确定联邦成员对对象类/交互类的公布/发布能力。SOM 开发者要考虑其公布类的所有可能的行为，确定这些行为是否与其他的联邦成员有关。
- （3）确定联邦成员对对象类/交互类的订购/接收需求。从中识别联邦成员希望从其他成员中输入的数据及其类型。

(4) 确定联邦成员对属性、参数的公布能力。识别公布的对象类中在将来联邦以可能有用的属性和公布的交互类中的参数。SOM 开发者应负责预测交互的接收方需要哪些参数。

(5) 确定联邦成员对属性/参数的订购需求。这可以细化可订购类的描述。可订购的属性应对仿真应用有确切的语义。

(6) 采用对象模型开发工具软件生成 SOM 对应于 OMT DIF (数据交换方式) 格式的文件。

按照 HLA 标准建立了各个联邦成员的 SOM 之后, 才开始 FOM 的开发, 其开发过程如下: (1) 分析联邦的需求, 定义想定计划; (2) 确定联邦中包含的客观对象和交互, 用 SOM 充分描述联邦中要用到的成员; (3) 确定联邦成员的公布能力及订购需求; (4) 确定联邦的公布责任; (5) 确定联邦对属性和参数的需求; (6) 采用 OMDT 软件生成 FOM 对应于 OMT 格式的文件。

BOM 明确定义于 IEEE Std 1516.3-2003 HLA FEDEP 中, 其被当做构建 HLA 对象模型的便利技术和提供可重用模型组件用于快速构建和修改联邦和联邦成员。它是开发和扩展仿真与互操作环境的基本构建模块。

BOM 的目的是为便利的互操作, 重用和组合能力提供一个关键的机制。BOM 产品开发小组提出了两种 BOM 构建的方法, 根据设计需要构建 BOM 和从 FOM、SOM 或 BOM 集合中提取互操作“模式”来构建 BOM。由于 BOM 具有基于组件开发的特性, 而目前我们可以获得很多现成的 FOM 和 SOM。所以采取从 FOM 和 SOM 中进行提取构建 BOM 的方法, 在目前比较实用。采用这种方法, 首先要分析原有 FOM 的交互类表, 选择出联邦交互相对独立的交互类和与此交互类相关的对象类。然后分析这些相对独立的联邦交互相关性, 避免提取的 BOM 过于单一或相似。最后给每个 BOM 加入元数据补充, 从而方便其今后的重用。

BOM 符合组件化的软件开发思想, 推动了建模与仿真资源的重用。BOM 重用思想最大获益者是利用 BOM 构建 FOM 或 SOM 的终端建模用户。利用 BOM 创建 FOM 或 SOM 的方法有两种: 从空白开始开发的方法和在有联邦 FOM 或成员 SOM 的基础上添加 BOM 开发的方法。

3.4 仿真模型 (实现模型)

现在的仿真建模, 主要是一些专业人员在同时具备对所建模型有着专业性的知识和对计算机编程相当熟练的基础上完成的, 最终完成仿真软件, 而实际情况下并非如此, 一些对所建模型有着深入了解的专业研究人员对计算机编程并不熟练, 基于 HLA 的仿真系统开发的优势在于使专业研究人员专注于各自领域的仿真应用开发, 尽可能地利用本领域的最先进的技术, 可以使各部分相对独立开发, 与底层支撑环境功能分离开。仿真模型层建立仿真的具体应用程序, 与对象模型层和概念模型层都有联系。仿真模型有对外界的信息交换, 需要 SOM 对象模型、FOM 对象模型和 BOM 库的支持, 同时它是实体活动的容器, 涉及那些实体及实体的具体活动过程, 需要实体模型和剧情模型的支持。仿真模型的建立过程是联邦各个成员的开发过程, 主要步骤包括: (1) 创建模型运行环境; (2) 选择仿真算法; (3) 进行程序设计; (4) 运行调试。

4 结论

利用面向对象的思想 and 标准的开发规范给出了 HLA 仿真系统的建模过程, 提出并分析了一个改进的基于 HLA 的仿真系统中的模型体系。概念模型的开发为联邦的设计与开发打下了坚实的基础, 在此基础上的对象模型的设计与开发并转换成各联邦成员的仿真实现是 HLA 联邦开发中的模型问题的本质所在。使用基于 HLA 的仿真系统建模方法能够缩短开发周期, 提高建模与仿真的效率。

参考文献

[1] IEEE Std 1516-2000. IEEE Standard for Modeling and Simulation High Level Architecture (HLA)-Framework and

Rules[s]. SISC of the IEEE Computer Society.

- [2] 张琦, 尹全军, 黄柯棣. 基本对象模型概念研究[J]. 2005, 17(7):1667-1669.
- [3] 邱晓刚, 黄柯棣. HLA 对象模型的概念、描述与建立[J]. 1998, 15(1):34-37.
- [4] 黄健, 黄柯棣, 邱晓刚. 任务空间概念模型研究[J]. 系统仿真学报, 2000, 12(1):1-5.
- [5] FKUHL,R WEATHERLY,JDAHMAN. 付正军, 王永红译. 计算机仿真中的 HLA 技术. 北京: 国防工业出版社 [M]. 2003.
- [6] CALI FGREGORIE GREGORIE,Dynamic tuning of IEEE 802.11 protocol to achieve theoretical throughput limit. IEEE/ACM transaction network[J].2000.

红外小目标背景抑制和检测方法研究

秦兴桥，郎士宁，王 辉

(防空兵指挥学院，河南 郑州，450052)

摘 要：针对红外图像序列中的小目标检测问题，提出了一种新的基于 Robinson guard 滤波和双向链表的检测方法。首先使用 Robinson guard 滤波抑制背景，然后运用自适应阈值分割图像，并累积图像序列形成目标运动轨迹。最后基于目标运动的规律性、连续性，利用双向链表来检测目标。实验结果表明，所提出的方法可以得到较高的检测性能。
关键词：背景抑制；目标检测；Robinson guard 滤波器；双向链表
中图分类号：TP391.4 **文献标识码：**A **文章编号：**1006-7043 (2010) xx-xxxx-x

Research on IR Small Target Detection and Backgroud Suppression

QIN Xingqiao, LANG Shining, WANG Hui

(Air Defense Forces Command Academy, Zhengzhou 450052,Henan, China)

Abstract: A new detection method is presented for the detection of dim small tar gets from infrared image sequences is pre - sented. Firstly, Robinson guard filter suppresses the background clutter in single frame. Secondly, a auto-adaptive threshold algo- rithm is applied to segment the potential targets from background. Finally, the bi-direction chain is used to identify the real target base on the continuity of moving and the consistency of track of the tar get in sequential images. Experimental result shows tha t proposed method is feasible.
Keywords: background suppression; target detection; robinson guard filter; bi-direction chain

1 引言

红外图像小目标检测与识别是近程主动拦截武器系统的重要任务。由于来袭目标通常为高速运动的小目标，在成像背景中往往只占几个像素，没有形状和纹理等特征可以利用，而且由于大气云层和光照等影响加之传感器本身的系统噪声，图像中包含较强的背景起伏和杂波干扰，使得信噪比降低、目标检测困难^[1]。红外小目标检测算法一般先要经过预处理进行背景抑制，而后分割出目标。背景抑制的结果直接决定能否从图像中分割出目标。常见的典型目标检测方法有：

(1) 采用基于图像均值和直方图的自适应门限背景抑制与目标分割方法^[1]，并用流水线管道结构进行运动轨迹估计。但是，该方法需要预先估计目标在图像中的面积，并且，当如果原始图像的直方图较为均衡时，该方法需要进行多次迭代，算法的运行时间也会随之增加。

(2) 基于温度场非线性分布，利用当视场角较小时同一行的温度差较小，并且相邻行间有较强的相关性的特点，将红外图像的像素灰度减去上一行或同一行的灰度均值，来抑制背景噪声^[2]。但这种方法不适用于宽视场角且背景有较大起伏的情况。

(3) 在时间剖面上，采用一维中值滤波和形态学的方法滤除由于运动小目标经过时引起的灰度扰动^[3]。但是，该方法如果局部背景和目标灰度接近，中值滤波器就失去其作用。

(4) 采用 tophat 算法进行单帧背景抑制^[4]，自用奇偶差分多帧叠加来增强目标点。但是，所需帧数较多，不利于实时检测。

作者简介：秦兴桥（1976—），男，讲师，硕士；
郎士宁（1973—），女，讲师，硕士；
王辉（1974—），男，讲师，硕士。

(5) 采用小波变换进行图像分割^[5]，该算法的缺点是复杂度高，实时性较差。

(6) 基于移动式加权管道滤波进行弱小目标的检测^[6]，该算法的不足之处在于，有多个候选目标需要建立多个管道，且管道的直径和数量不能随目标的变化而变化。

本文提出了一种采用 Robinson guard 滤波器算法，首先对单帧图像进行背景抑制，得到由目标和背景残值等少数孤立噪声点形成的预处理图像，再将这些图像进行帧间累加。由于运动目标在帧间运动的连续性，目标会在帧中保持一定的运动方位的趋势，而噪声则不会有这种特点。在累积帧上利用双向链表构造目标航迹，以链表长度判决航迹的起始，从而可以快捷地检测出真实的目标。

2 用 Robinson guard 滤波器进行背景抑制

红外目标成像的数学模型为^[3]：

$$f(x, y, k) = f_T(x, y, k) + B(x, y, k) + N(x, y, k) \quad (1)$$

式中， f 为红外图像序列中第 k 帧图像的灰度值； f_T 、 B 、 N 分别为目标、背景和噪声。

原始图像 I 可以看做背景图像 I_b 叠加上目标和噪声的细节图像 I_d ，即

$$I = I_b + I_d \quad (2)$$

目标信噪比，定义如下：

$$\text{SNR} = \frac{G_T - G_B}{\sigma} \quad (3)$$

式中， G_T 为目标的灰度均值； G_B 为背景的灰度均值； σ 为背景方差。

Robinson guard 空间滤波器是一种典型的非线性非参数型滤波器，它也是一种边缘增强滤波器。该滤波器对起伏背景与杂波有很好的抑制性能。与普通的边缘增强滤波器相比，该滤波器最大的优点是在待检测像素周围设置保护带，当目标尺寸小于保护带时，比背景亮或暗的目标都可以得到很好的保留，信息损失较小。

Robinson guard 算法采用的自适应模板 Z 完全以图像像素的灰度值作为模板取值：

$$Z = \begin{bmatrix} f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7 \\ f_{24} & 0 & 0 & 0 & 0 & 0 & f_8 \\ f_{23} & 0 & 0 & 0 & 0 & 0 & f_9 \\ f_{22} & 0 & 0 & 0 & 0 & 0 & f_{10} \\ f_{21} & 0 & 0 & 0 & 0 & 0 & f_{11} \\ f_{20} & 0 & 0 & 0 & 0 & 0 & f_{12} \\ f_{19} & f_{18} & f_{17} & f_{16} & f_{15} & f_{14} & f_{13} \end{bmatrix} \quad (4)$$

式中， f_i ($1 \leq i \leq 24$) 为滤波窗口内相应阵元上的图像函数值。

其滤波器算法表示为：

$$X = \begin{cases} X - \max(f_i) & \text{if } (X \geq f_i) \\ \min(f_i) - X & \text{else if } (X \leq f_i) \\ 0 & \text{else } (\min(f_i) < X < \max(f_i)) \end{cases} \quad (5)$$

3 自适应阈值分割

对于已经去除背景的图像，还要从中进行阈值分割，来提高信噪比。采用小波变换进行图像分割时，采用自适应的阈值法对图像做二值化处理^[7]，阈值 T 被定义为：

$$T = m + c\sigma \quad (6)$$

式中， m 为图像的均值； σ 为图像的标准差； c 为常数，可以通过试验确定，通常取值为 10~20。

采用这种方法，最明显的局限性在于 m 和 σ 描述的是图像的全局特性，当目标较暗，尤其是当目标穿越云层灰度接近周围邻域的杂波灰度均值时，按照式（6）进行分割时，目标灰度 G_T 就会小于阈值 T ，这样容易造成目标漏检。

定义局部均值 m_l ，对于 5×5 的模板，有：

$$m_l = \frac{1}{25} \sum_{i=1}^5 \sum_{j=1}^5 f(i, j) \quad (7)$$

重新定义阈值如下：

$$T' = \begin{cases} m_l + c\sigma & m_l \leq m \\ T & m_l > m \end{cases} \quad (8)$$

利用新的阈值，对已经去除背景且只包含潜在目标和少数噪声点的图像 I_d 进行二值化处理：

$$I_{bw}(x, y) = \begin{cases} 1 & \text{if } (I_d(x, y) > T') \\ 0 & \text{dse} \end{cases} \quad (9)$$

4 用帧间相关性进行目标判决

利用帧间候选目标的相关性是提高检测概率、降低虚警概率的重要手段。假设目标的运动方向是任意的，帧间最大速度为 $V_{\max}$ （单位像素）。由于目标运动的连续性，如果在第 i 帧中像素 (x_0, y_0) 处有目标，则该目标在第 $i+1$ 帧中必然会出现在该像素的一个小邻域 $N(x_0, y_0)$ 内， $N(x_0, y_0)$ 的大小由目标的最大速度 $V_{\max}$ 确定^[8]。目标灰度在帧间的波动很小，由于噪声的出现呈现随机性，所以不具有上述性质。

构造一个长度为 K 帧的与管道，将原始图像序列和阈值分割后的图像从第一帧开始，分别进行与操作，得到保留原始灰度的新的细节图像序列 $S = \{I'_{di} \mid i=1, 2, \dots, K\}$ 。然后将图像序列 S 进行累积得到目标轨迹图像 I_{track} 。

$$I_{\text{track}} = \sum_{i=1}^K I'_{di} \quad (10)$$

为了找到 I_{track} 上的目标轨迹，将图像中灰度不为 0 的像素坐标按灰度等级从左至右，从上至下进行整理存放在一个数组序列中，数组定义形式如下：

$$\text{array}_1 = \bigcup_{\substack{i, j \in I_{\text{track}} \\ f(x_i, y_i) = l}} (i, j), \text{ 其中, } 0 < i \leq f_{\max}$$

目标的运动轨迹必然存在于数组序列中，可以通过链表轨迹相关性进行判决。

定义如下双向链表节点：

```
typedef struct DNode
{
    int x;
    int y;
    int value;
    DNode* prior;
    DNode* next;
};
```

判决步骤主要包括链表搜索和目标标定。

步骤 1：链表搜索

按数组序列顺序读入一个灰度数组数据；从数组的第一个点开始构造链表的第一个节点，在该数

组内以 $r=V_{\max}$ 为半径，以某一搜索方位进行半圆周邻域搜索。如果在邻域内找到候选点，则将该候选点加入链表，继续向前搜索；如果没有候选点，则以搜索方位的反方向作为新的搜索方位进行半圆周搜索。当找到候选点且不是当前节点的前驱节点时将该候选点加入链表。如果在圆周邻域找不到候选点则将该点视为孤立的噪声点并从链表中删除，链表后退。当链表为空或者链表长度不再增长时，从数组的下一个点开始重新构造链表。

重复上述搜索过程，同时判断链表长度，当链表长度大于指定门限 T_{list} ($\frac{1}{2}K < T_{\text{list}} \leq K$)，则认为该链表存放的是候选目标航路。

步骤 2：目标标定

按候选链表上的像素位置，在原始图像序列的当前帧上标定目标位置；如果没有候选目标航路则认为搜索失败，判决结束。

该算法采用先进先出的排队管理方式进行图像的更新，继续进行上述检测过程，直到处理完整个图像序列为止。

5 实验结果

单目标穿越云层时的图像序列如图 1 所示。

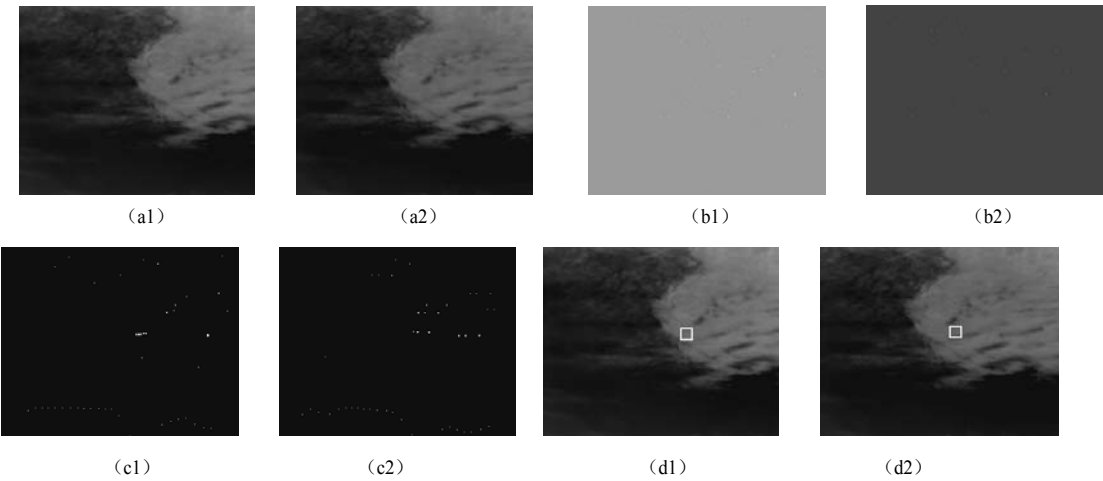


图 1 原始图像的检测过程及结果

a1 为第 640 帧，a2 为第 648 帧；图像尺寸为 320×256，信噪比分别为 SNR=1.585 和 SNR=1.367，目标几乎被云层淹没。b1 和 b2 为经模板大小为 7×7 的 Robinson guard 滤波后的结果。c1 和 c2 为经自动阈值分割后的 5 帧运动轨迹累积（为增强视觉效果图片做了亮度和对比度调整）。d1 和 d2 为标记的结果。从图 1（b2）上可以看到经过 Robin guard 滤波后，大面积的云层背景被充分滤除，剩余的是一些点状随机噪声。自适应阈值分割，使得在云层里的暗目标得到保留，并且在累积帧上形成轨迹。

6 结论

大量的试验结果表明，Robinson guard 滤波器在背景抑制方面具有良好的表现；由于使用了自适应阈值分割，和周围背景较接近的暗目标也能检测；帧累积可得到目标的大致轨迹，可以克服少数帧目标缺失的影响。采用双向链表构造目标航迹，方法简便易行运算速度快。算法的不足之处在于

Robinson guard 对斑状目标和线状目标保护性差，因此，本文所提的方法对慢速运动的线状目标或背景内有长斑状物体干扰的情况不能取得很好的检测效果。

参考文献

[1] 彭嘉雄, 周文琳. 红外背景抑制与小目标分割检测[J]. 电子学报, 1999, 27(12):47-51.

[2] 张弘, 赵保军, 毛二可. 低信噪比下抖动的红外弱小目标的实时检测[J]. 激光与红外, 2001, 31(4):225-227.

[3] 苏新主, 姬红兵, 高新波. 一种基于数学形态学的红外弱小目标检测方法[J]. 红外与激光工程, 2004, 33(3):307-310.

[4] 张文超, 王岩飞, 陈贺新. 基于 Tophat 变换的复杂背景下运动点目标识别算法[J]. 中国图像图形学报, 2007, 12(5):871-874.

[5] 李燕苹, 谢维信, 裴继红. 基于小波变换的红外弱小目标检测新方法[J]. 红外技术, 2006, 28(7):419-421.

[6] 刘靳, 姬红兵. 基于移动式加权管道滤波的红外弱小目标检测[J]. 西安电子科技大学学报 (自然科学版). 2007, 34(5):743-747.

[7] WEN Pei-zhi, SHI Ze-lin, YU Hai-bin. A Detection Method for IR Point Target on Sea Background Based on Morphology [J]. Opto-Electronic Engineering, 2003, 30(6):55-58.

[8] 李吉成, 沈振康. 红外起伏背景下运动点目标的检测方法[J]. 红外与激光工程, 1997, 26(6):8-13.

基于马尔可夫链模型的软件可靠性测试研究

何 焱¹, 张来顺¹, 石荣刚²

(1.解放军信息工程大学电子技术学院, 河南 郑州, 450004; 2.西安通信学院, 陕西 西安, 710106)

摘 要: 基于马尔可夫链模型的可靠性测试方法是软件可靠性测试的经典方法之一, 本文就是在基于马尔可夫链模型的基础上, 把软件测试问题转化为一个经典的数学问题。通过马尔可夫链模型构建使用链, 根据使用链进行序列抽样, 产生测试用例, 从而将软件测试结果的分析问题转化为一个经典概率问题。最后运用实例证明基于马尔可夫链的软件可靠性测试方法具有较好的实用性和有效性。

关键词: 可信软件; 软件可靠性测试; 马尔可夫链模型; 测试用例生成

中图法分类号: TP391 文献标识码: A

The Research of Software Reliability Testing Based on Markov Chain Model

HE Yan¹, ZHANG Laishun¹, SHI Ronggang²

(1.Institute of Electronic Technology of the PLA Information Engineering University, Zhengzhou 450004, Henan China
2.Xi'an Communication Institute Xian 710106, Shanxi China)

Abstract: The Markov model-based method is a classical and important software reliability testing method. In this paper ,we actually turned test problem into mathematics problem based on the Mark ov Chain model.Based on Markov Model , We build using chain, and then carry on in to sequen ce sampling and generate test case.at last, it turned test result analysis problem i nto classical probability problem . At the end of the paper ,we approve the method of software reliability testing based on Markov Chain Model is of practicability and availability by instance.

Keyword: trusted software; software reliability testing;markov chain mode;test cases generation

软件测试过程是软件工程领域必不可少的过程, 在软件生存周期中占有非常重要的位置。据统计, 软件开发总成本中, 用在测试上的开销要占 30%~50%, 特殊情况下, 对可靠性要求很高的软件, 其测试费用甚至高达所有其他软件工程阶段费用总和的 3~5 倍^[1]。本文在对传统测试用例的生成方法学习的基础上, 探索出了一种实用性强的方法, 该方法在马尔可夫链模型的基础上, 提出了一种软件可靠性测试模型, 并通过实例证明了这种模型的实用性和正确性。

1 马尔可夫链模型 (Markov Chain Usage Model) ^[2]

马尔可夫链是满足下面两个假设的一种随机过程:

假设 1: $t+1$ 时刻系统状态的概率分布只与 t 时刻的状态有关, 与 t 时刻以前的状态无关;

假设 2: 从 t 时刻到 $t+1$ 时刻的状态转移与 t 的值无关。一个马尔可夫链模型可表示为 $M=(S, P, Q)$, 其中各元素的含义如下:

(1) S 是系统所有可能的状态所组成的非空的状态集, 有时也称为系统的状态空间, 它可以是有限的、可列的集合或任意非空集。本文中假定 S 是可数集 (即有限或可列)。用小写字母 i, j (或 S_i, S_j) 等来表示状态。

作者简介: 何焱 (1984—), 男, 硕士;
张来顺 (1963—), 男, 教授, 硕士;
石荣刚 (1982—), 男, 硕士。

(2) $P = [P_{ij}]_{n \times n}$ 是系统的状态转移概率矩阵, 其中 P_{ij} 表示系统在时刻 t 处于状态 i 而在下一时刻

$t+1$ 处于状态 j 的概率, N 是系统所有可能的状态的个数。对于任意 $i \in S$, 有 $\sum_{j=1}^n P_{ij} = 1$ 。

(3) $Q = [q_1, q_2, \dots, q_n]$ 是系统的初始概率分布, q 是系统在初始时刻处于状态 i 的概率, 满足

$$\sum_{j=1}^n q_{ij} = 1。$$

在软件测试过程中, 对于被测软件, 基于逻辑覆盖和基本路径测试方法生成的测试用例, 覆盖了程序的所有测试, 保证了在测试中程序的每个可执行语句至少执行一次。由此可知, 这些测试用例构成了一个完整的测试实验样本空间; 生成的每个测试用例一定发生并且发生的概率相同。因此, 可以把测试用例的生成及其发生概率和马尔可夫链模型联系起来。把一个完全正确的逻辑结构或模块对应的状态集看成在时刻 t 处于的状态集, 则状态集中的每一个状态对应一个测试用例, 从而构成一个完整的满足马尔可夫链模型的测试用例集; 同样, 把接下来要测试的逻辑结构或模块所处的状态看做下一时刻 $t+1$ 处于的状态。通过以上对模型的分析, 可以初步构造一个该模型的软件可靠性测试的结果分析准则, 描述如下:

(1) S 是被测逻辑结构或模块所对应的测试用例集, 它应该是有限的、可列的集合或任意非空集;

(2) $Q = [q_1, q_2, \dots, q_n]$ 是时刻 t 测试用例的发生概率分布, q 是在 t 时刻测试用例 i 的发生概率, 满足

$$\sum_{j=1}^n q_{ij} = 1。其中, q_1 = q_2 = \dots = q_n, 即在 t 时刻, 测试用例集中的每个测试用例发生概率相同。$$

结果分析准则的概率表示形式为 $P_z = (D, H)$ 。

(1) D 是被测逻辑结构或模块所对应的测试用例集, 它应该是有限的、可列的集合或任意非空集;

(2) H 是时刻 t 测试用例的发生概率分布。其中 $H = [h_1, h_2, \dots, h_n]$, $P(\{h_i\})$ 表示 t 时刻第 i 个测试用例发生的概率, 满足:

$$P(\{h_1\}) = P(\{h_2\}) = \dots = P(\{h_k\}) = 1/k \quad (1)$$

$$P(\{h_1\}) + P(\{h_2\}) + \dots + P(\{h_k\}) = 1 \quad (2)$$

即在 t 时刻, 测试用例集中的每个测试用例全都发生, 并且发生概率相同。

2 马尔可夫链的基本概念^{[3]~[5]}

2.1 状态和状态转换

基于马尔可夫链的软件使用模型是由软件的状态和边组成。状态表示软件在使用过程中的内部环境。状态转换是指当软件在某一状态经输入激励, 从该状态转换到另一个状态。

2.2 输入激励与状态转换概率

软件处于某一稳定的内部状态, 外界环境有相应的输入激励, 激励可以是不同的输入变量或者相同的输入变量取不同的值。不同的输入激励将导致软件不同的状态转换。当软件在稳定的使用环境下, 不同的输入激励的出现是遵循一定的统计分布的, 因而导致软件状态转换间也存在相应的概率分布, 这个概率就称为状态转换概率。如果遵循软件的状态转换概率分布抽样产生测试输入序列, 则体现了统计意义上的软件使用方式。

2.3 软件的使用链和测试链

软件的使用链是用马尔可夫链描述的软件载誉期使用环境中的状态转换模型, 用 U 表示。定义为: $U = \{S, ARC\}$, 其中 S 代表软件的状态集, 有 $S = \{s_1, s_2, \dots, s_n\}$; 而 ARC 表示软件状态之间转换关

系，有：

$$ARC=\{arc_{11}, arc_{12}, \cdots, arc_{1n}, arc_{21}, arc_{22}, \cdots, arc_{2n}, \cdots, arc_{n1}, arc_{n2}, \cdots, arc_{nn}\} \quad (3)$$

而其中的每个状态转换关系是个二维向量。用 D 表示引起软件状态转换的输入激励域， P 表示软件状态转换率，则有：

$$arc_{ij}=(d_{ij}, p_{ij}) (d_{ij} \in D; i=1, 2, \cdots, n; j=1, 2, \cdots, n) \quad (4)$$

$U=\{S, ARC\}$ 中，应用第 1 部分中的结果分析准则，要使使用链的状态转换概率为 1，必须满足第 1 部分中的①和②，即使用链在时刻 t 的测试用例集 D 应满足完全覆盖。

软件的测试链是用于记录软件测试历史情况的软件状态转换链，用 T 表示。定义为： $T=\{S', ARC'\}$ ； S' 是软件状态和软件故障状态的集合，有

$$S'=\{s_1, s_2, \cdots, s_n, s_{n+1}, \cdots, s_m\} \quad (5)$$

其中，从 s_1 到 s_n 是软件的正常状态，也就是软件使用链中包含的状态；从 s_{n+1} 到 s_m 是软件的故障状态。 ARC' 代表软件状态之间的转换关系，即

$$ARC'=\{arc_{11}, arc_{12}, \cdots, arc_{1m}, arc_{21}, arc_{22}, \cdots, arc_{2m}, \cdots, arc_{m1}, arc_{m2}, \cdots, arc_{mm}\} \quad (6)$$

而其中的每个状态转换关系也是个二维向量。用 D 表示引起软件状态转换的输入激励域， P' 表示软件状态转换率，这其中不仅包含了软件正常状态之间的转换率，还包含了正常状态和故障状态之间的转换率：

$$arc_{ij}=(d_{ij}, p'_{ij}) (d_{ij} \in D; i=1, 2, \cdots, m; j=1, 2, \cdots, m) \quad (7)$$

从上面的介绍的测试链 T 和使用链 U 的概念可以看出，二者的主要区别如下：

(1) 测试链 T 中状态集 S' 除了包含使用链 U 中包含的正常状态外，如果测试中发现故障，那么还包含软件的失效状态。所以，测试链 T 的状态转换关系也比使用链 U 多一些失效状态的转换关系。

(2) 测试链 T 中的 p' 值记录的是测试过程中软件状态转换概率，是测试概率而不是使用链 U 中所代表的使用概率。

3 基于马尔可夫链使用模型的软件可靠性测试方法

从上看出，马尔可夫链使用模型也是一种很好的软件使用方式的方法。它将软件连续的运行序列看做一个马尔可夫链，状态之间的关系以转换概率来表达。因此，只要有足够的时间，马尔可夫链不同状态的转换概率将到达一个稳定状态，从而可以用稳定状态概率来从统计学意义上描述软件的使用方式。

在基于马尔可夫链使用模型的软件统计测试过程中，首先要根据软件需求规范建立软件的使用链：然后根据使用链进行序列抽样，产生测试用例；最后，执行测试用例，建立相应的测试链，通过测试链和使用链的比较，可以判断测试充分性。利用测试用例集驱动程序运行，获得输出数据后，利用评判准则进行评判；如果所有测试用例都发生，并且测试数据与预期数据相符，则说明测试成功，可进行下一时刻的测试。在进行每一时刻的测试中，都应满足评判准则。

在进行测试过程中，如果测试结果显示，说明被测软件逻辑结构不满足完全覆盖性，测试无须再进行检查，再进行第二步判定；如果测试结果数据与预期数据不相符，说明被测软件的逻辑结构或结构中的语句本身不正确，则对其重新进行检查。

当然，在软件可靠性增长测试过程中，利用基于马尔可夫链使用模型的统计测试所产生的失效数据作为软件可靠性增长模型的输入，可以获得相应的可靠性估计或预计。在软件可靠性验证测试过程中，直接利用马尔可夫链使用模型产生一定量的测试用例，通过对失效数据的观察，便可以获得对软件的可靠性验证。所以这里只关注如何建立相应的马尔可夫链软件使用模型，以及相应的测试用例的抽取方法。

4 测试方法示例

本文以一个菜单软件在某时刻 t 的使用链构建过程作为实例，说明软件使用链和测试用例的构建方法。

待测试的目标软件启动运行后，通过下拉菜单界面与用户交互。用户可以用向上键、向下键和 Enter 键控制软件。当软件打开文件时，“保存文件”和“打印文件”这两个菜单选项才可使用，否则两个菜单项不可使用。其中菜单选项（Menu Chosen-MC）；是否已经打开（File opened-FO）。

设这个软件的使用链为 $U=\{S, ARC\}$ ，则它的构建步骤如下：

● 确定软件状态集合

前面已经提到，软件状态指的是软件相对稳定的内部环境，这种内部环境决定了软件的某个输入激励是否可能或者合法。在确定软件状态时，需要仔细分析软件的每个输入激励及使用相关输入激励的信息。就该简单菜单选项软件来说，其输入激励是鼠标单击或者光标移动键与 Enter 键的组合，显然有两个因素决定着这些输入的作用（软件内部环境）：选中的菜单选项和是否已经打开文件。很明显，当选中的菜单为“打开文件”和“保存文件”时，软件的状态不同，鼠标单击的输入的结果不同，而且，当选中的菜单同为“保存文件”时，已经打开文件 and 没有打开时软件在输入激励作用下的反应也不一样。所以，这两个内部环境变量决定了软件的 8 个状态，再加上软件的起始状态和结束状态，软件有 10 个状态，最后，进一步分析，会发现软件在执行“打开文件”、“保存文件”和“打印文件”功能时，存在额外的 3 个软件状态，于是软件共 13 个状态，则该简单菜单软件的状态集合为：

$S=\{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}\}$ ，对这 13 个软件状态的描述如表 1 所示。

表 1 菜单软件状态表

MC		FO	MC		FO
s_1	打开文件	否	s_5	打开文件	是
s_2	保存文件	否	s_6	保存文件	是
s_3	打印文件	否	s_7	打印文件	是
s_4	退出	否	s_8	退出	是
	功能			功能	
s_9	打开文件		s_{10}	保存文件	
s_{11}	打印文件		s_{12}	软件未被调用状态	
s_{13}	软件结束状态				

● 状态转换关系的图形表达

软件状态的转换是由输入激励引起的，所以应该首先找到软件的输入域。然后分析软件状态之间的关系，特别是分析软件状态之间是否存在转换关系。

对于这个菜单软件，用户可以用向上键↑、向下键↓和 Enter 键控制软件，所以软件的输入域为 $D=\{\uparrow, \downarrow, Enter\}$ 。而后对于 11 个软件状态（除去软件结束状态和起始状态），可以把它们分成三组：第一组是 $s_1、s_2、s_3、s_4$ ；第二组是 $s_5、s_6、s_7、s_8$ ；第三组是 $s_9、s_{10}、s_{11}$ 。前面两组状态之间不存在直接的转换关系，而第三组只是一种暂时的软件内部处理状态，它只是与第二组软件状态发生交互。用椭圆表示出软件的状态，然后分析软件每一状态在相应软件输入激励作用下的状态转换，并用弧线表示，最后标示出输入激励符号。这样，一个用图形表示的软件使用链就完整地构建出来了，如

图 1 所示。

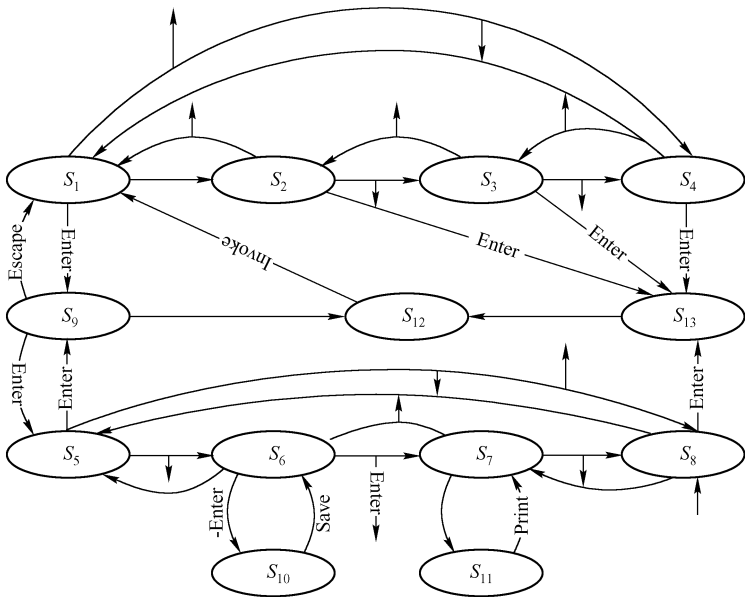


图 1 菜单示例的使用链

● 确定软件状态转换概率

软件的状态转换概率表现了软件统计意义上的使用方式，用 p_{ij} 表示源状态 s_i 到目的状态 s_j 的状态转换概率，有 $0 \leq p_{ij} \leq 1.0$ ， $\sum_{j=1} p_{ij} = 1$ 也验证了本文提出的测试结果评判准则。

在基于马尔可夫链使用模型的软件统计测试过程中，转换概率决定了对不同输入激励的选取，进而决定了不同软件状态出现的相对次数，可见，精确地估计出软件状态转换概率意义重大。估计的依据来源于相关软件的使用日志、目标软件的需求规范、使用规范、测试目标、使用和测试限制，甚至在必要时通过对建立的原型进行使用统计。在对软件状态转换概率估计非常困难时，可认为其是均匀分布的。至此，该简单菜单软件的使用链 U 已经构建完成。至此，该简单菜单示例的使用链 U 已构建完成。

5 基于使用链产生统计测试用例

测试用例生成问题直接决定了测试用例集的效率，对软件可靠性测试工作的实际展开起着重要的作用。目前，软件可靠性测试的测试时间长、费用高、资源消耗大，对加速测试提出了迫切的要求。这就需要有更好的测试用例生成过程。

可以对可靠性测试的测试用例生成作如下形式化描述：可靠性测试的测试用例生成可以描述为具有如下性质的过程 $F: T=F(D)$ ， T 的统计特征与 D 的统计特征一致， $Reliable(T) \Rightarrow Reliable(D)$ 成立。也就是说，测试用例生成是从输入域中找出满足可靠性要求的有限子集的过程。

在本菜单示例中当软件的使用链构建完毕后，就可以手动或者利用开发的 CASE 工具自动生成需要量的统计测试用例。基于马尔可夫使用链产生的测试用例是根据软件状态转换概率抽样产生的，由初态开始并经过若干中间状态到达终态的导致状态转换的输入激励序列。不同的测试用例的序列长度不一定相等。产生测试用例时，从初态开始，在每一个状态都生成一个 $0 \sim 1$ 之间的随机数，根据这个随机数选择这个状态的一条出边，转移到下一个状态，周而复始，直到终态。于是，测试用例便生成了。由于这样产生的测试用例是严格遵循软件的使用链并按照状态转移概率随机生成的，所以能够

很好地体现软件的真实使用方式。

6 结论

本文基于马尔可夫链模型，提出一种软件可靠性测试模型：根据软件需求规范建立软件的使用链：然后根据使用链进行序列抽样，产生测试用例进行测试，利用提出的评判准则对测试结果进行分析。通过实例证明，这种方法可以降低测试的复杂度，解决了简化测试难度的问题。

参考文献

- [1] 覃志东. 高可信软件可靠性和防危性测试与评价理论研究[D]. 成都: 电子科技大学, 2005.
- [2] 张德平, 聂长海, 徐宝文. 软件可靠性评估的重要抽样方法[J]. 软件学报, 2009, 20(10), 2859-2866.
- [3] Whittaker J.A. ACM Trans.Software Engineering and Methodology. 1993,2(2):93-103.
- [4] 赵亮, 王建民, 孙家广. 软件易测性和软件可靠性关系研究[J]. 计算机学报, 2007, 30(6), 986-991.
- [5] Whittaker J.A. Stochastic software tesing. The Annals of Software Engineering. 1997,4:115-131.
- [6] Whittaker J.A, Thomason M G.. A Markov Chain model for statistical Software testing . IEEE. Transactions on soft-ware Engineering. 1994,20(10):812-824.
- [7] Kirk sayre. Improved techniques for software testin g based on Markov chain usage models[D] . Knoxville:University of Tennessee, 1999.
- [8] John D Musa. Software reliability engineering[M]. New York: The McGraw-Hill companies, Inc. 1999.
- [9] Whittaker J.A, Markov Chain techniques for software testing and reliability analysis . PhD Dissertation, Department of Computer Science, University of Tennessee, Knoxville,TN. 1992.

VXI 总线在军事装备检测系统中的应用

王书伟¹, 杨 静², 丁彦芳¹

(1. 防空兵指挥学院, 河南 郑州, 450052; 2. 河南省电子产品质量监督检验所, 河南 郑州, 450003)

摘 要: VXI 总线技术是当今计算机测控技术发展的主流, 是自动检测设备 (ATE) 标准化技术的核心。将 VXI 总线技术在武器装备测试系统中推广应用具有很大的现实意义。本文结合研发工作实际, 论述了 VXI 总线自动测试系统的架构和开发 VXI 专用模块功能的一般方法和过程; 介绍了某型号武器装备的 VXI 总线测试系统; 并对 VXI 自动测试系统的发展做了展望。

关键词: VXI 总线技术; 武器装备; 测试系统; 应用展望

中图分类号: TP311.56 文献标识码: A 文章编号: 1006-7043 (2010) xx-xxxx-x

Application of the VXI Bus in the Military Equipment Detection System

WANG Shuwei¹, YANG Jing², DING Yanfang

(1. Air Defense Forces Command Academy, Zhengzhou 450052, Henan China;

2. Henan Supervision & Testing Institute for Electronic Products Quality, Zhengzhou 450003, Henan China)

Abstract: VXI bussing technique is now the computer observation technological development mainstream, is the automatic checkout equipment (ATE) the standardized technology core. The application VXI bussing technique in the weaponry test system to have great practical significance. This combination of the research and development work reality, elaborated VXI bus automated test system's construction and develops VXI special-purpose module function the general method and the process; Introduced some model weaponry VXI bus test system; And has made the forecast to VXI automated test system's development.

Keywords: VXI bussing technique; weapons; test system; application prospect

1 引言

VXI (VMEbus Extensions For Instrumentation) 总线代表着当今自动测试领域的发展方向。它是当前比较成功的新一代仪器测控平台, 它集中了 VME 总线的高速通信、GPIB 总线的易组合和 CAMAC 总线的模块化等诸多特点, 具有严格的系统同步、高可靠性和良好的模块互操作性^[1]。

VXI 总线促进了整个测试系统向开放式、集成化方向发展, 推动了测试仪器的标准化、模块化和通用化进程, 使系统的软件和硬件资源获得共享。它吸取了上述三种总线的全部优点, 并结合仪器测量系统的特点, 又增加了许多新的性能, 如“零槽”模块功能、资源管理器、配电、冷却和电磁兼容等。

2 开发 VXI 总线测试系统的一般过程

开发 VXI 总线测试系统的过程通常按图 1 所示的步骤进行。

作者简介: 王书伟 (1955—), 男, 副教授, 学士;
杨静 (1975—), 女, 工程师, 硕士;
丁彦芳 (1979—), 女, 讲师, 硕士。

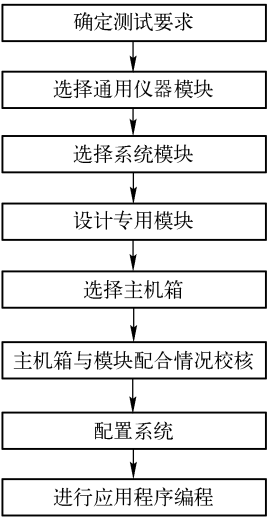


图 1 开发 VXI 总线测试系统的一般方法

3 VXI 总线在武器装备测试系统中的应用

3.1 某型号武器装备的 VXI 总线测试系统

由主控制计算机、基于 VXI 数据采集总线机箱、VXI 功能模板和应用数据采集软件组成数据采集单元，完成对被测设备的数据测试测量^[3]。VXI 测试系统功能图如图 2 所示。

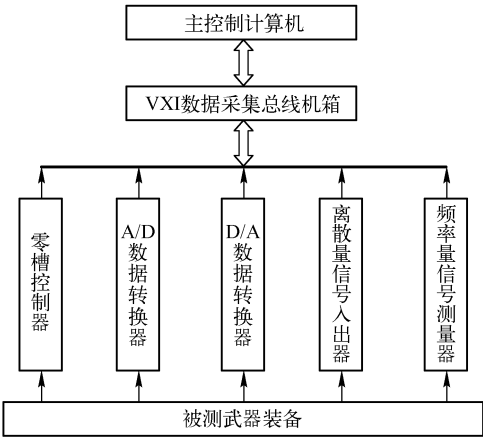


图 2 VXI 测试系统功能图

依据某型号武器装备综合测试系统要求，设计了基于 VXI 总线系列常规的装备数据信息采集系统。主要测试测量的参数包括：模拟量信号、离散量信号、频率量信号。主要功能是实现模拟量信号测试、离散量信号测试、频率量信号测试。

3.2 测试方法

数据采集（模拟量测试），采用传感器完成数据采集与转换。被测装备模拟信号输出的采用直接信号输出的方式，结构简单，能够满足常规数据采集系统要求。频率量信号测量器，完成对被测设备的频率信号、周期信号的测量。

3.3 设备选择

常规数据采集系统，由主控计算机、VXI 机箱及零槽模块、A/D 模块、D/A 数据转换器、离散量信号入出器、频率量信号测量器模块组成。

3.4 测试系统软件平台的选择

测试软件是为有效地运用硬件系统资源、实现各种测控功能而提供的程序系统及有关资料库的集合。除了 Windows 下通用的编程环境外，目前比较流行的开发平台有以下三种：NI 公司的 LabWindows/CVI、NI 公司的 LabView、HP 公司的 HP VEE。

VXI 测试系统软件结构关系示意图如图 3 所示。

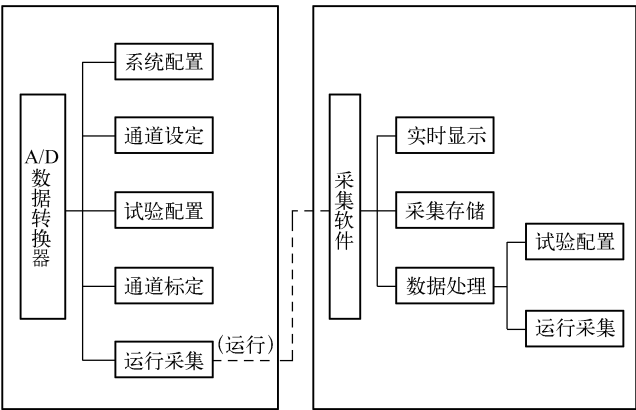


图 3 VXI 测试系统软件关系结构示意图

4 VXI 总线雷达自动测试系统总体构成

4.1 测试对象特性及主要测试任务

选择了国产某新型飞航导弹上的导引头（末制导雷达）作为待组建的 VXI 总线系统的测试对象。这种雷达需要测试的参数可分为四大类、23 个项目、107 个参数。四大类测试分别是低压测试、高压测试、抗干扰测试和向微波暗箱测试^[2]。这些项目中既有众多低频参数，又有大量高频参数、除了有 DC、AC、电流共 20 路参数外，还有频率参数、功率参数、时间参数（或距离参数）；既有模拟量测量（且动态范围宽），又有开关量（向雷达发送指令，检查雷达上的各种指令）测量。选用末制导雷达作为待组建的 VXI 总线系统的测试对象，可充分检验 VXI 总线系统的优越性。

4.2 系统构成

根据测试对象的特性及主要测试任务，用 VXI 总线组建的雷达自动测试系统的硬件原理方框图如图 4 所示。硬件主要由三大部分组成，分述如下：

4.2.1 外部主控制器

选用 PC，该微机配置有：120Gbyte 硬盘驱动器一个，512Mbyte 内存，还有 VGA 显示器及其外设与接口。

4.2.2 测试（控制）机柜

测试机柜中安装五个机箱，从上到下依次是：程控干扰信号源、程控目标信号源、VXI B 尺寸主

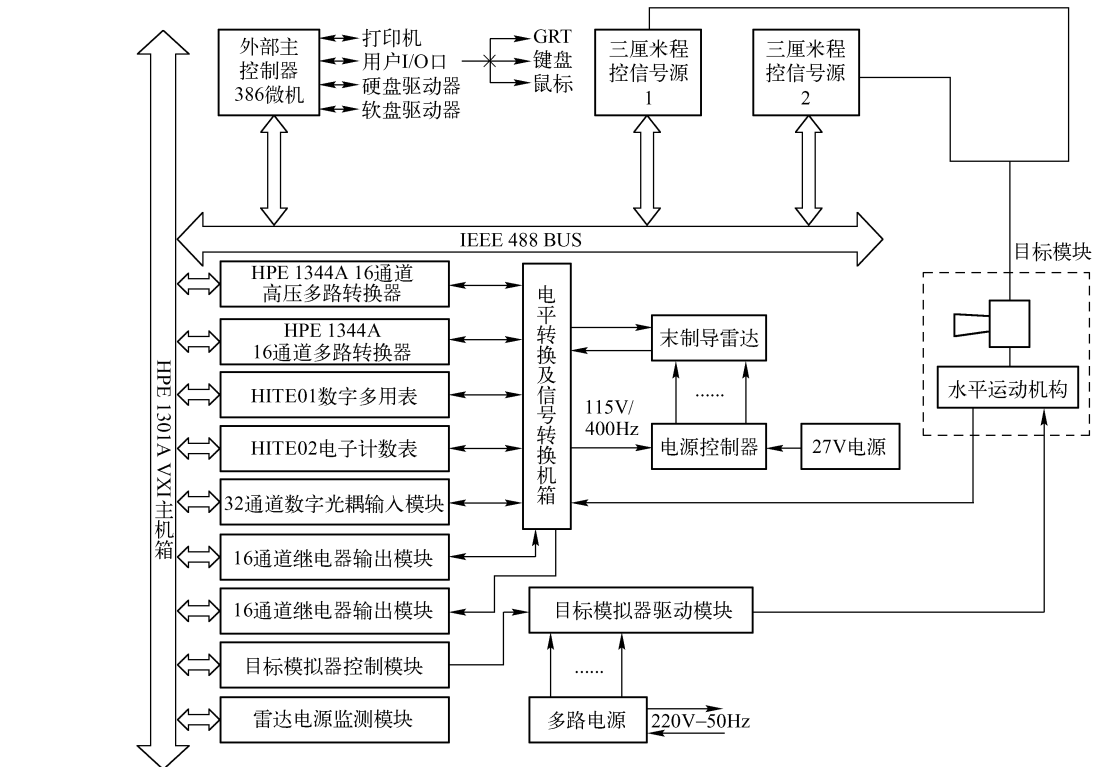


图4 雷达自动测试系统的硬件原理方框图

4.2.3 目标模拟器

目标模拟器由水平臂、垂直臂组成，喇叭天线通过卡盘波导固定在水平臂和垂直臂的运动丝杠上。丝杠一端安装有直流力矩马达、直流力矩电动机带动丝杠旋转时，喇叭天线就可在水平或垂直方向上运动、从而模拟海上舰艇或陆地上的车辆运动。

外部控制器通过 GPIB 电缆与 VXI 总线主机箱、程控目标信号源、程控干扰信号源相连。整个系统的测控功能在外部主控制器作用下, 由 VXI 主机箱内部的仪器模块完成。该系统是一个以 VXI 总线为主, 辅以 GPIB 总线的混合系统。

寄存器基专用模块分别是：32 通道数字光电耦合输入模块，设计的专用模块要求：既可使测试系统与雷达在电气上完全隔离，又能用一种模块测试具有多种电平的开关量。

16 通道继电器开关量输出模块（两块）。继电器选用 DS—2Y，有双刀、两位。利用它既可向雷达发“高压”通/断、“距离装定远近”、“自调频”自动、手动等开关量指令，又能进行回读，以诊断本模块是否有故障，该模块具有故障自诊断能力。

目标模拟器控制模块是一种带有智能芯片的寄存器基器件, 模块中用 AT89C52 单片机的高速输入/输出检测目标模拟器光电码盘送来的 A、B 两相数字脉冲信号的相位关系和脉冲个数、经积分分离的增量式 PID 算法, 去控制脉宽调制信号的脉冲宽度。

该脉宽调制 (PWM) 信号经 AT89C52 单片机的高速输出端输出, 送至目标模拟器驱动模块。经模拟多路开关依次检测上述电压、电流。检测值与上、下报警值相比较, 若超过, 则发出声、光报警信号; 若再超出上、下切断界限, 则输出控制信号, 分别自动切断 +27V、115V /400Hz 的继电器, 从而确保雷达安全。

4.2.4 系统测控软件

系统测控软件采用四级树状下拉式菜单结构设计。最低层一级对用户完全透明，使用 C 语言编程。为使系统具有“手动”和“自动”工作方式，软件中为“手动”方式设计了“开关量输入”、“开关量输出”、“目标模拟器”、“数字多用表”、“通用计数器”五种全汉字化虚拟仪器软面板。为“自动”方式设计了数据输入和雷达参数及性能显示的汉字化软面板。

5 测试系统的发展

随着微电子、计算机及数字信号处理（DSP）等先进技术越来越多地应用到测试技术中，未来测试系统发展有如下两种趋势。

5.1 集成仪器

仪器与计算机技术的深层次结合将产生全新的仪器结构概念，包括现有的虚拟仪器、卡式仪器及以 VXI 总线和 MMS 为基础的模块式仪器和新出现的集成仪器。集成仪器将基于“信息的数据采集（ADC）、信号的分析与处理（DSP）、输出（DAC）及显示”的结构模式。利用这个通用的硬件平台，调用不同的测试软件就可构成不同功能的仪器，因此“软件就是仪器”^[5]。由于硬件平台是通用的，可将多种测试功能集于一体，实现多功能集成仪器。

采用以软件替代硬件的设计路线，将废除由硬件积木单元实现的激励和响应的监测，而采用由测试系统中的计算机从数学上合成所希望的激励波形。响应信号则利用高速数据采集技术进行采集，然后将采集的数据由计算机进行数字处理和分析，从而得到测试结果。

5.2 集成测试环境

测试软件不管是对单台仪器，还是对测试系统都是十分重要的，而且也是未来发展竞争的焦点。可以预言“测试设备的未来属于软件”。未来的测试环境除生成测试程序外，还将应用其他领域的技术，如人工智能测试技术。人工智能测试技术除大量用于复杂测试的修正因子处理外，还将应用于现代装备系统的故障检测与维修。

充分利用通用集成测试仪器和集成测试环境，建立通用的仪器平台和测试系统平台，为各种功能的测试仪器和测试系统的二次开发打下硬件和系统软件的基础。最终用户只需在这个高水平平台的基础上开发一定的应用软件就能构成实用仪器和实用测试系统，从而缩短研制周期，降低研制成本，提高产品质量。其中，高速、高分辨率的数据采集和数字信号处理技术是未来测试仪器平台和测试系统平台的关键技术。

参考文献

[1] 李玉柏，彭启琮，管庆. 基于 VXI 总线的虚拟仪器平台[C]. 测控技术，1997，16(3):45-47.
[2] National Instruments Corporation，abWindows/CVI Reference Manual[M].
[3] 现代测量与控制技术词典编委会. 现代测量与控制技术词典[M]. 北京：中国标准出版社，1999.
[4] 陈光蹊. 现代电子测试技术[M]. 北京：国防工业出版社，2000.
[5] 于功敬，张韬. VXI 通用测试软件框架结构的研究[C]. 计算机自动测量与控制，1999(3).

基于嵌入式的车载式压实度检测系统

普 邑, 王新勇

(河南科技大学 电子信息工程学院, 河南 洛阳, 471000)

摘 要: 路基压实的质量是影响高等级公路使用寿命的重要因素。本文在分析振动压路机工作原理的基础上, 结合“振动轮—土壤”的二自由度振动系统数学模型, 提出了车载式路基压实度检测方法, 设计了基于 ARM 的嵌入式系统、加速度信号的采集、信号处理和传输等模块, 编写了在 $\mu\text{C}/\text{OS-II}$ 环境下系统软件应用程序, 实现了车载式压实度检测系统。实验结果表明, 该检测系统能满足对路基压实质量进行全面实时监控的要求。

关键词: 振动压路机; 压实度; 加速度传感器; 激振信号; 振动测试

中图分类号: TP216 文献标识码: A 文章编号:

Design of In-vehicle Compact Degree Detection System Based on Embedded System

PU Yi, WANG Xinyong

(Henan University of Science and Technology Luoyang 471003, Henan China)

Abstract: This article analyses the theory of compaction firstly, considering the theory of vibratory roller. By analyzing the roller-soil system, the mathematical model with the two-freedom degree system has been established. The relation between the degree of soil compaction and vibratory acceleration is the foundation of measure methods. Secondly, the system on the basis of ARM embedded standard modular has been designed, also transfer and disposal of vibratory signals and velocity signals. Combine with the digital signal processing algorithm, while write system software applications under $\mu\text{C}/\text{OS-II}$ environment. The scheme has been proved feasible by the testing and experiments. The system also has been tested in the laboratory. It can achieve basically the task of checking the compaction degree of the soil. There will be great practical significance in monitoring the quality of soil compaction.

keywords: vibrator roller; compaction degree; acceleration sensor; excitation signal; vibration test

随着我国公路等级的普遍提高, 对路基压实质量的要求更加严格。压实度是反映压实质量好坏的重要指标, 其检测的准确性及实时性对工程质量的控制至关重要。随着路基施工机械化水平的大幅度和先进的装运、摊铺、压实机械的使用, 路基填筑速度不断提高, 采用传统的压实质量检测方法^[1, 2]往往难以满足及时指导施工的要求。由于振动压路机的压实功能强、激振力大、工作效率高、有效的压实程度大等特点, 振动压路机已广泛用于路基填土和其他多种路基、路面材料的压实作业, 可以说振动压路机是目前在路基压实作业中应用最为广泛的压实设备之一, 但与之相匹配的压实检测系统尚显落后。

为此, 国内外都在研究如何进行连续、准确的压实度检测设备, 达到对路基压实质量的进行全面监控的目的。本文基于振动压实与加速度检测的原理, 开发了一种新的检测压实度的仪器, 它可以通过测量振动压实轮加速度幅值的变化情况来反映出路基的压实质量, 即压实度的高低。这样施工部门就可以在线地、实时地检测出压实度的大小, 既节省时间, 减少了人力、物力的投入, 又能有效地避免欠压和过压。

1 车载式压实度检测的理论基础

1.1 压实度检测基本原理

压实度计是根据在压路机——土壤模型中，振动轮的动力学参数的变化与压实材料的压实度有密切关系，特别是利用振动轮上激振器产生的激振信号与被压实材料的压实度密切相关的关系，来反映压实材料的压实状态的好坏。压实作业中，通过在振动轮上安装精密传感器采集激振信号，经处理器进行分析处理后，得出实时压实度值。对压实度数值的信号反映方法的选择和确定是研究压实度计的关键。

1.1.1 集成系统示意图（见图 1）

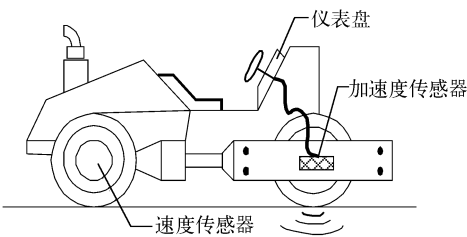


图 1 集成系统示意图

1.1.2 工作原理

振动压路机在振动压实过程中，压实度测量仪通过安装在压路机振动轮上的加速度传感器（如图 1 所示）采集激振信号，激振信号经放大电路放大，传感器微弱电荷信号转换为有用的电压信号，经滤波器滤波，消除加速度信号中多余的噪声信号，得到所需的电压信号，A/D 转换装置将采集的电压信号转换成数字信号，存放在微处理器中，并对原始信号进行数学处理，最终在显示器上输出压实度值。检测系统的总体结构如图 2 所示。

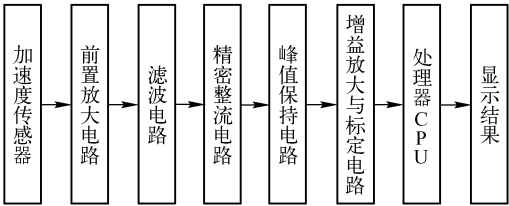


图 2 系统结构图

由分析可知，通过采用压实度仪检测方法，可以实现对压实材料压实度的连续实时检测，充分控制压实过程，从而提高公路建设效率及公路的使用寿命，其中振动轮上的激振信号是我们重点研究的对象。

1.2 振动压路机激振信号分析^[3]

在工程实践中，路基填土后先用光轮压路机碾压两遍，以确保路面平整，然后再利用振动压路机进行反复冲击压实。如果以 E 表示土的压实度， E 与振动压路机的振动参数和工作参数有下列函数关系：

$$E = f_1(P_L) + f_2\left(\frac{A\omega}{v}\right) \tag{1}$$

式中, P_L ——振动压路机振动轮的线载荷, N/cm; A ——振动压路机工作振幅, mm; ω ——振动压路机工作频率, rad/s; v ——振动压路机的工作速度, m/s。

为了克服土颗粒之间的黏聚力和吸附力, 振动压路机必须有足够大的线载荷 P_L 和振幅 A 。正常振动压实时, 振动轮与土始终接触在一起, 振动轮振幅的大小反映了土体变形及土体压力的大小。动压力越大, 土体承受的剪应力越大, 土壤压实度越大。可见, 振动轮振幅的大小间接反映压实状况。

1.2.1 振动轮——土壤数学建模

为了使振动压路机的数学模型尽可能与实际工况相吻合, 数学模型应力求简化, 使数学计算方法简单易行。在分析数学模型之前, 要对模型中有关参数和条件进行假设:

假设被压实土壤具有一定刚度的弹性体, 其刚度为 k_2 , 阻尼为线性阻尼 c_2 ; 振动压路机的机架、振动轮的质量简化为具有一定质量的集中质量块, 机架为 m_1 , 振动轮为 m_2 ;

振动压路机工作在任何一瞬间, 振动轮都保持与地面的紧密接触。
经过以上简化, “振动轮——土壤” 系统的数学模型建立如图 3 所示。

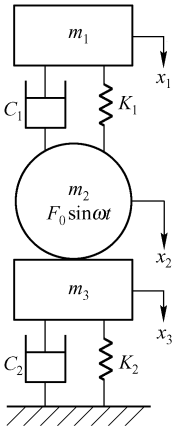


图 3 “振动轮——土壤” 系统的数学模型建立

x_1 —机架瞬时位移 (瞬时振幅); x_2 —振动轮瞬时位移 (瞬时振幅); m_1 —机架质量;
 m_2 —振动轮质量; k_1 —减震器刚度; k_2 —土的刚度; c_1 —减震器阻尼;
 c_2 —土的阻尼; F_0 —激振力; ω —工作频率 (角频率)

1.2.2 振动轮——土壤系统运动方程^[4,5]

(1) 根据振动模型, 图 3 的数学模型的运动方程是:

$$M\ddot{X} + C\dot{X} + KX = F$$

(2) 式中:

$$M = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \quad C = \begin{bmatrix} c_1 & -c_1 \\ -c_1 & c_1 + c_2 \end{bmatrix} \quad K = \begin{bmatrix} k_1 & -k_2 \\ -k_1 & k_1 + k_2 \end{bmatrix} \quad F = \begin{bmatrix} 0 \\ F_0 \sin \omega t \end{bmatrix} \tag{2}$$

式中

$$F_0 = M_e \omega^2 \quad M_e \text{——偏心块的静偏心力矩, } M = J\ddot{\theta} + F_0 R;$$

$$M_e = m_f r; \quad m_f \text{——偏心力; } r \text{——偏心块的偏心距。}$$

对于线性非时变系统, 激振为谐振力时, 微分方程式的解是

$$x_2 = F_0 \left[\frac{(A_2^2 + B_2^2)}{(C^2 + D^2)} \right]^{\frac{1}{2}} \quad x_1 = F_0 \left[\frac{(A_1^2 + B_1^2)}{(C^2 + D^2)} \right]^{\frac{1}{2}} \tag{3}$$

(3) 式中:

$$A_1 = k_1 \quad B_1 = c_1 \quad \omega = B_2 \quad A_2 = k_1 - m_1 \omega^2$$

$$C = m_2 m_1 \omega^4 - m_2 k_1 \omega^2 - m_1 k_2 \omega^2 - c_1 c_2 \omega^2 + k_1 k_2 - m_1 k_1 \omega^2$$

$$D = k_2 c_1 \omega + k_1 c_2 \omega - m_2 c_1 \omega^3 - m_1 c_2 \omega^3 - m_1 c_1 \omega^3$$

无阻尼状态下振动系统的一阶、二阶固有频率（角频率） ω_1 ， ω_2 分别为：

$$\omega_1 = \left\{ \left[(m_2 k_1 + m_1 k_2 + m_1 k_1) - \sqrt{(m_2 k_1 + m_1 k_2 + m_1 k_1)^2 - 4(m_1 m_2 (k_1 k_2))} \right] / 2 m_1 m_2 \right\}^{\frac{1}{2}}$$

$$\omega_2 = \left\{ \left[(m_2 k_1 + m_1 k_2 + m_1 k_1) + \sqrt{(m_2 k_1 + m_1 k_2 + m_1 k_1)^2 - 4(m_1 m_2 (k_1 k_2))} \right] / 2 m_1 m_2 \right\}^{\frac{1}{2}}$$

因加速度 $a_2 = \omega^2 x_2$ ，故加速度的幅值为

$$a_{2A} = \omega^2 F_0 \left[\frac{(A_2^2 + B_2^2)}{(C^2 + D^2)} \right]^{\frac{1}{2}}$$

根据式（4）用数值法可以拟合出压实度随 K_2 和 C_2 的变化趋势。

首先给定除 K_2 外其他参数的值，而给 K_2 由小到大一系列值，这样就可以得到加速度幅值随 K_2 增大的变化趋势。用同样方法可以计算出加速度幅值随 C_2 的变化趋势，计算结果如图 4 所示。

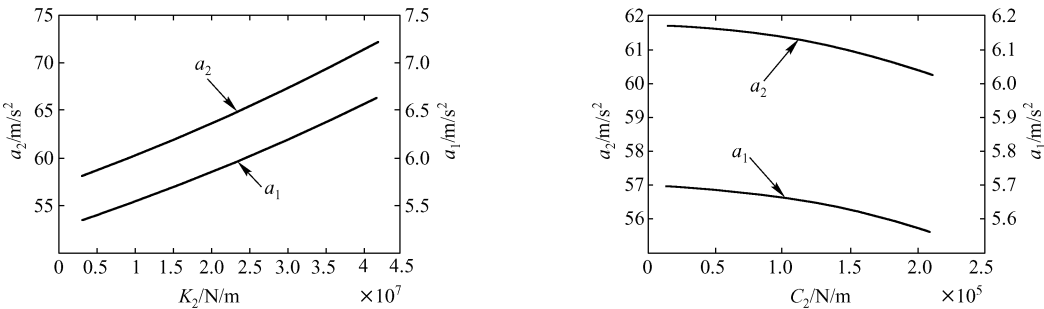


图 4 加速度随 K_2 、 C_2 的变化趋势曲线

在实际压实的过程中，随着压实的进行，土的刚度 K_2 增加，土的阻尼 C_2 减小。随着土刚度的增加，压实轮加速度幅值增加，随着土阻尼的减小，压实轮加速度幅值增加，也就是说，随着压实工作的进行，加速度幅值在增加。

2 车载式压实度检测系统硬件设计

2.1 硬件系统总体设计

车载式压实度计的硬件系统总体框图如图 5 所示。主要由传感器电路，信号处理电路及 ARM 微控制器接口电路组成。

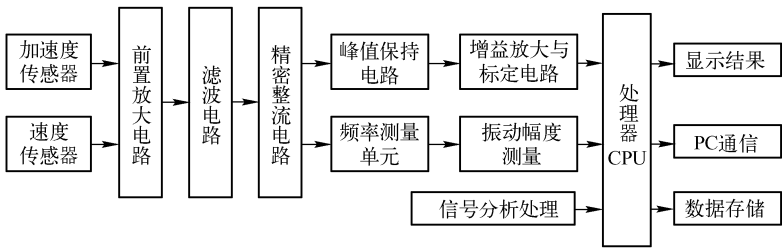


图 5 系统原理框图

2.2 传感器模块

加速度传感器选用美国 AD 公司 ADXL150 型加速度传感器。ADXL150 是一个完整的加速度测量系统，ADXL150 的测量范围为 $\pm 50g$ ，5V 电源供电时，灵敏度为 $35mV/g$ ，而振动压路机在压实路面时振动幅值为 $\pm 4\sim 7g$ ，在压实路基时振动幅值为 $\pm 5\sim 10g$ 。如果按最大 $10g$ 来考虑，加速度传感器的电压变化范围是 $350mV$ ，这个值相对来说是比较小的，因此需要对传感器信号进行增益放大。具体电路如图 6 所示。

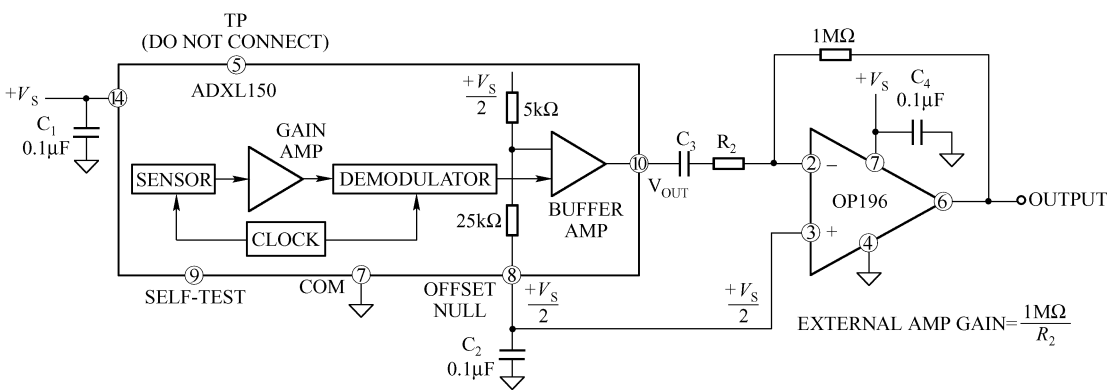


图 6 传感器模块电路图

图 6 中 C_3 电容为 $0.022\mu F$ ，目的是为了消除重力加速度对仪器的影响。电阻 R_2 为 $300k\Omega$ ，增益放大倍数为 3.3 倍，OUTPUT 端的电压变化范围为 1.7V。加速度传感器输出电压的计算公式：

$$V_{out}=V_s/2-(Sensitivity\times V_s/5\times a)$$

2.3 精密全波整流模块

从带通滤波器出来的正弦周期信号会随着土壤压实度的逐渐增加，激振信号产生畸变，不再呈现原来的正弦规律，而这种畸变随着压实遍数增加变得越发明显。然后在经过精密全波整流电路把下半周期的信号转成正信号（如图 8 所示），因此选用由运算放大器组成精密全波电路，如图 7 所示。

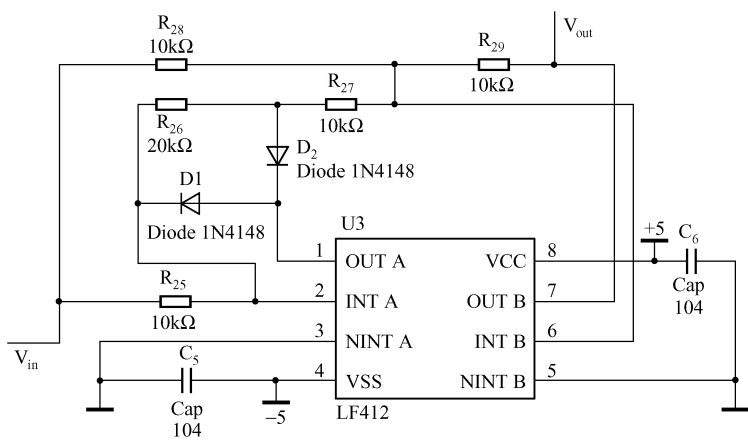


图 7 精密全波整流电路

2.4 峰值保持电路模块

峰值保持电路的输出能跟踪输入信号的峰值，并保持峰值直到复位信号到来为止，或输入信号终止后，通过放电电阻缓慢放电。

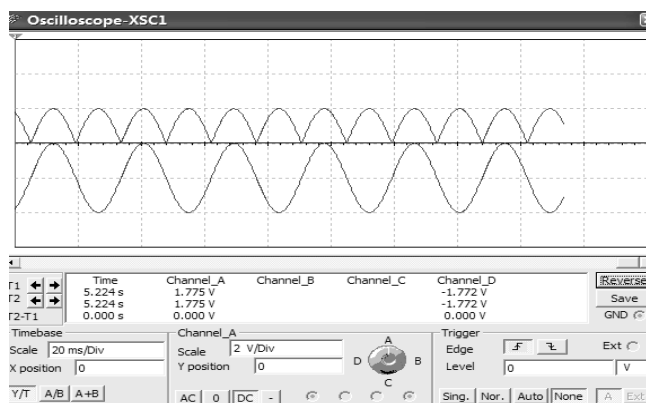


图 8 精密全波整流电路仿真波形图

随着土壤的压实度的逐渐增加，激振信号产生畸变，信号呈现出上半周幅值变尖且窄，因为峰值信号好像尖脉冲，计算机采样很难捕捉，为了能够准确地检测出信号的最大幅值，我们采用峰值保持电路模块，延长峰值信号的过渡时间。具体电路如图 9 所示。

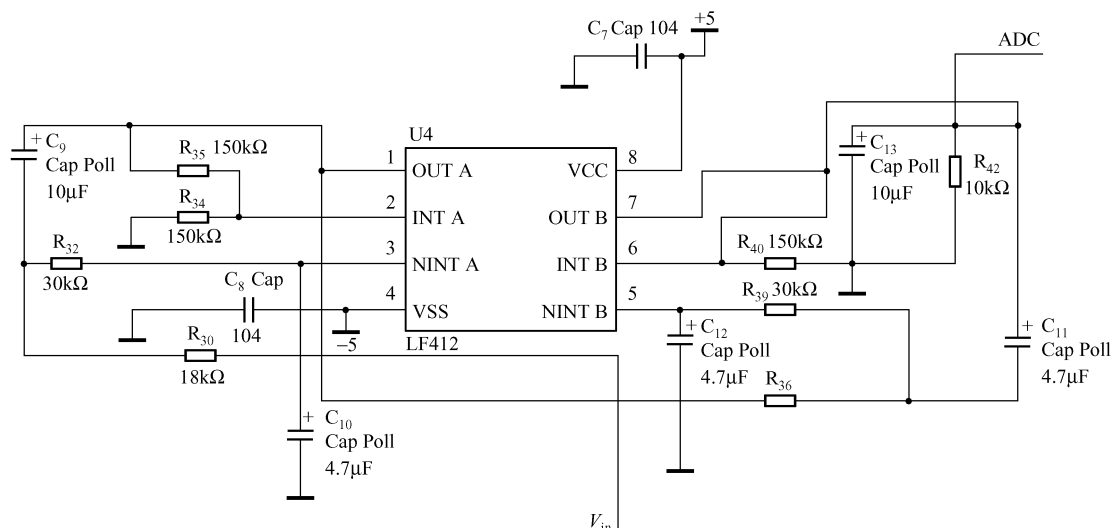


图 9 峰值保持电路

2.5 ARM 接口电路设计

ARM 微控制器采用 Philip 公司的 LPC213X 系列作为主控芯片，它是一块采用 ARM7TDMI-S 内核的 32 位微处理器，具有高性能低功耗的特点。主要负责整个系统中数据的处理和存取，通过串口与外界进行通信，实现相互间数据传送。主控制电路系统框图如图 10 所示。

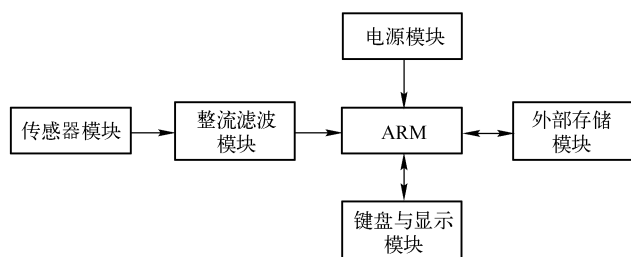


图 10 ARM 主控制电路系统框图

3 车载式压实度检测系统软件设计

3.1 动态连续检测模块的算法思想

振动压路机在碾压较松软的弹性路基面时，振动加速度信号呈有规律的正弦波状态，随着碾压遍数的增多，路基的密实度、承载力等指标也相应提高，路面逐渐坚硬，压路机振动辊加速度发生畸变，产生高次谐波。比较基波和二次以上谐波的成分，并计算两者的比值，以此找出与动态变形模量值（Evd）的相关关系。

计算机在进行数据处理时要对采集的加速度信号进行频谱分析，以求得动态信号中的各频率成分的幅值分布。频谱分析的原理时基于傅里叶变换（Fourier）原理，即一个复杂的周期振动波形可以分解成许多不同频率的正弦信号和余弦信号之和。在压路机的振动棍上安装加速度传感器，压路机在振动碾压工作时，加速度传感器连续检测振动辊的振动信号，通过压实度检测仪采集并放大振动信号、分析振动信号的振幅和频率，由计算机根据建立的数学模型计算出相应的模量值，然后显示或打印、存储测试结果。

动态连续检测的算法基本核心是将时域内的信号进行频谱分析，是基于傅里叶变换（Fourier）原理。由于系统实时性较强，因此算法上采用快速傅里叶变换（FFT）。FFT 的本质在于把长序列的 DFT 计算适当地分解为短序列的 DFT 计算。

3.2 基于 $\mu\text{C}/\text{OS-II}$ 的控制器软件设计

嵌入式操作系统的选择主要考虑所选的操作系统是否支持所使用的硬件平台、可移植性、开发工具的支持程度、使用成本以及设计本身对操作系统性能的要求（如实时性和可靠性等）。

本设计选用的 LPC213X 系列微控制器芯片没有内存管理单元（MMU），并且为了节省 I/O 资源和降低成本，希望能够将操作系统和应用程序全部下载到片内 Flash 运行，所以最终选用了 $\mu\text{C}/\text{OS-II}$ 操作系统。

车载式压实度检测系统软件程序主要包括任务、驱动程序，密实度、振幅测量程序，去极值平均滤波法程序，线性处理与任务报警处理程序，码制转换与显示程序及通信模块编程。

本系统中，首先调用 OSInit()，初始化 $\mu\text{C}/\text{OS-II}$ 所有的变量和数据结构，再调用 arm_init() 初始化微控制器的定时器及串口等硬件，通过调用 OSTaskCreate()，依次创建各个任务，最后调用 OSStart() 启动系统，开始多任务调度。整个系统控制程序模块及流程如图 11 所示。

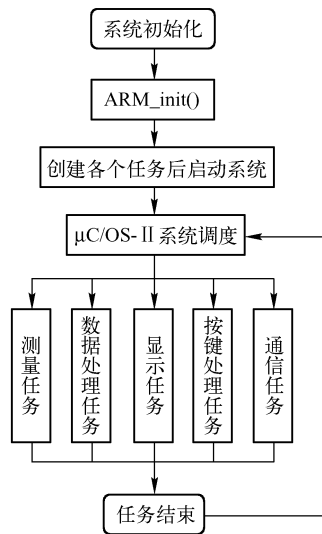


图 11 系统软件流程图

4 结论

本文通过对振动压路机振动轮土壤系统进行分析研究，得出振动加速度信号与土壤的压实程度存在直接关系，从而通过检测振动压路机的振动信号间接测出压实基础的压实度值。

对检测振动压实的系统原理进行分析设计，在此基础上开发设计了基于 ARM 的车载式压实度检测系统，可以测量土壤的密实度，压路机的振动频率及振动幅值。通过将振动压路机压实度仪应用于 LT220 型振动压路机及其进行的振动压实检测，实验表明开发的压实度仪具有一定的实用价值。

参考文献

- [1] 徐刚, 盛安连. 新颖的路基压实度测定方法的研究[J]. 农业机械学报, 2001(3):20-23.
- [2] 居彩梅. 车载式压实度检测仪[D]. 西安: 长安大学, 2001.
- [3] 盛安连. 路基路面检测技术[D]. 北京: 人民交通出版社, 1999.
- [4] 黄耀志, 杨梦华. 基于振动信号分析的地基密实度在线实时监测方法研究[J]. 振动测试与诊断, 1996(4):52-56.
- [5] 易斌, 戴瑜兴, 朱江. 基于 ARM 和 FPGA 的振动信号采集系统的实现[J]. 微计算机信息, 2007(2):137-139.

数据库设计中 SQL 优化策略和技巧

秦 军

(河南省政法管理干部学院, 河南 郑州, 450002)

摘 要: 在设计数据库应用系统过程中, 数据库的优化调整至关重要, 本文以 Oracle 数据库作为平台, 着重分析了 SQL 应用过程中影响其性能的一些因素并提出相应的解决方法。最后, 通过个例实验对一些优化方法进行了测试, 体现出性能的差距。

关键词: 数据库应用系统; ORACLE 数据库; 调整级别; SQL 优化

中图分类号: (作者本人填写) **文献标识码:** A **文章编号:** 1006-7043 (2004) xx-xxxx-x

Some Optimization Strategies and Techniques of SQL in Designing Database System

QIN Jun

(Henan Administrative Institute of Politics and LAW, Computer Science Department, Zhengzhou 450002, Henan China)

Abstract: In a design of database application system processing, it was very important that the database would be optimized and adjusted. In this article, Author, ORACLE database as the platform, Analyzed the affecting factors of the database system in designing, and put forward the corresponding solutions. Finally, author used some optimization methods to tests through the example experiments, which reflected the disparity of performance.

Keywords: database application system; ORACLE database; the adjustment of databases; optimization of SQL

在数据库应用系统的设计中, 数据库系统和操作系统一样, 在计算机上安装成功后, 还需要进一步配置和优化, 从而使其具有更强大的功能和运行在最佳状态。数据库应用系统的性能受多方面的限制, 如操作系统、数据库管理系统及前端开发工具等, 对于数据的存取, 优化方案中主要有四个不同的调整级别: 第一级调整是操作系统级包括硬件平台, 第二级调整是 ORACLE RDBMS 级的调整, 第三级是数据库设计级的调整, 第四级是 SQL 级。其中, 数据库和 SQL 调整级别是我们人为能够解决的。通常依此四级调整级别对数据库进行调整、优化, 数据库的整体性能会得到很大的改善^[1]。下面主要介绍 ORACLE 数据库和 SQL 查询优化设计策略。

1 充分利用系统全局区域 SGA (System Global Area)

SGA 是 Oracle 数据库的心脏。用户的进程对这个内存区发送事务, 并且以这里作为高速缓存读取命中的数据, 以实现加速的目的。正确的 SGA 大小对数据库的性能至关重要。SGA 包括以下几个部分:

- (1) 数据块缓冲区 (Data Block Buffer Cache) 是 SGA 中的一块高速缓存, 占整个数据库大小的 1%~2%, 用来存储从数据库重读取的数据块 (表、索引、簇等), 因此采用 Least Recently Used (LRU, 最近最少使用) 的方法进行空间管理。
- (2) 字典缓冲区。该缓冲区内的信息包括用户账号数据、数据文件名、段名、盘区位置、表说明

作者简介: 秦军 (1979—), 男, 讲师, 硕士学位。

和权限，它也采用 LRU 方式管理。

(3) 重做日志缓冲区。该缓冲区保存为数据库恢复过程中用于前滚操作。

(4) SQL 共享池。保存执行计划和运行数据库的 SQL 语句的语法分析树。也采用 LRU 算法管理。如果设置过小，语句将被连续不断地再装入到库缓存，影响系统性能。

另外，SGA 还包括大池、Java 池、多缓冲池。但主要是由上面 4 种缓冲区构成。对这些内存缓冲区的合理设置，可以大大加快数据查询速度，一个足够大的内存区可以把绝大多数数据存储在内存中，只有那些不怎么频繁使用的数据，才从磁盘读取，这样就可以大大提高内存区的命中率。

2 合理设计和管理表

(1) 利用表分区。分区将数据在物理上分隔开，不同分区的数据可以制定保存在处于不同磁盘上的数据文件里。这样，当对这个表进行查询时，只需在表分区中进行扫描，而不必进行 FTS (Full Table Scan，全表扫描)，明显缩短了查询时间，另外，处于不同磁盘的分区也将对这个表的数据传输分散在不同的磁盘 I/O，一个精心设置的分区可以将数据传输对磁盘 I/O 竞争均匀地分散开^[2]。

(2) 避免出现行连接和行迁移。在建立表时，由于参数 `pctfree` 和 `pctused` 不正确的设置，数据块中的数据会出现行链接和行迁移，也就是同一行的数据不保存在同一的数据块中。如果在进行数据查询时遇到了这些数据，那么为了读出这些数据，磁头必须重新定位，这样势必会大大降低数据库执行的速度。因此，在创建表时，就应该充分估计到将来可能出现的数据变化，正确地设置这两个参数，尽量减少数据库中出现行链接和行迁移。

(3) 控制碎片。碎片 (Fragmentation) 是对一组非邻接的数据库对象的描述。碎片意味着在执行数据库的功能时要耗费额外的资源 (磁盘 I/O，磁盘驱动的循环延迟，动态扩展，链接的块等)，并浪费大量磁盘空间。当两个或多个数据对象在相同的表空间中，会发生区间交叉。在动态增长中，对象的区间之间不再相互邻接。为了消除区间交叉将静态的或只有小增长的表放置在一个表空间中，而把动态增长的对象分别放在各自的表空间中。在 `create table`、`create index`、`create tablespace`、`create cluster` 时，在 `storage` 子句中的参数的合理设置，可以减少碎片的产生。

(4) 别名的使用。别名是大型数据库的应用技巧，就是表名、列名在查询中以一个字母为别名，查询速度要比建连接表快 1.5 倍。

(5) 回滚段的交替使用。由于数据库配置对应用表具有相对静止的数据字典和极高的事务率特点。而且数据库的系统索引段、数据段也具有相对静止，并发现在应用中最高的负荷是回滚段表空间。把回滚段定义为交替引用，这样就达到了循环分配事务对应的回滚段，可以使磁盘负荷很均匀地分布。

3 索引 Index 的优化设计

(1) 管理组织索引。索引可以大大加快数据库的查询速度，索引把表中的逻辑值映射到安全的 RowID，因此索引能进行快速定位数据的物理地址。但是有些 DBA 发现，对一个大型表建立的索引，并不能改善数据查询速度，反而会影响整个数据库的性能。这主要是和 SGA 的数据管理方式有关。ORACLE 在进行数据块高速缓存管理时，索引数据比普通数据具有更高的驻留权限，在进行空间竞争时，ORACLE 会先移出普通数据。对一个建有索引的大型表的查询时，索引数据可能会用完所有的数据块缓存空间，ORACLE 不得不频繁地进行磁盘读/写来获取数据，因此在对一个大型表进行分区之后，可以根据相应的分区建立分区索引。如果对这样大型表的数据查询比较频繁，或者干脆不建索引。另外，DBA 创建索引时，应尽量保证该索引最可能地被用于 `where` 子句中，如果对查询只简单地制定一个索引，并不一定会加快速度，因为索引必须指定一个适合所需的访问路径。

(2) 聚簇的使用。Oracle 提供了另一种方法来提高查询速度，就是聚簇 (Cluster)。所谓聚簇，简单地说就是把几个表放在一起，按一定公共属性混合存放。聚簇根据共同码值将多个表的数据存储在同一个 Oracle 块中，这时检索一组 Oracle 块就同时得到两个表的数据，这样就可以减少需要存储的 Oracle 块，从而提高应用程序的性能。

(3) 优化设置的索引，就必须充分利用才能加快数据库访问速度。ORACLE 要使用一个索引，有一些最基本的条件：

- ① where 子名中的这个字段，必须是复合索引的第一个字段；
- ② where 子名中的这个字段，不应该参与任何形式的计算。

4 充分利用数据的后台处理方案减少网络流量

(1) 合理创建临时表或视图。所谓创建临时表或视图，就是根据需要在数据库基础上创建新表或视图，对于多表关联后再查询信息的可建新表，对于单表查询的可创建视图，这样可充分利用数据库的容量大、可扩充性强等特点，所有条件的判断、数值计算统计均可在数据库服务器后台统一处理后追加到临时表中，形成数据结果的过程可用数据库的过程或函数来实现^[3]。

(2) 数据库打包技术的充分利用。利用数据库描述语言编写数据库的过程或函数，然后把过程或函数打成包在数据库后台统一运行包即可。

(3) 数据复制、快照、视图，远程过程调用技术的运用。数据复制，即将数据一次复制到本地，这样以后的查询就使用本地数据，但是只适合那些变化不大的数据。使用快照也可以在分布式数据库之间动态复制数据，定义快照的自动刷新时间或手工刷新，以保证数据的引用参照完整性。调用远程过程也会大大减少因频繁的 SQL 语句调用而带来的网络拥挤。

5 使用最优的数据库连接和 SQL 优化方案

(1) 使用直接的 OLE DB 数据库连接方式。通过 ADO 可以使用两种方式连接数据库：一种是传统的 ODBC 方式；另一种是 OLE DB 方式。ADO 是建立在 OLE DB 技术上的，为了支持 ODBC，必须建立相应的 OLE DB 到 ODBC 的调用转换，而使用直接的 OLE DB 方式则不需转换，从而提高处理速度^[4]，如图 1 所示。

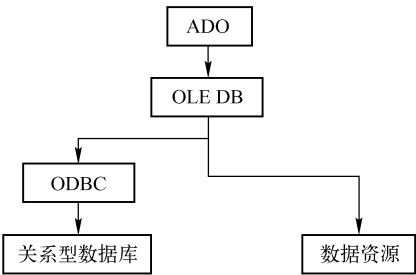


图 1 使用 ADO 访问数据信息源

(2) 使用 Connection Pool 机制。在数据库处理中，资源花销最大的是建立数据库连接，而且用户还会有一个较长的连接等待时间。解决的办法就是复用现有的 Connection，也就是使用 Connection Pool 对象机制。

Connection Pool 的原理是：IIS+ASP 体系中维持了一个连接缓冲池，这样，当下一个用户访问时，直接在连接缓冲池中取得一个数据库连接，而不需重新连接数据库，因此可以大大地提高系统的响应速度。

(3) 高效地进行 SQL 语句设计。通常情况下，可以采用下面的方法优化 SQL 对数据操作的表现：

① 减少对数据库的查询次数，即减少对系统资源的请求，使用快照和显形图等分布式数据库对象可以减少对数据库的查询次数。

② 尽量使用相同的或非常类似的 SQL 语句进行查询，这样不仅充分利用 SQL 共享池中的已经分析的语法树，要查询的数据在 SGA 中命中的可能性也会大大增加。

③ 限制动态 SQL 的使用，虽然动态 SQL 很好用，但是即使在 SQL 共享池中有一个完全相同的查询值，动态 SQL 也会重新进行语法分析。

④ 避免不带任何条件的 SQL 语句的执行。没有任何条件的 SQL 语句在执行时，通常要进行 FTS，数据库先定位一个数据块，然后按顺序依次查找其他数据，对于大型表这将是一个漫长的过程。

⑤ 如果对有些表中的数据有约束，最好在建表时的 SQL 语句用描述完整性来实现，而不是用 SQL 程序中实现。

⑥ 可以通过取消自动提交模式，将 SQL 语句汇集一组执行后集中提交，程序还可以通过显式地使用 COMMIT 和 ROLLBACK 进行提交和回滚该项事务。

⑦ 检索大量数据时费时很长，设置行预取数则能改善系统的工作表现，设置一个最大值，当 SQL 语句返回行超过该值，数据库暂时停止执行，除非用户发出新的指令，开始组织并显示数据，而不是让用户继续等待。

6 查询优化的必要性

数据库应用系统通常为 B/S 结构，采用 Oracle 数据库系统作为后台数据库。作为一个四层应用系统，对于用户事务，在结果被返回客户端之前，在中间件层、数据库层和物理存储器之间通常需要多个回合的信息交换，同时依赖于各层之间的网络^[5]。由于用户终端操作频繁，数据量较大，且查询多为跨表查询，在很多情况下查询过程是非常复杂的。因此为了获取更好的响应时间与系统性能，查询性能优化的作用就越显重要。为了在开发阶段能成功的进行 SQL 优化，需要采用“随时测量所有信息”的方法。其目标是使每个语句的执行时间最短。如果确保每个 SQL 语句被开发成具有最小的执行时间，那么用所有这些 SQL 组合成的任何系统就都将运行良好。

6.1 性能优化的策略

1) 统计量和事件的测量

在测量性能之前优化 SQL 语句是没有意义的。为了简化性能的测量，Oracle 提供了可以通过 SQL 视图得到的性能量度指标，这些量度指标是以计数器（即统计量）和等待时间（即事件）的形式存在的。我们可以通过 V\$SYSSTAT 视图查看全数据库范围的统计量，V\$SESSTAT 视图查看单个会话的统计量等。而等待时间则出现在 V\$SYSTEM_EVENT 和 V\$SESSION_EVENT 中。若要显示事件时间，应在 init.ora 文件里设置参数 TIMED_STATISTICS=TRUE。而在众多的统计量中，我们需要选择关注其中的某些。我们可以让 Oracle 自己进行选择统计量，这可通过在 SQL*Plus 里使用 SET AUTOTRACE ON STATISTICS 这特性即可在每个 SQL 语句执行完成后，针对一些所选择的统计量提供即时的反馈。例如，db block gets （在当前模式下读取的数据块数目），physical reads （物理数据块读取数目），CPU used by this session （占用的 CPU 时间）等。通过参考这些统计量的值，就可以了解如何解决性能问题，即采取何种措施来降低统计量和事件等待所确定的资源使用。

2) 设置缓冲区缓存

通常从内存中读取缓存数据的速度，要比读取物理文件的速度快几个数量级，则了解单个语句的缓存行为有助于提高 SQL 的性能，确保最频繁访问的数据块能最长时间地保留在高速缓存中。Oracle 维护着一个 SQL 语句所请求的“近期最常使用（MRU）”的数据块的高速缓存。我们可以通过

DB_CACHE_SIZE 设置数据库默认块大小的缓冲区缓存大小^[5]。设置过小会降低高速缓存命中率，而设置过大又会占据过多的内存容量。因此，缓冲区缓存大小需要我们仔细均衡事务操作效率，以及内存大小来设置。Oracle 使用 LRU 算法把最频繁使用的数据块保留在缓冲区缓存中，正确的设置缓存配置可以明显的改善性能^[6]。

3) 使用并行操作

物理 I/O 通常比逻辑 I/O 在操作时间上大好几个数量级，而全表扫描会涉及许多物理 I/O，因此具备检测全表扫描的能力是重要的。当优化器判定没有合适的索引来执行一个 SQL 语句以满足优化器目标时，Oracle 就会进行全表式扫描来执行该 SQL 语句。而当客户有条件置备多个 CPU 的服务器时，就可以考虑 Oracle 的并行查询特性。可以把一个扫描分成多个进程同时进行，每个进程并发地分别使用一个 CPU 来扫描一部分数据。但当执行时没有足够的 CPU 资源时，系统会默认将该 SQL 语句串行执行。因此，可以通过设置 ORACLE 的自动并行优化功能来实行并行操作，从而改善性能。

4) 改善 SQL 语句的运行速度

调整 SQL 语句，通常一个 SQL SELECT 语句指定的是所要求的结果，而具体如何执行则由 Oracle 在生成执行计划时决定。因此，几个等价的 SQL 语句通常可以等到一个特定的结果集，而每条语句执行时都会有不同的成本和消耗时间。下面通过学生健康管理系统中一个实例来说明调整 SQL 语句的必要性。例如：

有以下一条查询要求，用 SQL 语言表达如下：

```
select bwmc.jgmsfrom b_tj_tjbwb,b_tj_tjjgzdmsb wh     ere b_tj_tjjgzdmsb.ssbwbm=b_tj_tjbwb.bwbm andb_tj_tjjgzdmsb.xmbm="+ lb.Text +";
```

数据库系统可以用多种等价的关系代数表达式来实现这一查询，如可以先用笛卡儿积，再做选择操作，最后做投影，也可以先对表 b_tj_tjjgzdmsb 进行选择操作，然后做连接操作，最后做投影。采用不同的关系代数表达式来完成查询操作时，其运算次数相差很大，查询效率当然也相差很远。

我们来讨论一下，第一种方案先用笛卡儿积 b_tj_tjbwb×b_tj_tjjgzdmsb，这时意味着 b_tj_tjbwb 和 b_tj_tjjgzdmsb 的所有元组都要进行内存组合连接，由于内存有限，使得被连接的元组要反复进入内存多次才能够完成全部连接过程；而在第二种方案中，由于对 b_tj_tjjgzdmsb 首先进行了选择，只要在选择时各进入一次内存，选择完成后参加连接的只是其中被选中的很少一部分元组，故无须反复进入内存多次。

这个简单的例子充分调整 SQL 语句在系统中的必要性。

5) 使用内嵌视图代替表

在查询语句的 from 子句中，使用内嵌视图来代替表会使查询的目的性更易理解。这可以使我们的显式控制语句的执行次序，从而使被传送的行数在执行早期就被减少，在后期阶段就会仅有较少的行处理，从而降低了资源使用的时间和语句执行的时间。我们可以通过系统中一条查询语句来比较两种方法的差距。

6) 正确合理的使用索引

索引是数据库中一个常用而重要的数据库对象，而优化查询的一个最重要的方法就是合理地建立索引。在关系数据库系统中，通常对达标进行扫描时，应避免必要的全表扫描^[7]。而在表上建立合适的索引，从而改变了对数据的访问路径，我们就可以通过访问索引的方式获得记录的物理位置，从而达到访问表的目的。可以避免因全表扫描而造成的 I/O 开销，从而提高数据库数据查询的速度，改善数据库性能。但创建索引也会增加系统的时空开销，因此必须要与实际查询需求紧密结合，才能达到我们所需的查询优化的目的。建立索引也需遵循一些原则：

- (1) 两表间的关联字段上常建索引。
- (2) 对于多列排序的，可在列上建立复合索引，但复合索引应尽量少用。
- (3) 避免高重复率的字段建索引。

(4) 对于频繁进行 GROUP BY 或 ORDER BY 操作列上建立索引。

(5) 对不同值较少的列上不需建立索引。

(6) 对于大批量数据记录的插入及删除，首先删除索引，再进行数据的操作，操作完成后再重建必要的索引。

(7) SQL 网络性能的优化。通常信息应用管理系统中，客户端和服务端是通过广域网来连接的，因此，网络延迟时间也成为影响应用系统的客户端响应时间的因素之一^[8]。则需要我们在客户端尽量采用批量处理请求来满足信息往返服务器的时间最小化。Oracle 中引入了 BULK COLLECT 特性用于阵列操作。例如，当用户想查看过去某日期的全部客户的信息，我们就为这个查询需求生成一个批处理的请求，基于行变量设置，我们可以在 BULK COLLECT 的代码中，对 FETCH 的 LIMIT ROWS 限定符用于提出请求，使行以 1000 行形式批量返回表变量中。从而保证从远程服务器传送来的网络包充满记录行。因此，比每一行都用一个单独的网络包大大的缩小了网络开销。由此可知，任何能用来批量请求和减少网络往返时间的技术都可能改善性能。

7 小结

总之，所有的数据库性能优化问题，没有一个统一的解决方法，但 Oracle 提供了丰富的选择环境，可以从 Oracle 数据库的体系结构、软件结构、模式对象及具体的业务和技术实现出发，进行统筹考虑。而且 Oracle 的 SQL 性能优化也远远不止这些。比如，操作系统的影响，Oracle 的优先级，第三方软件及服务器整合，资源管理等都在不同程度上影响了数据库数据存取的性能。

本文中所提到的性能优化方法都是作者在查阅了大量文献资料及实际经验总结出来的。事实也证明，这些优化策略的确可以在一定程度上提高 Oracle 数据存取的执行效率。只有在具体使用时要根据实际应用环境，实际情况选择合理的优化策略，这样才可以达到充分利用数据库管理系统提供的高性能服务使应用系统充分发挥高效功能。

参考文献

- [1] 盖国强, 冯春培, 叶梁, 等.《Oracle 数据库性能优化》[M]. 北京: 人民邮电出版社, 2005.6.
- [2] Geoff Ingram. High-performance Oracle——Proven Methods for Achieving Optimum Performance and Availability [M]. 张建明, 英字. WILEY Publishing, Inc. 北京: 清华大学出版社, 2003.4.
- [3] Raghu Ramakrishnan, Johannes Gehrke, Database Management Systems Third Edition[M]. 周立柱, 张志强, 李超, 等. The McGraw-Hill Companies, Inc. 北京: 清华大学出版社, 2004.3.
- [4] [美] Joe Greene, Advanced Information Systems, Inc. et al. Oracle 8 服务器技术精粹 [M]. 北京: 清华大学出版社, 1999.
- [5] 胡欣杰.《Oracle 9i 数据库管理员指南》[M]. 北京: 北京希望电子出版社, 2002.4.
- [6] 瓮正科, 王新英. Oracle 8.X For Windows NT 实用教程[M]. 北京: 清华大学出版社, 2009.
- [7] Kevin Loney. Oracle 8i 数据库管理员手册[M]. 北京: 机械工业出版社, 2000.
- [8] 周渝斌. 基于 Oracle 8i 的大型数据库技术讲座之一数据库优化篇[J]. 电脑编程技巧与维护, 2008, (4):5-9.

.Net 控制 Excel 自动生成表格的应用研究

王 辉, 杨 凯, 郎士宁, 冯少华, 王月蓉

(防空兵指挥学院 河南 郑州, 450052)

摘 要: 通过把 Office 集成到所开发的应用系统中, 能够提高系统开发效率, 增加系统的稳定性, 扩展系统的功能。本文以教务管理系统开发为例, 提出了在 .NET 环境下控制 Excel 自动生成复杂表格的基本思路和实现方法。

关键字: .NET; Excel; 对象

中图分类号: TP393 **文献标识码:** A **文章编号:** 1006-7043 (2010) xx-xxxx-x

Application of Automatic Table Excel controlled by .NET

WANG Hui, YANG Kai, LANG Shining, FENG Shaohua, WANG Yuerong

(Journal of Air Defense Forces Command Academy, Zhengzhou 450052, Henna China)

Abstract: The application software of the integr ating office can boost efficiency of software, enhance the stability of system , and enlarge the function of system. By taking the example of Educational Administration Management System, this paper brings forward the methods ,with which it can completes the transform between Excel and complicated table in the .NET circumstance.

Keywords: .NET; Excel; object

办公自动化是当前最为广泛的计算机应用领域, 目前, 世界上绝大多数用户都是使用微软公司的 Office 组件, 根据实际工作的需求, 很多时候用户希望能够把数据库中的数据以表格的形式直接输出到 Excel。本文结合已实现的教务管理系统, 介绍了在 C#.Net 环境下, 调用 Excel 的基本方法和基本操作, 并给出了 Excel 表格的自动生成、单元格的合并及表格数据填充的基本方法。

1 调用 Excel 的 COM 组件

1.1 Excel 对象模型

Excel 对象模型包括了 128 个不同的对象, 从矩形、文本框等简单的对象到透视表、图表等复杂的对象。其中用的最多的四种对象为^[1]:

- (1) Application 对象。Application 对象处于 Excel 对象层次结构的顶层, 表示 Excel 自身的运行环境。
- (2) Workbook 对象。Workbook 对象处于 Application 对象的下层, 表示一个 Excel 工作簿文件。
- (3) Booksheet 对象。Booksheet 对象包含于 Workbook 对象, 表示一个 Excel 工作表。
- (4) Range 对象。Range 对象包含于 Booksheet 对象, 表示 Excel 工作表中的一个或多个单元格。

1.2 添加 Excel 的引用

.NET 要操纵 Excel 对象, 首先应向程序中添加一个对 Excel 对象库的引用。Office 程序以 COM

作者简介: 王辉 (1974—), 男, 讲师, 硕士;
杨凯 (1969—), 男, 讲师, 学士;
郎士宁 (1973—), 女, 讲师, 硕士;
冯少华 (1983—), 男, 助教, 学士;
王月蓉 (1985—), 女, 助教, 学士。

组件的形式对外部开放。在.NET 环境下一个组件事实上就是一个.NET 下的动态连接库（DLL），它包含运行程序本身和 DLL 的描述信息，而一个 COM 组件是用其类库（TLB）储存其描述信息，所以用 Visual C#调用 Excel 表格前，必须添加一个相关的引用。

添加 Excel 的 COM 组件的引用是个很简单的操作，在 Visual C#的主菜单中选择“项目”，然后选择“添加引用”选项，再选择“COM”标签，最后再选择“Microsoft Excel 11.0 Object Library”，（由于本机安装的 Office 版本的不同，组件的版本号有所不同，这里用的是 Office2003）这样就完成了 Excel 的 COM 组件的引用添加^[2]，如图 1 所示。

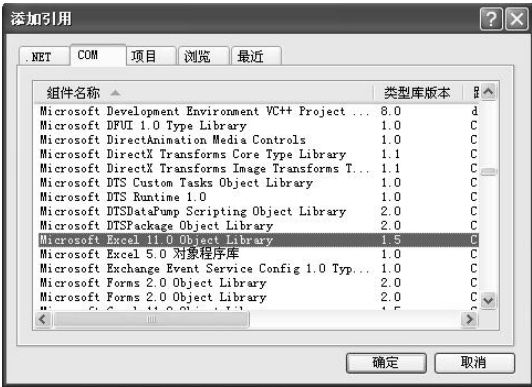


图 1 添加 Excel 的 COM 组件

2 调用 Excel 自动生成表格

2.1 一般方法

首先，打开 Excel 表格。在 Visual C #打开 Excel 表格通常只需很少的代码就可以完成打开 Excel 表格的工作^[3]，具体代码如下：

```
Excel.Application excelApp= new Excel.Application();
excelApp.Application.Work.Add(true);
excelApp.Visible=true;
```

接下来往 Excel 表格中写入数据。在命名空间 Excel 中定义了一个“Cell”类，这个类表示的是 Excel 表格中的一个单元格，通过对“Cell”赋值从而实现对 Excel 表格的数据填充，具体代码如下：

```
Excel.Application excelApp= new Excel.Application();
excelApp.Application.Work.Add(true);
excelApp.Cells[1,1]="1 行 1 列";
excelApp.Cells[1,2]="1 行 2 列";
excelApp.Cells[1,3]="1 行 3 列";
.....
excelApp.Visible=true;
```

这样就完成了一个 Excel 表格的生成和数据填充。

这种方法在处理数据时的效率很低，在数据量很小情况下，基本可以满足要求，而且表格生成后必须再通过手工对表格进行相应的调整。例如，相关数据的合并、数据的统计、表格标题的添加和格式的调整等。在处理的数据量较大的情况下，这种方法执行效率低的问题就会突显出来，如果再需要对数据进行手工调整，就会大大的增加工作量，而且手工操作过程中，数据的完整性和正确性无法得到保障，还有它的通用性也比较差，往往只针对某一个数据表。

基于此，在教务管理系统中，我们采用模块化设计与存储过程相结合的方法解决了这些问题。

2.2 模块化方法

当反复从数据库读数据并写入 Excel 表格时，都要使用相同的代码，我们通过封装一个 ToExcel 类，实现对 Excel 数据的输入、合并和统计等，整个过程都是由应用程序自动实现，这样大大降低了手工调整的工作量，保证了数据的完整性和正确性。

在 ToExcel 设计中，通过设置一个数据表名称参数 dt，来实现对不同数据表的操作，提高代码的通用性。对于数据的读取，利用存储过程从数据库中读取所需数据以提高执行效率。

在教务管理系统中，存储过程的关键代码：

```
CREATE PROCEDURE [dbo].[ProSchoolTaskSpring] (@staffroomID int)
AS
INSERT INTO _授课预告表 (课程名称, 课程简称, 学时, 区队, 人数, 考试类型)
(SELECT _课程.名称, _课程.缩写, _课程.学时, _编队情况.区队名称, _编队情况.计定额人数+ _编队情况.不计定额人数 AS 人数, _课程.考试类型 FROM _教研室 INNER JOIN _课程 ON _教研室.标识号= _课程.教研室标识号 INNER JOIN _课程设置 ON _课程.标识号= _课程设置.课程标识号 INNER JOIN _编队情况 ON _课程设置.培训类别= _编队情况.培训类别 AND _课程设置.专业标识号= _编队情况.专业标识号 WHERE ( _教研室.标识号= @staffroomID AND _课程设置.开课学期 IN (2,4,6,8)))
```

RETURN
ToExcel 模块的关键代码：

```
class ToExcel
{//对象实例

public Microsoft.Office.Interop.Excel.Application ExcelApp = null;
public Microsoft.Office.Interop.Excel.Range ExRange = null;
public Microsoft.Office.Interop.Excel.Workbooks WkBooks = null;
public Microsoft.Office.Interop.Excel.Sheets Sheets = null;

.....
//主体函数
public ToExcel(DataTable dt, ref ArrayList col, ref ArrayList hbcol, string title)
{int n = dt.Rows.Count;
ExcelApp = new Microsoft.Office.Interop.Excel.Application();
ExcelApp.Visible = true;

.....
if (col.Count != 0)
CombineColumn(dt.Rows, ref hbcol,ref col);//数据行合并函数
FillExcel(dt, title, ref col);//数据写入函数
Sheet.Cells.Columns.AutoFit();//自动调整单元格函数

.....
}
//数据合并函数
public void CombineColumn(DataRowCollection rows, ref ArrayList list,ref ArrayList col)
{ //合并数据

try
{ int listcount = list.Count;
int rowcount = rows.Count;
```

```

        for (int j = 0; j < listcount; j++)
        {
            int index = (int)list[j];
            int itemidx = 0;
            for (int i = 1; i < rowcount; i++)
            {
                .....
            }
        }
    }
    catch (Exception e)
    {
        //异常处理
        throw (e);
    }
}

//数据写入函数
public void FillExcel(DataTable dt, string title, ref ArrayList col)
{
    try
    {
        int columncount = dt.Columns.Count; //字段数
        ExRange = Sheet.get_Range(Sheet.Cells[1, 1], Sheet.Cells[1, columncount]);
        ExRange.Merge(0);
        Sheet.Cells[1, 1] = title;
        Sheet.get_Range(Sheet.Cells[1, 1], Sheet.Cells[1, 1]).Font.Size = 22;
        for (int j = 0; j < columncount-1; j++) //字段名
        {
            ..... //确定包含的字段名
        }
        int rowcount = dt.Rows.Count;
        for (int i = 0; i < rowcount; i++)
        {
            for (int o = 0; o < columncount - 1; o++)
            {
                ..... //写入数据
            }
        }
    }
    catch (Exception e)
    {
        //异常处理
        throw (e);
    }
}
}

```

其中，dt 指定要输入到 Excel 的具体的数据表，col 指定在 Excel 表格中要显示哪些列，hbcol 指定要对哪些列进行合并处理，title 指定 Excel 表格的标题。

3 结束语

本文介绍了 Visual C# 处理 Excel 表格的基本方法，文中描述的方法对 Office 其他组件的使用，如 Word、Powerpoint 等，也有很强的借鉴意义，处理方法非常相似。

参考文献

- [1] 刘柏峰, 陈伟, 陈晓军. C#中操纵 Excel 的几种方法. 微型电脑应用[J]. 2006(11). 60-64.
- [2] 张跃廷, 许文武, 王小科. C#数据库系统开发完全手册[M]. 北京: 人民邮电出版社, 2006.
- [3] 邹建峰, 周山峰, 项细威. C#企业级开发案例精解[M]. 北京: 人民邮电出版社, 2006.

.NET 平台下材料管理系统的设计模式研究

孟 军, 王 辉, 杨 凯, 秦兴桥

(防空兵指挥学院, 河南 郑州, 450052)

摘 要: 本文论述了基于 C/S 和 B/S 相结合的材料管理系统的开发设计, 结合工程实践, 阐述在.NET 平台下材料管理软件及其设计模式的相关问题, 对开发类似系统具有一定的参考价值。

关键词: 材料管理; C/S; B/S; MD5

中图分类号: TP393 **文献标识码:** A **文章编号:** 1006-7043 (2010) xx-xxxx-x

Study on Design Pattern of Material Management Base on .NET Platform

MENG Jun, WANG Hui, YANG Kai, QIN Xingqiao

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: This paper introduces the designation of material managemen t system ,which is base on C/S and B/S. According to the practice of the project, it expatiates so me problems of software of material management base on .NET, It's helpful for othe r developments of analogous systems.

Keywords: material management; C/S; B/S; MD5

水务行业是一个业务逻辑关系复杂, 数据采集量巨大, 同时对安全性和准确性要求较高。随着社会经济和信息技术的发展, 一方面, 依托于网络技术的在线办公业务模式不断地发展变化; 另一方面, 为用户服务的意识也在不断地提高。因此, 充分利用网络这种快捷、方便、高效的渠道为管理提供服务, 进行信息的发布与交流、数据的上报等, 实现业务的信息整合, 因而开发符合本公司需求的管理软件是势在必行。

本系统对材料的入库、出库等日常工作实施全面的信息化管理, 及时反映各种材料的库存、使用情况, 对材料管理提供了准确的参考依据。使管理部门能够及时掌握各种材料的使用情况及经费使用情况, 从而提高企业信息化建设水平, 有效增强企业内部管理, 降低管理成本, 为社会公众提供更好的服务。

1 材料管理的功能设计

.NET 架构下的材料管理的主要功能模块结构如图 1 所示。

各子模块的具体功能如下:

材料管理模块: 包括库存管理和材料基本信息两部分, 是此管理系统的主要部分, 材料管理中涉及的内容较为全面, 主要包括材料基本信息管理, 材料入库管理, 材料出库管理, 材料退货管理, 材料计划管理, 查询统计与定制报表等。

系统管理模块: 主要是针对该系统的安全性而设立的, 主要包括用户管理, 角色管理。系统登录时需要验证身份, 只有合法的用户才可以进入系统, 不同的用户具有不同操作权限。系统管理员具有

作者简介: 孟军 (1968—), 男, 讲师, 硕士;
王辉 (1974—), 男, 讲师, 硕士;
杨凯 (1969—), 男, 讲师, 学士;
秦兴桥 (1976—), 男, 讲师, 硕士。

最高权限。另外该模块还包括数据备份（提供手工备份与定期自动备份两种系统备份的方式）和数据恢复（可自动恢复到最近自动备份数据时的状态，或者手动恢复到过去任意时刻用户保存时的数据库状态）。

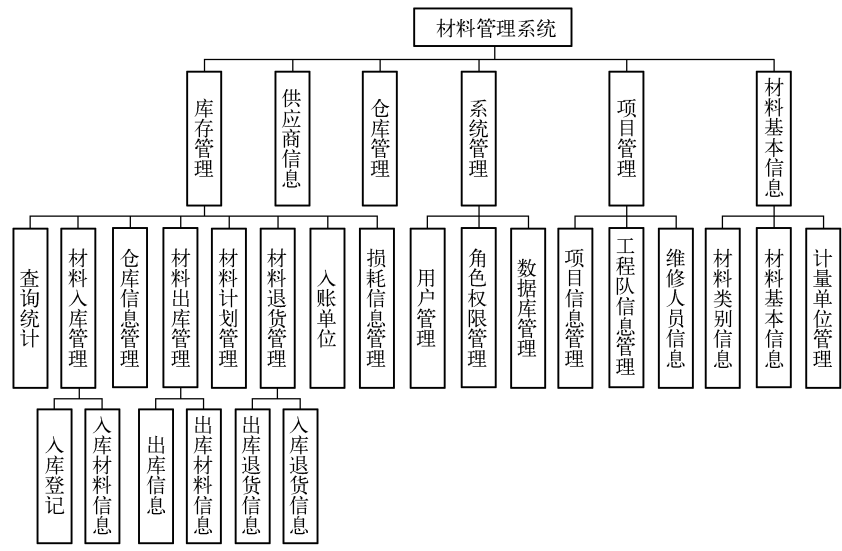


图 1 材料管理系统

项目管理模块：主要包括维修人员信息管理，工程队信息管理，项目信息等。
仓库管理模块和供应商信息模块两个模块主要是提供相关的基本信息，为其他模块提供数据支持。

2 软件详细设计

此系统开发时采用的开发工具是 Microsoft Visual Studio.Net 2008，采用的结构是 C/S 和 B/S 混合结构，其中，仓库保管员和系统管理员工作部分采用 C/S 结构，其他部分采用 B/S 结构。采用此种结构的好处是：一方面加强了数据的安全性，另一方面也为用户浏览数据提供了方便。该系统在开发过程中使用了 .NET 技术，SQL Server 2005 技术，水晶报表技术，XML 技术和 ASP.NET 技术。

2.1 数据库设计

该系统的数据库中共设计了 17 个表：供应商信息表、材料类别表、项目信息表、计量单位表、材料基本信息表、仓库信息表、工程队信息表、维修人员信息表、入库登记表、入库材料信息表、入库退货信息、库存信息表、入账单位表、出库信息表、出库材料信息表、出库退货信息表、损耗信息表等。

2.2 公共类设计

Open()函数：此函数的功能是打开数据库连接。其代码如下：

```
public void Open()
{
    if (con == null)
    {
        .....
    }
    if (con.State == System.Data.ConnectionState.Closed)
```

```
con.Open();
```

```
}
```

Close()函数：此函数的功能是关闭数据库连接。其代码如下：

```
public void Close()
{ if (con != null)
    con.Close();
}
```

Dispose()函数：此函数的功能是释放数据库连接资源。其代码如下：

```
public void Dispose()
{ if (con != null)
{ con.Dispose();
    con = null;
}
}
```

MakeInParam()函数：此函数的功能是传入参数并且转换为 SqlParameter 类型。其部分代码如下：

```
public SqlParameter MakeInParam( string ParamName, SqlDbType DbType, int Size, object Value)
{ ..... }
```

MakeParam()函数^[1]：此函数的功能是初始化参数值。其部分代码如下：

```
public SqlParameter MakeParam(string ParamName, SqlDbType DbType, Int32 Size, ParameterDirection Direction,
object Value) { ..... }
```

RunProc(string procName, SqlParameter[] prams) 函数：此函数的功能是执行参数命令文本且不需要从数据库中返回数据。主要用于执行添加、修改和删除，其部分代码如下：

```
public int RunProc(string procName, SqlParameter[] prams)
{
    SqlCommand cmd = CreateCommand(procName, prams);
    .....
    this.Close();
    //得到执行成功返回值
    return (int)cmd.Parameters["ReturnValue"].Value;
}
```

RunProc(string procName) 函数：此函数的功能是执行参数命令文本且需要从数据库中返回数据，主要用于直接执行 SQL 语句。其部分代码如下：

```
public int RunProc(string procName) { ..... }
```

RunProcReturn(string procName, SqlParameter[] prams, string tbName) 函数：此函数的功能是执行参数命令文本且需要从数据库中返回数据，主要用于查询。其部分代码如下：

```
public DataSet RunProcReturn(string procName, SqlParameter[] prams, string tbName) { ..... }
```

RunProcReturn(string procName, string tbName) 函数：此函数的功能是执行参数命令文本且需要从数据库中返回数据，直接执行 SQL 语句，其部分代码如下：

```
public DataSet RunProcReturn(string procName, string tbName) { ..... }
```

CreateDataAdapter()函数：此函数的功能是将命令文本添加到 SqlDataAdapter。其部分代码如下：

```
private SqlDataAdapter CreateDataAdapter(string procName, SqlParameter[] prams) { ..... }
```

CreateCommand()函数：此函数的功能是将命令文本添加到 SqlCommand。其部分代码如下：

```
private SqlCommand CreateCommand(string procName, SqlParameter[] prams)
{ ..... }
```

2.3 报表设计

本系统报表输出使用的是水晶报表，根据用户的要求，月底要打印材料汇总报表，分别按照不同的类别打印。使用报表可以减少一些烦琐的输入/输出设计，增强系统的健壮性，提高运行速度，加强系统的完善性。对于用户来说，可以让用户更加方便地使用本系统，提高用户对本系统的信赖度。

水晶报表生成：

```
TestSortReport billsortreport = new TestSortReport ();
try
{
    .....
    TestSortReport.SetDataSource(this.TestDataSet.Tables["TestTable"]);
    .....
}
catch (System.Exception ee)
{ MessageBox.Show(ee.ToString());}
this.crystalReportViewer1.ReportSource = TestSortReport;
```

2.4 安全设计

系统采用 C/S 与 B/S 混合模式中的安全问题主要表现在 B/S 模式上，由于 B/S 部分与公网有接口，系统为了达到系统资源的安全、数据安全和通信安全的目的，在系统设计时，B/S 部分使用了三级安全机制以防止信息的泄漏和非法用户对数据的修改和破坏。一是设置了防火墙作为材料管理系统的第一级防线，隔离了外界对服务器的直接访问；二是在防火墙和 Web 服务器间设置网关，通过 NAT 转换屏蔽访问其他端口的服务，只对访问 80 端口的 WWW 服务开放，保证了服务器数据的安全；三是在用户登录时，通过对用户密码的验证，保证不同用户的访问权限和服务权限；四是在 C/S 部分首先对用户进行了权限分配，权限分配的单位是菜单的命令项，每个不同权限的用户登录时动态生成各自不同的系统菜单，对连接字符串采用“恺撒”加密以防止通过 SQL Server 直接注册登录数据库，对用户密码进行“MD5”算法进行加密，以防止通过数据库直接获取别人的注册密码，主要代码如下：

```
Public string Md5password(string strPwd)//MD5 加密算法[2]
{
    MD5 md5=new MD5CryptoServiceProvider();
    //将字符编码为一个字节序列；
    //计算字节数据组中的哈希值；
    String str="";
    //把数组中的值写入字符变量 str；
    Return str
}

Public string Caesar (string str)
{ .....
String strCaesar="";
For(设置循环条件)
```

```
        { //对字符串进行加密 }  
    Return  strCaesar;  
}
```

3 系统特点

1) 简洁、灵活和方便的操作

软件操作简单、界面清晰美观、流程设计科学，主要功能一目了然，简单的操作，智能化的提示，方便用户的使用。

2) 完善的功能设计

例如，报表功能对材料可以按时间进行分类统计（记账单位、供应商、工程等）、盘点等操作。有库存预警功能，可以对客户设置上下限库存，当客户的库存达到设置值后，系统会有相应的提示。通过系统会使得仓库管理更加科学化，缩短货物出入库的时间，降低管理人员的工作量。

3) 良好的可扩充性和功能前瞻性

系统的设计着重于将空间和时间均离散化的物流操作，协调成顺畅的流水线式的作业流；在整个系统辖域内，统筹配置各种信息资源，减少人力物力的浪费，节约成本，以期获得最大的规模效益；系统具有友好的人性化图形用户界面、灵活的查询及统计能力。

4) 强大的安全性

材料管理流程复杂，安全要求较高，系统根据材料管理工作的特点，对用户进行合理的权限划分与管理，并提供全面的安全策略。每个用户在使用系统之前，必须进行身份验证，并根据其身份与角色配置相应的控制、访问权限。同时内嵌了加密模块，对所传输的敏感信息进行加密。

5) 较高的可靠性

平台的架构采用基于多层结构的组件开发技术，充分利用已成熟的组件，减少了系统出现错误的概率，增强了软件系统的可靠性。同时，对于用户录入的数据进行必要的检查和限制，提高了系统对异常的处理能力。

6) 易维护性

由于系统采用组件开发技术，充分体现了现代软件工程对于模块“高内聚，低耦合”的要求，因此大大降低系统维护的工作量。同时，采用了数据一致性校验、定期冗余数据检测、人工与自动相结合的数据库管理等技术，使得系统具有很强的数据维护功能^[3]。

7) 智能化

对于收料汇总、材料单项核算、材料汇总、材料使用明细等采用传统的手工或半手工方式需要耗费很大工作量才能产生的报表，使用本系统只需要输入一些简单的参数即可迅速自动生成，大大提高了工作效率。

4 结束语

本系统充分考虑到了材料管理工作的特点、功能和应用范围，选择了基于 C/S 与 B/S 相结合的混合模式的体系结构，更好地适应了日常材料管理和用户的需求，具有较强的实用性。

系统的设计、开发、运行并顺利实施，是基于系统工程与软件工程的思想，是在充分考虑到材料管理过程中的各个环节及影响因素的基础上，优化并集成相关数据，最大限度的实现数据共享，促进材料管理工作的科学化、网络化、信息化建设。

参考文献

- [1] Solid Quality Learning. SQL Server 2005 从入门到精通[M]. 北京：清华大学出版社，2006.
- [2] 张跃廷，许文武，王小科. C#数据库系统开发完全手册[M]. 北京：人民邮电出版社，2006.
- [3] 孟宪会，张慧妍. ASP.NET 2.0 应用开发技术[M]. 北京：人民邮电出版社，2006.

基于客户满意度的第四方物流 多属性指派决策机制

周宏宇¹,张战峰², 栗青生¹, 葛彦强¹

(1.安阳师院 计算机与信息工程学院 河南 安阳, 455002; 2.重庆国虹科技发展有限公司 重庆, 400000)

摘要: 本文首先深入分析第四方物流客户多样化需求,建立了客户满意度属性体系及量化标准;其次,对物流客户和第三方物流进行需求和供给的多属性分析,给出了属性需求和供给值矩阵;再次,分析属性需求和供给值之间的关系,建立了基于客户满意度的第四方物流多属性指派决策机制;最后,利用 Microsoft Excel 中的规划求解对一个实例来求解,分析和验证这一多属性指派决策机制的可行性和实用性。

关键词: 第四方物流; 客户满意度; 多属性指派

中图分类号: **文献标识码:** A

The 4th Party Logistics Multi-attribute Assignment and Decision Mechanism Based on Customer Satisfaction

ZHOU Hongyu¹, ZHANG Zhanfeng², LI Qingsheng¹, GE Yanqiang¹

(1. School of Computer , Anyang Normal University, Anyang 455002, Henan China; 2. Chongqing Guo Hong
Tech.Development Co.,Ltd, Chongqing 400000, China)

Abstract: Firstly this paper analyses diverse need s of customers of The 4th party logistic s, and establishes a system of customer satisfaction attributes and quantitative crite ria; and then analyses multi-attribute de mand and supply of l ogistics customers and third-party logistics, and gives the attrib utes demand and supply of matrix; meanwhile analyses the difference between attribut e supply value and attribute demand value, and builds the 4th pa rty logistics Multi-attribute Assignment and Decision Mecha - nism Based on Customer Satisfaction; at last through one example this paper solves、validates and analyses the feasibility and practicability of the 4th party logistics mu lti-attribute assignment and decision mech anism Based on Customer Satisfaction by using Microsoft Excel.

Keywords: fourth party logistics; customer satisfaction; multi-attribute assignment

1 引言

随着现在的市场越来越趋向于买方市场,客户的要求越来越多样化、个性化,客户也偏向于追求自身满意度最大化。作为物流服务集成商的第四方物流^[1~8],不仅考虑成本或时间等单属性指派决策问题,更要考虑安全可靠、服务水平、信息跟踪等多属性问题。因此,第四方物流在提供物流任务的最佳解决方案时,也应该以客户满意度为标准,满足客户物流的多样化、个性化需求。目前,有关第四方物流多属性指派决策研究还很少,陈建清^[9]等人研究了基于多维权的有向图的第四方物流中的优化决策模型。在此研究中,他们将第四方物流在选择第三方物流服务商时,需要考虑的价格、时间、运输能力及服务质量四个方面指标,通过多维权的有向图理论统一起来,建立了第四方物流集成

基金项目: 国家自然科学基金(60973051)、河南省重点科技攻关计划项目(092102210112)资助

作者简介: 周宏宇(1980—),男,讲师,硕士,研究方向: 科学工程计算与计算机模拟。

商选择第三方物流服务商的优化决策模型。有关基于客户满意度方面的研究，如周俊^[10]等人研究了基于顾客满意度最大化的生产指派问题，给出了基于客户满意度的生产指派模型；张建勇^[11]等人给出了基于顾客满意度的多目标模糊车辆优化调度问题的研究；姜继娇^[12]等人给出了基于顾客满意度的项目评价模糊技术的研究等。

本文主要针对第四方物流多属性指派决策机制问题进行深入研究，在分析前人有关多属性指派决策研究成果，以及第四方物流运作的基础上，提出基于客户满意度的第四方物流多属性指派决策机制。它是从客户满意度出发，深入分析第四方物流客户多样化需求，建立了客户满意度属性体系及量化标准；对物流客户和第三方物流进行需求和供给的多属性分析，给出了属性需求和供给值矩阵；以属性供给值与属性需求值之差越大为满意度越大标准，来建立的第四方物流多属性指派决策机制。

2 客户满意度属性体系

2.1 属性体系

在物流服务中，客户对第三方物流服务商的要求可能是多样化的、个性化的。例如，有些客户在看重物流成本的情况下，还会考虑服务商的服务水平、信息反馈能力、安全性能等；而另一些客户在看重时间响应能力的同时，会考虑成本、安全性能、公司信誉等。本节根据一些客户实际调查的分析，给出了一般物流客户要求的参考属性，如表 1 所示。

表 1 客户满意度属性体系

属 性	属 性 描 述
成本优化能力	客户对第三方物流服务商在成本降低方面的能力要求
时间响应	客户对第三方物流服务商在任务完成时间方面的要求
安全性能	客户对第三方物流服务商在物流过程中货物安全性的要求
服务水平	客户对第三方物流服务商服务水平的要求
物流信息反馈	客户对第三方物流服务商提供多大程度的物流信息反馈的要求
企业信誉	客户对第三方物流服务商企业信誉水平的要求
EDI 数据交换	客户对第三方物流服务商能够提供 EDI 数据交换要求

2.2 属性量化方法

从上面的属性体系中看到，大多数都属于定性属性，为了能够客观的评价，需要对所有属性进行量化和标准化。对于时间响应和安全性能，能够利用准点交货率和货物损坏率来实现，见表 2；对于其他如成本优化能力等，在本模型中采用十等级划分法来达到，见表 3。最终，通过这些方法将属性量化为[0.0, 10.0]范围内的值。同时也得到了属性的一致性，即无论哪一个属性所得的值越大表明物流客户在此方面的要求越高。

表 2 属性量化方法

属 性	属性量化方法	备 注
成本优化能力	十等级划分	见表 3
时间响应	准点交货率×10	保留小数点后一位
安全性能	货物损坏率×10	保留小数点后一位

属 性	属性量化方法	备 注
服务水平	十等级划分	见表 3
物流信息反馈	十等级划分	见表 3
企业信誉	十等级划分	见表 3
EDI 数据交换	十等级划分	见表 3

表 3 属性量化标准

客 户 要 求	优 秀	良 好	中	一 般	不 要 求
量化值	9 7 5 3 1				
	2, 4, 6, 8, 10 表示介于以上客户要求的中间值				

3 建立模型

3.1 前提假设

- ① 本模型所研究的对象是有多项任务的物流客户，模型要达到的目的是第四方物流集成商为其物流客户选择最为满意的第三方物流服务商任务指派方案。
- ② 第四方物流集成商拥有多家实力雄厚的第三方物流服务商，每家第三方物流服务商都能够独立完成客户的某一项任务。
- ③ 为了简化问题，本文假设客户每项任务的任务量都为单位 1。

3.2 客户分析

对于一个具体物流客户而言，设客户有 m 项任务，而每一项任务有 n 条任务属性。每条任务属性表示一项任务的不同任务要求，如本文属性体系中所提到的属性。每条任务属性都有一个任务属性值，它表示客户对第三方物流服务商在这一任务属性上的基本要求状况，值越大就说明客户在此属性上的基本要求越高。根据客户在不同任务上的任务属性值，可以构造出客户各任务的任务属性值需求矩阵 $\mathbf{X}$ ：

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & x_{ik} & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

(1)

其中 x_{ik} ($i=1,2,\cdots,m;k=1,2,\cdots,n$) 表示客户对第 i 项任务的第 k 条任务属性的需求值。矩阵中的第 i 行即向量 $(x_{i1}, x_{i2}, \cdots, x_{in})$ 表示客户第 i 项任务的 n 条任务属性需求值向量；而矩阵中的第 k 列即向量 $(x_{1k}, x_{2k}, \cdots, x_{mk})^T$ 表示第 k 条任务属性在 m 项不同任务下的任务属性需求值向量。客户任务要求的属性值 x_{ik} ，可以根据客户任务的具体基本要求，由第四方物流集成商利用本文 2.2 节中的量化方法进行量化和标准化，得到任务属性需求值矩阵 $\mathbf{X}$ 。客户任务属性的需求矩阵 $\mathbf{X}$ ，一旦确定就不再改动。

同时，客户对于各任务间及同一任务内各任务属性间存在价值偏好，在本文中，用 b_i ($i=1,2,\cdots,m$) 表示第 i 项任务对客户重要程度及优先级的相对权重，且 $\sum_{i=1}^m b_i = 1$ ；用 a_{ik} ($i=1,2,\cdots,m$ ； $k=1,2,\cdots,n$) 表示客户的第 i 项任务的第 k 条任务属性在该任务中的相对权重，且 $\sum_{i=1}^m a_{ik} = 1$ 。

3.3 第三方物流服务商分析

对于第四方物流集成商而言，它下面拥有 r 家第三方物流服务商，且各家第三方物流服务商的最大任务量 u_j ($j=1,2,\cdots,r$)。为了能够更好地满足客户个性化、多样化的需求，达到使客户综合满意度最大化，第四方物流集成商也为每家第三方物流服务商提出了 n 条任务属性。第四方物流集成商根据以往各家第三方物流公司任务响应的历史数据分析，利用本文 2.2 节中的量化方法进行量化和标准化，得到各家第三方物流服务商的任务属性供给值（表示第三方物流服务商在此任务属性下能够提供的物流服务能力状况）。这样，第四方物流集成商也可以建立关于 r 家第三方物流服务商的任务属性供给矩阵 $\mathbf{Y}$ ：

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & y_{jk} & \vdots \\ y_{r1} & y_{r2} & \cdots & y_{rn} \end{bmatrix} \tag{2}$$

其中 y_{jk} ($j=1,2,\cdots,r; k=1,2,\cdots,n$) 表示第 j 家第三方物流服务商的第 k 项任务属性的供给值。 $\mathbf{Y}$ 矩阵中的第 j 行即向量 $(y_{j1}, y_{j2}, \cdots, y_{jn})$ 表示第四方物流集成商下的第 j 家第三方物流服务商所能输出的任务属性供给值向量；矩阵中的第 k 列即向量 $(y_{1k}, y_{2k}, \cdots, y_{rk})^T$ 表示在第 k 条任务属性下，第四方物流集成商中各第三方物流公司所能提供的任务属性供给值向量。

3.4 模型建立

对于客户将 m 项任务提供给第四方物流集成商，其目的是希望第四方物流集成商能够为其提供最满意的第三方物流服务商指派方案，即客户满意度最大的方案。由客户分析和第三方物流服务商分析可知，对于客户某一任务的单个任务属性而言，哪家第三方物流服务商的任务属性供给值与客户任务属性需求值之差越大者，则它便是在此任务属性上使客户满意度最大化的选择。对于拥有 m 项任务的客户，应该取其综合满意度最大化，其方法也可以用类似的属性差值比较的方法来求解，但应该同时考虑到以下两点：一是任务之间及任务属性之间存在着相应的权重；二是在某一任务任何任务属性上，第四方物流集成商所选择的第三方物流服务商的任务属性供给值必须不小于客户的任务属性需求值。据此，给出的客户综合满意度最大化的目标函数为：

$$\max \theta = \sum_{i=1}^m b_i \sum_{j=1}^r \left\{ \frac{z_{ij} \left[\prod_{k=1}^n g(y_{jk} - x_{ik}) \right]}{\frac{1}{\sqrt[n]{\sum_{k=1}^n a_{ik} (y_{jk} - x_{ik})^2}}} \right\} \tag{3}$$

变量 z_{ij} 表示为：

$$z_{ij} = \begin{cases} 1, & \text{表示客户的第} i \text{项任务由第} j \\ & \text{家的第三方物流服务商完成；} i=1,2,\cdots,m; j=1,2,\cdots,r \\ 0, & \text{否则？} \end{cases}$$

$g(x)$ 为符号函数，满足 $g(x) = \begin{cases} 1, & x \geq 0; \\ 0, & x < 0. \end{cases}$ ； $\prod_{k=1}^n g(y_{jk} - x_{ik})$ 为了满足第二个条件，当存在任意 $y_{jk} - x_{ik} < 0$ ($k=1,2,\cdots,n$) 时，即 $\prod_{k=1}^n g(y_{jk} - x_{ik}) = 0$ ，则第三物流服务商 j 为客户第 i 项任务的不可行方案。

在客户的某一项任务上，为了表示第三方物流服务商所提供的多元化服务优于客户任务服务基本需求的程度，本文利用第三方物流服务商所提供的任务属性供给值与客户任务属性需求值之间的 Minkowski 距离来表示。由于欧氏距离（Euclid）具有很多优良的性质，本文应用欧氏距离（Euclid）来表示，欧氏距离（Euclid）越大说明客户满意度就越大，同时也考虑了客户任务各任务属性间的相对权重，则欧氏距离（Euclid）表达式为^[13,14]：

$$d_{ji} = \sqrt[2]{\sum_{k=1}^n a_{ik} (y_{jk} - x_{ik})^2}, \quad j=1,2,\cdots,r; i=1,2,\cdots,m \tag{4}$$

式中， d_{ji} 表示第三方物流服务商 j 所提供的多元化服务优于客户任务 i 服务基本需求的程度， d_{ji} 值越大，即第三方物流服务商 j 所提供给客户任务 i 的满意度越大。

目标函数的含义： θ 值越大，则客户综合满意度越大，当 θ 值最大时所对应的方案，即为第四方物流集成商为客户选择的最满意方案。

由此，所建的客户综合满意度最大化模型为：

$$\max \theta = \sum_{i=1}^m b_i \sum_{j=1}^r \left\{ \frac{z_{ij} \left[\prod_{k=1}^n g(y_{jk} - x_{ik}) \right]}{\frac{1}{\sqrt[2]{\sum_{k=1}^n a_{ik} (y_{jk} - x_{ik})^2}}} \right\}$$

s.t:

$$\sum_{i=1}^m z_{ij} \leq u_j, \quad j=1,2,\cdots,r \tag{5}$$

$$\sum_{j=1}^r z_{ij} = 1, \quad i=1,2,\cdots,m \tag{6}$$

$$\sum_{k=1}^n a_{ik} = 1, \quad i=1,2,\cdots,m \tag{7}$$

$$\sum_{i=1}^m b_i = 1 \tag{8}$$

$$x_{ik} \in X, y_{ik} \in Y, z_{ij} = 0 \text{或} 1$$

$$g(x) = 0 \text{或} 1$$

约束条件（5）表示第四方物流集成商最终所分配给第 j 家第三方物流服务商的总任务量不大于该物流服务商的最大任务量 u_j ($j=1,2,\cdots,r$)；约束条件（6）表示保证对于客户的每项任务最终只有一家第三方物流服务商为其服务；约束条件（7）表示对于第 i 项任务的各任务属性权重之和等于 1；约束条件（8）表示客户各项任务之间权重之和等于 1。

4 实例分析

为了更好地说明问题，下面以一个实例来进行分析。现有一个物流客户 X 有 10 项任务，对于其各任务提出了一个多样化、个性化的服务需求。为了简化问题，本例假设这些任务是双属性的，即成本优化能力和时间响应两个属性。其客户各任务的属性基本需求值如表 4 所示，计算方法此处不在细述。

表 4 成本优化能力与时间响应属性基本需求值

客户任务	任务 1	任务 2	任务 3	任务 4	任务 5	任务 6	任务 7	任务 8	任务 9	任务 10
成本优化能力	6.5	6.7	7.0	7.5	8.0	6.4	7.1	7.0	6.0	7.5
时间响应	7.1	7.8	8.8	7.2	8.0	6.5	8.2	7.6	8.5	7.0

对于第四方物流集成商 Y 而言，它拥有四家实力雄厚的第三方物流服务商。根据第三方物流服务商历史数据，给出一个与上面客户属性一致的各第三方物流服务商属性供给值，如表 5 所示。

表 5 成本优化能力与时间响应属性供给值

第三方物流服务商	成本优化能力供给值	时间响应供给值	最大任务量
1 8.3		8.2	4
2 7.5		7.5	3
3 7.9		7.9	4
4 9.1		9.1	3

为了简化问题，本例不考虑任务重要程度与优先性的相对权重 b_i ($i=1,2,\cdots,m$)，以及任务属性间的相对权重 a_{ik} ($i=1,2,\cdots,m;k=1,2,\cdots,n$)。这样，本例的求解复杂难度就大大减少。在本文中，采用了 Microsoft EXCEL 中规划求解的程序来实现，则决策模型计算结果如表 6 所示。

表 6 决策模型计算结果

客 户 任 务	物流服务商 1	物流服务商 2	物流服务商 3	物流服务商 4
1 0 0 1 0				
2 0 0 0 1				
3 0 0 1 0				
4 0 1 0 0				
5 0 0 0 1				
6 0 0 1 0				
7 1 0 0 0				
8 0 0 1 0				
9 0 0 0 1				
10	0 1 0 0			

由此可知，当第四方物流集成商按照上表把客户任务分配给第三方物流服务商时，可以得到的综合满意度最大 θ 值为 20.20670321。同时，通过其他方案的验证，所得值都小于 20.20670321。所以能够使客户综合满意度最大化的最优指派方案如表 6 所示，即第四方物流集成商将任务 1、3、6 和 7 分配给第三方物流服务商 1，将任务 5 分配给第三方物流服务商 2，将任务 2、10 分配给第三方物流服务商 3，将任务 4、8 和 9 分配给第三方物流服务商 4。

5 结语

随着现在的市场越来越趋向于买方市场，客户的要求越来越多样化、个性化，客户也偏向于追求

自身满意度最大化。目前，在第四方物流指派决策研究中，更多是针对单属性指派决策问题进行研究，而第四方物流多属性指派决策研究还没有。本文通过分析客户满意度属性，建立客户满意度属性体系，提出一种基于客户满意度的第四方物流多属性指派决策机制；并通过一个实例来验证模型的可行性和实用性，得到很好的指派结果。这一第四方物流多属性指派决策机制将使第四方物流能够进行多属性指派，选择最满意的第三方物流服务商给物流客户，满足其物流客户多样化、个性化的需求。

参考文献

[1] GATTORNA J. Strategic supply chain alignment[M]. Aldershot, Hants, England: Gower Pub Co., 1998: 45-60.

[2] Foster Tom, 4PLs: the new generation for supply chain outsourcing, logistics management and distribution report V. no4 (apr,1999) p 35. ISSN:1089-537x.

[3] 杨宝军, 李华增. 第四方物流剖析[J]. 工业工程与管理, 2003(3): 49-52.
Baojun Yang, Huazheng Li. An Analysis of the Fourth Party Logistics[J]. Industrial engineering and management, 2003(3): 49-52.

[4] 文海旭, 冯兰杰. 第四方物流在中国施行的现实性论证[J]. 软科学, 2003, 17(5): 31-35.
Haixu Wen, Lanjie Feng. The reality Prove about the Fourth Party Logistics in China[J]. soft science, 2003, 17(5): 31-35.

[5] 唐斌, 唐万生. 第四方物流及其在我国的发展[J]. 工业工程, 2003, 6(2): 42-46.
Bin Tang, Wansheng Tang. The Fourth Party Logistics and its Development in Our Country[J]. IE industrial engineering, 2003, 6(2): 42-46.

[6] 陈久梅. 第四方物流及其业务流程研究[J]. 科技进步与对策, 2004, 3: 109-110.
Jiumei Chen. Distributed Data Mining System of the 4th Party[J]. Sci-tech Progress and Suggestions, 2004, 3: 109-110.

[7] Xiu Li, Wenhuan Liu, etc. The Design and Realization of Four Party Logistics[C]. Proceedings of the 2003 IEEE International Conference on Systems, Man and Cybernetics. Washington DC, 2003, 1: 838-842.

[8] Hoong Chuin UAU Yam Guan GOH, An Intelligent Brokering System to Support Multi-Agent web-Based 4th-party Logistics[C], IEEE International Conference on Tools with Artificial Intelligence. Washington DC, 2004, 2: 741-762.

[9] 陈建清. 第四方物流中基于多维度的有向图模型及算法[J]. 工业工程与管理, 2003, 3: 45-47.
Jianqing Chen. The Directed Graph Model with Multi Dimensions in the Fourth Party Logistics and Its Algorithm [J]. In -dustry System and Management, 2003, 3: 45-47.

[10] 周俊, 梁樑, 余玉刚. 基于客户综合满意度最大化的生产指派问题[J]. 管理工程学报, 2004, 18(4): 1-5.
Jun Zhou, Liang Liang, Yugang Yu. Maximizing production of the question of appointed Based on Customer Satisfaction. Journal of industrial engineering and engineering management, 2004, 18(4): 1-5.

[11] 张建勇, 郭耀煌, 李军. 基于顾客满意度的多目标模糊车辆优化调度问题研究[J]. 铁道学报, 2003, 25(2): 15-17.
Jianyong Zhang, Yaohuang Guo, Jun Li. Research of multi-objective fuzzy vehicle scheduling problem based on Customer Satisfaction. Journal of the China Railway Society, 2003, 25(2): 15-17.

[12] 刘树林, 邱苑华. 多属性决策的 TOPSIS 夹角度量评价法[J]. 系统工程理论与实践, 1996, 16(7): 12-16.
Shulin Liu, Wanhua Qiu. The TOPSIS Angle Measure Evaluation Method for MADM[J]. Systems Engineering —Theory & Practice, 1996, 16(7): 12-16.

[13] A. Shanian, O. Savadogo. TOPSIS. multiple-criteria decision support analysis for material selection of metallic bipolar plates for polymer electrolyte fuel cell[J]. Journal of Power Sources, 2006, 159: 1095-1104.

[14] Hung-Tso Lin, Wen-Ling Chang. Order selection and pricing methods using flexible quantity and fuzzy approach for buyer evaluation[J]. European Journal of Operational Research, 2008, 187: 415-428.

面向用户和领域本体的 Web 信息采集系统

张素智，李宝燕，樊得强

(郑州轻工业学院 计算机与通信工程学院，河南 郑州，450002)

摘 要：针对传统搜索引擎不能满足用户个性化专业化需求的特性，提出了一种既面向用户又面向领域本体的搜索策略，设计了一个本体支持的 Web 信息采集系统。该系统通过在网站模式的网页配置文件中记录网页本体信息，来阐明网页如何与领域本体相关。为了满足用户的个性化需求，在爬虫内部设计了用户定义的优先对列。实验证明该系统提高了页面查询的精确率和召回率。

关键词：爬虫；本体；个性化技术；网站模式

User and Domain Oriented Web Information Collection System

ZHANG Suzhi, LI Baoyan, FAN Deqiang

(College of Comp. and Com. Eng., Zhengzhou Univ. of Light Ind., Zhengzhou 450002, Henan China)

Abstract: Based on the character that traditional search engines can't satisfy the individuating and specialization demands of users, this article proposed a both User-Oriented and Domain-Oriented way in the search for the strategy, and designed an ontology-supported web information collection system. This system annotates how a webpage is related with the domain through collecting the ontology information of the webpage in the configuration webpage files. In order to satisfy the users' personalization, this system designed a user-defined priority List in the crawler's internal. Experiments indicate the system improved the accuracy and recalled rate of the query of webpage.

keyword: crawler; ontology; personalisation technology; website model

0 引言

通用搜索引擎由于其针对整个 Web 资源，面向全体网络用户需求的特点，越来越不能满足用户个性化专业化的需求，特定领域的搜索引擎利用查询扩展技术帮助用户缩小了搜索范围。例如自动分类技术和主题爬行技术等，但是它们几乎完全忽略了用户的个性需求。如何让用户快速，精确地找到所需信息已成为用户搜索的重要组成部分。

本体建模语言 OIL、DAML 及 DAML+OIL 和本体概念的提出为信息的语义归属提供了一种可能。本文在支持本体的信息采集系统中加入用户个性化技术，使信息采集既面向领域本体又面向用户，从而提供更加快速，准确和稳定的查询结果。

1 相关概念

1.1 爬虫

爬虫的概念主要出现在 Web 信息采集系统，主要用来进行信息的采集和集成，改善信息的收集过程及不同资源的搜索结果。例如，Dominos^[1]可以每秒抓取数千个网页，包括一个透明配置、平台独立的高性能故障处理器，因而不会产生额外的硬件开支。Ganesh, Jayaraj, Kalyan and Aghila^[2]提出的关联十进制基于领域本体估计 URL 的语义内容，加强了列入优选网址队列的度量。UbiCrawler^[3]是一个可扩展的分布式 WebCrawler，是平台独立，线性可扩展的，优化简化了故障问题的存在，采用一

个非常有效的分配函数来进行对领域的分割检索，很大程度上，对每一个任务的完成进行权力下放。在本文中，我们设计了一个本体爬虫使用本体和网站模式作为核心技术，它可以解决搜索范围和用户兴趣等问题。

1.2 本体的概念

本体的定义有很多种，目前最著名并被引用得最为广泛的定义是由 Gruber 提出的“本体是概念化对象的明确规范说明”^[4]。Ontology 提供了一种明确的定义语义的方式，通过本体定义的语义，使机器能够进行互操作，使机器能够理解数据的语义，从而达到语义 Web 的数据是机器可理解的要求。本体中体现的是共同认可的知识，反映的是相关领域中公认的概念集，它所针对的是团体而不是个体；本体的目标是捕获相关领域的领域知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇术语和词汇之间相互关系的明确定义。

1.3 个性化技术

个性化技术^[5]，即对不同的用户根据用户的个性行为采取不同的、有针对性的服务策略，提供符合用户个性化需求的服务内容。目前出现的支持个性化的三类技术有：（1）人工决策支持系统，由人工方式寻求用户的个性化需求；（2）基于内容的过滤系统，即首先为用户建立一个兴趣总集，然后根据用户的历史访问记录建立用户的兴趣子集，在搜索资源时根据用户兴趣子集与待访问资源的相似度来进行过滤；（3）协同过滤系统，先将用户分类为不同的兴趣群，然后为相同兴趣群的人提供相似的个性化服务。

2 系统设计

2.1 系统框架

本文提出的框架结构主要包括三部分，如图 1 所示：（1）爬虫；（2）本体管理；（3）用户兴趣管理。相比普通本体爬虫框架^[6]增加了用户兴趣管理模块。其中，（1）爬虫是按照页面抓取——页面预处理——主题过滤——链接分析来不断地进行循环工作的；（2）在本体的管理中，首先构建本体，我们采用斯坦福大学开发的本体编辑和知识获取软件 Protégé 软件进行本体的构建，再按照本体管理——链接分析——更新本体这个循环对本体进行更新与维护；（3）用户兴趣管理，通过收集和更新用户个性化信息形成用户兴趣向量，然后通过用户的兴趣决定系统的搜索范围。

2.2 本体支持的爬虫内部设计

我们提出了一个本体支持的主题爬虫如图 2 所示，采用一种改进的获取领域相关信息的检索策略来获取信息。在结构的内部，爬虫从网络上收集数据；文本池存储所有采集回来的 Web 信息；文本抽取器用于网页配置文件的建设，还存储搜索引擎的查询结果，通常包含一个网址列表，其中网址只提取查询结果中领域相关网址并对这些网址进行调度，但不在网站模式中进行存储；面向用户的网页扩展在网站模式下通过用户查询来扩展用户感兴趣的网站网址；用户优先级队列存储用户的搜索字符串和在网站模式下从面向用户的网页扩展得到的网址；网站优先队列存储网站模式下从自主网址程序和网址提取器得到的网址；分类器通过给每个

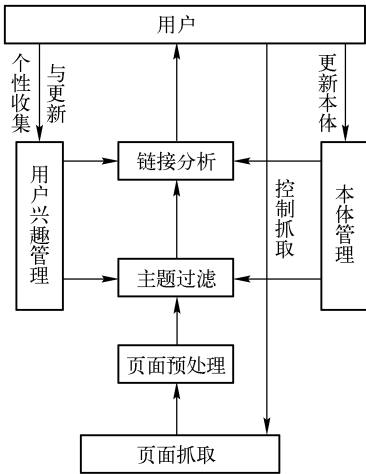


图 1 面向用户的本体爬虫框架

网址（或搜索字符串）一个相关的优先评分，来控制网络搜索并把每个网址放置在一个适当的优先级别队列里。我们把每个 URL（或者字符串）的评分定位为它的优先级别，其中搜索字符串优先级最高，网站模式处理的网址优先级次之，由 URL 抽取器提取的 URL 优先级最低。

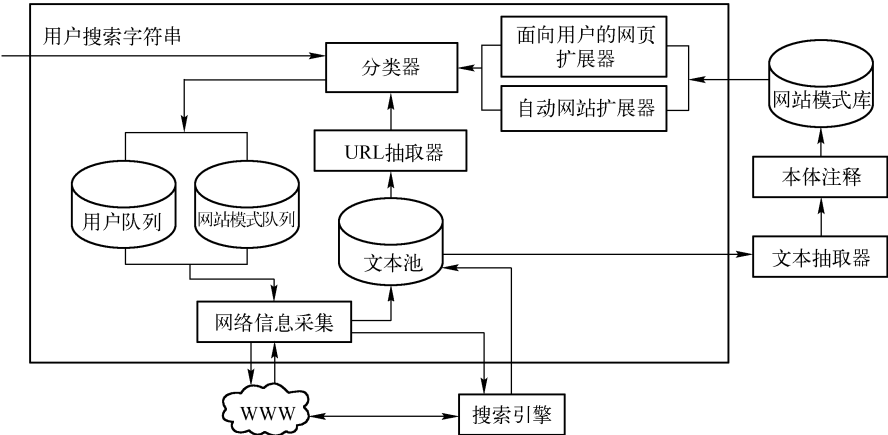


图2 面向用户的本体支持的爬虫系统

这种设计主要是维护面向用户的 Web 资源检索、面向用户的查询和网页扩展, 兼顾用户的兴趣和领域的限制, 可以更好地满足我们的设计目标。

2.3 网站模式的设计

网站的模式结构如图 3 所示,包括网站配置文件和网页配置文件。这种模式结构有助于解释通过网站收集的信息的语义,还有助于网页信息和网络资源快速地自主搜索和检索。

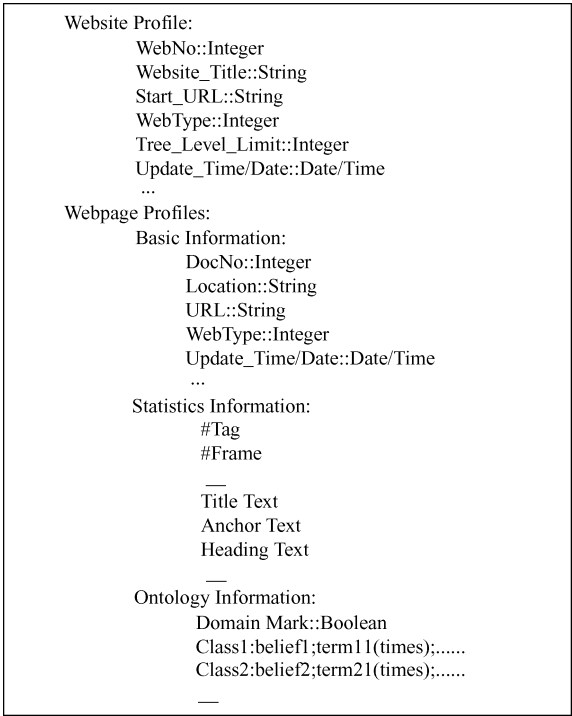


图 3 网站模式

在网站配置文件中 WebNo 标识一个网站，通过这个号码，我们可以访问这个网站配置文件描述

的这些网页。Website_Title 记录该网站的主页<TITLE>标签之间的文本。Start_URL 存储了 Website 的启动地址。WebType 确定了在网页配置文件中的使用类型之一。Tree_Level_Limit 限制了搜索的深度。Update_Time/Date 为网站被修改最后时间。

网页配置文件中包含三个部分：基本信息、统计信息和本体的信息。第一部分和第二部分为网页基本信息，最后为网页的语义领域。其中 DocNo 为自动识别所产生的一个网页的结构体系指标。Location 记录了网页在网站模式中存储的路径，我们可以用它来响应用户查询。URL 是网页在互联网中的路径，与用户查询返回的 URL 相同，它有助于超链接分析。WebType 确定了以下 6 个网站类型之一：com(1), net(2), edu(3), gov(4), org(5), and other(0)，每个编码为括号中的一个整数。WebNo 表示为该网站包含此网页。Update_Time/Date 记录网页被修改的最后时间。Statistics Information 部分统计 HTML 标记属性，具体来说，记录标题、锚及网页分析相关的文本，也记录面向用户的网页扩展 Outbound_URLs。最后是网页本体信息记录部分，记录网页是如何由领域本体解释的。Domain_Mark 用于标记网页是否属于某个特定的领域。这一部分说明了网页是如何与领域相关的，可以作为网页的语义，有助于网页的正确检索。

在构建和扩展一个网站模式的过程中，我们需要提取原始网页的资料并执行统计。网站模式包括三个模块，本体抽取模块、本体注释模块和本体分类模块^[7]。简单来说，我们使用本体抽取模块提取网页的基本信息并执行统计，然后注释本体信息。由于本体信息包含很多类型的网页，本体注释模块需要调用本体分类器来执行网页分类，为了使分类更加明确，把本体结构重组为两层结构（超类和参考类）^[8]，强调了概念属性与类定义的联系。每个超类包含一系列有代表性的特定功能的本体概念集，而每个引用类包含本体特征之间的相关特性，这种设计明确构建了本体类之间的语义结构和关系，可以作为一种网站扩展的快速语义定义工具。本体分类有两个阶段，第一阶段通过计算网页/网站出现的特定类中实体的特征在网页中出现的数量来测量网页/网站与一个特定类的相关性，把相关性高的归为一类。如果第一阶段不能返回一类网页/网站类型，我们进入分类的第二阶段，它通过对另一组相关的本体特性采用一种加权机制对相关的网页/网站进行分类。

2.4 Web 信息采集运行系统的运作与技术

- （1）开始：把内部查询转换为 URI 代码，然后嵌入到谷歌、百度的网址查询接口。
- （2）谷歌/百度搜索器：首先通过谷歌、百度搜索器把一个声明好的 URL 对象添加到查询接口中，从而得到更加良好的 URI 编码，然后用一种迭代循环的方式逐行读取其内容。最后，把网页的 HTML 源文件内容以文本形式输出，为最终分析作参考。
- （3）检索链接：研究表明，由于谷歌、百度的网页是通过用户输入关键字，从本地数据库中动态生成的，因此利用正则表达式寻找用户兴趣相关的 URL，不能及时返回检索到所有链接。所以利用一个迭代循环与双重运行的方法，更加完全地抽取到与用户搜索条件相符合的超链接，最后，把返回的所有超链接统一输出到文本文件中，提供给系统进行下一步处理。
- （4）检索内容：利用步骤（3）中输出的文本文件，以迭代循环的方式读取其内容，并确保获取的每一个 URL 都能真正链接到网页。然后通过读取网页源文件判断网页的类型码，并给予正确的标记，最后统一输出到文本文件中，以便系统进行下一步处理。在完成上述所有程序后，我们可以用搜索匹配方法来判断该网页是否位于我们查询的期望距离；假定答案为“是”，我们将执行 Remove HTML Tags，在源文件中删除 HTML 标签，只保留文本内容，以便系统进一步的分析和处理。把收集到的与查询相关的网页数与网页总数相比，得到查询处理的百分比。
- （5）搜索匹配：支持（4）中的内部通话服务，判断该网页是否属于我们的搜索范围。比较网页提取内容与数据库中的本体内容，如果有返回与我们设定值相对应的值，系统将返回“true”布尔变量，这意味着该网页符合查询条件，相反，如果返回“false”则表示不符合我们的网页查询条件。

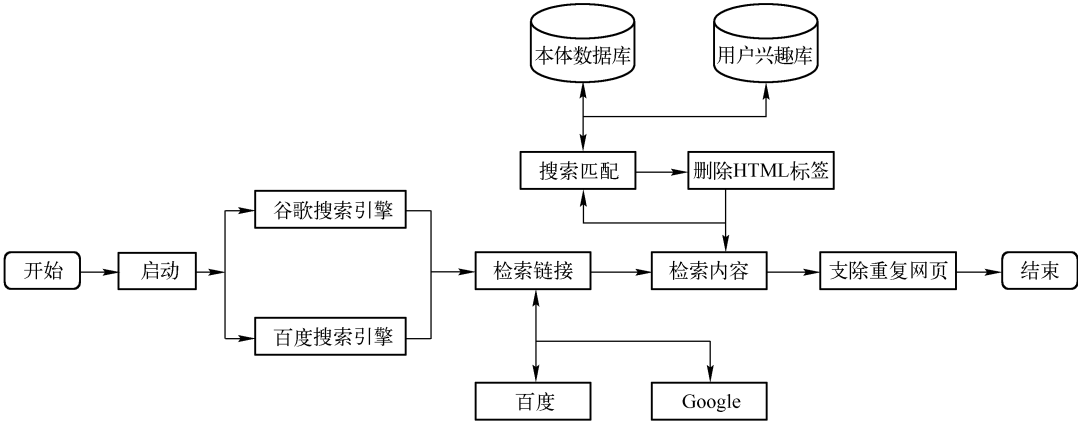


图 4 Web 信息采集的运作体系结构

(6) 删除 HTML 标签：和搜索匹配一样，它支持 (4) 的内部响应服务，删除 HTML 源文件中的 HTML 标签。

(7) 去除重复网页：对谷歌和百度返回的页面进行完全的交叉对比以删除重复的网页，避免系统后端的重复操作，提高其性能。

3 实验与分析

3.1 本体的构建

本体的构建可以分为两个阶段：第一阶段用 Protégé 定义作者本体，图 5 所示为作者领域的本体结构。第二阶段将 Protégé 定义的作者领域本体转移到 MS-SQL 数据库中。具体步骤如下：

- (1) 用 Protégé 定义作者本体。
- (2) 导出一个在 Protégé 知识上构建的 XML 文件，然后导入到 MS Excel 中纠正。
- (3) 最后，把 MS Excel 导入到中完成本体建设。

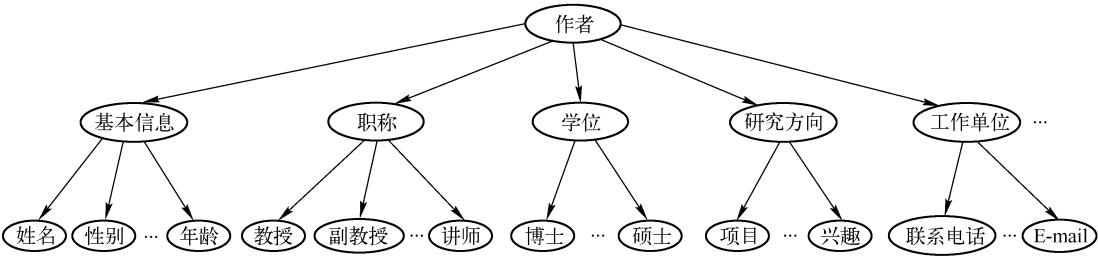


图 5 作者本体领域的本体结构图

3.2 用户兴趣信息的获取

用户相关兴趣信息采用协同过滤系统来获取：首先将用户分类，形成多个用户兴趣群，为相同兴趣群的人提供相似的个性化服务。其中用户的个性化偏好用兴趣向量方式来存储与表示。将各种兴趣爱好划分成若干兴趣主题， $I=(I_1, I_2, I_3, \dots, I_i, \dots)$ 每一个兴趣主题为向量模型中的一个项，用户对兴趣主题的兴趣度用 H 来表示，则所有兴趣项对应的兴趣度也可以用向量表示为： $H=(H_1, H_2, H_3, \dots, H_i, \dots)$ 。因此用户的兴趣可以表示成为一个序偶对向量 $C: \langle I_1, H_1 \rangle, \langle I_2, H_2 \rangle, \dots, \langle I_i, H_i \rangle, \dots$ 。向量的初始化通过用户向系统提交兴趣偏好得到，之后通过记录用户的行为动态进行更新。我

们通过用户兴趣向量形成待查询关键词集，然后调用元搜索引擎获取 Web 信息，计算返回页面与用户兴趣的相关度，最后以倒排索引文件方式返回检索结果。

用户兴趣向量与页面摘要之间的相关度采用式（1）来计算，其中 H_1 代表用户兴趣向量， H_2 代表页面摘要向量， W_{1n} ， W_{2n} 分别为 H_1 ， H_2 中的项， $\text{Sim}(H_1,H_2)$ 的取值越大，说明两者越相似；反之则越不相似。通过分析这些结果 URL，可以发现很多 URL 来自同一站点，则把该站点定义为目标站点。

$$\text{Sim}(H_1,H_2)=\cos\theta\frac{\sum_{i=1}^nW_{1i}*W_{2i}}{\sqrt{\left(\sum_{i=1}^nW_{1i}^2\right)\left(\sum_{i=1}^nW_{2i}^2\right)}}\tag{1}$$

3.3 结果分析

在面向用户与领域本体的爬虫系统（UOCrawler）中输入关键字“张素智”后得到的返回界面，如图 6 所示，在这个实验中，我们与谷歌和百度搜索引擎做比较，首先对百度、Google 返回的前 100 个网页逐一与领域本体进行比较，然后使用式（2）和式（3）确定精确率 R_p 和召回率 R_R ，其中 N_T 是指返回的网页的总数； N_C 是指返回网页的正确数； N_R 是指返回的相关网页数量。

$$R_p=\frac{N_C}{N_T}\tag{2}$$

$$R_R=\frac{N_C}{N_C+N_R}\tag{3}$$



图 6 UOCrawler 返回的界面

通过比较得到的结果如表 1 所示，其中谷歌的 R_p 和 R_R 分别为 14%和 48.3%，百度分别为 9%和 40.9%。面向用户和本体的爬虫（UOCrawler）在搜索网页上提供了比谷歌和百度更高的精确度和召回率。

表 1 比较结果

	N_C	N_R	N_T	R_p (%)	R_R (%)
Google 19		72	100	19	20.9
Baidu 17		65	100	17	20.7
UOCrawler	21 4		25	84	84

4 结语

本体论与定题信息采集相结合使得搜索引擎有了一定的语义理解能力，本文提出了本体支持的一种网站模式，通过该网站模式设计和应用，说明了网页是如何与领域本体相关的，有效地支持了网络搜索。该方法提高了网页查询的精确度和召回率。由于我们的本体构建并不完善，因而不能达到信息更加精确的返回。本文中本体建设仅是为所要收集的网页设置一个特定的领域范围，还不能实现领域本体的自动构建及对多领域本体信息的采集。另外，返回的结果中存在大量重复网页等问题。

参考文献

- [1] Hafri Y., & Djeraba, C.. Dominos: A ne w Web Crawler’s design. In Proceedings of the fourth international web archiving workshop, Bath, UK. (2004).
- [2] Ganesh, S., Jayaraj, M.,Kalyan, V., & Aghila, G. (2004). Ontology-based Web Cr awler. In Proceedings of the interna - tional conference on information technology. Coding and computing (p337–341), Las Vegas, NV, USA.
- [3] Boldi,P., Codenotti,B., Samtini,M., &Vi gna,S. (2004). UbiCrawler: A scalablefu lly distributed Web Crawler. Software : Practice and Experience, 34(8), 711-726.
- [4] 蒋子龙. 基于本体的专题性搜索引擎的研究与实现. 武汉理工大学, 2009.
- [5] 王忠, 程磊. 基于元搜索引擎的个性化 Web 信息采集. 计算机工程与设计. 2009, 30 (13):3117-3119.
- [6] 郑健珍, 林坤辉, 周昌乐, 康恺. 基于本体语义的定题爬虫. 山东大学学报（理学版）2006 年 6 月, 第 41 卷, 第 3 期. 90-94.
- [7] Sheng-YuanYang. OntoCrawler: A focused crawler with ontology-supported website models for information agents. Expert Systems with Applications,37 (2010), 5381-5389.
- [8] Yang, S.Y. (2006a). An Ontology-Direct ed Webpage Classifier for Web Services. In Proceedings of Joint the 3rd Interna - tional Conference on Soft Computing and Intelligent Systems and the 7th International Symposium on advanced Intelligent Systems, Tokyo, Japan (pp. 720-724).

ERP 项目中成本管理子系统的分析与设计

申 康

(河南省政法管理干部学院, 河南 郑州, 450002)

摘 要: 成本管理涉及的内容很多, 主要包括成本计算、成本计划、成本日常控制、管理与成本分析。采用面向对象的分析方法和统一建模语言 UML, 描述成本管理系统主数据部分和内部订单子系统的用例图, 设计出面向对象的系统模型。在实例中使用顺序图描述了用例的实现; 使用类图来描述系统各个类及其关系。在研究成本管理的功能层次及各功能层次的具体解决问题的基础上, 设计了系统的总体结构模型、功能模块和数据库结构。整个设计和开发过程中使用面向对象的程序设计思想, 实现了内部订单会计、主数据的维护、计划、预算、成本计算单查询等功能。通过本系统的实现, 明显提高了企业成本核算的效率, 使成本核算的时间有很大的缩短, 并使成本核算的准确性有很大提高。

关键词: 企业资源计划; 面向对象; 成本管理; 统一建模语言

Analysis and Design of the Subsystem of the Cost Control System in the ERP Project

SHEN Kang

(HeNan Administrative Institute of Politics and LAW, Zhengzhou 450002, Henan China)

Abstract: The Cost management related to a lot of content, including cost, the cost of the scheme, the cost of day-to-day control, management and cost analysis. Object-oriented analysis of the Unified Modeling Language and UML, described the cost of data management system for the main part of the sub-system and internal order of use case diagram, object-oriented design of the system model. In the instance of the use of the order described plans to use the case to achieve; plans to describe the use of various types of systems and their relationship. In the study of cost management functions and levels of functional level to address the specific issues on the basis of the design of the system's overall structure, function and structure of the database. The whole process of design and development using object-oriented programming ideas, the realization of the internal accounting orders, master data maintenance, planning, budget, costing a single query, and other functions.

Through this system, significantly improved the efficiency of the business cost accounting, cost accounting so that there is a lot of the time shortened, and cost a lot to improve the accuracy.

Keywords: Enterprise Resource Planning; Object-oriented; cost management; unified modeling language

近年来, 随着我国经济的快速发展, 计算机和网络技术的不断发展, 我国企业的信息化建设也开始蓬勃发展, 越来越多的企业认识到信息化建设对企业发展的重要性。

ERP的实施对企业的影响大致可以分为以下几个方面^[1]。

(1) ERP体现了先进的生产管理思想。

ERP 的核心管理思想就是实现对整个供应链的有效管理, 主要体现在以下三个方面: 体现对整个供应链资源进行管理的思想, 体现精益生产、同步工程和敏捷制造的思想, 体现事先计划与事中控制的思想。

(2) ERP 中蕴涵先进的成本管理思想。

(3) 企业在实施 ERP 的过程中提升自己的经营创新能力。

(4) ERP 系统为企业提供全方位的创新工作方式。

(5) ERP 系统能创造性地应用各种信息资源。

作者简介: 申康, 男, 1983 年 6 月出生, 河南省政法管理干部学院计算机科学系, 助教, 硕士学位。研究方向: 数据库系统应用。

近年来，随着先进企业经营方式和管理模式的改革，越来越多的企业使其管理系统升级为 ERP 系统。这些企业的 ERP 系统的功能更加强大，集成化程度越来越高，从整体上提高了企业的市场竞争力。

本文给出制造业企业资源计划（ERP）成本管理子系统的分析、设计和实现工作。所做工作主要有以下几个方面的内容。

- （1）分析和研究成本管理模块的系统结构和功能层次。
- （2）研究各功能层次的具体解决问题的和提供的各项功能。
- （3）使用用例图描述系统的需求并在此基础上建立系统的总体结构模型。
- （4）以定单管理主数据模块为例进行详细阐述，给出顺序图来描述用例的实现。
- （5）使用类图来描述系统各个类及其关系。

1 成本管理系统的需求分析

需求分析作为软件工程的第一阶段，是整个软件开发项目进行设计和实现的基础，决定了一个项目的成败。现在的软件项目中返工开销几乎占了总开发的一半，而导致返工的主要原因是需求分析不明确。本章介绍了成本管理子系统的需求分析，用用例图的形式表示出来。

1.1 需求的获取

清晰、正确的需求分析对整个系统来说是至关重要的。基于 UML 的软件需求分析，通过使用用例图，避免了文字描述的弊端，为建立系统正确的需求分析提供了保证。因此，我们将按照 UML 标准的用例视图来组织这些需求^[2]。

获取需求的工作步骤如下。

1) 获取开发所需信息

通过与企业的工作人员的交流，了解他们对这个系统的总体需求是什么。通过学习成本管理的相关知识，理解 ERP 成本管理的基本原理。消化国外优秀 ERP（如 SAP R3）软件的基础上，结合企业的需求特点进行开发^[3]。

2) 从信息中识别角色

角色是指在系统外部与系统进行交互的人或物。

这个模块中所涉及的角色有系统管理员、成本中心业务管理员、成本会计。角色描述如下：

- （1）系统管理员：用户信息，角色信息的维护。
- （2）成本中心业务管理员：负责对成本中心会计中的所有需要一次性定义的数据项进行设计，并维护到系统中。
- （3）成本会计：负责将自己的会计业务用计算机系统来完成。

3) 从信息中识别用例

用例是一种需求技术，一种基于用户目标的需求组织技术，一种有层次的需求组织技术。对已确定的每个角色回答下列问题可以确定用例^[4]。

- （1）该角色需要从系统获取哪些功能，该角色需要做什么。
- （2）该角色是否需要在系统中阅读、建立、修改或保存信息。

通过回答这些问题，基本可以确定系统的用例了。

在这个系统中，对前面提到的角色获取用例，简单介绍如下：

系统管理员：用户信息的增加、删除、修改；角色信息的增加、删除、修改。

成本中心业务管理员、成本会计：通用模块、主数据模块、手工计划模块、分配计划模块、实际记账模块、期末结算模块的操作。

通用模块包括维护成本控制范围，维护成本版本，维护控制范围相关参数，维护成本版本年度相

关参数，维护货币和评估参数文件，设置控制范围，设置期间锁，显示期间锁。

主数据模块包括创建初级成本要素，创建次级成本要素，创建成本中心，创建作业类型，创建分配因子，修改成本要素，删除成本要素，维护成本要素组，显示组使用点，修改成本中心，删除成本中心，维护成本中心组，修改作业类型，删除作业类型，维护作业类型组，修改分配因子，删除分配因子，维护分配因子组，显示成本要素，显示成本要素有效期间连续性，显示成本要素组，显示成本中心，显示成本中心有效期间连续性，显示成本中心组，显示作业类型，显示作业类型有效期间连续性，显示作业类型组，显示分配因子，显示分配因子组，维护成本中心类型。

软件系统的需求分析，使用用例图来描述是比较合适的。值得注意的是，大多数的用例能够在需求分析时确定，但随着系统的进展，可能又会发现更多的用例需要添加进去，也可能会发现前面定义的用例有错误需要重新修改。因此，在系统的发展中，应时刻注意用例的变化，以便随时进行修改。

1.2 用例图的设计原则

在设计过程中，主要遵循以下几个原则设计用例图。

- 1) 先进性和实用性
开发出来的系统是真切地解决实际问题、应用于现实生活当中的，所以，它必须具有实用性和先进性，而不能是一个功能和技术都陈旧的系统。
- 2) 标准化和可扩展
已经开发完成的系统在今后可能会增加一些功能，这就要求系统必须具有可扩展性，同时，在系统开发时所用到的各种接口必须保证是标准化的，这也为今后系统的扩展提供了方便。
- 3) 可管理与可维护
可管理与可维护指用户需要能够很方便地对开发出来的系统进行管理和维护，同时在管理和维护的成本上也能够得到有效的控制。
- 4) 安全性与保密性
安全性和保密性不仅指模块内的数据，同时还包括用户所执行的各种操作的安全性。模块必须对这数据和操作同时提供安全保证。

1.3 成本管理系统的模型描述

通过以上的分析，根据系统的实际需求，得出系统中相关的参与者和他们所发起的执行动作。成本中心会计总的活动图如图 1 所示。

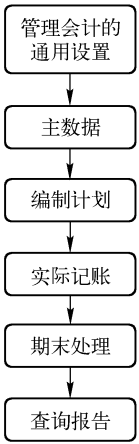


图 1 成本中心会计总的活动图

下面以成本中心业务管理员主数据模块为例来进行展开说明。
成本中心业务管理员在主数据模块的功能包括：成本要素管理，成本中心管理，作业类型管理，

分配因子管理。

成本要素管理：创建初级成本要素，创建次级成本要素，修改成本要素，删除成本要素，维护成本要素组，显示成本要素，显示成本要素有效期间连续性，显示成本要素组，显示组使用点，如图 2 所示。

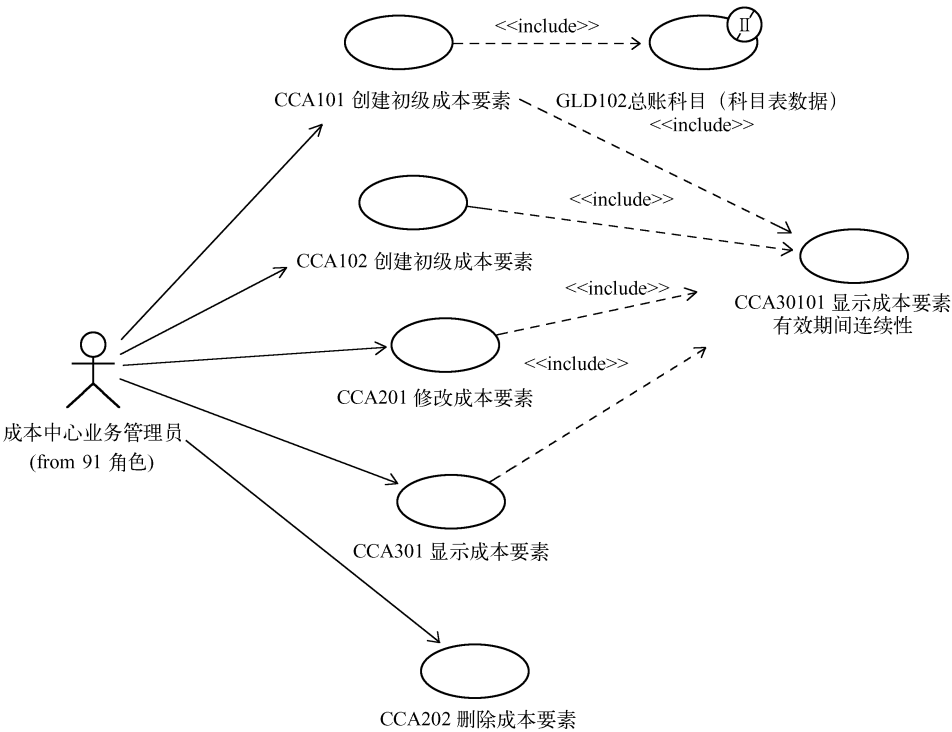


图 2 维护成本要素用例图

成本中心管理：创建成本中心，修改成本中心，删除成本中心，维护成本中心组，显示成本中心，显示成本中心有效期间连续性，显示成本中心组，显示组使用点，如图 3 所示。

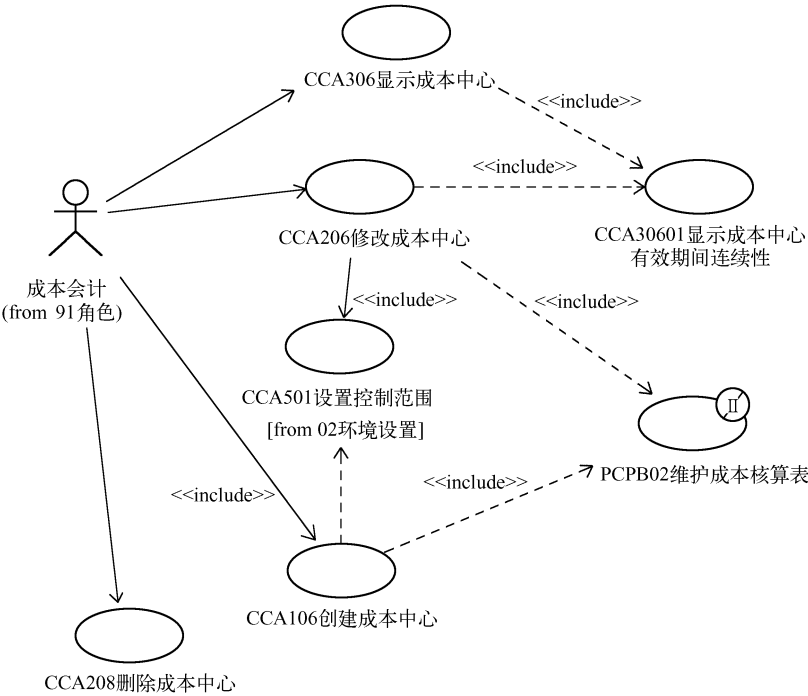


图 3 维护成本中心用例图

作业类型管理：创建作业类型，修改作业类型，删除作业类型，维护作业类型组，显示作业类型，显示作业类型有效期间连续性，显示作业类型组，显示组使用点，如图 4 所示。

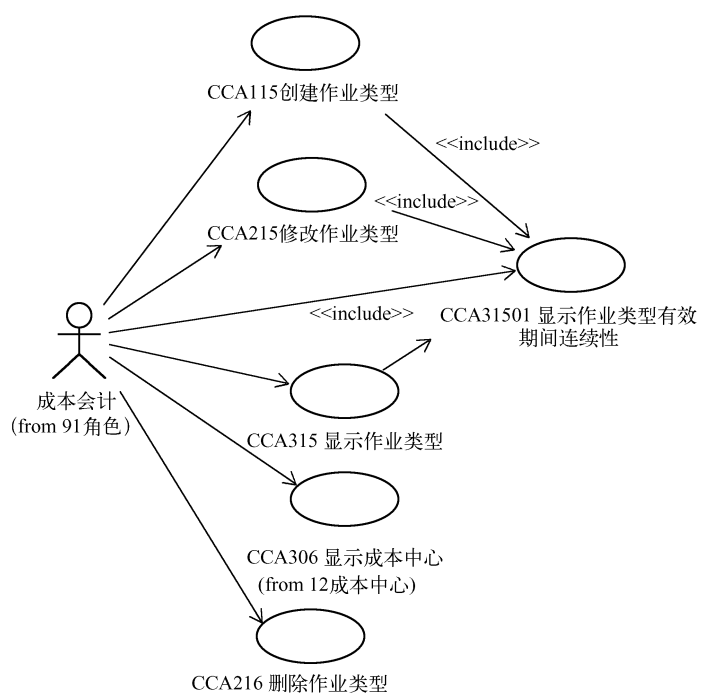


图 4 维护作业类型用例图

分配因子管理：创建分配因子，显示组使用点，修改分配因子，删除分配因子，维护分配因子组，显示分配因子，显示分配因子组，如图 5 所示。

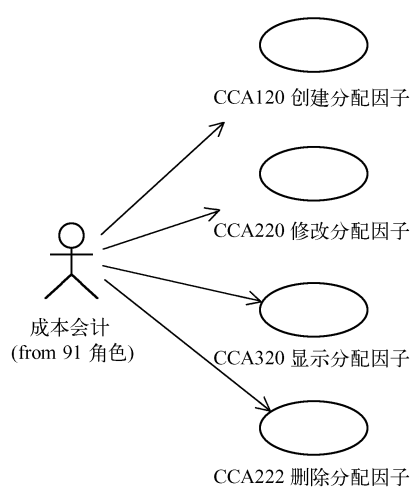


图 5 维护分配因子用例图

2 成本管理系统的的设计

基于上一点对需求的获取，本章将主要集中于分析阶段和设计阶段。在系统的分析和设计过程中，采用 UML 为系统建模的语言^[5]，以 RUP 的开发过程借助于 Rational Rose 建模工具^[6]，对系统进行了分析和设计。为了说明问题，仅以其中一部分的设计来阐述开发过程。

2.1 系统总体流程

成本管理是该 ERP 项目的一个子系统，它与其他子系统如账务系统、车间管理系统及存货（库存）系统有着密切的联系。其中，成本中心的费用信息可以从账务系统中的总账里获取，也可以手工录入；成本系统的完工产品和材料消耗信息可以从车间管理系统或存货系统中获取，也可以手工录入。其系统流程如图 6 所示。

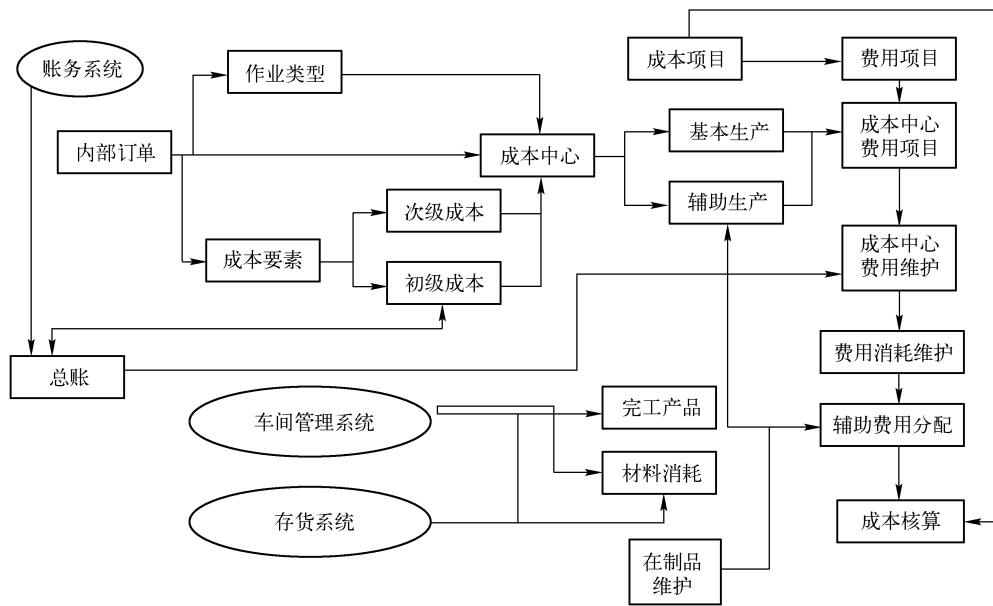


图 6 系统总体流程图

2.2 系统总体模块设计

成本中心会计的主要业务流程如图 7 所示。

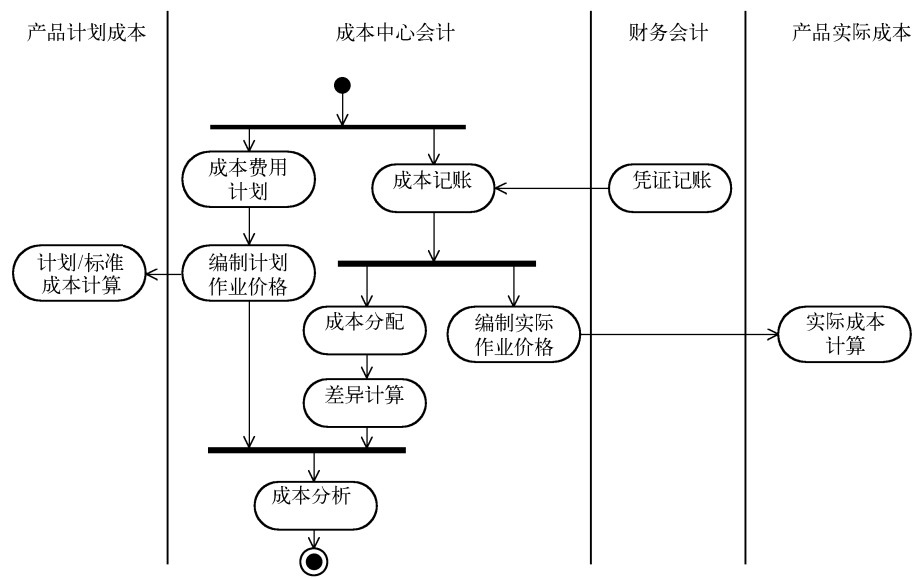


图 7 成本中心会计业务流程

基本业务功能为描述如下。

1) 计划

成本要素/输入作业计划、作业类型/作业价格计划、分配因子计划、计划复制、计划分配、计划分摊。

2) 实际记账

维护分配因子、维护实际作业价格、分配、分摊、作业分配。

3) 期末处理

进度管理器定制与监控、获取分配因子值、间接作业分配、作业差异计算。

4) 查询与报表

计划报告、计划与实际对比分析报告、累计成本报告、平均成本报告等。具体包括以下部分：管理会计通用性需求，主数据，运行环境设置，成本计划，实际过账，期末结账，查询报告。

2.3 主数据模块模型设计

设计阶段的主要任务有以下几个方面。

(1) 定义系统的软硬件环境，构架系统的具体配置，这可通过 UML 的部署图来表示。

(2) 识别子系统及其接口，这可从分析包的跟踪完成。

(3) 识别中间件和系统软件子系统。包括操作系统、数据库管理系统、通信软件、对象分布技术、图形用户界面设计工具、事务管理技术等^[21]。

(4) 定义子系统间的依赖关系。

(5) 利用分析类识别设计类，同时定义类中的操作，这可由分析模型中类图的协作图获取。

设计阶段是技术性非常强的一个过程，重点是类的提取，以及它们相互关系的描述，这包括许多面向对象技术、设计模式等技术的应用。根据面向对象的语言定义，一个对象类由类名、属性和操作组成。下面主要是针对成本控制主数据模块提取出的若干类及类的定义^[31]。在这些类中，主要用到类之间的三种关系，关联（Association）、聚合（Aggregation）、泛化（Generalization）。这三种关系在前面已经进行了介绍，这里就不再赘述。

2.3.1 成本要素

成本要素分为初级成本要素和次级成本要素两大类。初级成本要素除了记录所发生成本数据之外，还需要依据它与财务会计进行数据传输与匹配。初级成本要素的编码与选定科目表的会计科目相同，在创建之前首先要在财务会计中将要作为成本要素应用的会计科目定义完毕；在定义时只要输入初级成本要素编码，系统到财务会计对应科目表中将相对应的会计科目传到管理会计中，并带回相应的说明等信息，作为初级成本要素。如果财务会计对应科目表中没有相对应的会计科目，系统提示。

在管理会计内部次级成本要素用于记录、分配所发生成本数据，这些成本数据与财务会计没有直接的对应关系，因而不需要将次级成本要素与财务会计中的会计科目相对应，并且在创建时还要检查不能与财务会计中的会计科目编码相同。

按照成本要素的应用目的，还需要使用成本要素类别来对成本要素进行分类。这些分类主要用于控制的目的，因此是系统内置的，不需要用户来定义。

创建成本要素要求有时间有效性控制，在应用成本要素时还要按照时间有效性进行检查。

对于管理会计来说，成本要素是一项非常重要的数据，它的变化对系统处理会产生重大的影响，因此需要对创建、更改、显示、删除功能具有不同的权限控制，同时也需要对变更历史进行记录。图 8 所示为维护成本要素类图。

2.3.2 成本中心

成本中心是成本控制范围内的一个组织单位，它与控制范围一起共同构成完整的成本核算组织结构。在新创建成本中心时，要将成本中心匹配到成本中心标准层次的一个节点上。因此，在创建成本

中心之前，一定要首先创建完控制范围的标准层次。

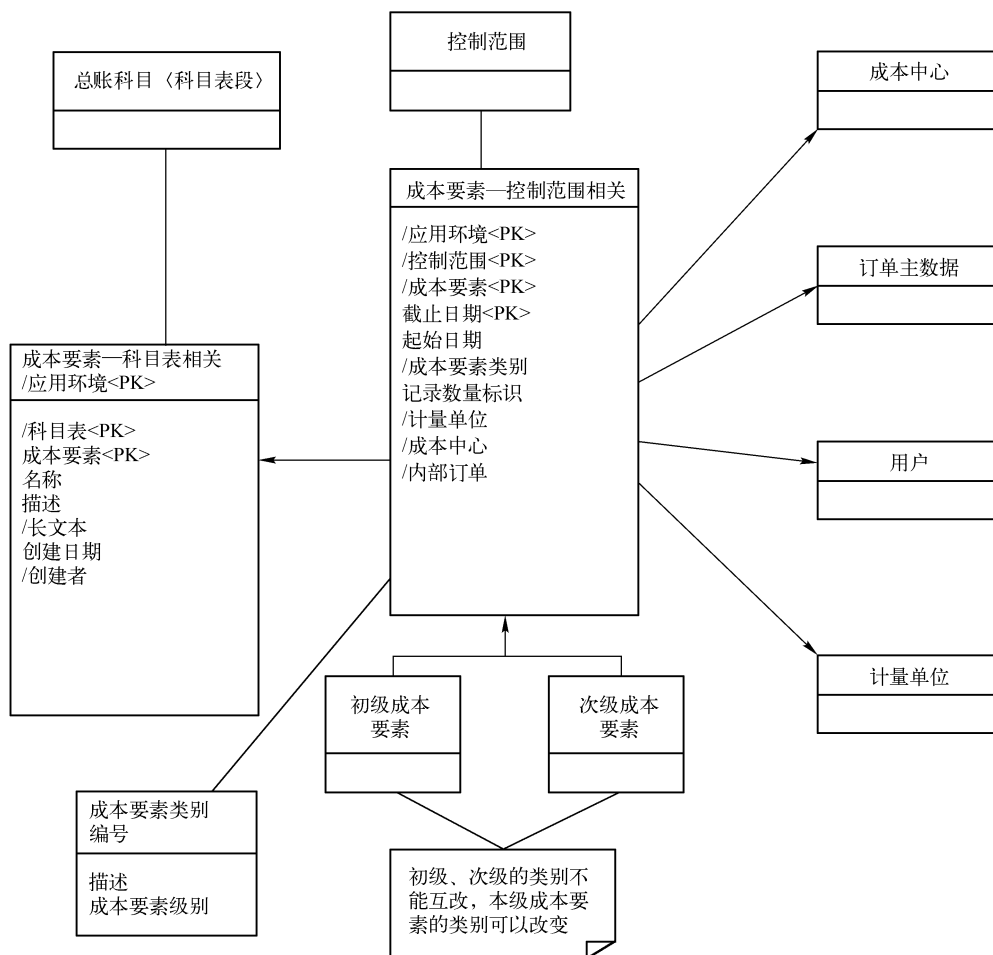


图 8 维护成本要素类图

如果需要，可以将成本中心与财务会计的业务范围、功能范围、与利润中心等组织相关联，作为默认值保存。

为了与作业类型相匹配、以及其他的应用目的，需要使用成本中心类别来对成本中心进行分类。这些分类主要应用的目的，需要用户在后面的功能中定义。

创建成本中心要求有时间有效性控制，系统在应用成本中心时要按照时间有效性进行检查。

对于管理会计来说，成本中心是一项非常重要的数据，它的变化对系统处理会产生重大的影响，因此需要对创建、更改、显示、删除功能具有不同的权限控制，同时也需要对变更历史进行记录。

成本中心有一个状态管理：未激活、激活，新创建的成本中心是未激活状态，在使用成本中心之前，必须将其激活。

改变状态、更改、删除成本中心要有内部控制。如果成本中心已经被应用，就不能再将其变为未激活状态；激活状态的成本中心不能删除；更改成本中心要保证不能引起数据混乱。图 9 所示为维护成本中心类图。

2.3.3 作业类型

作业类型是在特定控制范围之内定义的。

必须为作业类型输入确定的计量单位，以便对作业量进行方便、准确地计量和分配。

作业类型必须分配相应的成本中心类型，在作业计划和分配过程中允许对匹配的成本中心进行操作。

作。分配的时允许选择一个、多个或全部的成本中心类型。

作业类型必须分配一个成本要素，用于对发生的作业进行成本记账。要求成本要素类别为内部作业分配的成本要素。

按照作业类型的用途，需要使用作业类别来对作业进行分类。这些分类主要用于控制的目的，因此是系统内置的，不需要用户来定义。

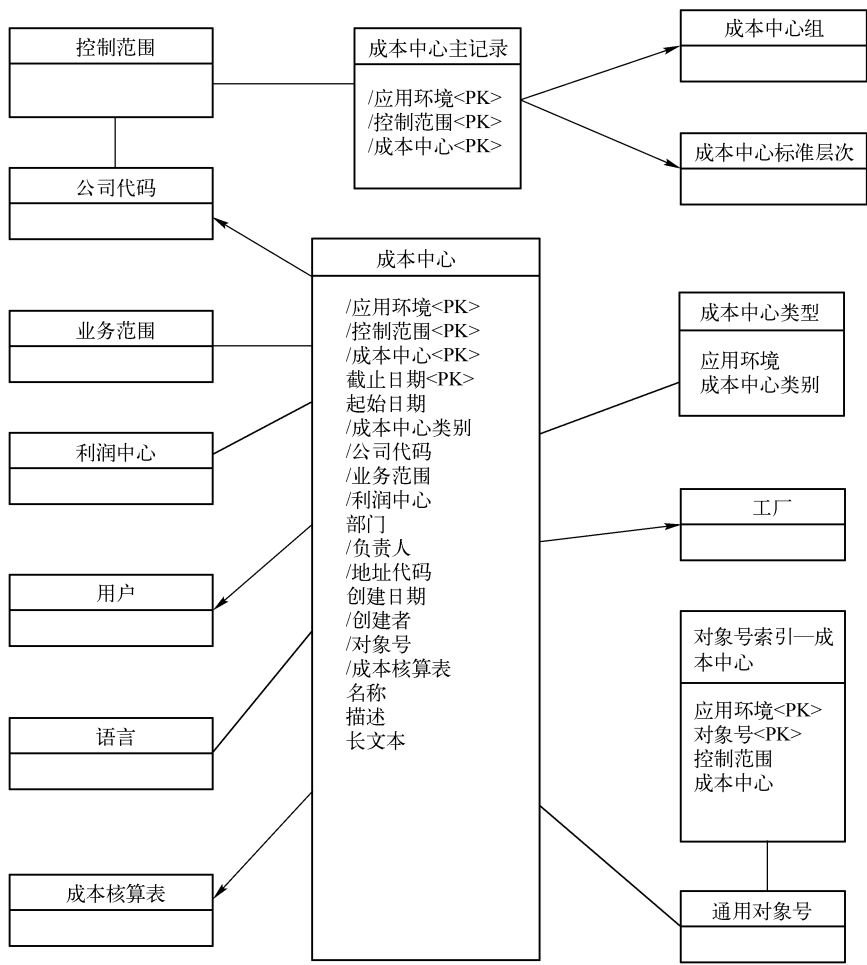


图 9 维护成本中心类图

创建作业类型要求有时间有效性控制，系统在应用作业类型时要按照时间有效性进行检查。

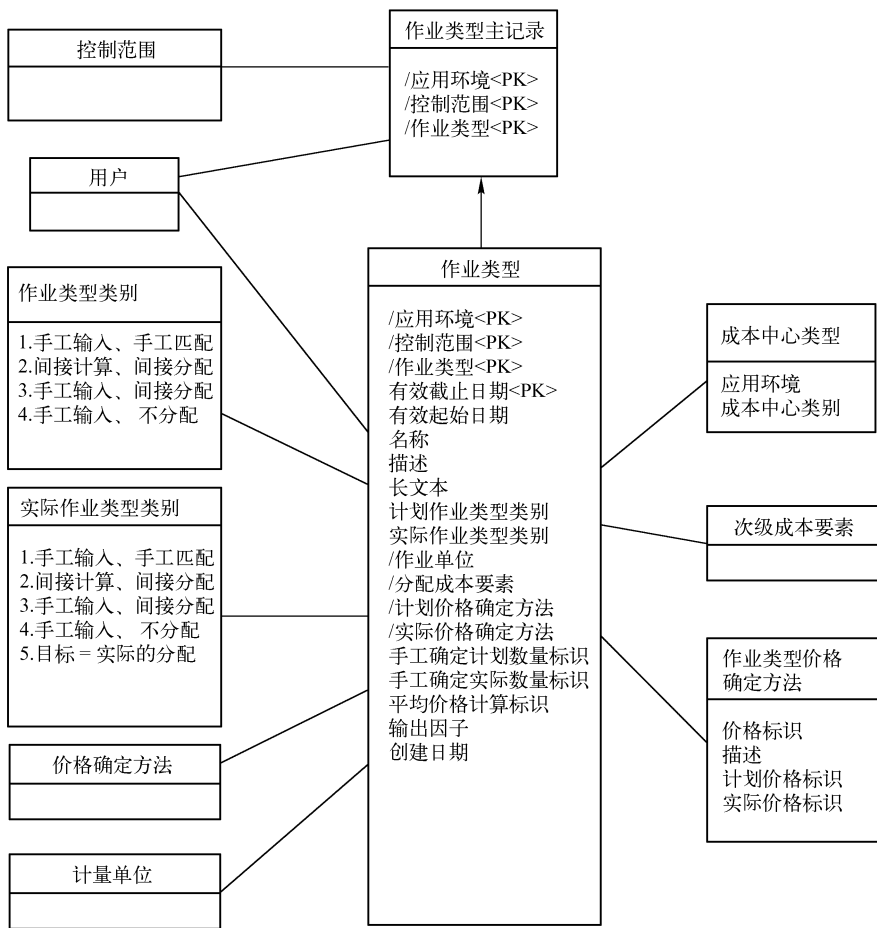
维护作业类型需要有创建、更改、显示、删除功能。对于管理会计来说，作业类型是一项非常重要的数据，它的变化对系统处理会产生重大的影响，因此对创建、更改、显示、删除功能需要有不同的权限控制，同时也需要对变更历史进行记录。图 10 所示维护作业类型类图。

3 成本管理系统的实现

本系统采用 C/S（客户-服务器）体系结构进行设计和开发，前端开发工具选择 PowerBuilder10.0，后台使用 MS SQL Server 关系型数据库。整个软件系统充分利用了面向对象编程技术和软件复用技术，大大缩短了产品的构建周期。

3.1 系统参数设置

定义成本系统的会计期间、系统参数、编码结构，主要内容描述如下。



启用期间：启用成本管理系统的时间。

是否与车间联动：确定成本核算的相关数据是从车间系统读取还是从库存系统读取。当与车间管理系统联用时，产品数量、原料消耗数量、工时、台时等信息可以直接从车间系统中取数。

编码结构定义：增加或者减少级数并且输入每级的编号长度，系统默认所有的结构编号为“222”，即分为三级，每级的编号长度为“2”则编号的总长度为“6”位。

成本核算对象：如果按“类别”核算对象，应该维护“成本对象字典”，即将一类物料按一个成本核算，选择“产品”和“产品批次”，则不必维护成本核算对象。进行成本核算前，可以根据用户需要，选择其中的一种，大大增加了成本核算的灵活性。

核算方式：按成本区间或者按照会计期间核算成本。

其主要界面如图 11、图 12 和图 13 所示。

3.2 成本中心维护

成本中心既可以是基本生产车间、辅助生产车间等成本归集的单位，还可以是基本生产车间、辅助生产车间之外的成本费用的消耗部门，主要内容描述如下。

中心编号：编号长度不允许超出在“编码结构定义”中定义的当前级数的编号长度，不可重复。

中心名称：成本中心的名称，不可重复。

中心类别：确定该成本中心是“基本生产”、“辅助生产”、“外部单位”或者“其他”。

辅助费用：如果成本中心类别为“辅助生产”时，需要指定该辅助生产车间所生产的产品或劳务，从费用项目字典选择。

部门编号、部门名称：指定成本中心对应的部门，在部门字典中选择，一个成本中心可以对应多个部门。

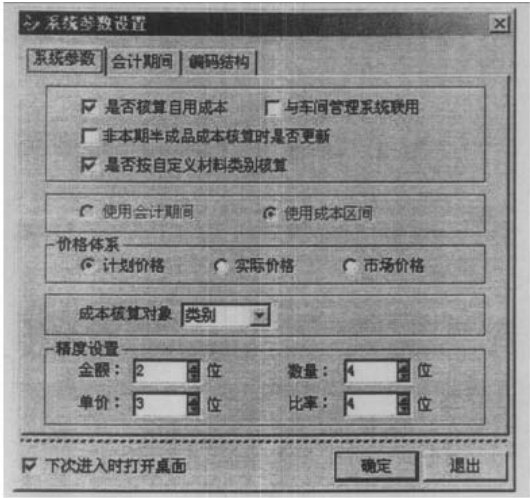


图 11 系统参数设置——系统参数

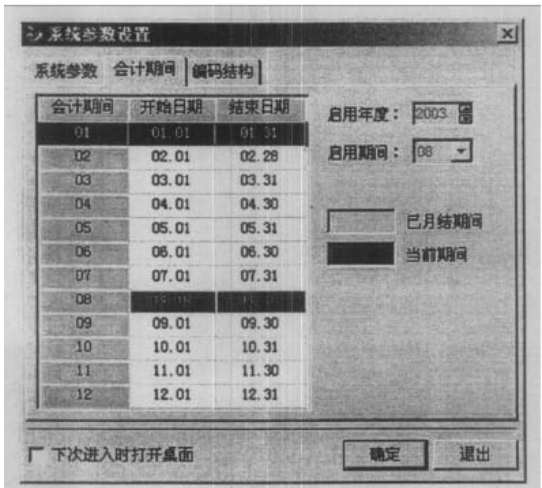


图 12 系统参数设置——会计期间

其主要界面如图 14 所示。

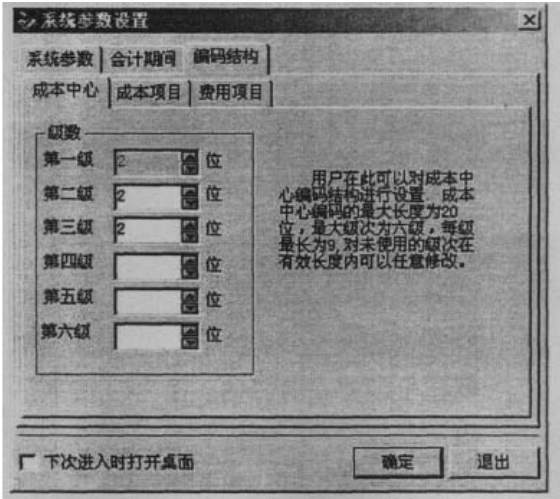


图 13 系统参数设置——编码结构

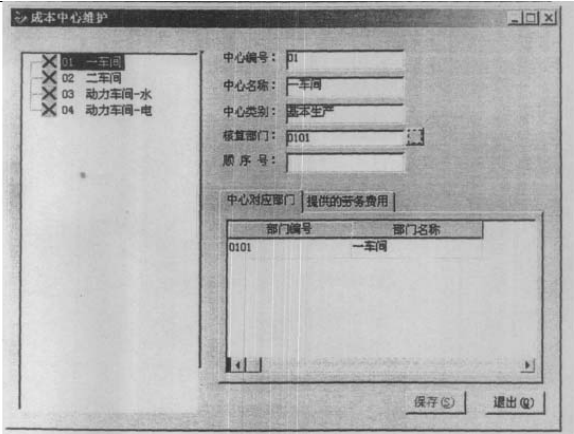


图 14 成本中心维护

3.3 成本对象维护

成本对象是成本核算的对象，也就是核算什么。在系统参数中果选择“成本核算对象为按类别核算”，必须为生产的各种半成品、产成品、联副产品、废品等指定一个成本核算对象。如果系统参数中指定按成本对象核算，没有在成本对象维护中录入，在成本计算时自动按产品核算。

主要界面如图 15 所示。

3.4 成本项目维护

反映构成成本核算对象的成本构成项目，是对成本的细化。构成成本的一切都应属于某一个成本项目，成本项目的设置应根据企业的实际需要，成本项目一旦使用不允许删除。主要内容描述如下。

项目编号：成本项目的编号，系统默认，没有存盘前可以修改，但是不允许重复输入。

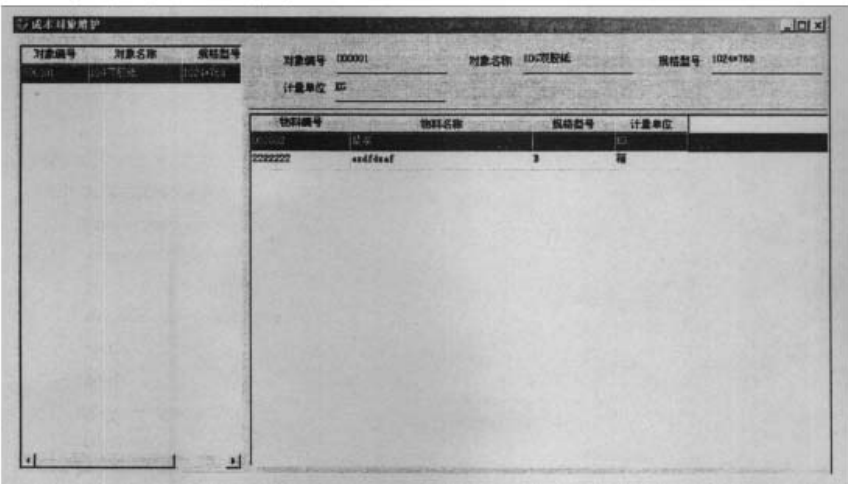


图 15 成本对象维护

项目名称：当前成本项目的名称，不允许重复输入。
成本类别：该成本项目所属类别，成本类别包括直接材料、直接人工、制造费用、燃料动力。
分配方式：成本项目费用的分配方式，包括定额费用、实际工时、实际机时、实际产量。
主要界面如图 16 所示。

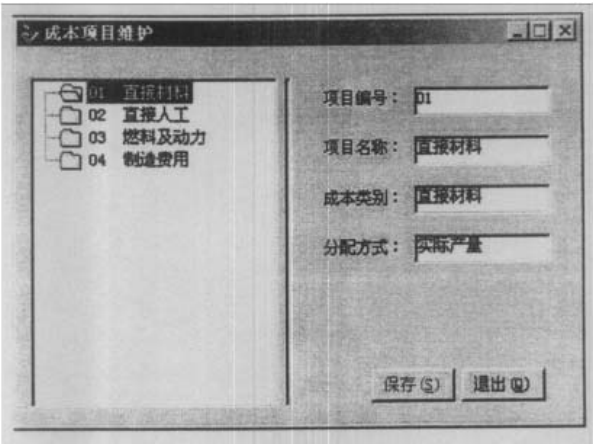


图 16 成本项目维护

4 总结

本文在分析 ERP 系统的国内外应用现状及其成本控制模块功能的基础上，系统地研究了如何使用统一建模过程对系统建模，并分析了几个关键问题，主要研究内容有：

(1) 通过对 ERP 的发展现状和发展趋势的分析和总结，系统地介绍了企业资源计划系统，并提出了企业资源计划系统与成本中心会计的开发思想。

(2) 使用 UML 描述系统。借助于 ROSE2003 的强大功能和 UML 对模型的完美的描述，对系统开发各个阶段进行了分析和建模。得出了系统的用例图、类图、顺序图等重要分析设计结果，为系统的编程实现和测试提供了巨大的帮助。在节约了系统分析员的时间的同时，也在系统设计和用户之间建立了一座信息交流的桥梁。

(3) 给出了成本管理模块的分析与设计。在设计过程中，本系统采用面向对象的设计方法，由用例驱动，抽取实现所需的类和对象，在类图中清晰地描述了各种类之间的关系。

参考文献

[1] 杨先功. ERP 与企业现代化管理. 湖北社会科学出版社, 2002, 6:34-35.

[2] Ivar Jacobson, Grady Booch, James Rumbaugh. The Unified Software Development Process [M]. Addison Wesley Longman, Inc.1999,56-78.

[3] Foster G, Swenson D. W.Measuring The Success of Activity-based Cost Management and Its Determinants. Journal of Management Accounting Research, 1997, 9 (3): 31-36.

[4] Kehoe. D., Boughton. N. Internet based supply chain management, a classification of approaches to manufacturing planning and control.International Journal of Operations&Production Management, 2001, 21(4): 516-524.

[5] Priestley, M. 面向对象设计的 UML 实践（影印版）（Practical Object-Oriented Design with UML）. 清华大学出版社, 2000, 38-72.

[6] 张海梅. 基于 UML 的模型转换及一致性验证: [硕士学位论文] 武汉: 海军工程大学计算机学院, 2005.

[7] 刘鹏, 刘亚彬, 杜梅先. 管理信息系统. 武汉: 武汉大学出版社, 2004, 4: 23-31.

[8] Alan Shalloway, Jim Trott. 设计模式精解: 面向对象设计的新视角. 透明译. 清华大学出版社, 2002, 44-53.

分布式缓存策略模式在高考网上报名系统设计中的应用

杨浩杰, 宋 涛, 刘 刚

(河南大学计算机与信息工程学院, 河南 开封, 475000)

摘 要: 河南普招网上报名系统是由河南省招生办公室开发的面向全省考生的在普招信息网上采集系统。在系统的开发中, 为了提高系统的性能和降低服务器的压力, 合理并且高效地使用缓存技术。对数据量小的数据采用微软企业库的缓存应用程序块进行本地缓存, 对数据量大的数据采用 Memcached 技术进行分布式缓存。在程序实现上, 采用了分布式缓存策略模式, 该模式借鉴软件工程中设计模式的思想, 把本地缓存和远程分布式缓存采用策略模式进行封装, 达到了在程序中灵活调用不同缓存模块的目的, 提高了系统的可扩展性, 收到了极好的效果。

关键词: 分布式缓存策略模式; 缓存; 企业库; 缓存应用程序块; Memcached 分布式缓存; 设计模式; 策略模式
中图分类号: TP311.5 **文献标识码:** A **文章编号:**

The Distributing Cache Mechanism by Strategy Patterns and Implementation of Higher Education Admission Information Collection System Based on Web

YANG Haojie, SONG Tao, LIU Gang

(HeNan University, KaiFeng, 475000,Henan China)

Abstract: The higher education admission information collection system based on web of HENAN province is designed by higher education admission office of HENAN province. In the development of system, for raising the capability of system and lowering the pressure of server, using cache technique with reason and efficiently. using cache application block of enterprise library to store the data of small capacity, to the data of great capacity, using Memcached technique. In the implementation of programing, adopted The Distributing Cache Mechanism by Strategy Patterns, make reference to the thought of design patterns in software engineer, Distributed to local cache and remote cache mode using strategies and procedures for implementation of the package, to achieve the goal of call different cache module flexible,raised the extendibility of the system, received an excellent result.

Keywords: the distributing cache mechanism by strategy patterns; cache; enterprise library; cache application block; memcached distributing cache; designer patterns; strategy pattern

河南普招网上报名系统是服务于全省考生, 提供普通高招网上报名服务并采集相关网报信息的系统。近年来, 随着网络技术和信息技术的不断发展和普及, 各省陆续推出了基于 Web 的信息采集系统, 基于 Web 的信息采集系统实现了将分散而繁重的招生工作集中在省级招生办公室或教育考试院统一处理, 从而大大减轻了各地市、县区招生单位的工作强度, 使其能够专注于面向考生的服务, 极大地提高了招生工作的效率和质量。河南是一个考试大省, 普通高招考生数量达到九十多万人, 网报期间平均每天的网络负荷能达到数万人。因此, 我们在系统开发时重点考虑了系统的稳定性、安全性和性能。系统的开发基于 .NET 平台, 采用 C#语言编写, 使用了 JQuery、SQL 分布式视图, 缓存, Membership 等新技术, 整套系统具有极高的稳定性、可靠性和安全性, 通过一年的试运行, 系统能够在考生并发访问量极大的情况下稳定运行, 达到了预期的效果。

作者简介: 杨浩杰, 男, 系统分析师, 河南大学计算机与信息工程学院在读研究生;
刘刚, 男, 硕士生导师, 教授, 博士。

在 Web 项目中，系统的稳定性和响应速度是一个很重要的指标，响应速度差的系统可能使用户失去信心。为了提高系统的性能，基于 Web 的开发要能够找出影响系统性能的瓶颈并在技术上进行合理的解决。在 Web 应用中，最大的瓶颈往往是数据库吞吐量，即用户频繁访问数据库读取数据造成的网络阻塞，为了解决这个问题，将用户经常访问而其内容又不经常变动的数据进行缓存，用户访问这些数据时从本地缓存或数据库缓存中进行读取，而不是从数据库中进行读取，这样就大大减轻了数据库服务器的压力和网络负荷，从而提高了系统的性能。在河南普招网上报名系统中，我们采用了分布式缓存策略模式，将分布式缓存机制及其算法采用策略模式进行了封装和实现。

1 缓存简介

缓存是在内存中保存创建代价高的信息副本的一种技术。例如，你可以缓存复杂查询的结果，这样后续的请求根本就不需要再访问数据，相反，它们直接从服务器内存抓取适当的对象（这是一个能快很多的方式）。缓存的真正之美在于它和其他多数提升性能的技术不同，它可以同时提升性能和可扩展性。性能更好是因为用于获取信息的时间被显著缩小。扩展性被提升是因为你打开了诸如数据库连接之类的瓶颈。这样，应用程序可以用更少的数据库操作服务更多的并发页面请求。^[1]

1.1 缓存适用场合

- （1）必须重复访问的静态数据或极少更改的数据。
- （2）在创建、访问或传输方面，数据访问的开销很高。
- （3）即使在源（如服务器）不可用时，数据也必须始终可用。

1.2 缓存设计原则

1) 关键问题

缓存设计需要考虑以下问题：（1）缓存的数据同步问题；（2）缓存的数据更新问题。

2) 数据同步

对于数据同步，必须考虑与多线程相关的若干技术：（1）lock 关键字；（2）ReaderWriterLock/ReaderWriterLockSlim；（3）InterLocked；（4）Mutex；（5）Monitor。

3) 数据更新

对于数据更新，要考虑以下问题：（1）自动更新（包括有效期的使用）；（2）手动更新（包括代码直接调用，时间通知）；（3）WeakReference（如果要考虑空间因素）。^[5]

2 微软企业库缓存机制

2.1 企业库简介

企业库（Enterprise Library）是微软的模式与实践（Patterns & Practices）的下一代应用程序块（Application Blocks）。其设计思想是为了协助开发商解决企业级应用开发过程中所面临的一系列共性的问题，如安全（Security）、日志（Logging）、数据访问（Data Access）、配置管理（Configuration Manage）等，并将这些广泛使用的应用程序块集成封装至一个叫企业库的程序包中。通过这些程序块，可以解决共性的企业级开发过程中所面临的问题。使用新的设计理念整合应用程序块，使得各应用程序块具有重用性、一致性、扩展性、易用性、集成性。

2.2 利用 Cache Application Block 在服务器集群缓存数据

缓存应用程序块（Cache Application Block）是 Enterprise Library 的一部分，支持多种方式的缓存

过期方式，对于多服务器同步缓存效果最好的还是使用文件依赖方式。即 Web 服务器和数据库服务器共同依赖一组共享的文件，由于数据的变化最终要反映到数据库中，所以各个 Web 服务器对依赖文件一般只是读操作，而由数据库服务器对依赖文件进行写操作。

由于 CAB 对缓存依赖文件会进行频繁的检查，所以如果依赖文件存储在数据库服务器上，而 Web 服务器远程访问这些依赖文件，将会很容易造成 Web 服务器极高的 CPU 负载，所以，尽管依赖文件设置在数据库服务器上最省事，还是需要把依赖文件设置在每台 Web 服务器上，并共享给数据库服务器访问。

2.3 缓存应用程序块的主要操作

1) 创建 CacheManager

```
CacheManager myCacheManager = CacheFactory.GetCacheManager();
```

2) 添加缓存项 (CacheKey:缓存 Key、Object:缓存的内容)

```
myCacheManager.Add("CacheKey",Object);
```

3) 缓存的读取

```
IDataReader toBeDisplay = (IDataReader)myCacheManager.GetData("MyDataReader");
```

4) 依赖文件的创建 (DependencyFile.txt)

```
myCacheManager.Add("FileKey", "String: Test Cache Item Dependency", CacheItemPriority.Normal, null, new FileDependency("DependencyFile.txt"));
```

5) 移除缓存条目

```
myCacheManager.Remove("FileKey");[6]
```

3 Memcached 分布式缓存机制

3.1 Memcached 简介

Memcached 是由 Danga Interactive 开发的，高性能的，分布式的内存对象缓存系统，用于在动态应用中减少数据库负载，提升访问速度。Memcached 通过在内存里维护一个统一的巨大的 Hash 表，用来存储各种格式的数据，包括图像、视频、文件及数据库检索的结果等。Memcached 的缓存是一种分布式的，可以让不同主机上的多个用户同时访问，因此解决了共享内存只能单机应用的局限，更不会出现使用数据库做类似事情时，磁盘开销和阻塞的发生。Memcached 的最初实现基于 UNIX 系统，目前在 Windows 下也有相关的实现。其实 Memcached 服务可以运行在任何操作系统下，而且在 .NET 下均可访问这些服务。

3.2 Memcached 的适用场合

Memcached 是“分布式”的内存对象缓存系统，也就是说，那些不需要“分布”的，不需要共享的，或者干脆规模小到只有一台服务器的应用，Memcached 不会带来任何好处，相反还会拖慢系统效率，因为网络连接同样需要资源，即使是 UNIX 本地连接也一样。在测试数据中，Memcached 本地读/写速度要比直接 PHP 内存数组慢几十倍，而 APC、共享内存方式都和直接数组差不多。可见，如果只是本地级缓存，使用 Memcached 是非常不划算的。

Memcached 在很多时候都是作为数据库前端 Cache 使用的。因为它比数据库少了很多 SQL 解析、磁盘操作等开销，而且它是使用内存来管理数据的，所以它可以提供比直接读取数据库更好的性能，在大型系统中，访问同样的数据是很频繁的，Memcached 可以大大降低数据库压力，使系统执行效率提升。另外，Memcached 也经常作为服务器之间数据共享的存储媒介，例如，在 SSO 系统中保存系统单点登录状态的数据就可以保存在 Memcached 中，被多个应用共享。^[7]

需要注意的是，Memcached 使用内存管理数据，所以它是易失的，当服务器重启，或者 Memcached 进程中止，数据便会丢失，所以 Memcached 不能用来持久保存数据。很多人认为 Memcached 的性能非常好，好到了内存和硬盘的对比程度，其实 Memcached 使用内存并不会得到成百上千的读/写速度提高，它的实际瓶颈在于网络连接，它和使用磁盘的数据库系统相比，好处在于它本身非常“轻”，因为没有过多的开销和直接的读/写方式，它可以轻松应付非常大的数据交换量，所以经常会出现两条千兆网络带宽都满负荷了，Memcached 进程本身并不占用多少 CPU 资源的情况。

3.3 Memcached 客户端访问代码（C#）

1) 创建 Memcached 客户端

```
MemcachedClient mc = new MemcachedClient();
```

2) 在 Cache 中存储一个字符串

```
mc.Store(StoreMode.Set, "HEAOSample", tempData);
```

3) 从缓存中取出一个缓存数据条目

```
DataSet tt = (DataSet) mc.Get("HEAOSample ");
```

4 分布式缓存策略模式

4.1 河南普招网上报名系统缓存机制的选择

在使用.NET 进行开发时，对数据进行缓存最简单的办法就是直接利用.NET Framework 提供的缓存功能，但是其实现比较简单，有一定的不足：（1）不能限制缓存的数据个数或者占用的内存大小，对于普通高招这样的应用，用户高达百万，如果不加限制的使用缓存，可能服务器的内存将会被耗尽；（2）不能对达到限制条目数的缓存进行清除；（3）SQL Dependency 技术对原始 SQL 语句要求过严，不利于实际过程中的使用。

基于以上的不足，河南普通高招网上报名系统的缓存机制采用 Microsoft Enterprise Library 中的 Cache Application Block 和分布式的内存对象缓存系统 Memcached 进行实现。我们把数据量小的数据用企业库的缓存应用程序块进行缓存，数据量大的数据用进行分布式 MemCached 技术进行缓存，在实际运用中，根据当前数据的缓存位置（客户端必须知道）调用相应的缓存实例进行读取及相关操作。

4.2 什么是分布式缓存策略模式

分布式缓存策略模式的提出基于以下两个目的：（1）系统开发中能够灵活地采用不同的缓存机制；（2）能够在不影响用户使用的前提下扩展已有的缓存机制。

针对以上要求，结合河南普招网上报名系统对缓存机制的选择，我们将系统缓存机制采用了设计模式中的策略模式进行架构设计，同时进行了程序实现，即将分布式缓存的算法实现和扩展同调用者彻底分开，以适应以后系统对缓存机制需求的变化。

5 设计模式简介

5.1 什么是设计模式

在生活和工作当中的各个方面，不断地重复一些事件和事件的方法，可以看做是一种设计模式。程序设计模式没有一个统一的定义，是开发者在开发中不断积累、总结的一种可复制的方案。一般认为，设计模式记录并描述了在设计面向对象软件中的设计经验，是一套被反复使用、多数人知晓的、

经过分类编目的、代码设计经验的总结^[2]。

5.2 学习使用设计模式的优点

1) 复用解决方案

通过复用已经建立的设计，可以受益于学习别人的经验，不必再为普通、复用的问题重新设计解决方案。

2) 在分析和设计上给予更高的视角

对于问题、设计过程和面向对象，模式给你一个更高层次的视角。这样的视角将你从“过早处理细节”的误区中解放出来。

3) 代码的可修改性得到改善

大多数设计模式还让软件更具可修改性。因为它们都是经受时间考验的解决方案。所以，它们发展成为的模式结构，可以比首先在脑海中浮现的解决方案更容易处理变化。

4) 设计模式阐述了基本的面向对象的原则

设计模式的使用可以大大增加学习者对基本面向对象设计原则的理解^[4]。

6 对象行为型设计模式的 Strategy（策略）模式

6.1 策略模式简介

策略模式是设计模式中用来处理变化的一种模式，属于对象行为型，其用意是针对一组算法，将每一个算法封装到具有共同接口的独立的类中，从而使得它们可以相互替换。策略模式使得算法可以在不影响到客户端的情况下发生变化。

使用策略模式可以把行为和环境分割开。环境类负责维持和查询行为，各种算法则在具体策略类（ConcreteStrategy）中提供。由于算法和环境独立开，算法的增减、修改都不会影响环境和客户端。^[2]

6.2 策略模式的应用场景

- （1）多个类只区别在表现行为不同，可以使用 Strategy 模式，在运行时动态选择具体要执行的行为。
- （2）在不同情况下使用不同的策略（算法），或者策略还可能在未来用其他方式来实现。
- （3）对客户隐藏具体策略（算法）的实现细节，彼此完全独立。

6.3 策略模式的优点

- （1）提供了一种替代继承的方法，而且既保持了继承的优点（代码重用）还比继承更灵活（算法独立，可以任意扩展）。继承本身可以处理多种算法或行为。如果不是用策略模式，那么使用算法或行为的环境类就可能会有一些子类，每一个子类提供一个不同的算法或行为，但是，这样一来算法或行为的使用者就和算法或行为本身混在一起。决定使用哪一种算法或采取哪一种行为的逻辑就和算法或行为的逻辑混合在一起，从而不可能再独立演化。继承使得动态改变算法或行为变得不可能。
- （2）避免程序中使用多重条件转移语句，使系统更灵活，并易于扩展。
- （3）遵守大部分 GRASP 原则和常用设计原则，高内聚、低耦合。

6.4 策略模式的缺点

- （1）因为每个具体策略类都会产生一个新类，所以会增加系统需要维护的类的数量。
- （2）客户端必须知道所有的策略类，并自行决定使用哪一个策略类，这就意味着客户端必须理解这些算法的区别，以便适时选择恰当的算法类。也就是说，策略模式只适用于客户端知道所有的算法

或行为的情况^[8]。

6.5 Strategy 类图

Strategy 类图如图 1 所示。

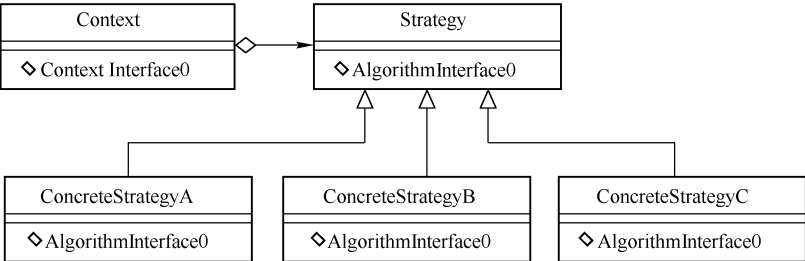


图 1 Strategy 类图

7 分布式缓存策略模式在高考网上报名系统中的运用

在河南普招网上报名系统的缓存机制中，如何根据当前要读取数据的缓存位置创建相应的缓存实例是一个关键问题。通过对设计模式的应用研究，结合系统对缓存的使用场合，我们选用了策略模式来实现网上报名系统的缓存机制，将基于企业库缓存应用程序块和分布式缓存 Memcached 的两套算法采用策略模式进行了程序实现。

7.1 缓存策略类图

7.1.1 缓存接口类图（派生类必须实现接口中的所有方法）

如图 2 所示。

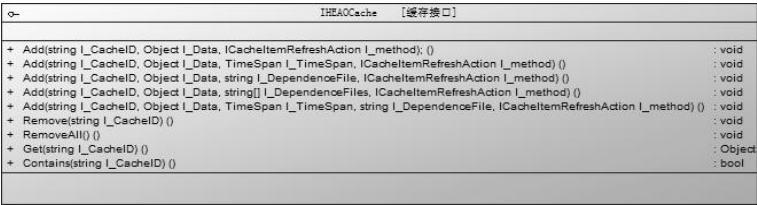


图 2 缓存接口类图

7.1.2 缓存策略类图和主要方法的实现

如图 3 所示。

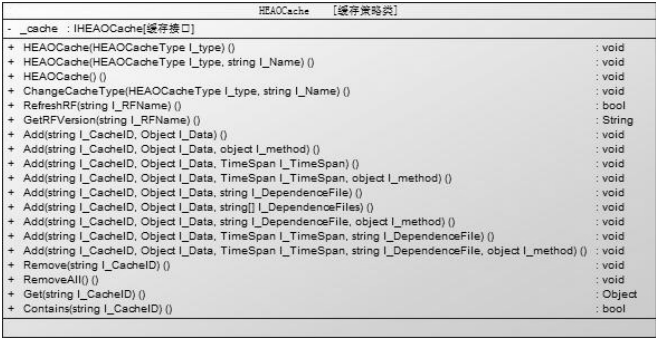


图 3 缓存策略类图

1) 添加数据到缓存并设置文件依赖

```
public void Add(string l_CacheID, Object l_Data, string l_DependenceFile, object l_method){
    _cache.Add(l_CacheID, l_Data, l_DependenceFile, (ICacheItemRefreshAction)l_method);}
```

2) 移除指定 ID 的缓存中的数据

```
public void Remove(string l_CacheID){_cache.Remove(l_CacheID);}
```

3) 从指定 ID 的缓存区中读取数据

```
public object Get(string l_CacheID){ return _cache.Get(l_CacheID);}
```

4) 查询缓存中有无指定的 ID

```
public bool Contains(string l_CacheID){return _cache.Contains(l_CacheID);}
```

7.1.3 本地缓存类图 and 主要方法的实现

如图 4 所示。

HEAOCacheLocal [本地缓存类]	
- primitivesCache : ICacheManager	
+ HEAOCacheLocal() ()	: void
+ HEAOCacheLocal(string l_Name) ()	: void
+ Dispose() ()	: void
+ Add(string l_CacheID, Object l_Data, ICacheItemRefreshAction l_method) ()	: void
+ Add(string l_CacheID, Object l_Data, TimeSpan l_TimeSpan, ICacheItemRefreshAction l_method) ()	: void
+ Add(string l_CacheID, Object l_Data, string l_DependenceFile, ICacheItemRefreshAction l_method) ()	: void
+ Add(string l_CacheID, Object l_Data, string[] l_DependenceFiles, ICacheItemRefreshAction l_method) ()	: void
+ Add(string l_CacheID, Object l_Data, TimeSpan l_TimeSpan, string l_DependenceFile, ICacheItemRefreshAction l_method) ()	: void
+ Remove(string l_CacheID) ()	: void
+ RemoveAll() ()	: void
+ Get(string l_CacheID) ()	: Object
+ Contains(string l_CacheID) ()	: bool

图 4 本地缓存类图

1) 添加数据到缓存并设置文件依赖

```
public void Add(string l_CacheID, Object l_Data, string l_DependenceFile, ICacheItemRefreshAction l_method){
    primitivesCache.Add(l_CacheID, l_Data, CacheItemPriority.Normal, l_method, new ICacheItemExpiration[] { new
FileDependency(l_DependenceFile) });}
```

2) 移除指定 ID 的缓存

```
public void Remove(string l_CacheID){primitivesCache.Remove(l_CacheID);}
```

3) 从指定 ID 的缓存区中读取数据

```
public object Get(string l_CacheID){ return primitivesCache.GetData(l_CacheID);}
```

4) 查询缓存中有无指定的 ID

```
public bool Contains(string l_CacheID){ return primitivesCache.Contains(l_CacheID);}
```

7.1.4 分布式缓存类图 and 主要方法的实现

如图 5 所示。

HEAOCacheRemote [分布式缓存类]	
- primitivesCache : MemcachedClient	
+ HEAOCacheRemote() ()	: void
+ HEAOCacheRemote(string l_Name) ()	: void
+ Dispose() ()	: void
+ Add(string l_CacheID, Object l_Data, ICacheItemRefreshAction l_method) ()	: void
+ Add(string l_CacheID, Object l_Data, TimeSpan l_TimeSpan, ICacheItemRefreshAction l_method) ()	: void
+ Add(string l_CacheID, Object l_Data, string l_DependenceFile, ICacheItemRefreshAction l_method) ()	: void
+ Add(string l_CacheID, Object l_Data, string[] l_DependenceFiles, ICacheItemRefreshAction l_method) ()	: void
+ Add(string l_CacheID, Object l_Data, TimeSpan l_TimeSpan, string l_DependenceFile, ICacheItemRefreshAction l_method) ()	: void
+ Remove(string l_CacheID) ()	: void
+ RemoveAll() ()	: void
+ Get(string l_CacheID) ()	: Object
+ Contains(string l_CacheID) ()	: bool

图 5 分布式缓存类图

1) 添加数据到缓存并设置文件依赖

```
public void Add(string l_CacheID, Object l_Data, string l_DependenceFile, ICacheItemRefreshAction l_method )
{throw new Exception("远程缓存不支持文件依赖! ");}
```

2) 移除指定 ID 的缓存

```
public void Remove(string l_CacheID){primitivesCache.Remove(l_CacheID);}
```

3) 从指定 ID 的缓存区中读取数据

```
public object Get(string l_CacheID){return primitivesCache.Get(l_CacheID);}
```

4) 查询缓存中有无指定的 ID

```
public bool Contains(string l_CacheID){return Get(l_CacheID) != null;}
```

7.1.5 缓存类型枚举

如图 6 所示。

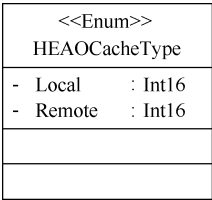


图 6 缓存类型枚举

7.2 缓存策略类间关系和失效机制

从以上的类图可以看出，待实现的缓存操作抽象算法描述都封装在“IHEAOCache”这个接口中，而具体算法则在“HEAOCache”类中实现。同时，系统定义了“HEAOCacheLocal”和“HEAOCacheRemote”两个派生类，均继承于“IHEAOCache”接口，也就是说，这两个类都要实现在“IHEAOCache”接口中定义的方法。在具体调用中，把要生成的缓存类型作为“HEAOCache”类的构造方法的参数，以生成一个具体的缓存实例，从而调用“HEAOCache”类中的具体缓存操作方法。

在缓存的失效机制方面，我们对本地缓存采用了文件依赖的方式，当需要更新缓存时，更新相应的依赖文件即可。对于远程分布式缓存，因为 Memcached 不支持文件依赖的方式，因此对远程缓存的清除采用了按缓存 ID 进行清除的办法。

7.3 缓存访问设计要求

访问缓存时，要满足以下要求：（1）数据能被多线程访问；（2）对数据的访问分为读和写；（3）当任一线程读数据时，其他线程不能写数据；（4）当任一线程写数据时，其他线程不能写数据，其他线程不能读数据；（5）由于读数据的频率远远高于写数据的频率，所以读数据线程的优先级更高；（6）不允许死锁的情况发生。

7.4 创建缓存实例的方法

根据系统对缓存使用的要求，在生成一个具体的缓存实例时（对缓存数据进行读/写），我们采用了 Lock 机制，并将生成缓存实例的操作封装到“HEAOCacheHelper”类中。

1) 类图

如图 7 所示。

2) Lock 机制的代码实现（以本地缓存为例）

```
public static HEAOCache GetCacheInstance(string l_CacheName){
```



```
if (!_localCacheList.Contains(l_Name)){  
    lock (oLock){  
        if(!_localCacheList.Contains(l_Name))  
            _localCacheList.Add(l_Name, new HEAOCache(l_type, l_Name));}}  
return (HEAOCache)_localCacheList[l_Name]; }
```

HEAOCacheHelper	
- oLock	: Object
- _localCacheList	: HashTable
- _RemoteCacheList	: HashTable
+ GetInstance(HEAOCacheType l_type, string l_Name) ()	: HEAOCache[缓存策略类]
+ GetInstance(HEAOCacheType l_type) ()	: HEAOCache[缓存策略类]

图 7 类图

3) 生成一个本地缓存实例

HEAOCache cache = HEAOCacheHelper. GetCacheInstance (HEAOCacheType.Local, "LocalConfig");

4) 生成一个远程缓存实例

HEAOCache cache = HEAOCacheHelper. GetCacheInstance (HEAOCacheType.Remote, "RemoteConfig")

8 结论

在河南普招网上报名系统的开发中，我们采用了分布式缓存策略模式，根据数据的特点，对数据分别采用企业库的缓存应用程序块技术和 Memcached 分布式缓存技术进行本地缓存和分布式缓存，达到了提高 Web 系统性能的目的。同时，我们结合设计模式的思想，将这两种缓存算法采用策略模式进行程序实现，满足了系统设计中灵活创建两种缓存实例的要求，使算法的实现和变化与客户端彻底分离，大大提高了系统的可扩展性和可维护性，实际效果令人满意。在以后的开发中，我们要使用更多的数据库新技术，继续深入研究设计模式，将其原理与实际应用更加紧密地结合在一起，以提高我们程序的质量，从而更好地为我省的招生服务工作。

参考文献

[1] [美]Matthew MacDonald, Mario Szpuszta 著, [译]博思工作室. ASP.NET 3.5 高级程序设计 (第 2 版). 人民邮电出版社, 2008.

[2] 明日科技.张跃廷, 王小科, 张宏宇. ASP.NET 技术方案宝典. 北京: 人民邮电出版社, 2008.

[3] [美] Robert C.Martin , Micah Marti n 著, [译] 邓辉, 孙鸣. 敏捷软件开发-原则、模式与实践 (C#版), 北京: 人民邮电出版社, 2008.

[4] [美] Alan Shalloway & James R.Trott 著, [译] 熊节. 设计模式精解. 北京: 清华大学出版社, 2004.

[5] <http://www.cnblogs.com/czy/archive/2010/05/05/1727616.html>

[6] <http://www.cnblogs.com/Terrylee/archive/2005/11/11/273731.html>

[7] <http://kb.cnblogs.com/page/42776/>

[8] <http://www.cnblogs.com/justinw/archive/2007/02/06/641414.html#1812547>

无线测控通信平台中间件的设计与实现

潘 磊, 张书钦, 郑秋生

(中原工学院计算机学院, 郑州, 450007)

摘 要: 随着无线网络技术、嵌入式计算技术的迅速发展, 利用无线多跳网络实现工业测控具有广阔的应用前景。本文结合无线多跳网络的特性及实时应用的特点, 提出了面向工业测控应用的中间件平台系统, 实现了多应用的业务模型、分布式的功能模型、基于策略代理的管理模型。通过本文的无线测控通信平台可以快速、方便地构建各类测控应用, 可大大降低测控应用系统设计和维护的成本, 提高测控系统开发和运行效率, 保证测控任务的高效实施。

关键词: 无线多跳网络; 测控系统; 中间件; 嵌入式计算

中图分类号: TP393

文献标识码: A

文章编号: 1006-7043 (2004) xx-xxxx-x

A Middleware for Wireless Multi-Hop Network Based Measurement and Control System

PAN Lei, ZHANG Shuqin, ZHENG Qiusheng

(School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, Henan China)

Abstract: With the continuing advances in wireless networks and embedded computing technology, wireless multi-hop network have many applications in industrial measurement and control systems. In the paper, with characteristics of wireless network and real-time application, a Middle ware Platform System for Wi reless Multi-Hop Network based Me asurement and Control is pro - posed, which implements a multi-applicatio n model, a distributed function model an d a policy agent based management model . Through adopting the proposed middleware platform, various industrial measuremen t and control application systems can b e built rapidly and conveniently, and the cost of design and maintain can be reduced . Therefore, the middleware platform can prompt the efficiency of measurement and control systems.

Key words: Wireless Multi-Hop; measurement-control system; middleware; embedded computing

近年来出现的无线多跳网络技术可以完成高度分散的测控节点之间的互连, 极大地方便了测控设备的部署, 以有效实现监视、测量和控制等任务目标。无线多跳网络一般由大量低功耗、多功能的微型节点组成, 具有自组织、多跳路由、动态拓扑等特点。通过对无线多跳网络技术低层次的封装, 并为测控业务提供高层抽象的中间件技术是一种崭新的设计方法, 可以充分解决无线测控通信网络设计和实现时所面临的一些关键问题, 方便复杂测控应用的开发与实现。

本文设计和实现了一种开放式的无线测控通信平台中间件, 可完成数据转发、资源管理、拓扑控制、任务调度、网络互连等功能。该中间件抽象了网络协议、节点操作系统和硬件等细节, 为测控业务开发提供了统一的支撑平台, 将原本独立发展的无线网络和测控应用融合为一体, 具有开放式、可扩展、轻量级的特征, 可为网络化的测控业务提供了监控、监测等支撑, 可以承担现场化、远程化、智能化、网络化、一体化的多类型测控应用。

1 无线测控系统模型

基于无线多跳网络技术的测控系统由现场网络、测控网关和管理客户端三部分组成, 测控通信平

基金项目: 河南省科技攻关项目 (092102210331)

作者简介: 潘 磊 (1975—); 男, 讲师, 硕士;
张书钦 (1978—); 男, 副教授, 博士;
郑秋生 (1965—); 男, 教授, 硕士。

台中间件实现了无线通信网络的数据、拓扑、资源的统一管理，屏蔽了无线网络的复杂性，加快了各类测控业务的开发，使整个测控系统互连更方便、部署更快速、运行更可靠。

1.1 无线测控系统结构

基于无线多跳网络的测控系统无须布线，具有快速展开、抗毁性强等特点而得到越来越广泛的应用，可以弥补有线测控系统存在的缺陷。本平台系统主要由现场无线网络、无线测控基站和管理终端三部分组成，可实现双向实时通信、测量和控制。如图 1 所示为无线测控系统网络体系结构示意图。

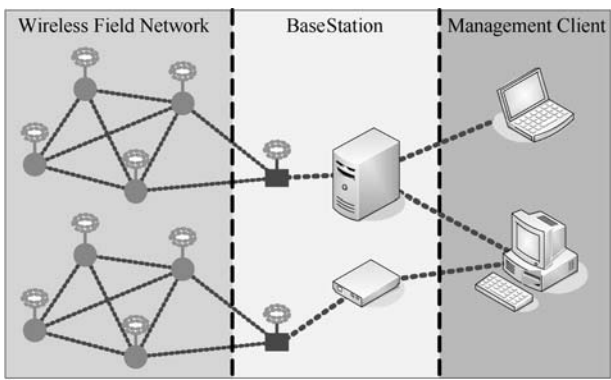


图 1 无线测控系统网络体系结构示意图

现场网络中的无线节点通过 I/O 口与测控设备相连，采集测控设备的电压、电流、温度等数据信息，并可向测控设备发送控制指令。无线节点具有有限的计算和通信能力，节点之间可通过多跳的方式形成网络进行通信，并可以通过基站与外部网络中的管理终端通信。

测控基站可以实现现场网络与外部网络的双向数据交换，并完成数据存储、网络管理等测控业务功能。测控基站充当了现场无线网络与外部以太网或 GPRS 网络之间的网关，同时，通过基站还可以实现不同测控区域的互连，外部网络中的管理终端相连，并可以通过本地服务器实时处理测控业务的处理。

管理终端担负着测控系统的网络管理、数据处理、操作管理和数据存储等功能。管理终端可以对收到的信息分析后进行决策，操作人员在测控管理中心即可实现远距离的测控工作。

1.2 无线测控通信平台中间件参考模型

由于基于无线多跳网络的测控系统的技术复杂度，使得其中的路由、拓扑、性能等方面的开发较为困难。为了便于测控应用的开发，利用中间件的方式来封装系统底层的网络结构、操作系统、通信协议、数据库和其他应用服务，完成对复杂多变的网络环境下数据分散处理的性能和效率、安全等共性问题的处理，为应用系统提供一个统一的、可操作的软硬件通信支撑。该平台中间件在异构的无线节点上构建一种共用的网络结构、通信接口和系统运行方式，以及一套测控业务接口，可以供不同的测控应用重复使用。该体系结构如图 2 所示。

测控通信平台的基本功能，一方面是各种测控设备的互连，保证无线多跳网络或其他网络上数据传输、数据处理的正确性，实现对测控设备的本地和远程操作控制；另一方面是与上层测控应用软件交互，通知上层软件终端设备的一些工作情况或传输业务数据，同时为上层软件提供终端设备的操作接口。

配置管理主要实现系统运行的配置信息管理。配置信息管理获取测控节点、基站等设备及其之间的逻辑关系等配置信息，并实现查询、修改、删除等。基站提供远程配置操作接口。配置信息变化有手动更新和自动更新两种管理方式。

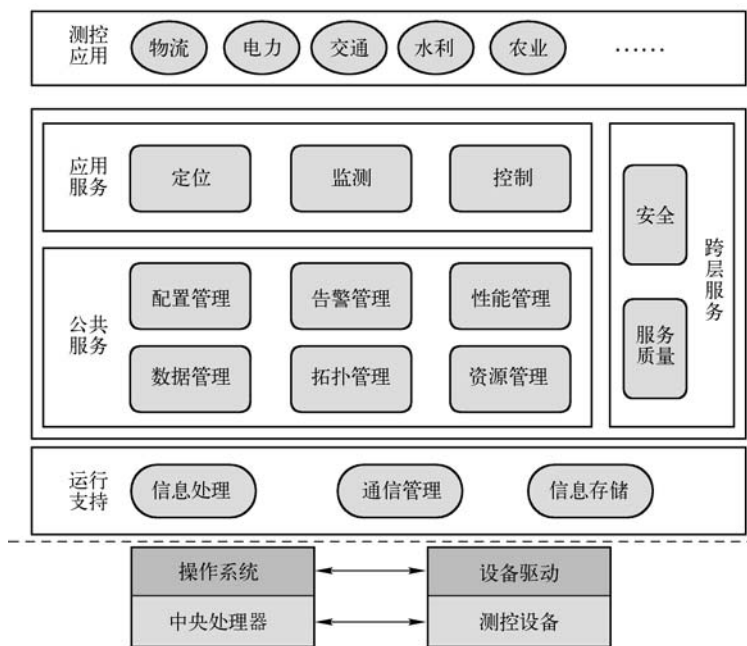


图2 中间件参考模型体系结构

数据管理为系统运行提供数据的存取、查询等数据服务。在测控节点仅实现少量、本地的节点运行、路由、算法参数的存储，在基站和管理中心则利用关系数据库提供更为复杂的数据管理功能，尤其是在管理中心更需要提供更为复杂的系统业务数据管理，能够有效、快速地存取、关联、查询数据服务。

拓扑管理需要完成网络链路管理。无线测控系统中节点数目众多，类型复杂，需要通过自动发现引擎实现拓扑信息处理，并在管理中心以 GIS 地图等形式呈现各种资源的分布、状态和关联关系，并为配置、性能、告警等管理功能提供操作界面，为无线测控业务管理提供有效手段。

性能管理是测试网络各个单元性能的过程，它包括测试网络连接和当前网络段利用率、识别可能发生拥塞域、杜绝高出错率和检测网络传输状态等。性能管理主要包括性能数据采集、性能分析。性能数据采集主要通过定期报告、实时查询和事件驱动等方式向基站和管理中心汇总流量、负载、丢包、延迟、测控数据等信息。性能分析通过对历史数据和实时数据的分析和计算，以图形等形式反映系统的性能状态。

2 中间件的实现

该平台中间件以无线网络作为主要测控数据承载，建立完整的、有效的测控业务框架，其实现主要采用了多应用域的测控业务模型、分布式的测控功能模型、基于策略代理的网络管理模型。

2.1 多应用域的测控业务模型

测控系统中可以由不同的应用子系统组成，各应用系统由于功能侧重、设计方法和开发技术有所不同，也就形成了各自独立的数据处理和决策控制。利用本通信平台可以为各应用子系统提供业务支撑，方便了不同应用之间的数据交换，且易于构建复杂的多应用测控系统，并实现测控系统的统一管理。

将一种应用子系统的业务范围定义为应用域，根据所涉及应用的不同可以将现场测控网络中的无线节点及设备划分为不同的应用域。应用域是系统的逻辑功能区域划分，应用域可以重叠，即一个物

理节点可以同时服务于多个应用域，如图 3 所示。

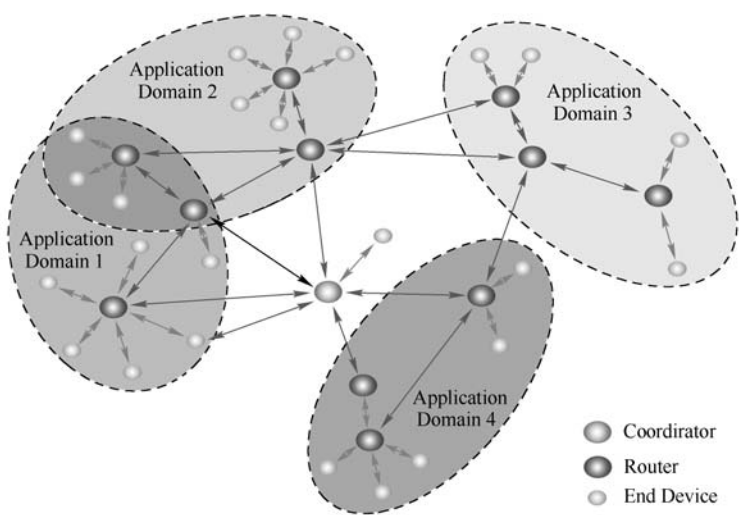


图 3 系统中的应用域划分示意图

例如，在无线智能停车系统中可以定义如下的应用域：

- 照明子系统。包括开关、灯等设备。
- 动力子系统。包括开关、风扇、空调等设备。
- 停车子系统。包括车位指示灯、车位锁、行车方向提示灯等。

每个应用域具有特定的标识，且可以用数据结构进行描述。应用域中包含有相关的测控设备，如转速传感器、电源开关等。测控节点间的消息交互仅在同一应用域内完成，并且这些消息具有相对固定的收发和处理模式。

实际上，一个应用域可以通过设备、消息来进行描述。通信平台系统定义了应用域、测控设备、应用消息等描述符数据结构，也定义了通用的应用消息处理函数库，一般的测控应用通过继承或重用这些函数库即可快速实现业务功能。

2.2 分布式的测控功能模型

测控系统需要实现三方面的任务：实时数据采集，即对被控量的瞬时值进行检测和输入；实时决策，对实时的给定值和被控量的数值按已定的控制规律进行运算，决定下一步的控制过程；实时控制，根据决策适时地对执行机构发出控制信号。

对于基于无线网络的测控系统，一般的运行设备都采用本地控制和远程控制相结合的方式实现管理，如系统中的照明灯，用户可以在通过本地的开关设备进行直接控制，也可以在管理中心进行远程控制。针对设备的测控实现，本平台提供了分布式的测控功能模型，其中包括远程管理、网络控制、本地设备三层。网络控制层根据测控设备状态等信息对设备进行控制，同时又将运行过程中的具体参数传送给远程管理层，使管理层能够对设备运行进行实时监视。

考虑到测控功能的实时性和可靠性，以及控制节点的处理能力、无线链路带宽、随机时延和丢包对控制功能的影响，网络控制层功能一般分布在邻近测控设备的无线节点。远程管理层可以通过网络控制层，或直接向测控设备下达具体的控制指令。对于实时性要求较高的设备，主要由智能控制层来完成，管理层只是起到了下达任务、对运行过程进行监视的作用。

在网络控制层和远程管理层的测控功能可以有不同的实现形式。如图 4 所示为简单的照明控制功能实现示意图，其中，位于管理层的用户可以通过网络控制层的开关控制开闭，也可以绕开网络控制层直接控制灯光的开闭。前者的优点是控制方式统一，控制操作一致，缺点是本地的网络控制节点一

一般都是终端节点类型，不适合控制“代理”的角色。后者更适合无线网络结构，远程控制直接作用于运行设备，性能较好，但易导致控制方式和控制操作的不一致。当然，可以采用可配置的方式来根据实际情况来实现这两种方式，可用技术包括移动代理和策略方式。

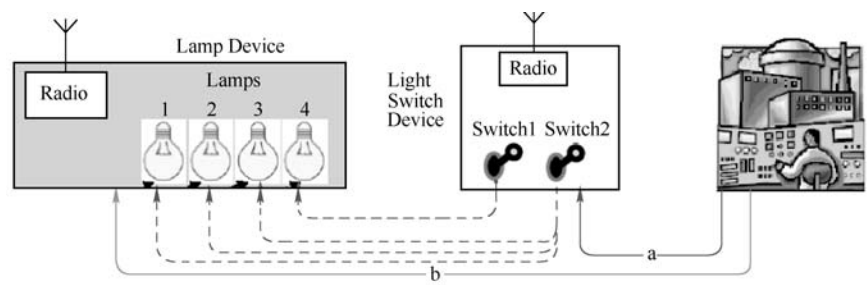


图 4 简单的照明控制功能实现示意图

系统中基本的通信和测控功能是通过无线节点的多任务单片机实现。节点系统通过主控制器程序来管理具有任务标识和状态标识的多任务，完成任务初始化、任务间的消息交换、任务同步等支撑功能，并利用中断等机制实现各类业务功能。为了实现具体的测控功能，需要将功能对象分解成多个彼此独立，相互关联独立的任务，并根据任务的驱动类型和实时性指标来确定任务的执行方式。

网络控制功能的具体实现依赖于具体工业测控系统的需求。一般来说需要为具体的测控系统建模，并考虑传输时延、丢包、信号误码率等因素对系统性能的影响，尤其是无线无线网络环境下测控系统的鲁棒性问题，即在系统存在不确定的情况下，设计具有鲁棒性的测控策略，优化测控系统参数和通信协议参数，实现测控和通信的平衡，保证系统的性能最优。由于无线网络中的节点具有可移动性，网络拓扑结构呈现动态，丢包的概率变大，导致通信会出现间断。

2.3 基于策略代理的管理模型

无线测控网络管理与传统网络的网络管理有相似的管理任务，但无线节点仅具有较低能量和处理能力，而且节点由于节点数量众多，节点间自组织，使得无线测控网络的管理和维护都面临着一些与传统网络管理不同特点。

无线测控网络管理旨在提供一个一体化的管理机制，有效地监视和控制远程的环境或被管实体，以较少的耗能对网络的资源配置、性能、故障、安全和通信进行统一的管理和维护。为了实现这个目的，对无线测控网络进行管理时要具有高效率的通信机制，轻量型的结构，智能而灵活的资源分配机制。本系统中采用了层次式的网络管理结构，测控系统网络管理功能体系结构如图 5 所示。

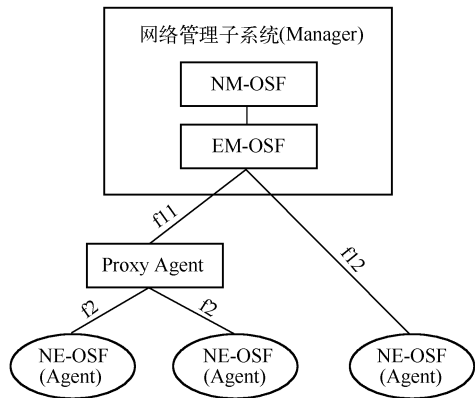


图 5 测控系统网络管理功能体系结构

网元管理级的运行系统功能（EM-OSF）通过 f11、f12 参考点向下直接或间接连接各被管设备上的网元级运行系统功能（NE-OSF），并进行网管信息的交互，向上与网络管理级运行系统功能（NM-OSF）相连。NE-OSF 由管理代理实现。

EM-OSF 和 NM-OSF 可以分别实现独立的管理系统，充当管理者（Manager）。在测控系统中，管理中心实现了管理者，基站和路由节点上驻留了中间代理（Proxy Agent），无线网络节点上驻留了终端代理（Agent）。管理者可以通过 f12 参考点连接终端代理，也可以通过中间代理间接连接终端代理，中间代理完成 f11 和 f12 参考点之间的转换。f11 参考点与 f12 参考点的区别表现在：f12 参考点对应的网管接口为 Manager—Agent 接口，而 f11 参考点支持中间代理（Proxy Agent）实现对多 Agent 的转发工作机制。

管理中心可以管理全局的管理策略。每个中间代理就是一个管理的关键控制点，网络中的其他节点分簇，并且簇的层次控制在两层内。一个终端代理管理一个节点，实施本地的配置、监控、过滤、性能、故障等管理任务。一个中间代理管理多个终端代理。上层的代理运用本级的策略规则来发现拓扑变化，监测网络故障，调整网络性能。

中间代理和终端代理中包含有策略转换、策略调度和策略信息库几个部分，实现策略的接收、决策和调度功能。在代理的结构中，有一个策略转换、策略调度、策略信息库等几个部分，还包括配置代理、监控代理、过滤代理、故障代理、性能代理。监控代理负责不断收集周围的环境数据。过滤代理可以对收集到的数据要进行一些融合处理，对冗余的信息进行过滤。来自上层的管理策略可能是不正确，在上层时没有及时发现，过滤代理可以进行筛选。故障代理是在节点或链路发生故障时，进行试探性的修复。性能代理主要就是对用户指定的性能指标进行监控。

3 结束语

无线测控通信平台具有高度模块化、集成化的软硬件架构，具有部署方便、覆盖区域广、测控对象丰富、安全智能、响应快速等优点，可实现测控现场与控制台的双向有效传输，并对现场设备进行实时控制，实现了真正的实时测控任务，克服了现有的监测系统部署和进行实时动态监测的困难。

通过本文的无线测控通信平台可以快速、方便地构建工业控制、环境测量、精细农业、畜牧养殖、家居自动化等方面的各类测控应用，可大大降低测控应用系统设计和维护的成本，提高测控系统开发和运行效率，保证测控任务的高效实施。平台的可重用性增强了新测控系统的开发，避免类似的故障和错误频繁出现，从而使整个测控系统更可靠、更可信、互连更方便。对本平台的深入研究与推广应用将提高我国的无线测控技术和综合竞争能力，提高工业生产效率，促进测控产品的升级换代，对社会生产力的可持续发展具有重要意义。

参考文献

[1] 胡侃，刘云生，李坚. 计算机科学，一种保证传感器网络实时服务的中间件机制[J]，2007.34 (12):56-60.

[2] 李建中，高宏，无线传感器网络的研究进展，计算机研究与发展[J]，2008 年 45 卷 1 期：1-15.

[3] 刘昌鑫，谭云兰，唐卫东. 无线传感器网络中间件的研究，传感器与微系统[J]，2008 年 27 卷 11 期: 41-43.

[4] 石为人，尤浩，唐云建，张建. 中间件在无线传感器网络节点设计中的应用，传感器与微系统[J]，2008 年 27 卷 7 期：111-113.

[5] 顾寄南，顾志刚，张荣标. 温室无线传感器网络中间件设计方法的探讨[J]，制造业自动化 2008 年 30 卷 10 期：47-49，53.

[6] 黄海平，王汝传，王翠. 基于移动 Agent 的无线传感器网络中间件[J]，南京大学学报：自然科学版，2008 年

[7] Wang MM, Cao JN, Li J et al. Middleware for wireless sensor networks: A survey[J]. Journal of Computer Science and Technology 23(3): 305-326 May 2008.

[8] Karen Henriksen, Ricky Robinson. A Survey of Middleware for Sensor Networks: State-of-The-Art and Future Directions[C], International Workshop on Middleware for Sensor Networks (MidSense 2006), Melbourne, Australia, November, 2006.

云计算及其在移动学习模式下应用初探

赵 萌

(河南大学计算机与信息工程学院, 河南 开封, 475000)

摘 要: 移动学习是继网络学习后又一种新型的学习模式, 是教育技术研究领域的一个新热点。随着 3G 网络的发展及移动终端设备的普及, 移动学习将会作为一种新型的学习模式迅速地发展起来。而云计算是继网格计算后兴起的一种新型计算模式, 是下一代互联网发展的趋势, 其技术特点及应用模式可以在移动学习模式下发挥很好的作用。本文主要介绍了云计算的定义、特点及其应用; 移动学习的含义及特征; 最后讨论了如何利用云计算解决移动学习模式中存在的问题。

关键词: 云计算; 移动学习; 学习模式; SaaS

中图分类号: G40-057 文献标识码: A 文章编号: 1006-7043 (2004) xx-xxxx-x

Preliminary Study of Cloud Computing and Its Application Under the Mode of Mobile Learning

ZHAO Meng

(Henan University Computer And Information College , Kaifeng 475000, Henan China)

Abstract: Mobile Learning is another new learning mode after Network Learning, which is also a new focus in the research field of Education Technology. With the quickening pace of 3G network and popularity of Mobile terminal equipment, Mobile Learning as a new kind of learning mode will rapidly development in the near future. Cloud Computing, one kind of new Computing Mode which emerged after Grid Computing, is becoming the trend of the development of NGN(next generation network). Based the mode of Mobile Learning, the Technical Characteristics and Application Mode of Cloud Computing can play the important role. This paper mainly discuss the definition of the Cloud Computing, the characteristics and the application ;then analyzes the meaning and the features of Mobile Learning; finally discuss how to use Cloud Computing to solve the problems of Mobile Learning.

Keywords: Cloud Computing; mobile learning; learning Mode; SaaS

随着网络通信技术及移动计算技术的飞速发展, 随时随地的移动学习已经成为了可能。移动学习(M-learning)作为网络学习(E-learning)模式中的一个分支, 是一种新型的学习模式。移动学习在自主性、便捷性、灵活性、互动性及延续性等多个方面比传统的学习模式更具优势。因此, 移动学习将吸引更广泛的学习者参与到学习中, 并对于构建终身学习体系及学习型社会具有积极的推动作用。移动学习被认为是一种未来的学习模式, 或者说是未来学习不可缺少的一种学习模式。随着移动学习的发展, 移动学习模式中也出现了一系列亟待解决的问题, 对于这些问题, 可以利用云计算得到解决。云计算是一种新型的计算模式, 对移动学习模式的发展具有促进作用。通过云计算提供的服务能够方便学习者随时随地通过移动终端设备接入“云”中进行移动学习。下面将分别介绍下移动学习与云计算。

作者简介: 赵 萌 (1984—); 男, 硕士研究生, 在读硕士, 研究方向: 计算机网络教育。

1 移动学习相关概念简介

1.1 移动学习的定义

移动学习是现代远程教育的一种形式的变种，由于是一种新型的学习模式，因此对其没有一个统一的严格的界定。Chabra 和 Figueiredo 结合了远程教育的思想，对移动学习做了一个较宽泛的定义：移动学习就是能够使用任何设备，在任何时间、任何地点进行学习；Harris 对移动学习的定义是：移动学习是移动计算技术和 E-learning 的交点。它能够为学习者带来一种随时随地学习的体验。而 Alexander Dye 等人对移动学习做了较具体的定义：移动学习是一种在移动计算设备帮助下的能够在任何时间任何地点开展的学习，移动学习所使用的移动计算设备必须能够有效呈现学习内容并提供教师与学习者之间的双向交流^[1]。本文认为，移动学习的意义不仅在于随时随地随心的学习，实现学习者个性化的需求，更重要的是移动学习拓展了教育领域的范围，并对促进终身学习体系及创建学习型社会的形成具有重要的现实意义，因此移动学习在将来具有广阔的发展前景。

1.2 移动学习的特点

(1) 便捷性的学习工具。随着移动设备的飞速发展，小型化的、智能化的移动终端已经相当普及，功能也相当强大，这就为学习者提供了多种学习工具，这里所指的移动终端主要是智能手机 PPC、掌上电脑 PDA、小型上网本及其他体积小、携带方便的移动终端设备。这类设备突出特点是便捷、灵活，与网络学习模式中的台式主机不同。

(2) 灵活性、自主性的学习环境。移动学习使得学习者能在任何时间、任何地点开展学习，这使得学习者彻底摆脱地理位置的束缚，随时在自己舒服的环境下灵活地学习。另外，由于每个学习者的学习目的与学习任务是不同的，移动学习使他们可以根据自身的学习需求与学习意愿自主的安排学习。

(3) 学习活动更具有情境性、资源丰富性，并以真实情境作为学习隐喻。移动学习为情境式学习提供了支持，学习活动发生在真实自然的社会情境中，学习者便于联系实际环境进行学习，这使得较抽象的学习内容具体化，便于学习者理解；另外，情景学习也有助于学习者把所学到的知识应用到实际生产、生活当中，即实现了真正意义上的“活学活用”。

(4) 广泛的学习群体及终身学习的过程。随着科技的发展及社会的进步，人们已经意识到学习的重要性及终身学习的意义。移动学习使得更广泛的人参与到学习中。广大学习者可以方便地通过无线网络去浏览自己想学的内容，快速地下载自己想要的资料。所以移动学习拓宽了教育的范围，让更多的人接受教育，实现当前倡导的终身教育体系。

1.3 移动学习的应用模式

目前移动学习的应用模式主要分为两类：在线类移动学习与脱机类移动学习。

1) 在线类移动学习

SMS (Single Message System) 模式：基于短消息的移动学习模式主要应用于通信数据少，简单文字描述的学习活动。是目前普遍的一种移动学习途径，技术也相对比较成熟，费用较低，用户数量也最多。通过短信系统，学校可以及时提供各种服务信息，但短信内容只能是文字，而且字数有限，所以应用范围限于通知的发送、简短信息的查询^[2]。

MMS (Multiple Message System) 模式：基于多媒体的移动学习模式主要应用于表达丰富信息，需要使用图像、声音、动画等多媒体信息的学习活动。

基于浏览、链接的模式：通过交流与协作的实时交互来进行移动学习。这种模式的资源形式丰富多彩包括：BBS、BLOG (博客)、图文资料的浏览、远程互动学习、流媒体课件点播、课程视频下载

等。这种类型又包括两类移动学习模式：一类是 WAP 及移动互联网业务；另一类是移动宽带业务。

2) 脱机类移动学习

存储携带模式：是指将电子书、多媒体课件、图文课件等数字化内容存储在便携式移动设备上，帮助学习者进行随时随地的学习。

上述几种应用模式各有优点，各自有其使用的范围及对象，但是在实际应用当中，这几种模式也暴露了各自的一些缺点及不足。因此，亟须一种解决方法来弥补上述移动学习模式的不足。随着云计算技术的发展及其应用的普及，云计算将能很好地解决这些问题。

2 云计算相关概念简介

针对上述讨论的问题，云计算将为移动学习的发展提供技术支持。基于云计算，移动学习的模式将更加完善，学习资源可以在“云”上无限制地增加，学习者可以用低端的移动终端设备随时进行学习，更重要的是云计算能够解决目前移动学习应用模式中存在的缺点及不足。

2.1 云计算定义

云计算（Cloud Computing）是在 2007 年第三季度才诞生的新名词，但仅仅过了半年之多，其受到关注的程度就超过了网格计算（Grid Computing）。近年来，云计算更是发展势头迅猛，逐渐成为未来互联网发展的热点之一。然而，对于什么是云计算，有很多种解释，目前还没有公认的定义。Hewitt^[3]认为云计算是将数据信息存储在云端服务器上，用户在使用信息时在客户端进行缓存，其中客户端包含 PC、笔记本电脑、智能手持终端等。Wang Lizhe^[4]等人从云计算具备的功能角度给出了云计算的定义，指出云计算能够向用户提供 HaaS（Hardware as a Service）硬件即服务、DaaS（Data as a Service）数据资源即服务、SaaS（Software as a Service）软件即服务及 PaaS（Platform as a Service）平台即服务。用户可以按需向云服务器提交自己的软件安装与维护、硬件配置、数据访问、数据存储等需求。Buyya^[5]等人从面向市场的角度认为云计算是由许多内部结构相似的虚拟机互连而成的并行的分布式计算系统，系统能够根据服务提供商和客户之间协商好的服务等级协议动态的提供计算资源。UC Berkeley^[6]的观点认为云是指数据中心的硬件和软件，云分为公有云（对普通大众开放）和私有云（业务组织者自己使用）。在公有云的基础上，云计算是指终端用户请求的应用软件通过互联网以服务的形式由 SaaS 提供商交付，而云提供商向 SaaS 提供商提供数据。本文认为云计算是一种以虚拟技术为核心的新型计算模式，其继承与发展了分布式处理、并行处理和网格计算，并且把基础设施、开发平台、软件作为一种服务按需交付给用户使用（如图 1 所示）。

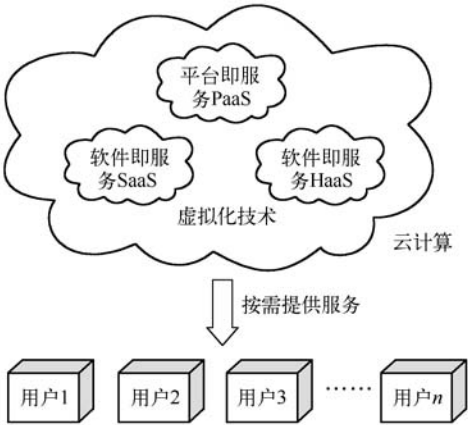


图 1 云计算定义

由此可见，云计算是一种动态的易扩展的且通常是通过互联网提供虚拟化资源的计算方式。用户不需要了解云内部的细节，也不必具有云内部的专业知识，只需接入云端服务器，便能使用相关资源。在云计算时代，我们只需要一个移动终端设备，就可以通过网络服务来实现我们需要的一切，甚至包括一些个人计算机无法应对的超级计算任务，这就是云计算的魅力。云计算的实现原理是通过使计算分布在大量的分布式计算机上，而非本地计算机或远程服务器中，企业及用户的数据中心运行将更与互联网相似。这使得用户能够将资源切换到需要的应用上，根据需求访问计算机和存储系统。

2.2 云计算的主要特征

(1) 虚拟化技术。是云计算最强调的特点，包括资源虚拟化和应用虚拟化。每个应用部署的环境和物理平台是没有关系的。通过虚拟平台进行管理达到对应用进行扩展、迁移、备份，操作均通过虚拟化层次完成。

(2) 动态可扩展。通过动态扩展虚拟化的层次达到对应用进行扩展的目的。可以实时将服务器加入现有的服务器机群中，增加“云”的计算能力。

(3) 按需部署。用户运行不同的应用需要不同的资源和计算能力。云计算平台可以按照用户的需求按需部署资源和计算能力。

(4) 高灵活性。现在大部分的软件和硬件都对虚拟化有一定的支持，各种 IT 资源，如软件、硬件、操作系统、存储网络等所有要素通过虚拟化，放在云计算虚拟资源池中进行统一管理。同时，能够兼容不同硬件厂商的产品，兼容低配置机器和外设而获得高性能计算。

(5) 高可靠性。虚拟化技术使得用户的应用和计算分布在不同的物理服务器上面，即使单点服务器崩溃，仍然可以通过动态扩展功能部署新的服务器作为资源和计算能力添加进来，保证应用和计算的正常运转。

(6) 高性价比。云计算采用虚拟资源池的方法管理所有资源，对物理资源要求比较低。可以使用廉价的 PC 组成云，而计算性能却可以超过大型主机。

(7) 安全性。云计算采取中央集权的数据管理模式，数据的安全性大大提高。因为，供应商能够把资源用于进行安全审计并解决安全问题，而一般的客户能力或者资金有限，无法有效的保障本地数据资源的安全。

(8) 自治性。云计算系统是一个自治系统，系统的管理对用户来讲是透明的，不同的管理任务是自动完成的，系统的硬件、软件、存储能够自动进行配置，从而实现对用户按需提供^[7]。

2.3 云计算的体系结构

云计算的体系结构分为四层：物理资源层、资源池层、管理中间件层和 SOA（Service-Oriented Architecture，面向服务的体系结构）构建层^[8]。物理资源层包括计算机、存储器、网络设施、数据库和软件等。资源池是将大量相同类型的资源构成同构或者接近同构的资源池，如计算资源池、数据资源池等。构建资源池更多的是物理资源的集成和管理工作。管理中间件层负责对云计算的资源进行管理，并对众多应用任务进行调度，使资源能够高效、安全地为应用提供服务。SOA 构建层将云计算能力封装成为标准的 Web Service 服务，并纳入 SOA 体系进行管理和使用，包括服务接口、服务注册、服务查找、服务访问和服务工作流等。管理中间件层和资源池层是云计算技术的最关键部分，SOA 构建层的功能更多依靠外部设施提供。具体云计算的体系结构如图 2 所示。

2.4 云计算的应用形式

(1) SaaS（软件即服务）：这种应用形式的云计算，是把软件当做服务提供给用户使用，用户只需通过浏览器访问在云端的软件即享受到以前只有安装到本地计算机才能使用的各种软件。从用户角度而言，这样会省去在服务器和软件授权上的开支；从供应商角度而言，只需要维持一个程序即可，

这样能够减少成本。

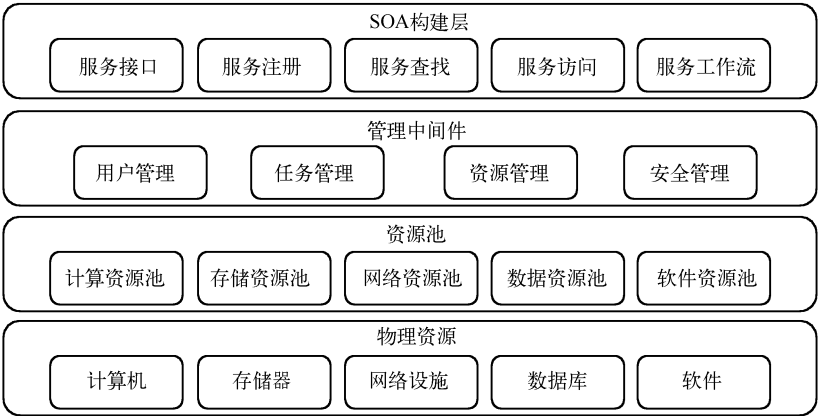


图 2 云计算的体系结构

(2) PaaS（平台即服务）：即另一种形式的 SaaS，这种形式的云计算，是把软件开发环境作为一种服务来提供。软件开发人员可以使用各种云计算中间商的设备开发自己的程序，并通过互联网和其服务器传输到用户手中使用。

(3) Utility Computing（实用计算）：这个技术原先就有，但是直到近几年才在 Amazon、Sun、IBM 其他提供存储服务 and 虚拟服务器的公司的推广中得到重生。这种云计算是为 IT 行业创造虚拟的数据中心，使得其能够把内存、I/O 设备、存储和计算能力集中起来，成为一个虚拟的资源池为整个网络提供服务。

(4) Business Platform（商业服务平台）：SaaS 和 MSP 的混合应用，该类云计算为用户和提供商之间的互动提供了一个平台。比如用户个人开支管理系统，能够根据用户的设置来管理其开支并协调其订购的各种服务。

(5) MSP（管理服务提供商）：是最经典的云计算运用之一，这种应用更多的是面向 IT 行业的从业者本身，而不是普通的终端用户，常用于邮件病毒扫描、程序监控等。

(6) Net Service（网络服务）：网络服务提供者提供 API，让开发人员能够开发更多基于互联网的应用，而不是单机程序。

(7) Internet Integration（互联网整合）：互联网上提供的服务众多，不便于用户的查找和使用，把提供类似服务的公司整合起来，以使用户能够更方便地比较和选择自己需要的服务供应商。

3 利用云计算解决移动学习模式中存在的问题

1) 整合移动学习资源，实现资源利用最大化，弥补移动学习资源的匮乏

移动学习模式中传统的学习资源比较匮乏，而互联网中的教育资源却十分丰富，但是互联网上的学习资源分散度高、聚合度低，这个特点不利于在配置不高的终端设备上搜索使用。由于学习者不能从浩瀚的网络中快速地找到学习资源，从而使学习效率下降。因此，利用云计算对学习信息进行融合、存储并通过网络服务进行共享，通过云计算的虚拟化技术，使得网络中分散的学习资源得到最大限度的整合，以服务的形式提供给学习者使用，省去了搜索的中间环节，便于学习者通过移动终端设备进行移动学习。

2) 丰富教学软件数量，减少移动学习模式下程序开发与维护的成本

移动学习相关功能需要各种教育应用程序的支持，开发与维护这些程序需要付出大量的成本，而移动学习模式处在发展初期，大部分人还对其认识不足，因此，投入到这些程序的开发者相对较少，

进而软件也较少。在云计算时代，应用程序的开发者可以利用云服务提供商提供的 PaaS 开发各种各样移动学习所需的应用程序，并且将开发的程序上传到云服务器当中并通过网络提供给学习者使用。对于开发者而言，只开发与维护一个程序从而减少了成本，这样能促进更多的开发者投入到移动学习的程序开发，丰富移动学习的软件数量。对于学习者而言，只需根据需要输入关键词，即可获取各种教学相关的应用程序。

3) 弥补移动学习传统模式的不足，促进移动学习的普及

在传统的移动学习模式下基于短信的 SMS 模式是其主要模式之一，这种模式下学习者通过定制短信息进行学习，服务端短信息一般都是编辑完成后通过短信平台发布给学习者的，学习者无法与教学者进行实时的交流，即使将问题以短信息方式回复给教育者，教育者也很难在第一时间进行回答。况且，短信息的发送、接收还受信息堵塞、网络是否通畅等诸多因素的影响。另外，基于多媒体 MMS 模式和基于 WAP 站点的模式由于 WAP 教育站点不足，以及 WAP 协议的传输速率不足等因素，使得教学效率不高。

云计算很好地弥补了这些不足，将大量的教育资源存储在“云”服务器中解决了资源不足的问题；云计算整合了多方计算机、互联网新技术，能使移动学习突破单一 WAP 协议的限制，使得传输速率大大提高，学习者不必再通过短信或者 WAP 站点点播的方式，只需要使用带有浏览器的移动终端设备进入云服务器就可以自行选择适合自己的教学资源进行学习。

4) 降低移动学习模式对终端设备的配置要求

在移动学习模式下移动终端设备配置不高、运算速度不快等问题制约着移动学习的发展。主要表现在：

- (1) 移动终端设备（智能手机 PPC 及掌上电脑 PDA）的计算速率及信息处理能力不及计算机。
- (2) 大部分移动终端只支持 WAP 协议，不支持 HTTP 协议，这使得多媒体资源的传输受到限制。
- (3) 移动终端设备的接口不统一，链接不同类型设备时有一定困难。

云计算技术的出现为解决这些问题提出了很好的方法，云计算把所有的数据存储在云服务器当中，移动设备不再需要具备太大的存储空间进行资源的存储；另外，云计算把所有的数据处理及计算都放在云端的计算机群中进行，其强大的计算能力远远高于普通的个人计算机，还能完成个人计算机无法完成的处理任务，从这一点来说，云计算降低了对移动设备 CPU 计算能力的要求。另外，目前云计算提供的服务已经很好地支持了 WAP 协议及 HTTP 协议，即使使用各种不同的终端都能轻松访问云端服务器，这样就避免了各种终端平台分别开发自己的应用程序而导致不同终端设备互访时兼容性低的问题。由此可见，一切处理均在“云端”进行，学习者只需要一个浏览器接入云端服务器，便可完成像个人计算机上才能完成的工作。对移动设备来说，除了具备运行浏览器本身所需计算能力以外，无须任何别的数据处理要求，目前大部分智能移动终端设备就完全可以胜任。

当然，移动学习中还存在其他方面的问题，例如，网络传输速度的发展及 3G 网络的普及；学习者在移动学习过程中的自我约束问题；移动学习模式下教师角色的转变等问题，但这些问题并不在本文的讨论范围之列。

4 结束语

云计算是一种全新的很有发展前景的计算模式。云计算对很多学科领域都具有促进作用。对于移动学习，应当充分研究云计算的教育应用，发挥云计算在移动学习模式下的技术优势，创建一种新型的移动学习模式，进而弥补传统移动学习模式的缺点及不足。下一步研究工作主要在开发与设计一种新型的以云计算为基础的移动学习模式。本文只是做了一些基础性的研究，希望对今后相关研究起到抛砖引玉的作用。

参考文献

[1] 刘豫钧, 高淑芳. 移动学习——国外研究现状之综述[J]. 现代教育技术, 2004(3).

[2] 李晓丽, 王晓军. 移动学习模式探讨及系统架构设计[J]. 北京邮电大学学报(社会科学版), 2007(5).

[3] HEWITT C. ORGs for scalable, robust privacy-friendly client cloud computing [J]. IEEE Internet Computing, 2008, 12(5): 96-99.

[4] WANG Lizhe, TAO Jie, KUNZE M. Scientific cloud computing: early definition and experience[C] // Proc of the 10th IEEE International Conference on High Performance Computing and Communications. 2008: 825-830.

[5] BUYYA R, YEO C S, VENUGOPAL S. Market-oriented cloud computing: vision, hype, and reality for delivering IT services as computing utilities[C] // Proc of the 10th IEEE International Conference on High Performance Computing and Communications. 2008: 5-13.

[6] ARMBRUST M, FOX A, GRIFFITH R, et al. Above the clouds: a Berkeley view of cloud computing[R/OL]. (2009-02-10)[2009-05-15]. <http://www.grid.pku.edu.cn/cloud/Berkeleyabovetheclouds.pdf>.

[7] 张建勋, 古志民, 郑超. 云计算研究进展综述[J]. 计算机应用研究, 2010(2).

[8] 刘鹏. 云计算[M]. 电子工业出版社, 2010.

网络统计及软件研究

贺学剑，孟光胜

(河南科技大学林业职业学院，河南 洛阳,471002)

摘 要：Internet 作为一种计算机通信网络，正在迅速地改变着人类的生活、工作和思维方式，对社会生活的各个方面也都产生着深刻的影响，同时也影响着统计工作进行的组织手段和行为方式。全球网络化对现行的统计报表制度、统计调查方法及数据处理技术等统计信息的生产方式产生了巨大的影响，同时也对传统的统计观念产生了冲击，迫使统计观念更新和重构。在网络时代传统统计已经难以适应网络时代发展的要求，新形势下统计改革势在必行。在此基础上进一步探讨网络环境下统计软件的变化和开发工作的新发展，对于网络时代统计理论研究具有一定的积极意义。

关键字：网络统计；网络环境；网络统计系统

Reach on Internet Statistic and Software

HE Xuejian, MENG Guang sheng

(Forestry Vocational College Attached to Henan Science and Technology University, Luoyang 471002, Henan China)

Abstract: Internet as a computer communication network, is rapidly changing the human life, work and ways of thinking in all aspects of social life are also exerting a profound impact, but also affect the organization of statistical work carried out by the means and patterns of behavior. Network to the existing statistical reporting system, survey methods, information transmission means, data processing techniques such as statistical information on production methods produce a revolutionary effect, but also on the traditional statistical concepts produced a strong impact, forcing the concept of update statistics and reconstruction. In the Internet era has been difficult to adapt to the traditional statistical requirements of the development of the Internet age, the new situation, statistical reforms. On this basis, further explore the statistical software under the network environment changes and development of new developments in statistical theory for the Internet Age has some positive significance.

Keywords: Internet statistics, Internet environment, Internet statistic system

1 网络统计的近期研究

网络统计是传统统计在新的信息传播媒体上的应用，具体指为研究总体特征而利用计算机国际互联网络进行的统计数据资料收集、处理、展示、发布等活动的总称。网络统计的发展是计算机科学（特别是其中的网络科学技术）与统计活动相结合而产生的新的领域之一。作为理论研究，目前主要集中于两个方面：一是考证传统的统计方式与利用网络进行统计的方式之间的优缺点问题；二是如何有效地完善利用网络进行统计活动的新方式，研究内容包括统计网站的建立、调查网页的设计、在线数据处理、统计信息系统等方面。

现有统计软件存在着以下一些问题：

- （1）易用性差：某些统计软件实际上是一门计算机语言，需要进行相当程度的学习才能掌握，至于熟练使用更是需要耗费大量的时间和精力。例如 SAS，功能强大，扩展性好，能适应绝大部分的统计计算工作的需要。但是，这种功能的强大在是以牺牲软件的易用性为代价的，除了还要学习其提供脚本语言，还需要具备基本的计算数学背景知识，对使用者要求过高。
- （2）扩展性差：某些软件的易用性很好，不需要使用者了解任何计算机语言、计算数学的背景知识，只要使用者懂得基本的计算机操作和相应的数理统计的知识即可。但是，这种软件往往是建立在某一或某几个统计模型基础上的，可扩展性较差，当需要应用超出模型适应范围的统计方法时就无能为力。

为力了，只能使用其提供的统计方法进行统计计算工作，无法适应快速发展地数理统计科学对新方法的需要。**SPSS** 是这种软件的一个代表。

(3) 维护性差：目前大多数应用软件系统都是 **Client/Server**（客户-服务器）形式的两层结构，这种系统存在着维护费用高、系统升级困难等诸多问题；同时在 **C/S** 模式下的应用系统无法实现在线实时统计及发布的功能。传统的 **C/S** 体系结构虽然采用的是开放模式，但这只是系统开发一级的开放性，在特定的应用中无论是客户端还是服务器端都还需要特定的软件，没能提供用户真正期望的开放环境。因此，现存软件在扩展性、易用性上往往顾此失彼，不能满足统计的需要。针对这种情况，研究基于 **Web** 的应用系统，来取代传统的客户-服务器模式，以便能够利用 **Internet/Intranet** 上丰富的信息资源，构建一个易于开发、易于维护、并具有良好可伸缩性的应用程序，就显得尤其必要。

2 网络环境的发展与网络时代的特征

2.1 现状和特征

2.1.1 网络环境的发展现状

Internet 诞生于 20 世纪 60 年代，是一种集通信技术、信息技术、计算机技术于一体的网络系统，是有史以来人类拥有的最大的网络空间。简单地说，它是使用 **TCP/IP** 协议（传输控制协议/网络间协议），将各个地区、各个国家的各种不同类型计算机网络联系起来的“国际网”，是目前计算机之间进行信息交换和资源共享的最佳方式。到 20 世纪 90 年代，随着 **WWW**（万维网）技术的应用，**Internet** 开始商用化，并取得了极大的成功，尤其是近几年取得了爆炸性发展，使 **Internet** 越来越成为当前世界不可缺少的重要工具。

2.1.2 网络时代的特征

由于 **Internet** 的存在，地球上不同角落的人们可以随时看到外面世界的瞬息万变。对许多人来说，一种崭新的网络生活方式出现了：你上网看新闻、学习、娱乐变成很自然的事，同时你可以上网购物、炒股、淘金，也可以上网收集或传送各种你需要的信息资料。**Internet** 强大的信息共享性、信息交流的主动性和信息交流的交互性，使得以国际互联网为基础的网络时代呈现出虚拟性、互动性、全球性的特点。

虚拟性：在网络中可以“相识不相见”。正如美国匹兹堡大学的心理学教授扬格所说，即使你是个足不出户的家庭主妇，也可以与世界各地的人交往、学习，于是平淡的生活中就有了不平凡的内容。也正是这种网络的虚拟性增加了网民之间的平等性。交谈者可以对对方的身份一无所知，使得交流变得更加自由轻松。网民中有句经典名言：“在 **Internet** 上，没人知道你是只狗。”

互动性：与普通媒介信息传递方式不同，网络媒介中从信息的发布、传输到接收，都是双向的、可逆的、互动的（如图 1 所示）。因此，网络用户既是信息的接收者，也是信息的参与者、制造者。

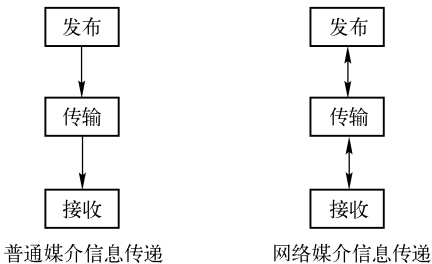


图 1 两种媒介信息传递比较

全球性：互联网络本身是无边无际的，网络资源又是全球人共享的资源，网络技术实现了从“天涯海角”到“近在咫尺”的突破，把全球联系起来。网络不再局限于一个地区，一个国家，而是扩展到全世界。坐在计算机前就可以定向抵达一点，也可同时到达多点，可以抵达最遥远的地方，超越时空距离，为人们在更大范围内获得信息提供了方便。互联网络自发明以来，以爆炸性的速度膨胀，每次发展，对用户而言，都变的更有价值，于是就会得到进一步的扩展。这一个个的不断扩展、一次次更有价值，给社会带来了深刻的影响。

2.2 网络环境对传统统计工作的冲击

通信技术发展，尤其是传播媒介的变革通常透视着统计的演变过程。文字发明之前的原始统计是结绳记事、结绳计量；人类第一次通信革命是发明文字，它的出现使知识克服了时空的障碍，因此，有了用文字进行人口、粮食和土地的统计记录。由此出现了纸介质统计，第三次通信革命是电信业的发展，出现电话、电报和传真等，并与纸介质统计并存；第四次通信革命计算机和网络技术的发展，把文字、图形和声音转化为二进制的数字语言，从而使传统统计发生了前所未有的变革，磁介质统计替代了纸介质统计，并迅速发展成无介质统计——电子统计。所以有人说，现代计算机技术和网络通信技术的发展是网络统计产生的技术基础。它不仅加速了统计技术现代化的步伐，而且将引起统计管理体制、组织机构、工作方式的根本变革。

2.2.1 传统统计调查方式受到挑战

1) 统计信息保存技术的变化给传统统计调查带来压力

在传统统计调查中，大量、繁杂的简单劳动占据了统计人员绝大部分时间，而且这些统计工作对统计人员的责任心要求高，对业务素质和知识结构的要求相对较低。然而，信息技术的发展和计算机的广泛应用，改变着被统计信息的保存方式和登记手段。统计信息的保存类型由传统单纯的数字和文字，发展为图像、声音、文字、数字多媒体等多种形式。统计信息的登记也由传统的手工登记转变为利用计算机技术和网络技术的统计信息系统自动完成。比如，现在英国的统计调查人员携带的不是纸和笔而是笔记本电脑，他们将调查获得的信息通过网络直接传给数据处理中心；在税收统计调查中，则是将纳税人的“刷片”信息直接传输到统计中心，由计算机直接进行汇总，并将调查表通过电子邮件传给被调查者，被调查者通过网络及时反馈有关信息。

同时，随着统计信息标准化的发展、自动分类技术的应用，统计信息的统计单位、统计分类也会发生变化。在这种情况下，如果再运用传统的统计调查方式开展统计调查，将很难收集到一线的信息，统计调查工作的效率将会是事倍功半。因此，如何适应统计信息保存格式、登记手段等的变化，在信息技术条件下收集统计信息给传统统计调查带来了一定的压力。

2) 快速的信息传输、庞大的信息量使传统统计调查面临挑战

统计调查最讲究的是快、精、准，Internet/Intranet 上信息传输速度的迅速及时正符合统计调查的要求。网络中的信息发布是即时、快速、全天候的，同样网络中信息的收集也是即时、快速、全天候的。因此，要想收集到及时的统计信息，就必须充分利用网络优势，寻求统计调查方法与网络技术的结合点，这将给传统统计调查方式带来冲击。

随着信息技术的日臻完善，信息的传输和交流更为迅速、更为简便，同时信息的内容也更为丰富，来源也更为广泛。网上信息资源丰富和浩瀚程度，已远远超过了传统的文献资源量。面对信息的海洋，只有针对不同统计服务对象对统计信息提出的需求，运用现代化的统计调查方法，才能在良莠不齐的统计信息中挖掘出有用的信息。这也将使传统统计调查面临挑战。

3) 多样的信息交流渠道冲击传统统计调查方式

网上信息交流拉近了人们之间的时空距离，增加了信息交流的渠道。统计调查作为社会经济信息的收集过程，必将受到这种影响的冲击，其直接冲击之一将是引起调查方式的重大转变，即以

E-mail 调查、主动浏览访问等方式为代表的网上调查成为统计调查的流行方式。我国也已有“零点”调查公司、中央电视台、人民日报社、163 电子邮箱和多家网站等进行了网上调查。这种调查方式将因其快速、便捷、经济和高效的优点受到越来越多人的欢迎，这对传统统计调查来说无疑是一种强大的压力。

2.2.2 传统统计工作方式面临挑战

1) 人们生产、生活方式的改变给传统统计工作带来深刻影响

网络时代将造就一些不同于传统的全新统计客体。比如目前的经合组织国家在工商界中出现了“虚拟商店”、“虚拟公司”、“虚拟企业”、“虚拟市场”等新型客体。“虚拟企业”利用网络、电子商务的手段将各种业务外包，其本身只以创新行为和名牌效应为龙头，突破了传统企业的界限，在全球范围内对企业内部和外部资源进行动态配置，优化组合，这样的企业没有完全的经营门面，没有现成模式的营销经营方式，是一种“全天候”的经营和一种无国界、无地域限制的经营。这种工商景象虽然刚刚露出萌芽，但有迅猛发展的网络作为肥田，加上人们观念的快速更新作为外部条件，它的成长将是茁壮的。统计工作必须未雨绸缪地看待这些统计客体的出现。要根据它们的特点，充分利用网络资源和先进的分析方法来推断分析这些客体的经营状况。

2) 传统统计信息存储、发布方式面临挑战

统计信息的存储是贯穿于整个统计工作过程的重要组成部分，包括统计调查资料的存储、统计整理资料的存储和统计分析资料的存储。存储统计工作各阶段的信息，是保存和积累统计历史资料的需要，也是用数据来编写人类社会发史的需要。那么该如何来存储统计信息呢？传统的纸载存储方式显然已越来越不适应新形势的需要了，因为它存在（1）成本高、速度慢，从编写到排版、再到印刷和装订，有很多环节，需要较多的时间和投入；（2）容量少、弹性小，如果要扩容或调整，需从头来过；（3）存储信息所要占用的空间大；（4）不便于管理，尤其是资料的携带和搬迁更为麻烦；（5）信息的查阅、引用、转换和重新组合比较麻烦，不方便等缺点。信息技术的发展将改变这一切，它为我们提供了体积小、容量大、性能优、速度快、成本低，几乎可以随心所欲地进行扩容、调整、引用、转换和重新组合的信息存储方式，那就是磁载和光载方式，它使我们实现无纸化信息存储完全有了可能。

与此同时，统计信息的发布方式也将随之而变，计算机网络方式将逐步取代报纸、杂志和年鉴等方式而成为主体。统计信息的这种存储和发布方式的改变，正是前面讲到的统计信息满足细分化、个性化社会需求的基本条件，也是把统计部门建设成为信息库、数据库、思想库和智囊库的基本要求。

3 网络时代给统计工作带来的新机遇

网络时代计算机技术、网络技术的发展给传统统计工作带来巨大的冲击，传统统计已经难以适应时代发展的要求，但我们应该看到，这既是挑战，也是机遇。由于网络是一个新生事物，前面没有规则可循，统计工作在抓住机遇发展自身的同时，不可避免地在前进中会遇到一些挫折，如何解决这些挫折带来的困惑，需要认清形势，才能摸索到出路，在发展中前进。

3.1 信息网络技术的发展增加了统计信息的需求

世界各国经济日益形成“你中有我、我中有你”的相互依存局面，这既深化了合作也加剧了竞争。各国必然会需要更多、更具体地反映世界各国经济情况的信息作为自己分析决策的依据。从宏观角度来看，各国必须核算 GDP，从而掌握整个国民经济发展态势，以期进行国际横向和本国纵向比较，这就要求统计外贸、投资、消费和财政（ $GDP=C+I+G+X$ ，当然这只是一个简化的公式）方面的数据资料。宏观经济目标除经济增长外还有通货膨胀、国际收支平衡、就业情况等。随着宏观调控的

面越来越大，各国政府对统计信息的需要将越来越迫切。从微观角度来讲，各国企业必须把握市场动向，分析市场结构，明确市场定位，了解竞争对手，才能制定正确的战略对策，这对统计信息有更大的需求。

此外，新出现的“新经济学”、“知识经济”，“知识”已经越来越成为主要的生产要素。但有关“知识”的量化、测定的统计工作进行得远远不够，日益增多的知识产品，如高新技术的投入、运作、产出都需要给予恰当的分类和估算，这是信息需求与统计供给之间的断层。因此，如何进行知识的界定，如何建立知识统计体系等一系列问题给统计信息也提出了更多的需求。

3.2 网络时代为统计工作提供高效的信息收集、处理能力

3.2.1 信息网络技术为统计信息的收集处理提供良好的支持

1) 信息技术创造了良好的外部环境

随着现代科学技术的迅猛发展，信息技术的开发和应用正以前所未有的速度遍及社会的各个领域，强烈地影响和改变着人们的生产、生活方式，改变着人们的生存观念。计算机（Computer）、通信（Communication）、信息内容（Content）“3C”的有机结合为数字化、网络化、信息化“三化”的实现提供了广阔的技术空间。无纸办公，网络营销数字化生存，数字化地球、自动化控制与管理等完全依靠信息技术支持的社会活动，成为人们谈论最多的话题之一。可以说，是信息化大潮推动人们迈入 21 世纪。也可以说，在不久的将来，信息技术对我们每个人来说，就像住房服装一样必不可少，人们对网络的要求如同对交通道路一样，不但要有，而且要有快速、高效、畅通无阻的能力。

2) 通信技术提供了先进的技术支持

随着我国国民经济的发展和科技的提高，我国信息化建设的脚步也越来越快，国家公众数据网和多媒体技术正不断改进和完善，计算机在各项生产、管理和机关办公自动化中得到广泛应用。统计人员如何在信息技术高度发展的今天，抓住机遇，加快发展，革新手段，不能不说是一项重要的基础性工作。

3) 电算化软、硬件的普及奠定了坚实的基础

我国 IT 产业经过十几年的探索，研究和发展，在硬件、软件的开发、管理及运用方面，已经拥有了比较成熟的技术，形成了一整套较为科学、先进的技术体系，在人员方面，已形成了一支有较高素质，对信息技术有一定了解和掌握的统计队伍，这些条件，为网络环境下统计系统的普及与推广提供了最基本的应用基础和人力资源。

3.2.2 高效的统计信息收集、处理能力

1) 多元化的信息提供方和需求方

统计信息提供方是指专门负责数据收集、加工、处理的单位或部门；统计信息需求方是指作为社会经济主体的政府、企业及广大社会公众。传统的统计信息提供方仅为国家及地方各级统计部门和企业统计部门，统计的信息需求方仅为国家及地方各级政府。在网络经济时代，统计信息的需求方和提供方将是多元化的，除原有的信息提供和需求方外，社会各微观经济主体均将成为统计信息需求方，信息咨询公司或民间统计组织等均将成为统计信息提供方。

2) 先进的信息收集、处理能力

在网络时代，统计调查员将不再风里来、雨里去地挨家挨户收集统计数据，而是充分利用先进的网络技术开展网上普查和抽样调查，或直接访问统计信息中心，收集、获取大量一线原始信息资料。同时利用计算机技术，对取得的原始信息进行定性、定量分析，生成大量的再生信息。网络技术和计算机技术的发展极大提高统计信息搜集处理的效率，大量的统计信息直接在网络中收集得到，大量数据的整理汇总运算可以在几秒甚至瞬间完成。目前，世界上运算速度最快的计算机 IBM 最新开发的“走鹃”占据了榜首宝座。“走鹃”的运算速度达到了 1.026 petaflop，即每秒可进行 1026 万亿次

浮点运算。美国国际商用机器公司（IBM）研制的“蓝色基因/L”超级计算机系统可在每秒最高运算478.2 万亿次浮点运算。因此，网络时代由于统计信息收集和处理能力的提高，将带动统计工作效率的提高，降低统计工作的劳动强度。

3) 专业的信息配送机制

当前统计信息中存在着一个久久不能解决的问题，即各专业自成体系，各自为政，造成专业间数据不统一，不能实现资源共享。在网络时代，统计工作将运用先进的设备、复合型的人才、完备的网络等软硬件条件，利用现代化的手段进行统计数据的收集、分配，建立“统一采集、集中处理、数据共享”的新型的专业的统计信息配送机制。这种专业的统计信息配送机制有利于提高统计信息的利用率，有利于建立统一的统计信息数据库。

3.3 信息网络技术与统计工作的结合促使信息质量进一步提高

众所周知，统计工作的基本要求是准确、及时、全面、系统地提供统计信息，而其中统计信息的准确性更是统计工作的生命线和灵魂。随着网络时代统计信息需求方和提供方的多元化、统计信息传播渠道的更加先进便捷，统计信息的质量将进一步提高。

3.3.1 统计信息可靠性增强

统计信息的可靠性是指统计信息能够反映客观实际，在收集、整理和传递的过程中客观、公正，没有人为干扰和篡改，这是统计信息质量的生命之本。如何保证统计信息“不出假数”、“真实可信”，有效地提高统计信息的质量，已成为当前统计工作的重要议题。

3.3.2 统计信息时效性加强

统计信息时效性是指统计调查基准期与统计发布时间间隔，这关系到统计信息自身的价值和效力，是统计信息质量的活力之源。统计信息作为决策的依据，时效性直接决定决策效果的好坏，过时的统计信息对决策将毫无实用价值。因此，统计信息的时效性如何是影响统计信息质量高低的重要因素之一。

在网络时代，统计工作将充分利用计算机技术和网络传输技术，实现统计资料编辑电子化，资料收集、交换网络化，加快了信息传播速度，保证了信息传播的及时性，全方位、多层次地开发和利用统计信息资源，有效地实现了统计信息的社会共享。

1) 国家统计信息网络的建立和完善

国家统计信息网络是以国家、省级和地市统计局为主要网络节点的中高速多媒体数据通道，带动建设全国统计系统国家、省、地、县四级网络系统，使信息交流网络化。各级统计机构要在现有基础上加大统计信息网络的硬件和软件建设，尽快建成国家统计系统全国社会经济信息网络。（1）建立县以上统计信息互连网络，并建立数据查询系统；（2）实行大中型企业与统计部门联网，建立数据直传制度；（3）统计部门与各级党委、政府联网，建立统计信息库，供决策查询；（4）开设信息客户咨询网络，建立有偿服务专线；（5）建立 Internet/Intranet 体系服务平台，为接入统计信息网络的大中型企事业单位及政府部门提供数据报送通道和各种信息访问服务。有了这个集查询、传输、服务为一体的国家统计信息网络，统计信息的时效性自然就会加强了。

2) 统计信息发布制度的建立和完善

在网络时代，国家统计局将负起主要的组织和协调工作，以求不断完善国家统计新闻发布制度，其主要内容包括以下几个方面：（1）负责向社会发布年度国民经济和社会发展统计公报，以及月度经济运行报告；（2）组织系列报道和专题性统计分析报道，针对社会经济生活中的一些重点、难点、焦点问题进行深入的调查研究和解剖分析，通过新闻媒体发表；（3）将主要统计信息在国际互联网上发布。此外，完善统计新闻发布制度，要密切统计与新闻单位的联系，努力争取新闻单位的合作，通过计算机网络向新闻单位提供信息，增强统计新闻发布的广度和深度。同时，应制定严格的新闻发布规

定，提高统计新闻发布的权威性，努力实现统计信息传播渠道的多样化。

3.3.3 统计信息适用性加强

统计信息的适用性是指统计信息的提供能够解决实际的问题，起到一种辅助决策的作用，是统计信息质量的核心。只有适用的统计信息才是最有价值、最能体现统计效益的信息，否则信息再及时、再准确可靠，也只能是毫无用处的“信息垃圾”。例如，在资料内容上，根据需要，向顾客提供比较性信息或分析性资料；在服务手段上，可增加统计信息 VCD、统计图文电视等形式；在时间安排上，充分利用发达的互联网技术和通信技术，建立和优化统计信息传播系统，提供快速、准确的传播统计信息服务。可见，统计信息传播服务方式的现代化极大地加强了统计信息的适用性。

4 网络统计体系结构

纵观统计数据处理手段的发展历史，经历了手工、机械、机电、电子等阶段，许多很好的统计方法，如 20 世纪 20 年代的多元统计方法对于处理多变量的种类数据问题有很大的优越性，但由于计算工作量大，使得这些有效的统计分析方法在一开始并没有在实践中很好地推广开，随着计算机技术的发展，一些相应的统计软件被开发和商品化，使得原本复杂的数据处理工作变得方便和迅速，一些非统计专业的科技工作者也可以凭借这些软件来处理有关统计数据。

4.1 网络统计系统的设计原则与要求

网络统计系统的设计是一项复杂的系统工程，它不是网络技术与统计系统的简单叠加。要求设计人员在具有很强的计算机开发技术的同时还要具有很强的统计知识，除具有统计功能以外，还要具有很强的网络功能，如网络通信，远程控制，超强兼容和远程教学等，所以网络统计系统在设计与实现上要比其他一般网络系统困难得多。

4.1.1 网络统计系统的设计原则

网络统计系统实际上是一个在网络系统支持下的计算机统计系统，网络统计系统既要具备网络功能，还必须具备强劲的统计功能，是一个运行在网络上的高级计算机软件和任务系统。它不仅执行网络协议负责计算机的信息交换，更对网络资源统一进行管理，对核算管理信息进行统计监督，因此，网络统计系统在设计上与其他一般网络信息系统相比，要求更严、更高，具体应把握以下几条原则。

- (1) 数据安全性原则。统计数据一般集中保存在专门的服务器上，坚持数据安全性原则是建立网络统计系统的首要原则，数据安全性原则要求网络统计系统为保护网络资源提供多级安全保护。例如文件级、网络目录级、用户组、用户组级的保护，以及对用户登录工作站点及时间的限制。
- (2) 可靠性原则。可靠性是衡量一个系统成熟、完善的基本指标，它要求网络系统的运行不仅能达到系统设计的目的，而且在遇到非法操作等情况下，不得给用户造成损失，可靠性原则要求网络系统，应能达到：①复式存储方式，能对硬盘目录和文件分配表进行保护；②热修复及写后经验证这个互补技术，能对硬盘表面损坏的数据进行修复；③硬盘损坏时，能采用磁盘镜像的方法对硬盘进行保护；④磁盘双式，能在磁盘通道或硬盘驱动损坏时起保护作用；⑤事务跟踪系统能及时对数据进行恢复等要求。
- (3) 健壮性原则。健壮性原则要求网络统计系统在设计上能够适应信息技术、市场经济的发展，提供便捷的修改空间及新的模块接口。

4.1.2 网络统计系统的设计要求

网络统计系统的设计要求主要包括两个方面：一方面是网络系统的总体要求；另一方面是对系统

的程序要求。对网络统计系统的总体要求包括:

- (1) 功能完善。一套完善的网络统计系统应具备基本的网络功能和统计功能,如电子邮件服务(E-mail),文件输出服务(FTP),终端仿真服务(TELNET),信息查询服务,数据采集、数据通信、分布处理交互操作等功能。
- (2) 方便易用。网络统计系统应考虑到系统用户多,使用频率高,操作人员水平不一等特点,网络统计系统在用户界面设计上应做到简洁合理、美观大方,力求使操作尽可能简单。
- (3) 技术先进。一套好的系统应采用开放式体系结构,以业务为主线实现即插即用,便于不断适应形势的发展进行功能扩充和系统集成。在设计上,应注意运用最近几年创造或流行的先进的网络技术和计算机技术系统集成技术,以适应市场经济不断发展的需要。对网络统计系统程序的设计要求包括结构清晰、容易理解、便于维护、性能强健和运行高效。

4.2 网络统计系统的结构

网络统计系统结构从总体上看主要包括统计数据处理,以及统计信息发布两大功能模块,在此基础上可以根据需要,建立如远程教学等其他的功能模块。

4.2.1 统计数据处理模块

- 统计数据处理模块主要完成对统计数据的收集、分析、处理,并能得到所需的统计信息。
- (1) 建立统计网站,进行网上统计数据收集和分析。要收集到良好的统计数据,统计网站应加大宣传力度,提高网站知名度和形象,得到被调查者的充分信赖。统计网站还要具有先进的网络技术和统计技术为保证,提高网站的公正性和权威性。
 - (2) 利用计算机抽样技术,确定抽样样本,采用包括电子邮件的多种网络调查方法,广泛收集统计资料。现阶段可以采用的网络调查方法主要有以下两种:①网上普查或抽样调查。以较完整的E-mail地址清单作为样本框,利用计算机的模拟随机功能进行随机抽样,发放E-mail调查表,调查表将自动生成并发往被调查者的电子邮件地址,被调查者通过电子邮件填报并回答表格。对回答的内容进行自动检查和编辑,编辑失败的内容再通过电子邮件原路返回被调查者以求得协调一致。在调查实施中访问者还可以采用多媒体技术,向被调查者展不包括问卷、图像在内的多种测试工具。②网上典型调查。运用NetMeeting或其他交互方式,直接征集与会者,在约定时间举行网上座谈。该方法适合于对某项专题进行深入细致的研究。
 - (3) 运用计算机技术对统计资料进行分析。由于网络统计数据的样本量都比较大,采用计算机的技术势必会取得良好的效果。但是关键的问题是要选择好准确性高、权威性强,并且能用于网上实时统计分析的计算机统计软件,才能提供让人信服的统计分析结果。目前可供选择的统计软件很多,如SAS、SPSS、TSP、EXCEL等,选择这些国外著名公司的统计软件进行统计分析,可以提供良好的具有国际对比性的权威统计信息。

4.2.2 统计信息发布模块

- 统计信息发布模块则是实现及时、高效地通过网络进行统计信息发布的功能。
- (1) 采用国际通行的统计指标,统一统计口径,提供具有国际可比性的统计信息。网络统计的统计口径与国际的一致,将有利于避免网络统计出现混乱,减少数据失真的可能性。同时,统一统计口径也是建立网络统计数据库的基础,是加快统计信息传播和扩大统计信息应用范围的必要条件。
 - (2) 利用网络资源共享的优点,采用先进的大型数据库技术,建立统计信息数据库,及时、高效地存储和发布统计信息。系统数据及生成数据可依据权限在网上发布与传输。并通过在内部数据处理系统与外部数据发布传输之间“一步到位”的信息发布数据库系统来产生出许多不同媒体的多样化的统计出版物,以满足不同信息使用者的需要。
 - (3) 建立专门的统计信息查询网站,为用户提供统计信息实时在线服务。统计信息查询网站能够

根据在线用户的实时需求，及时向用户提供符合要求的服 务，在最大程度上满足用户的需求。同时，通过与用户进行交流，也可以挖掘新的用户需求，促进统计信息查询网站自身的发展。

(4) 开发高效率的信息查询和浏览软件。不言而喻，开发高效的信息查询软件可以为信息使用者方便快捷的从世界上任何一个地方得到最新的和最有用的统计信息。此外，统计信息是具有一定的价值和使用价值的“商品”。这就要求对统计信息进行编码处理和加密，对信息进行保护。与之相对应的便是要开发出一套安全高效的信息浏览软件，通过软件对数据进行快速解码，以达到统计信息安全快速传播的目的。

4.3 网络统计系统开发

随着网络技术的兴起，人们开始将其广泛的应用到管理信息系统的开发中。基于网络技术，原来集中计算任务可以分布连接到网络上的不同计算机运行。软件结构随之开始发生变化，从而出现了新的管理信息系统的软件模式。

4.3.1 系统的模式选择

传统的统计软件系统往往是单机版或基于 C/S 模式。但采用 C/S 模式的数据库系统无论在设计开发还是在应用方面都具有一定的局限性。比如在应用中，操作人员必须学会本系统的操作方法、规程等，不具有普及性、易懂性等。随着 Internet/Intranet 技术的发展和普及，各种各样的信息都可以在 Word Wide We b 上发布，人们之间的信息沟通比以往变得更为高效快捷，Web 技术的应用已成为一种必然。

B/S 模式的数据库体系是利用 Web 服务器和动态服务网页（Active Server Pages）作为数据库操作的中间层，将 C/S 模式的数据库结构与 Web 技术密切结合，从而形成具有三层 Web 结构的 B/S 模式的数据库体系，具体结构如图 2 所示。

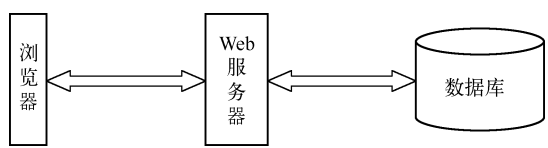


图 2 三层 Web 结构的 B/S 模式的数据库体系

系统的工作原理是：在前端采用 IE、Netscape 等浏览器将用户提交的操作信息向 Web 服务器发出 HTTP 请求，Web 服务器通过 ASP.NET 和一些中间组件访问后台数据库，并将操作结果以 HTML 页面的形式返回给前端浏览器，如图 3 所示。

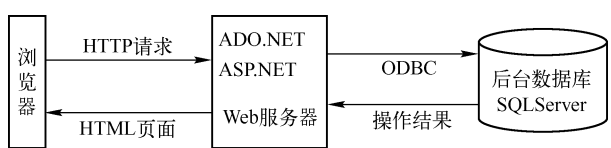


图 3 三层 Web 结构工作原理

4.3.2 基于.NET 的网络统计软件开发技术

信息统计系统（网络版）是基于.NET 的 ISS 系统。微软的.NET 被称为新一代 Internet 计算模型，能够适应新形势对软件技术的新要求，其核心是服务，即 Web S ervice。应用系统开发为三层架构。在客户端采用 IE 浏览器、中间层开发采用 Visusl S tudio.NET。后台数据库采用 Microsoft SQL Server 2000。

1) 客户端开发工具

在客户端使用网络浏览器运行程序 JavaScript 脚本程序。JavaScript 是运行在浏览器上的脚本程序，其功能虽然简单，但交互及时，而客户端并没有过多的处理任务，需要及时与用户交互，所以很适合在客户端运行。

2) 中间层开发工具

中间层的任务是实现系统的主要业务流程，有着繁重的处理任务。所以中间层选用的开发工具是微软为开发 Web 应用提供的 ASP.NET。ASP.NET 是一种全新的 Web 开发技术，在框架的架构基础、页面生成方式、面向对象的编程、代码执行的方式等方面都存在突破性的进展。ASP.NET 突破性的改进是提供了一个真正的面向对象的编程模型。.NET 支持的语言包括 C#, Visual Basic.NET 等。这些语言本身就是面向对象的编程语言，用户可以自定制类也可以随意使用所有的.NET 基类类库。ASP.NET 并不只是一种编程语言，而是一种用于创建交互式网页的框架，是微软公司最新发布的技术，ASP.NET 的选择与使用，即能够有效便捷地开发出本课题所需的各项功能，也可以使我们接触到计算机领域最新的开发技术，为我们今后的发展奠定了坚实的基础。

3) 后台数据库开发

Microsoft SQL Server 2005 是一个功能强大的关系型数据库管理系统。具有完全的 Web 功能，支持扩展标识语言 (XML) 并且拥有一个新的、集成的数据挖掘引擎。由于同为微软公司产品，SQL Server 2005 能够与 ASP.NET 有机地结合。.NET 框架类库提供了大量访问数据库的类，这些类在底层与 SQL Server 2005 有很好的交互功能。

5 结束语

网络是信息技术革命的结果，它使信息高速自由流动，使信息供给变得丰富而公平，它打破了传统的种种交流界限，容纳了各种各样的信息，形成一种新的生活方式，深刻地影响着人类。为适应网络带来的变化，统计必须进一步发展。网络时代，计算机将被广泛应用、网络将在全社会广泛普及，统计软件的开发将达到很高的水平，企业统计手段的现代化成为现实，通过网络可以收集、加工大量的统计信息资料，极大地提高了统计工作的质量和效率。随着统计方法制度改革的深化、信息技术的日益进步和统计组织机构的不断优化，统计工作在社会中的地位 and 作用就更加重要。因此，对网络时代统计工作和其软件开发进行研究是一项极具价值的课题。

网络时代是一个新鲜的时代，网络时代的统计工作如何开展尚属比较新的课题，加上 Internet 的发展一日千里，目前我们还无法预见它真正的完全的潜力，因而本文的撰写难度是比较大的。本文虽然论述了网络时代统计工作面临的机遇和发展，但对比即将出现的新情况，这些机遇和发展就又不复存在了，甚至还会变成统计工作适应时代发展要求的桎梏了。因此，广泛收集、阅读与信息技术、网络技术相关的资料，特别是注意收集最新的相关资料，第一时间全面了解掌握信息、网络的发展状况及特性，并将其统计工作相结合，仍然是今后很长一段时期内需要做的工作。

统计是世界通用语言，国际统计事业是我们共同的事业。中国统计事业面临的挑战非常严峻，改革和发展任务十分艰巨，我们要走的路还很长，需要得到各有关国际组织和各国同行的大力帮助和支持。我们应该加强同有关国际组织和世界各国同行的合作和交流，互相借鉴，努力推动国际统计事业的不断发展，更好地发挥统计在增进各国间的沟通了解，在描述、诠释和推动世界文明进步等方面的重要作用。

参考文献

- [1] 中国网. 统计信息化发展概况. www.china.org.cn, 2003-02-26.
- [2] 陈方明. 信息时代下的统计工作. 山西统计, 2003 (8): 50-51.

- [3] 赵启正. 靠什么决胜信息时代. 北京青年报, 2000-0_5-22.
- [4] 中国互联网络信息中心. 中国互联网络发展状况统计报告. www.cnnic.net.cn.
- [5] 黄燎隆. 试论网络统计及其对传统统计的影响. 广西统计, 2000(2): 10-12.
- [6] 方小梅. 对网络统计系统的若干思考. 华东经济管理, 2001, 15(3): 126-127.
- [7] 袁云峰, 范飞龙. 网络经济时代提高统计信息质量的几点思考. 统计教育, 2002(2):30-32.
- [8] 中国统计信息网. 四川德阳: 建立专业统计数据配送中心的探索. [www.Stats.gov.cn](http://www.stats.gov.cn), 2004-07-22.
- [9] 曹孟军. 制定统计数据质量标准, 强化全面质量管理意识. 湘潭师范学院学报(社会科学版), 2004, 26(1): 36-38.
- [10] The white house. The National Strategy To Secure Cyberspace. www.whitehouse.gov, 2003-02-14.
- [11] Information work. 2004 Global Information Security Survey. www.information-week.com, 2004-03-30.
- [12] 袁津生, 吴砚农. 计算机网络安全基础. 北京: 人民邮电出版社, 2004, 79-81, 261.
- [13] 褚盈. 关注信息安全. 科技成果纵横, 2005(2): 42-43.

基于 CBR 的人防城市人口应急疏散预案 辅助决策研究

左 军, 刘凤荣, 张 涛, 彭祥新

(防空兵指挥学院, 河南 郑州, 450052)

摘 要: 本文在对城市人口应急疏散体系及疏散预案的内容要素调查和研究基础上, 将应急疏散决策中所用的预案, 进行了改进、结构化处理和知识表示, 使之同计算机辅助决策过程相吻合, 便于计算机的存储处理, 并将 CBR 技术应用到预案辅助决策过程中, 建立了预案库部件, 采用归纳法和最近邻域法进行检索匹配, 最后得出最佳的人口应急疏散方案。

关键词: 人防; 人口疏散; 辅助决策; CBR

中图分类号: TP391

文献标识码: A

文章编号: 1006-7043 (2010) xx-xxxx-x

Urban Population Based on CBR's Civil Air Defense Decision on Emergency Evacuation

ZUO Jun, LIU Fengrong, ZHANG Tao, PENG Xiangxin

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: In this paper, on the urban emergency evacuation system and evacuation plan the contents of the elements of investigation and research, based on the emergency evacuation plans used in decision-making, were improved, the structure of processing and knowledge representation, so that decision-making process with computer-aided consistent, easy to deal with the computer's memory, and CBR technology to plan for supporting decision-making process, the establishment of a plan library components, by induction, and nearest neighbor method to retrieve the match, the final population for optimum emergency evacuation plan.

Keywords: population of the evacuation air defense; decision; support; CBR

1 引言

人防城市人口应急疏散指挥决策是一个复杂的系统工程, 在人口疏散过程中, 决策人员必须根据瞬息万变的环境, 快速、准确地做出科学决策。辅助决策是疏散指挥员分析判断情况、确定疏散方针、下定疏散决心和制订疏散行动计划的重要手段。通过辅助决策系统可以实现疏散预测, 评估疏散预案, 选择疏散预案, 优化疏散方案等功能, 并在平时以此完成疏散模拟和训练。

基于案例推理是从以往案例中搜索与当前问题类似的案例, 并选择一个或多个与当前问题最相似和相关的案例, 通过对所选案例的适当调整和改写, 从而获得当前问题求解结果和对这一新案例的存储以备使用的一种推理模式。从认识过程角度, 案例推理是基于记忆来指导问题的一种方法。CBR 推理过程可简述为“回想+修改+学习”, 也就是从以前的案例中检索与当前问题最相似的案例, 修改相似案例的解后用来解决新的问题, 并将这个新经验存储到案例库中, 用于解决以后的新问题。

作者简介: 左 军 (1965—), 男, 副教授, 硕士生导师;
刘凤荣 (1961—), 男, 教授, 硕士生导师;
张 涛 (1963—), 男, 副教授, 硕士生导师;
彭祥新 (1978—), 男, 讲师, 硕士。

2 人防城市人口应急疏散预案辅助决策

2.1 人口疏散预案的预处理

人口疏散预案作为基于案例推理的人口疏散辅助决策支持系统的知识基础，首先应该将其转化为便于计算机识别和处理的形式。这个转化过程，我们把它称为对人口疏散预案的预处理。

为了实现预案制定的“实用性、统一性、灵活性、扩展性”，提出了预案的模块化设计思想。预案模块化即预案体系的模块化，它首先需要对预案进行分类管理和标准化，确保各类预案的数据表达和存储具备一定的通用性。因此，预案模块化包含两个部分的内容：一是预案的分层和分类，二是预案的组件化实现。

由于人口疏散预案涉及国家许多部门、单位，因此，对人口疏散预案进行层次化处理就显得非常重要：一是可以使得各个部门与各个单位之间方便管理和协调；二是也可以使得本系统在进行预案的检索和匹配时的针对性更强，准确性更高；另外，对于分层后的预案进行检索时，检索的范围大大缩小，效率也就有所提高。

根据人口疏散预案的特点，将人口疏散预案大致分为两层：宏观型预案和微观型预案。其中，宏观型预案是高层次的预案，一般为国家、省人防部门制定的预案。一般此类预案涉及的范围广，总体要求明确，涉及内容丰富，但是具体内容不够详细，需要补充。上述层次中的国家级、省、自治区和直辖市的预案即属此类。微观型预案是具体执行疏散任务的详细预案，详细描述了人口疏散的数量、对象，疏散路线，集结地点，疏散方式，安置地域，负责人等，上述层次中的单位、社区级预案即属此类。

为了在进行方案检索时进一步的缩小检索范围，提高检索效率，同时也为了更加方便地管理每一层次的预案，这里还对各层的预案进行了分类。

经过这样的分层，分类处理后，预案变得容易计算机实现、管理和维护。

为了提高预案内容的结构化，便于计算机对预案本身进行处理，对预案的内容进行标准化，也就是将整个人口疏散预案分解成不同类型的模块，并针对上述提到的每一种类型的预案，对其每一模块的内容进行规范。

通过研究分析基本预案模块主要包括：

- 基本情况：主要包括预案编号，预案名称，城市的人口组成，敌情判断，任务决心，城市地位，人防设施等要素。
- 疏散原则：是整个疏散行动的总体要求，是其他各种保障预案的指导方针。
- 疏散时机：根据敌情判断和国际国内形势选择早期疏散、临战疏散还是应急疏散。
- 疏散类型：依据城市或社区、单位的地理位置来决定采取散射型、多方向或单一方向。
- 疏散对象：疏散对象是疏散的实质问题，是有选择的，不是随意的。
- 组织指挥：高技术局部战争条件下组织指挥难度大，疏散要按照“散得开，走得快，藏得好，责任明”的总要求，原则上按行政区划分而统一组织实施。
- 协作关系：疏散行动涉及单位部门多，必须明确各单位部门的协作任务和协作要求。
- 疏散保障：人口疏散是个非常复杂的行动离不开各种辅助保障行动，主要包括生活必需品保障、医疗卫生保障、通信保障等。
- 疏散安置：如何把疏散来的人口安置好也是疏散行动成败的关键，根据我国城市发展的特点应以城市近郊安置为主。
- 疏散路线：最佳疏散路线并不等于最短路线而是最优路线，因为战时或自然灾害发生后，各种情况瞬息万变，所以如何选择确定最佳疏散路线也是重点考虑的问题。
- 其他附件：各种图表等。

以上基本情况，疏散原则，疏散时机，疏散对象，组织指挥，疏散保障，疏散安置，疏散路线是人口疏散预案的通性共有的基本模块，各城市或单位社区可根据自己的情况灵活添加其他模块。

2.2 人口疏散预案的检索匹配

当一个新问题出现时，系统根据索引，从人防人口疏散预案库中检索出预案（集），并对其进行修改，使之达到当前疏散决策问题的要求，最后将新问题与求解策略当做一个新的案例存入案例库中，以备后用。

显然，如何在 CBR 中高速、有效地完成疏散预案的检索是十分重要的，对问题求解的性能有直接的影响。特别是当疏散案例库足够大时，检索的目标很难快速达到。现在的 CBR 系统的检索策略主要有以下几种：

（1）最近邻域法：这种方法是 CBR 系统中一个最基本的算法，由于它的逻辑和计算都非常简单，因此在很多 CBR 系统中都得到了应用。

通过最近邻域法，每个疏散预案中的每个属性都被赋予一个值，这些值用来反应各个属性间的相似程度。由于预案的属性将被作为案例间相互比较时的指标，为了计算各个疏散预案间的相似度，疏散预案中的所有属性值都会被相加。如果输入预案的属性值之和与人防疏散预案库中某个预案的属性值之和相近，则可以将预案库中该案例的解决方案应用到当前疏散问题。

在进行案例匹配时，案例中各个属性的重要程度是有差异的。为了体现这种差异，需要给各个属性确定权重，根据属性的重要性的不同，它们被设置的权值也不相同，权值越大，则该属性越重要。这样一来，在进行相似度计算时，重要的属性对于相似度的影响会更大。为了避免由于属性选择不准或者权值设置较小而造成比较好的解决方案被漏选，可以采取 k-最近邻域法。这种方法为用户提供 k 个与所求问题最相近的疏散预案。

（2）归纳法：提取疏散预案之间特征上的差异，并根据这些特征将这些案例组成一个类似判断网络的层次结构。检索是采用决策树搜索。适用于疏散预案特征相互独立或推理结果只是预案中的某一特征的情况，也适用于事例的内容易于分层组织，特征信息的选取不能太具体也不能太抽象。

（3）知识导引法：采用一套规则进行搜索控制，根据已知知识来决定预案的哪些特征在进行案例检索时是最重要的，并根据这些特征来组织检索，使预案的组织与检索具有一定动态性。对于大型系统的建设而言，建立完备的基于知识的检索很困难。适用于事例中包含有丰富的经验知识，内容不易分层组织。

在使用最近邻域法时，最重要的就是确定相似度函数，以确定疏散预案库中的预案和当前问题之间的相似度。在系统中，将采用如下函数来描述预案间的相似度：

$$\text{sim}(A, B) = \frac{1}{1 + d(A, B)}$$
$$d(A, B) = \sqrt[r]{\sum_{i=1}^n |A_i - B_i|^r \times \omega_i}$$

式中， A_i 和 B_i 分别为预案 A 和 B 的第 i 个属性的属性值； ω_i 为第 i 个属性的权重；r 为指示系数，当 $r=1$ 时即为海明距离；当 $r=2$ 时即为欧几里得距离，通常取 $r=2$ 。

系统将根据以上公式，检索出人防人口疏散预案库中相似的预案，并给出相似度最大的 3 个预案，经排序后，供用户选择、决策。用户可以逐个浏览候选疏散预案，对于基本满意的预案，还可以进行手动修改。

3 结束语

本文利用基于案例推理技术的特点，研究了 CBR 在人防城市人口应急疏散预案辅助决策中的应

用，对人防疏散信息化建设有一定的借鉴意义。

参考文献

[1] 于伟. 城市防空信息系统建设[J]. 郑州防空兵指挥学院学报, 2006.

[2] 岳超源. 决策理论与方法[M]. 北京: 科学出版社, 2003.

[3] 高洪深. 决策支持系统 (DSS) 理论 • 方法 • 案例[M]. 北京: 清华大学出版社, 1996.

[4] 王永军等. 基于预案库的智能决策支持系统的研究——CBR 技术的应用[J]. 控制与决策, 2003.

[5] 柳少军. 智能决策支持理论与智能决策知识系统研究[D]. 西安交通大学博士论文, 2005.

人防城市人口应急疏散路径选择模型研究

左 军, 刘凤荣, 黄欢欢, 王月蓉

(防空兵指挥学院, 河南 郑州, 450052)

摘 要: 人民防空是国防的重要组成部分, 具有强大的威慑作用, 而疏散对于保存战争潜力具有重大的战略意义。由于敌方空袭兵器的精确度与突防能力的不断增强, 未来人防城市人口应急疏散应立足于紧急疏散和近郊疏散。基于这种情况, 本文分析预警和非可预警下的应急疏散的交通流的特点, 并给出了交通调度策略。对已有的道路交通通行能力模型进行了优化, 并给出了疏散路径选择的具体算法。

关键词: 人民防空; 应急疏散; 路径选择; 模型

中图分类号: TP391

文献标识码: A

文章编号: 1006-7043 (2010) xx-xxxx-x

Urban Air Defense Emergency Evacuation Route Choice Model

ZUO Jun, LIU Fengrong, HUANG Huanhuan, WANG Yuerong

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: Civil air defense is an important component of national defense, a powerful deterrent effect, but the potential evacuation of the war for the preservation of great strategic significance. Since the accuracy of enemy air attack weapons capabilities and growing penetration in the future air defense emergency evacuation of the urban population should be based on emergency evacuation and evacuation suburbs. Based on this, this paper can be early warning early warning and non-emergency evacuation under the traffic flow characteristics, and gives the transportation scheduling. On the existing model of road traffic capacity has been optimized, and the evacuation route choice given the specific algorithm.

Keywords: civil air defense; emergency evacuation; route choice; model

1 高技术空袭背景下的人防城市人口应急疏散

人民防空是国防的重要组成部分, 具有强大的威慑作用, 而疏散对于保存战争潜力具有重大的战略意义。科索沃战争和伊拉克战争等几场高技术局部战争已凸显高精度空袭、大纵深作战、全方位制敌的作战理念。这充分说明了未来战争初期对城市居民和防空力量最大的威胁将来自空中。美国认为, 核时代威胁的可靠性不仅取决于国家的战略进攻能力, 而且取决于保存自己的能力。那么如何在高技术空袭背景下有效地开展人防城市人口疏散工作, 就成为未来实现国家防卫战略的重要手段。如果说城市道路网络和各类人防防护工程是组织城市人口疏散的基础硬件设施, 那么交通调度手段与策略则是保证人防城市人口应急疏散高效进行的重要软件保证。

2 人防应急疏散交通流的特点

2.1 空袭可预警下的人口疏散交通流特性

对于人防城市人口应急疏散预案来讲, 可以确定受空袭影响区域的疏散任务和疏散方式。在这种

作者简介: 左 军 (1965—), 男, 副教授, 硕士生导师;
刘凤荣 (1961—), 男, 教授, 硕士生导师;
黄欢欢 (1973—), 女, 助教, 学士;
王月蓉 (1973—), 女, 助教, 学士。

情况下可以分析出交通流的特点：疏散居民总数确定即交通需求时固定的，可以根据疏散预案进行交通管制，可以采取分区域分时段的交通疏散策略，疏散路径基本能够满足疏散交通流的需求，发生事故堵塞的概率较小。根据各分区域居民数量和疏散道路实际情况，合理制定疏散时序。人防指挥部门及交管部门通过各种媒体发布疏散目的地、疏散路径及实时路况信息，疏散居民在较为稳定的心理境况下自由选择疏散路径。

2.2 空袭可预警下的人口疏散交通策略

可预警情况下，当市级人防指挥接到空袭警报后，根据军队提供的空情，立即转入防空袭应急准备阶段，并根据拟疏散地区的疏散预案有序地开展工作。疏散方式为：步行+小汽车+公交运输；疏散道路为：城市快速路+主干路+次干路。步行疏散的群众可以就近疏散至人防工程内。对于敏感单位多采用交通运输的方式疏散至指定区域。因为疏散交通需求是固定的，而且有时间进行交通管理指挥、交通诱导、道路交叉口限制、应急救援力量的部署协调，所以疏散路径上的交通流量处于不饱和状态或次饱和状态。在具体实施交通调度策略中，在疏散路径方向的城市主干路和绝大部分次干路十字交叉口上实行全绿交信号，外车道可设立专用道供其他社会车辆驶出，人防疏散指挥车辆、应急救援保障车辆的出入。对于其他疏散路径方向上的次干路、连接道、支路，人防疏散指挥部门协同交管部门限制其分布密度，减少车辆出入对疏散交通的滞后影响。

2.3 空袭非可预警下的人口疏散交通流特性

非可预警空袭事件具有突发性、高危害性，而且随着敌空袭兵器突防能力的增强，紧急疏散，近郊疏散将成为未来人防城市人口应急疏散最主要的方式^[2]。因此，对城市人防人口应急疏散指挥部门和交通管理部门的压力非常大。

其交通疏散交通流具有以下特征：

- （1）敌方空袭兵器的杀伤力强，需要疏散和安置居民较多，短小时内交通需求突然较大。
- （2）敌方空袭兵器下一步的打击目标为确定，城市其他地区范围内的居民正处于或者即将处于危险中，而且还可能伴有其他突发事件发生。
- （3）疏散车流始终处于跟车状态，疏散道路上车流基本为饱和流或者过饱和流。
- （4）驾驶员生命和财产受到危害，想急于离开危险区易产生恐慌和急躁心理。
- （5）根据人防应急疏散预案的要求部署，疏散起止点之间的路径固定。
- （6）对疏散道路的通畅性要求较高。

2.4 空袭非可预警下的人口疏散交通调度策略

合理配置各疏散路径的车流流量，使得短时较大的交通压力得到均分，避免了居民无序流动所带来的交通拥堵。另外，固定的疏散路径便于人防疏散指挥和交管部门的交通组织措施实施，包括指挥人员、警员的布设，路障设置和可变信息板等信息发布设备的设置，便于疏散车流的疏导，能够及时地将实时路况信息发布给疏散人员和车辆，缓解了驾驶员恐惧和急躁心理，减少了次生灾害事件。在紧急疏散过程中，疏散路网中去往人民防空工程和安全区域的疏散车辆，在空间和时间上具有优先路权。国内外研究表明，在城市道路交叉口车流进出造成的延误占车辆总延误的 60%。因此，在人防应急疏散预案的指导下，主要交通疏散路径方向上应采用战时交通管制，将一些交叉口或连接路封闭，同时还可以提高疏散道路方向的绿灯时间。

3 城市人防人口应急疏散路径选择模型研究

在疏散预案的制定过程中，疏散道路主要从城市快速路和交通性主干路中选择，每个疏散方向会

选择一条或者几条道路条件较好、ITS（智能交通系统）设施较为完善的路径，作为主要的疏散路径提供给疏散车辆选择，同时疏散路径也要回避大型商业区、铁路枢纽等吸引大量人流的地区。对于步行疏散的群众应及时采取广播、手机通信、指示路牌等指挥调度手段就近疏散。因为这有利于人防疏散运送保障的指挥管理，交通管理部门警力部署、路障设置等各方面交通组织措施的实施，从而可以缓解紧急情况下车辆人群的混乱状态。

关于最大流的概念：

若有向图 $G=(V,E)$ 满足下列条件：有且仅有一个顶点 S ，它的入度为零，即 $d^-(S)=0$ ，这个顶点 S 便称为源点，或称为发点；有且仅有一个顶点 T ，它的出度为零，即 $d^+(T)=0$ ，这个顶点 T 便称为汇点，或称为收点；每一条弧都有非负数，称为该边的容量。边 (v_i,v_j) 的容量用 c_{ij} 表示，则称为网络流图，记为 $G=(V,E,C)$ 。对于网络流图 G ，每一条弧 (i,j) 都给定一个非负数 f_{ij} ，这一组数满足：所有的弧的流量 f_{ij} 不大于弧的容量 c_{ij} ；对所有的中间点，流入的流量和等于流出的流量和；发点流出的总流量 F 等于流进收点的总流量 F ，称为可行流。网络 G 中流值最大的流 f^* 称为 G 的最大流。

3.1 人防应急疏散网络道路的实际通行能力

人防城市人口应急疏散是以人防应急疏散预案为指导的，虽然疏散各要素在预案中都已明确，但现实中的疏散交通流不可避免地包含各类型汽车、疏散用大型公交车、徒步者。本文借鉴北京工业大学交通研究中心研究成果的基础上进行改进，把除公交车外其他类型的车换算成标准车，并设出标准车对通性能力的修正系数^[6]。

$$C_p = C * N * f_{wide} * f_{bus} * f_{car} * f_{driver} * f_{fault} * f_{people} \tag{1}$$

式中 C_p ——单向车道可能通行能力；

C ——单向车道的车道数；

f_{wide} ——车道宽度和侧向净宽对通行能力的修正系数；

f_{bus} ——公交车对通行能力的修正系数；

f_{car} ——其他车型的車輛换算成的标准车对通行能力的修正系数；

f_{driver} ——应急状况下驾驶员条件对通行能力的修正系数；

f_{fault} ——故障车辆对通行能力的修正系数；

f_{people} ——行人与自行车干扰对通行能力的修正系数。

对疏散道路承载能力的计算，可利用图论中的网络极大流理论，将疏散道路网络中的交叉口和疏散路径抽象成为由节点和弧组成的有向图，其中双向道路对应两条同源同汇且方向相反的弧，同理单向道路对应一条指向疏散方向的弧。通过城市交管部门的道路监控数据，可以确定每条疏散路径的交通容量。为了计算方便可将节点的交通容量根据专家意见折减为弧交通容量。疏散网络的承载能力即带有弧容量限制的最大流求解具体方法和过程，相关文献中已经有明确的答案。

3.2 人防城市人口应急疏散路径选择模型

人防城市人口应急疏散路径的选择模型，是以疏散时间最短为目标，将道路通行能力、交叉口个数、车辆在交叉口处的转向折减，ITS 设施的设置等主要因素，纳入其中。单个起止点之间的最优疏散路径为模型计算的对应于“当量疏散时间”的疏散时间最短的路径。

$$T = \min(T_i) = \min(K_{i1}K_{i2}T(L_i)) \qquad i = 1, 2, \dots, n$$

式中 T_i ——第 i 条候选疏散路径的疏散时间；

$T(L_i)$ ——第 i 条候选疏散路径，仅考虑道路通行能力的疏散时间；

K_{i1} ——第 i 条候选疏散路径，交叉口对疏散时间的惩罚系数值，这里取 1；

K_{i2} ——第 i 条候选疏散路径，ITS 设施对疏散时间的惩罚系数值，取值见表 1。

设置良好	设置一般	设置较差
疏散时间影响系数值1	1.1 1.3	

4 总结

本文通过对未来人防城市人口应急疏散的分析，把疏散情况分为可预警下的临战疏散和非可预警下的紧急疏散。以非可预警下的紧急疏散为主要研究对象，分析了其疏散交通流的特性和交通调度指挥策略。最后给出了疏散时间最短路径的算法。

参考文献

[1] 文国玮. 城市交通与道路系统规划[M]. 北京：清华大学出版社，2007.

[2] 宫建. 奥运应急交通疏散路径选择模型研究[D]. 北京：北京工业大学，2007.

[3] 张超. 城市地理信息系统——原理应用与项目管理[M]. 北京：科学出版社，2008.

[4] 石玉峰，彭其渊. 带有区间数弧容量上限的网络优化[J]. 交通运输工程与信息学报，2005，2 (3):34-38.

[5] 谢凡荣. 求解网络最大流问题的一个算法[J]. 运筹与管理，2004，13(4):37-40.

[6] 应急交通疏散预案仿真研究中期报告. 北京工业大学交通研究中心，2006.

基于 Pastry 算法的物联网信息发现服务

李占波¹, 刘冬冬²

(1.郑州大学信息工程学院, 河南 郑州, 450002; 2.郑州大学信息工程学院, 河南 郑州, 450002)

摘要: 近年来, 物联网的应用越来越广泛。信息发现服务是物联网中一种重要的服务机制, 它主要用来对物联网中的物品进行定位查找, 进而使用户可以对物品的详细信息进行访问。物联网信息发现服务的核心组件是对象名解析系统(ONS)及 EPC 信息服务器。然而现在的 ONS 系统存在安全性差、查找效率低等缺点。随着物联网的快速发展, 物联网的地址空间急剧增长, 因此信息发现的安全性和查找效率都亟待提高。

运用 DHT 网络的 Pastry 算法, 为物联网提出一种新的信息查询机制, 通过仿真实验, 证明该机制在网络负载均衡、查询效率方面都有了很大的改善。

关键词: 物联网; 对象名解析服务; DHT; Pastry

Discovery Service Based on Pastry for Internet of Things

LI Zhanbo¹, LIU Dongdong²

(School of Information Engineering, Zhengzhou University, Zhengzhou 450002, Henan China)

Abstract: In recent years, more and more widely networked applications. Discovery service is an important networking service, it is mainly used in networking to object orientation, which, in turn, is looking for a user to the details of the goods. The core components of discovery service for the Internet of things are the Object Naming Service(ONS) and EPC Information Server. Weak security and delay time exist in the ONS system. With the rapid development of the internet of things, the rapid growth of address space, so the security of information discovery and lookup efficiency need to be improved.

Using Pastry in DHT Network to advance a new information query mechanism. By simulation experiments to prove that the mechanisms has improved considerably in load balancing and query efficiency.

Keywords: Internet of things ; ONS; DHT ; pastry

EPC (Electronic Product Code) 系统是在计算机互联网和射频技术 RFID (Radio Frequency Identification) 的基础上, 利用全球统一标识系统编码技术给每一个实体对象一个唯一的代码, 构造了一个实现全球物品信息实时共享的实物互联网 “Internet of Things” (简称物联网)。EPCglobal 是一个非营利性组织, 其职能是为 RFID 的商业应用建立合适的标准体系。作为 EPC 物联网组成技术的重要一环, EPC 信息发现服务包括对象命名服务 ONS (Object Naming Service) 及配套服务。其作用就是通过电子产品码, 获取 EPC 数据访问通道信息。随着物联网的快速发展, 物品的地址空间也会越来越大, 运用现有的 ONS 系统进行信息发现服务会产生较大的时延, 查询效率比较低。因此, 在 DHT 网络基础上提出一种高效安全的查询机制, 这对加快物联网的发展有很重要的意义。

1 ONS

在 EPCglobal 网络中, ONS 主要是运用 EPC 码对 EPC 信息服务器 (EPCIS) 提供一种查找服务。ONS 中并不包含 EPC 码相关的物品的详细信息, 它只包含存储这些详细信息的服务器的 IP 地址。就像 DNS 把域名地址转换成 IP 地址一样, ONS 负责把 EPC 码转换成 EPCIS 的 URL 地址。鉴于

ONS 与 DNS 的相似性，目前的 ONS 就是基于 DNS 架构的。

1.1 ONS 基本结构

ONS 系统是一个类似于 DNS 的分布式的层次结构，具有自身的查询机制。整个系统主要由映射信息、根 ONS（ROOT ONS）、ONS 服务器、ONS 本地缓存（ONS Cache）、本地 ONS 解算器（Local ONS RESOLVER）五个部分组成。

1.2 ONS 系统存在的问题

根 ONS 服务器处于 ONS 层次结构中的最高层，拥有命名空间中的最高层域名。基本上所有的 ONS 查询都从根 ONS 服务器开始，所以根 ONS 服务器性能要求很高。物联网是一个物物相连的网络，可想而知，EPC 码的地址空间要比 IP 地址空间大得多。所以，根 ONS 会同时接收到大量的查询请求，这样就会对根 ONS 造成很大的查询压力，查询时延随之增大，查询效率降低，最终根 ONS 会成为物联网中信息发现服务的瓶颈问题。因此为了满足不断发展的物联网，必须找出一种新的高效率的查询方法。另外，ONS 系统的安全性不够，用户不加限制的访问可能会造成个人或公司的隐私泄露。

根据上面对 ONS 不足的分析，对新的信息发现服务提出以下两点要求：

- （1）对用户权限进行设置和管理。
- （2）解决可能会出现的问题，缩短访问时延，提高查询效率。

2 对等网络搜索技术

对等计算的核心思想是所有参与系统的节点处于完全对等的地位，没有客户端和服务端之分，也可以说，每个节点既是客户端又是服务端；既向别人提供服务，也从别人那里享受服务。

DHT 搜索网络中，将内容索引抽象为 $\langle K, V \rangle$ 对。K 是内容关键字的 Hash 摘要（ $K = \text{Hash}(\text{Key})$ ），V 是存放内容的实际位置，例如节点的 IP 地址等。所有的 $\langle K, V \rangle$ 对组成一张大的 Hash 表，因此该表存储了所有内容的信息。每个节点都随机生成一个标识（ID），把 Hash 表分割成许多小块，按照特定规则（即 K 和节点 ID 之间的映射关系，这些规则包括 Pastry、Chord、CAN 等）分布到网络中，节点按照这个规则在应用层上形成一个结构化的重叠网络。给定查询内容的 K 值，可以根据 K 和节点 ID 之间的映射关系在重叠网络上找到相应的 V 值，从而获得存储文件的节点 IP 地址。图 1 给出了 DHT 网络中索引发布和内容定位的过程。

2.1 Pastry

Pastry 在 2010 年由位于英国剑桥的微软研究院和莱斯（Rice）大学提出。它是 DHT 的一个变种。在 Pastry 网络中，节点 ID 分布采用环形结构，内容定位如图 1 所示，在此网络中，内容的存储发布过程如下：

- （1）Hash 节点 IP 地址得到 M 位的节点 ID，表示为 NID。
- （2）Hash 内容关键字得到 M 位的 K，表示为 KID。（NID 和 KID 是以 2^b 为基的数，其中 b 是一个配置参数，一般为 4）。
- （3）节点按 ID 从小到大顺序排列在一个逻辑环上。
- （4） $\langle K, V \rangle$ 存储在 NID 与 KID 数值最接近的点上。

Pastry 中的每个节点拥有一个路由表 R（Routing table），一个邻居节点集 M（Neighborhood Set）和一个叶子节点集 L（Leaf Set），它们一起构成了节点的状态表。当一个 $K=D$ 的查询消息到达节点 A 时：

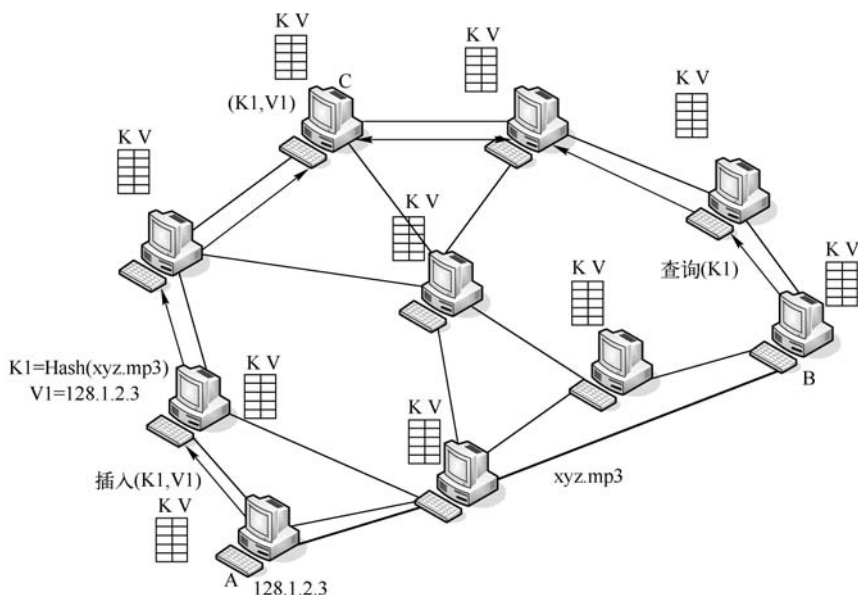


图 1 DHT 索引发布与内容定位的过程

(1) 节点 A 首先查看 D 是否在当前节点的叶子节点集中, 如果是, 则查询消息被转发到目的节点, 也就是叶子节点集中节点 ID 与 D 数值最接近的那个点 (有可能就是当前节点), 否则进行下一步。

(2) 在路由表中查找与 D 具有更长前缀的表项, 如果不为空, 则将查询消息转发。例如, 路由 D=0629 的查询消息 5324 → 0748 → 0605 → 0620 → 0629。

(3) 如果不存在这样的节点, 当前节点会从其维护的所有邻居节点集合中选择一个距离该键值最近的节点作为转发目标。

Pastry 引入了叶子节点和邻居节点集合的概念。在应用层能及时准确地获得这两个集合的节点信息时, 可以大大加快路由查找的速度, 同时降低因路由引起的网络传输开销。Pastry 的每一步路由和上一步相比都更靠近目标, 因此这个过程是收敛的。如果路由表不为空, 每步路由至少能够增加一个前缀匹配数位, 因此在路由表始终有效时, 路由的步数至多为 $\log_b N$ ($B=2^b, b=4$)。在地址空间日益庞大的物联网中, 运用 Pastry 进行信息查找定位, 能够很好地减小信息查询的时延。这种基于 Pastry 算法的信息查询运用了对等网络的思想, 保证网络上信息存储平衡, 避免各个节点信息访问量不均衡, 因此可以避免信息查询的瓶颈问题, 提高网络查询效率。

2.2 运用 Pastry 对物联网中的信息发现服务进行改进

在文章第二部分, 我们对现有的物联网信息发现服务的核心组件 ONS 系统进行了分析, 指出了它的缺点。在本文中, 我们运用 Pastry 实现物联网中资源的快速定位查找。同时为了提高访问的安全性, 我们对每个节点增加用户管理组件, 它主要实现访问者身份确定及权限分配。

在 Pastry 搜索网络中, 每件物品都会运用物品的 EPC 码经过哈希算法得到一个 K 值, 而 V 值就是存放物品详细信息的 IP 地址。得到 $\langle K, V \rangle$ 对之后, 根据 Pastry 规则将其分布存储于搜索网络中。当某个 EPC 码被查询时, Pastry 网络就会根据上文所述的查询过程, 对产品的详细信息进行定位查找。因为要保证查询的安全性, 我们对此查询过程稍加修改, 如图 2 所示。

在图 2 所示的过程中, 对产品 A 的信息注册, 以及客户节点对 A 的 IP 地址进行定位查找的过程都不变, 都是依据 Pastry 算法, 但是当客户节点获得产品 A 的 IP 地址后, 客户节点要依据查询到的 IP 地址对存储 A 详细信息的节点 6 进行产品信息查询, 此时, 客户节点不但要向节点 6 发送 IP 地址, 而且要向节点 6 发送它的身份信息。节点 6 中的用户管理模块会对客户的身份进行验证, 结果如

下：如果用户合法，则根据用户的权限（普通用户、公司内部员工、公司管理者等）显示不同的内容；否则不返回错误信息。

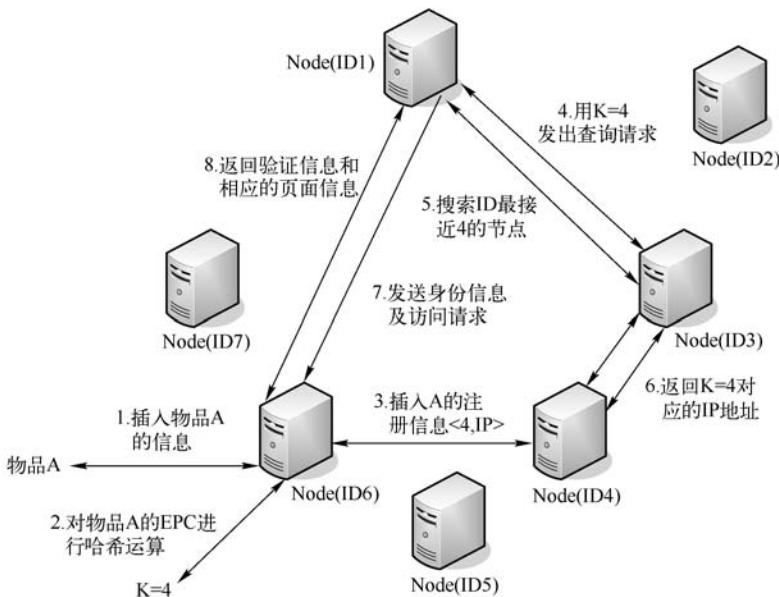


图 2 Pastry 发现机制的信息发现过程

3 实验与分析

OMNET++是开源的基于组件的模块化的开放网络仿真平台，它可以对 Pastry 进行很好的仿真。本文分别从网络延时、路由跳数两个方面，对 ONS 查询算法和 Pastry 路由算法进行了性能比较。其中：

网络延时计算公式为：

$$\text{delay} = \sum_{i=1}^n (\text{receive\_time} - \text{send\_time}) / n$$

其中，receive_time 表示接收包时间；send_time 表示发送包的时间；n 表示发包的数量，delay 的单位是秒（s）。

图 3 验证了 Pastry 算法的逻辑路由跳数是 $O(\log_B N)$ ， $B=2^b$ 。随着网络规模的增大，Pastry 搜索算法的路由跳数并没有急剧地增加，而是保持在一个相对稳定的数值区间。从图 4 可以看出 Pastry 算法的网络时延在 0~0.025 之间，这是因为在对等网络中节点之间可以直接进行数据交换，而不是通过中心节点等其他节点，所以传输时延比较小。从图 5 可知，旧的 ONS 发现服务的网络平均时延是 1.5s，是 Pastry 的 50 多倍，这是由于 ONS 系统是一种类似 C/S 的模型，根 ONS 相当于中心节点，当查询请求数量较大时，由于 ONS 的瓶颈问题，消息需要排队等候，所以网络时延大。

图 3、图 4 和图 5 所示为进行仿真实验后得到的图表数据。

4 结论

本文在分析 ONS 系统的基础上指出了现行的物联网信息发现服务的不足之处，然后用 Pastry 路由算法替代了现有的 ONS 查找算法，同时在 Pastry 查找过程中加入安全机制，在提高查询效率的同时，增强了信息查询的安全性。随后我们用 OMNET++进行了仿真实验，从网络延时、路由跳数两个

方面，对 ONS 查询算法和 Pastry 路由算法进行了性能比较。实验结果显示运用 Pastry 进行资源定位查找与 ONS 系统相比，查询时延大大降低，有效地提高了物联网的查询效率，这对物联网的快速发展有很重要的现实意义。但是改进后的信息发现服务仍然存在不足之处，Pastry 的算法复杂度较大，这也会带来查询时延，影响查询效率，因此，今后我的重点工作就是对 Pastry 算法进行改进，降低算法复杂度，从而使 Pastry 能够提供更好的查询服务。

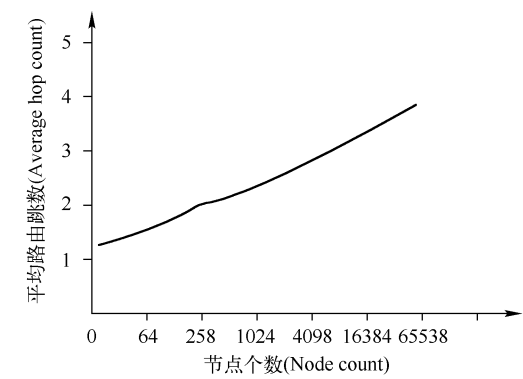


图 3 Pastry 搜索的路由跳数

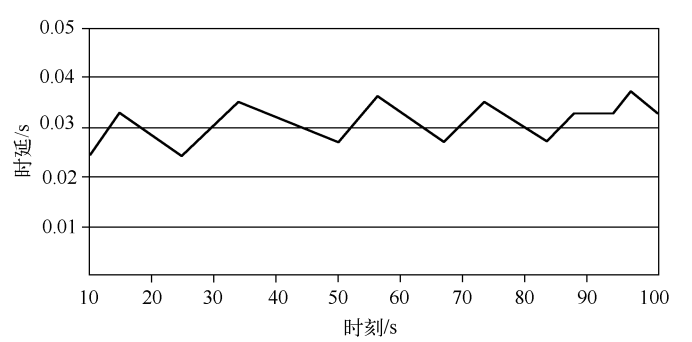


图 4 Pastry 的搜索时延图

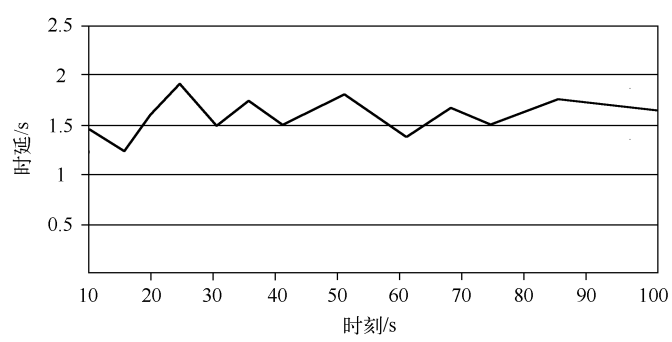


图 5 旧的 ONS 的搜索时延图

参考文献

[1] Dijiang Huang, Mayank Verma, Archana Ramachandran. A Distributed ePedigree Architecture [R]. Proceedings of the 2005 11th International Conference on Parallel and Distributed Systems (ICPADS'05)0-7695-2281-5/05.

[2] EPC-Global: <http://www.epcglobalinc.org/home>

[3] Benjamin Fabian and Oliver Gunther. Distributed ONS and its Impact on Privacy [R]. IEEE International Conference on Communications (IEEE ICC 2007).

[4] B.Fabian, O.Gunther, and S.Spiekermann, "Security Analysis of the Object Name Service," in Proceedings of the 1st International Workshop On Security, Privacy and Trust in Pervasive and Ubiquitous Computing (SecPerU2005), with IEEE ICPS 2005, Santorini, 2005.

[5] Fangli LIU, Huansheng NING and Zhiqiang XU. RFID-based EPC System and Information Services in Intelligent Transportation System [R]. 2006 6th International Conference on ITS Telecommunications Proceedings.

[6] Ching-Wen Chen, Phui-Si Gan and Chao-Hsiang Yang. A Service Discovery Mechanism with Load Balance Issue in Decentralized Peer-to-Peer Network [R]. Proceedings of the 2005 11th International Conference on Parallel and Distributed Systems (ICPADS'05) 0-7695-2281-5/05.

- [7] Benjamin Fabian. Implementing Secure P2P-ONS [R]. IEEE Communications Society subject matter experts for publication in the IEEE ICC 2009 proceedings.
- [8] Loïc Schmidt, Nathalie Mitton, David Simplot-Ryl .Towards Unified Tag Data Translation for the Internet of Things [R]. Wireless VITAE'09.
- [9] EPC Global Inc., EPCglobal Object Name Service(1.0.1)[S].
- [10] Garcia-Alfaro , J.Barbeau,M., Kranakis, E.Analysis of Threats to the Security of EPC Networks.[R]. 6th Annual Conference on Communication Networks and Services Research (CNSR'08), IEEE, Canada, May 2008.

看我国 SNS 社交网站现状与趋势

牛 星

(河南大学 计算机与信息工程学院, 河南 开封, 475004)

摘 要: 随着 Web2.0 技术的不断发展, SNS 社交网站成为了继 BBS、博客之后当今最为流行的社交网站模式。SNS 网站更是在 2009 年飞速发展, 已逐渐成为大家平时联系交流的互联网互动应用的集成平台。本文从其定义, 发展情况, 比较流行的几个社交网站, 主要用户群体, 未来发展的关键及趋势等几个方面来对 SNS 进行了介绍与分析。

关键词: Web2.0; SNS 网站;

中图法分类号: TP391

文献标识码: A

The Status and Trends of Social Networking Site in China

NIU Xing

(Computer and Information Engineering College of Henan University, Henan University, Kaifeng 475004, Henan China)

Abstract: With the development of Web2.0 technology, Social Networking Site has become the most prevalent social networking site model which following the bbs and blog. In 2009, Social Networking Site develop rapidly, it has gradually become the internet interactive application platform for us connection and communication. And this article introduce and analyze the Social Networking Site from the definition, development, several popular social networking site, major user groups, the key to the future development and trends.

Keywords: Web2.0; social networking site

1 引言

“今天, 你偷菜了吗?” 如果你不知道这句话, 只能说你 OUT 了。这是 “开心农场” 的游戏术语, 现在已经成为亿万网民见面的问候语。现在不管上班聚会还是外出都要想着 “偷菜”, 更有一些网民半夜起来偷菜屡见不鲜。随着开心网的开心农场的兴起, 各大社交网站纷纷效仿, 偷菜、好友买卖、抢车位、养宠物、入住公寓等更是成为一种时尚, 受到大众网民的钟爱。伴随着 SNS 社交网站对外开放 API, 国内的 SNS 社交网站的小游戏层出不穷, 每天都会有一定数量的增长, 不断创新, 不断让用户体验新感觉^[1]。

SNS 游戏为 SNS 社交网站吸引了大量的用户, 同时它也是伴随着 SNS 社交网站的发展而发展起来的。游戏只是 SNS 社交网站的一部分功能, 它的主要功能在于 “认识朋友并且与朋友保持联系”。你可以通过它来抒写心情, 发表日志, 发视频, 发照片, 送礼品, 聊天及现在非常的热门的转帖等。而伴随 2009 年何洁和许茹芸先后现身开心网, SNS 网站已成为名人拓展宣传渠道、加深与粉丝互动沟通的重要平台。之后如演员成龙, 李连杰, 小说家苍月, 各种娱乐明星, 体育明星, 商界及学术界各领域的精英名士纷纷加入, 利用社交网站与粉丝建立联系。中国的 SNS 社交网站正在从交友+互动游戏的单一模式, 向真正的、中国式的网络社交模式转变。强调 SNS 的社会化, 提供人与人、人与机构、人与社会的互动^[2]。

2 定义

SNS, 全称 Social Networking Services, 即社会性网络服务, 专指旨在帮助人们建立社会性网络的

• 298 •

互联网应用服务。1967 年，哈佛大学的心理学教授 StanleyMilgram 创立了六度分割理论，简单地说：“你和任何一个陌生人之间所间隔的人不会超过六个，也就是说，最多通过六个人你就能认识任何一个陌生人。”

SNS 社交网站，全称 Social Network Site，就是依据六度理论建立的网站，帮你运营朋友圈的朋友。如现在的 Facebook 等在基于这个理论的网站，它会把朋友的朋友列出来供你来认识，这样使每个人都可以认识更多新的朋友，社交圈都不断增大，最后成为一个大型网络。但通过朋友来认识新的朋友只是社交拓展的一种方式，而并非社交拓展的全部。现在所谓的 SNS，含义也已经远不止“熟人的熟人”这个层面，而是扩展到更广阔的范畴，比如根据相同话题进行凝聚（如贴吧）、根据学习经历进行凝聚、根据出游的地点相对凝聚。

3 现状

经过 1998—2001 年的初级发展，2002—2204 年的市场培育，2005—2009 年快速成长，尤其是 2009 年 SNS 社交网络的飞速发展，如今国内的 SNS 网站，个人区域，关系趋于已经发展到了一个较为成熟的阶段，且模式大同小异，于是，服务区域的功能差别成为了吸引用户，黏住用户的主要力量，也成为了 SNS 网站竞争的主要方向，如图 1 所示。而对于国外较国内的 SNS 网站产生的早，各国情况，文化也不一样，其发展情况也不一样。当今对于国外最为出名的当属 Facebook。

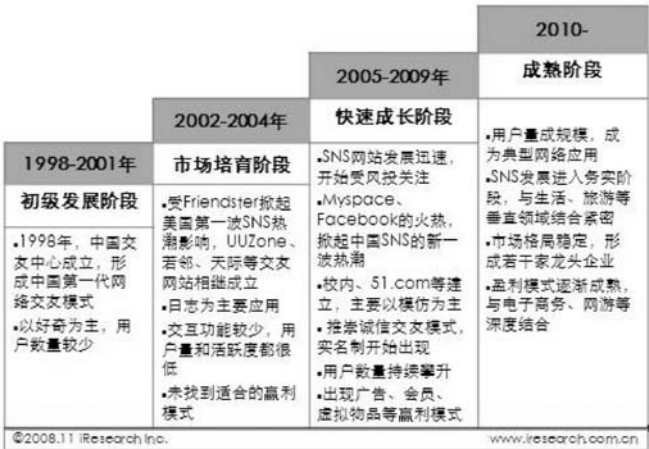


图 1 中国 SNS 发展历程^[3]

它与 2004 年 2 月 4 日由哈佛学生 Mark Zuckerberg 建立。2007 年取消学校限制，面向大众开放。大家可以在上面共享照片，加入感兴趣的组，发表日志，与全世界的朋友交流信息，还可以在上面做广告等，提供了一个供全世界各国朋友相互认识相互交流的平台。2009 年 12 月，Facebook 的独立人次达到了 4.69 亿，与上月相比整整增加了 3100 万人次。Facebook 一个月内增加的新用户量相当于雅虎一年所增加的用户量，也相当于 Digg 总用户量和 Twitter 的用户量的一半。截至 2010 年 4 月，据 comScore 的数据显示，谷歌目前是美国最大的网站，覆盖了 81% 的美国人口，Facebook 覆盖了 53% 的美国人口，落后于谷歌、雅虎和微软。如今，Facebook 的用户数量可谓是节节攀升。

LinkedIn.com，它是一个专为商业界设计的社交网站，它的商务性及一些特殊功能已被一些商业网站用来当做营销的渠道。像知名 B2B 电子商务平台阿里巴巴、tradekey、bytrade 等都在上面建立了各自的庞大的营销网络，Linkedin 真正地把社交关系变成了商业网络。同时国外目前比较出名的 SNS 社交网站还有 Twitter，Myspace.com 等。

在 Facebook 商业传奇的感召下，社交网站（SNS）已迅速蹿红为全球最炙手可热的互联网细分市场

场，掀起了新一波网络创业浪潮。国内各公司也纷纷效仿，分别推出自己的 SNS 网站，如图 2 所示为一张 2010 年 4 月国内 SNS 社交网站用户数量的调查图表。

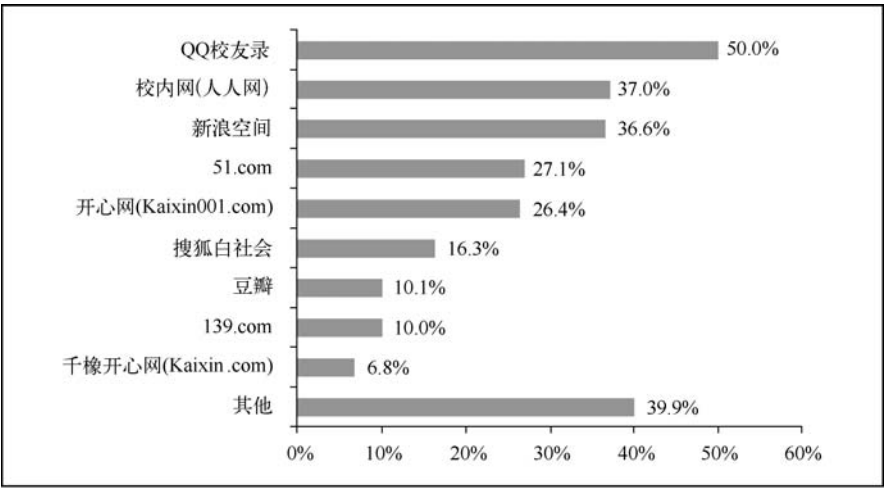


图 2 国内 SNS 社交网站用户数量统计^[5]

从图 2 中可以看出，QQ 校友在所有调查的网站中排行第一（占 50.0%），其次是校内网（占 37.0%），新浪空间（36.6%），51.com（占 27.1%），开心网（26.4%）。

QQ 校友，是腾讯开发的一款真实的互动交友社区，同时整合了 QQ 校友录、QQ 空间，QQ 群等若干服务。QQ 校友于 2008 年 6 月 6 日，QQ 校友正式内测，与 2009 年 1 月 6 日正式对外发布，凭借着其庞大的 QQ 用户群体迅速走红，在当今国内 SNS 社交网站中占据重要的席位。

校内网成立于 2005 年 12 月，是中国最早的校园 SNS 社区。2006 年 10 月，被千橡公司收购^[4]，2009 年 8 月 4 日正式更名为人人网，把仅仅面向每个高校改为面向大众网民全体。2009 年 10 月 27 日，人人网将通过人人连接技术实现与各垂直领域优秀网站的全面连接，据人人网预计，通过与第三方网站合作，人人网将获得 30%~200%的注册用户数量的增长，以及 15%~100%的用户自主生产内容的增长。

4 用户群体

SNS 社交网站已逐渐成为大家平时联系交流的互联网互动应用的集成平台现在，随着互联网用户的逐渐增加，越来越多的人加入社交网站，它已成为很多人生活中不可缺少的一部分。下面通过 IDC 评书网就 CNNIC 2010 年四月份发布的报告进来对当前网民的情况进行分析。

从图 3 中可以看出，社交网站年龄特征和职业特征非常突出。目前主要的用户群体在 20~29 岁之间，在整个网民中占有 52.6%。其次是 10~19 岁之间，占全体网民的 33.0%。而这两个年龄阶段的刚好是学生及刚工作的白领阶段，图 4 中学生占全体网民 31.7%，政治机关事业单位工作者占 10.5%，公司一般职员占 13.9%，与图 3 中学生与刚工作的白领占主要成分相符合。

从分析可以看出，我国目前的社交网站主要以学生群体和白领阶层为主，学生群体所占成分最多。而其他年龄阶段及其他职业的网民还待各 SNS 网站来发掘，这将为每个 SNS 社交网站增加庞大的用户数量。

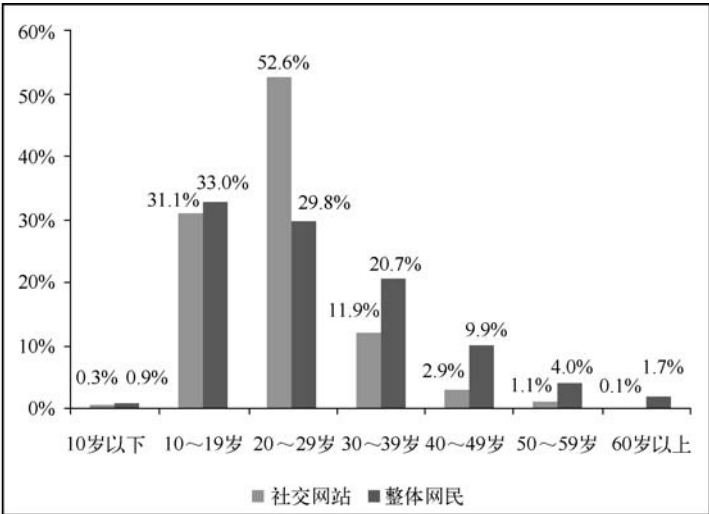


图 3 SNS 社交网站用户年龄统计

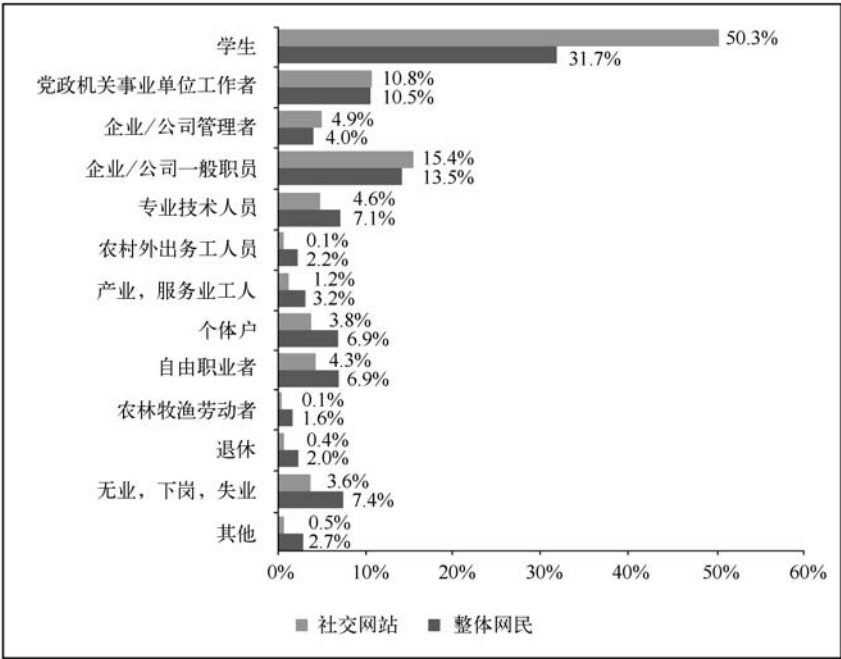


图 4 SNS 社交网站用户职业统计^[5]

5 趋势与存在问题

2009 年是国内 SNS 社交网站快速发展的一年，SNS 社交网站的发展代表了互联网应用的重要方向。SNS 作为新型的互联网人际交流方式，其具有独特的魅力，使得越来越多的人开始认识到 SNS 社交网站并加入到其中。例如，电子邮件一对一或者一对多，但却比较正式，一个人很少会通过邮件来更新自己的日志；自己喜欢的音乐等诸多信息发给每个好友比较麻烦；即时通信比较快捷，但本身却不能存储个人数据，也不能延时沟通；博客更多是以个人数据位中心，有评论互动，但信息的传递和互动性方面却不如 SNS 丰富和快捷。^[7]诸多优点明显表明了 SNS 网站有这美好的前景，它代表了

互联网将来发展的方向。而目前其发展的好坏关键在于其用户的数量，如何吸引新的用户加入和维持老用户的数量成为了影响其发展的首要问题。一些网站采用了传销式的经营方法，以及通过不断地增加新的特色来为网站吸引用户。

市场定位要明确，自己的目标用户是那些群体，目标用户有什么需求和上网习惯。明白了这些，才能更好了来吸引用户。如校内网精准的定位于大学生，MySpace 定位于音乐和娱乐领域，LinkedIn 定位于职场人士。而现在校内网，Facebook 随着知名度提高，已经向全民开放，从而其用户数量有明显增加。

对于国内市场，由于 QQ 的根深蒂固，其通过 IM 已经奠定了成熟的用户关系链，这样的核心竞争力，在目前的格局下很难被超越，使得 SNS 网站的发展扩张受到一定的阻力。QQ 早已深入人心，如从早期的 IM，QQ 空间，QQ 群，又到现在 QQ 校友的推出，使得外人很难进入。而后随着开心网“开心农场”受到大家的好评，QQ 又推出 QQ 农场，使得一些玩开心农场的用户转向了 QQ 农场。^[8]这一切只是由于 QQ 上沉淀了亿万用户的“关系”。

目前国内最为风光的当属 SNS 游戏，它火暴的同时更为各自社交网站带来了大量的用户。但是其缺乏创新性，起初用户可能会好奇，但慢慢的用户可能会厌倦^[9]。而且正在其发红发紫时，政府放出了一枚重磅炸弹。文化部曾在不久前推出“网络游戏未成年人家长监护工程”，并对几家大型 SNS 游戏公司提出整改内容。SNS 网站除了要申请新设立从事网络游戏经营活动的互联网文化经营单位除符合有关规定外，还应具备 1000 万元以上的注册资金。如此高的门槛，让国内开发商望而却步。从而导致一些 SNS 公司纷纷出海寻找机会^[3,10]。

6 结束语

SNS 社交网站带来的多样化社交网络应用正在改变着人们对互联网的使用习惯，也对我国互联网的发展和普及起到了积极的推动作用。随着各类社交网站在竞争中快速发展，用户规模迅速增长。随着用户数量的积累和滚雪球式的不断增大，社交网站渐渐走向大众化，并迅速融入社会生活。社交网站来源于海外，扎根于脚下，并将成为我国互联网应用的重要组成部分。^[5]但由于国情的不同，我国的 SNS 社交网站不能一味地模仿国外，根据自己国家的情况，文化，应该加入自己的特色，调整策略。其发展的关键还是在于如何吸引并维持住更多的用户。由于 QQ 在国内的根深蒂固，拥有着大量的用户，使得一些 SNS 网站遇到了不小的挑战。同时一些 SNS 网站要明确好自己的定位，才能更好地把握住自己的用户，以至于不被流失。而且面对的网民不够广泛，大多数是学生、白领阶层。还有更多其他年龄阶段和职业的网民有待各网站来发掘，这将是一笔可观的用户数量。虽然人人网等已经面向大众开放了大门，但是其对各个年龄阶段及职业的网民的吸引力还不够，还需要增加一定的特色。同样 SNS 游戏设计要体现出新意，不然久而久之用户会感到厌倦。日前政府提出的 1000 万元以上的 SNS 游戏注册资金，更是使目前正火的 SNS 游戏商纷纷抱怨，进退两难。所以这一切只能说明我国 SNS 社交网站发展还不成熟，还在逐渐探索中前进，还要迎接一定的挑战。

参考文献

[1] 今天，你偷菜了吗？——2009 网络社会潮流盘点，中广网，2010，1。
[2] 2010：中国社交网站转型年 偷菜并非 SNS 全部，新华网，2010，2。
[3] 许彬. SNS 社交网站的发展趋向于细分、务实、开放，联盟志，2009，2。
[4] 陈安迪. Web2.0 时代下 SNS 网站发展瓶颈 ——以校内网为例，人民网，2009，11。
[5] 洋风来袭 分析我国网民 SNS 应用发展现状，IDC 评述网，2010，9。
[6] boyd, d. m., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Media ted

Communication, 13(1), article 11.

- [7] 黄亮新, SNS 的存在价值, 艾瑞网, 2009, 3.
- [8] 王嫻, 谢弛, 荣雪, 范雯. SNS 网站运营的现状和未来趋势研究, 2008, 12.
- [9] 国内外 SNS 网站分析报告.
- [10] SNS 迎来生死抉择新准入门槛有望近日发布, 2010.6.

由 AT89S52 与 TC35i 实现的短信息处理系统

王书伟¹，张鸣¹，杨静²

(1. 防空兵指挥学院，河南 郑州，450052; 2. 河南省电子产品质量监督检验所，河南 郑州，450003)

摘 要：本系统采用嵌入式技术，由单片机 AT89S52 最小系统与 TC35i 模块实现 GSM 基带处理、GSM 射频的传输与控制。本文着重阐述短信息接收与发送及终端数据处理的信息流程。短信息接收与发送单元主要由单片机系统对 GSM 模块的实施信息交互控制，终端处理单元完成收、发信息的数据融合与处理。实现单片机控制 GSM 模块与移动台（手机）信息流的查询与控制，即由单片机控制手机的 SMS（Short Message Service）的收发。

关键词：GSM；SMS；嵌入式技术；数据融合；TC35i 模块

中图分类号：TP311.56 **文献标识码：**A **文章编号：**1006-7043 (2010) xx-xxxx-x

Short Information Management System Which Realize by AT89S52 and TC35i

WANG Shuwei¹，ZHANG Ming¹，YANG Jing²

(1. Air Defense Forces Command Academy, Zhengzhou 450052,Henan, China;

2. Henan Supervision & Testing Institute for Electronic Products Quality, Zhengzhou 450003,Henan, China)

Abstract: This system adopts technology of inlay, and the minimum system from chip microprocessors AT89S52 realizes process of basic belt and transmission and control of GSM radio frequency with TC35i mould. (Global System for Mobile Communication) This essay carefully explains information process of reception and sending of short message and terminal data process. The unit of reception and sending of short message is mainly controlled by information which is given from chip microprocessor system to GSM mould. Terminal processing unit finishes data combination and procession of short message of reception and sending. Realizing chip microprocessor to control GSM mould and to inquire and control mobile message flow is equal to reception and sending of mobile phone which is controlled by chip microprocessor.

Keywords: GSM; SMS; embedded technology; data fusion; TC35i chip

1 前言

SMS（Short Message Service）短信息服务是 GSM 系统中提供的一种 GSM 终端（手机）之间，通过服务中心（Service Center）进行文本信息收发的应用服务。其中，服务中心完成信息的存储和转发工作。

以 GSM 网络作为数据无线传输网络，目前已开发出各类应用，如无线数据的双向传送、无线远程检测和控制等。具体的事例有：变电站、电表、水塔、水库或环保监测点等监测数据的无线传输和无线自动警报；远程无线控制高压线路断路器、加热系统、防洪拦阻系统或其他机电系统的启动和关闭；车队交通管理和控制指挥系统等。

GSM 系统是目前基于多址技术（TDMA、SDMA、FDMA、CDMA）的移动通信体制中，比较成熟完善系统。在全国范围内实现了联网和漫游，用户无须另外组网，只要实现资源共享就可以节省昂贵的建网费用和维护费用，进行无线通信还具有双向数据传输功能，性能稳定，是远程数据传送与监控设备通信的一个强大的支持平台。

作者简介：王书伟（1955—），男，副教授，学士；
张 鸣（1984—），女，助教，在读硕士研究生；
杨 静（1975—），女，工程师，硕士。

2 系统总体设计思想

本系统包括短信息接收、发送部分和终端数据处理两部分。前部分实现单片机对 GSM 模块收发短信息的控制；后部分则完成对所接收到的信息进行处理、结果反馈，实现系统自动查询功能。

3 系统组成

系统由 GSM 网络、蜂窝通信引擎电路（TC35i 模块）、单片机控制电路（AT89S52 最小系统）、PC 通信接口电路（RS-232/RS-485）、客户终端（手机）构成，如图 1 所示。本文着重阐述短信息接收与发送和终端数据处理流程。

3.1 电路原理

3.1.1 TC35i 模块组成

TC35i 模块主要由 GSM 基带处理器、GSM 射频模块、供电模块（ASIC）、闪存、ZIF 连接器、天线接口六部分组成。通过 ZIF 连接器与单片机实现电路接口，通过该接口读取或发送 TC35i 模块中的数据，将是 TC35i 模块的应用核心。

3.1.2 TC35i 模块的主要特性与技术指标

- (1) 频段为双频 GSM900MHz 和 GSM1800MHz（phase 2/2+）。
- (2) 支持数据、语音、短消息和传真。
- (3) 高集成度（54.5mm×36mm×3.6mm）。
- (4) 电源 3.3~4.8V。
- (5) 可选波特率 300bps~115kbps，动波特率 4.8 ~115kbps。
- (6) 电流消耗：休眠状态为 3.5mA，空闲状态为 25mA，发射状态为 300mA（平均）
- (7) 温度范围：正常操作-20℃~+55℃，存放-30℃~+85℃。
- (8) SIM 电压为 3V/1.8V。

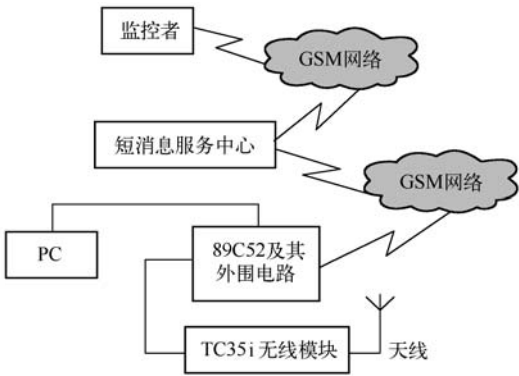


图 1 智能监控系统原理图

3.2 AT89S52 单片机的主要性能及特点^[1]（如表 1 所示）

表 1 AT89S52 单片机的主要性能及特点

1. 兼容 MCS-51 指令系统	2. 8kΩ可反复擦写（>1000 次）ISP Flash ROM
3. 4.5~5.5V 工作电压	4. 256×8bit 内部 RAM
5. 低功耗空闲和省电模式	6. 3 级加密位
7. 软件设置空闲和省电功能	8. 32 个双向 I/O 口
9. 3 个 16 位可编程定时器/计数器 10	，全双工 UART 串行中断口线
11. 2 个外部中断源	

3.3 TC35i 各引脚与单片机及各元件的连接 (如图 2 所示)

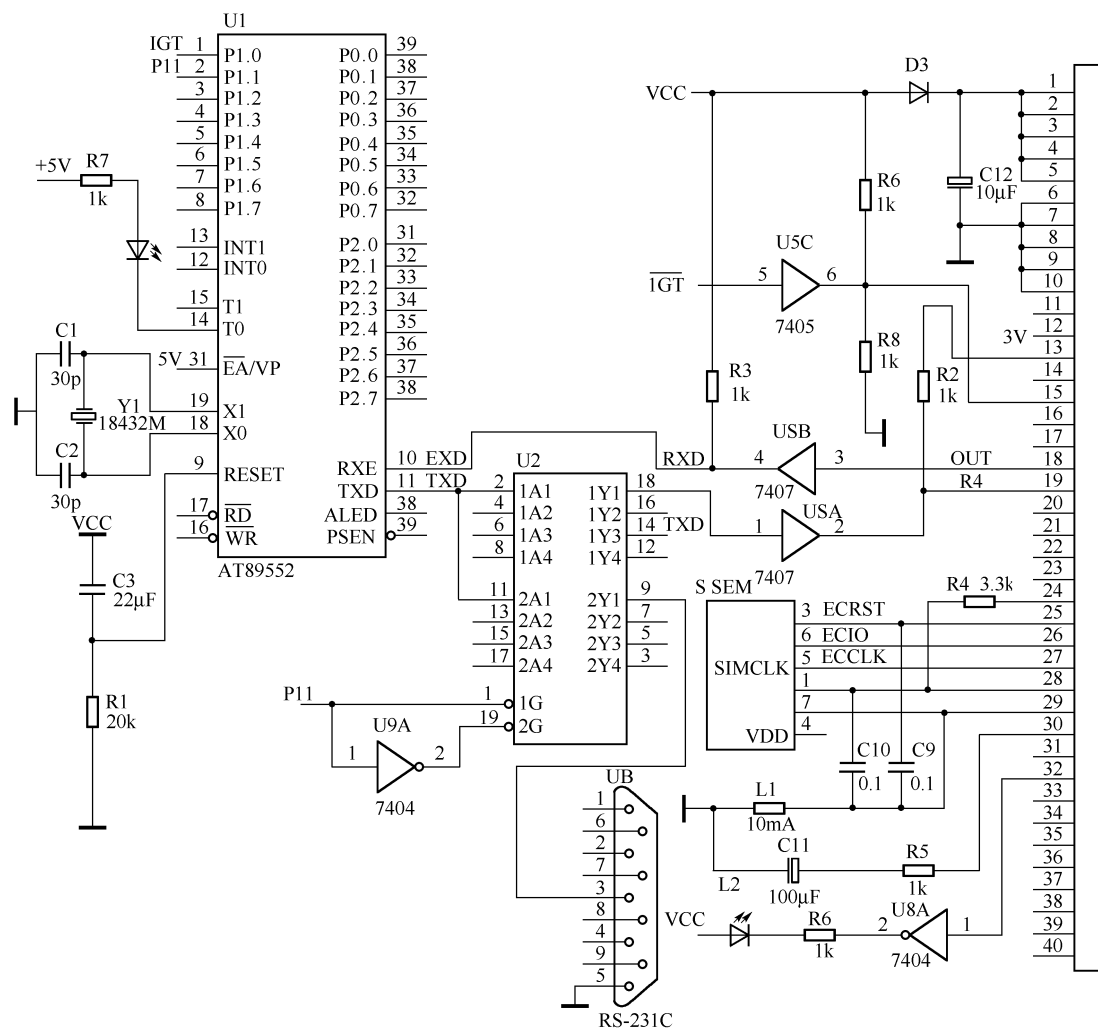


图 2 TC35i 各引脚与单片机及各元件的连接原理图

4 系统的工作流程

用户通过编辑短信息（关键字、状态字、查询信息），发送到指定的号码上，TC35i 接收到信息后，把所有信息通过串口发送给单片机，单片机首先判断所发的信息是否符合规定，如果不符和规定则单片机会返回给用户一条正确格式的信息，告诉用户按此正确格式发送信息。而单片机只有接收到符合规定的信息才把短信传送给 PC，PC 根据实际需要可以添加不同的数据库信息。比如，天气预报、车辆违章、公交车路线查询、产品真伪、股票信息查询、物流信息等，而这些查询只需要在 PC 上添加相应的数据库信息，即可满足用户的需要，为广大用户提供经济方便快捷的服务。

4.1 单片机与 TC35i 的软件接口设计思想^[2]

单片机与 TC35i 的软件接口，完成单片机通过 AT 指令，控制手机的短消息有关的 AT 指令有：

- (1) 单片机与 TC35i 模块由串口建立连接: AT。
- (2) 设置 TC35i 模块工作模式^[4]: AT+CMGF= n , $n=0$: PDU 模式; $n=1$: 文本模式, 通常要设置

为 PDU 模式，在这种模式下，能传送或接收透明数据（用户自定义数据）。

(3) 读 TC35i 模块短消息数据：AT+CMGR= n ， n 为短消息号（十进制）。

(4) 列出 TC35i 模块内的短消息：AT=CMGL= n ， $n=0$ ：未读的短消息； $n=1$ ：已读的短消息； $n=2$ ：未发送的端消息； $n=3$ ：已发送的端消息； $n=4$ ：所有的短消息。

(5) 删除 TC35i 模块短消息：AT+CMGD= n ， n 为短消息号（十进制）。

软件设计流程如图 3 所示。

4.2 短消息收发

根据设置不同，GSM 模块将收到的短消息保存在缓存单元或存入 SIM 卡，单片机从 GSM 模块中接收短消息实质上就是从 SIM 或缓存中读出信息^[3]。这主要利用 AT+CMGR 和 AT+CMGL 两条指令来完成，其工作过程如图 4 所示。由于不同的厂商对 AT 指令集的解释代码和响应信息不一样，所以单片机首先要确认能否与 GSM 模块建立起通信，一般用 ATE 指令完成此确认；然后用 AT+CMGF 指令选定短消息的数据格式；在收到 GSM 模块的正确回答以 AT 指令完成读出功能。一般用 AT+CMGL 读取以前的信息，在收到手机的 RING（振铃）数据时，用 AT+CMGR 读取实时信息。

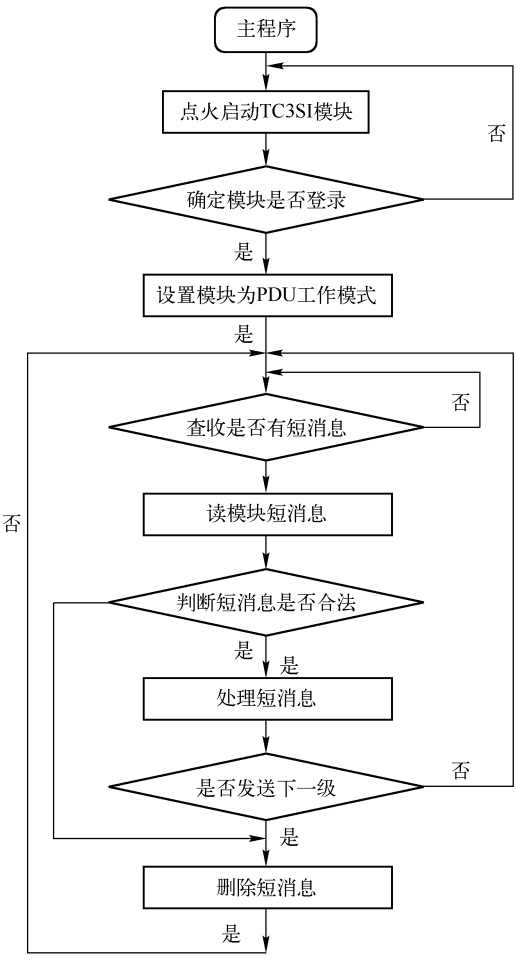


图 3 单片机与 TC35I 软件设计流程图

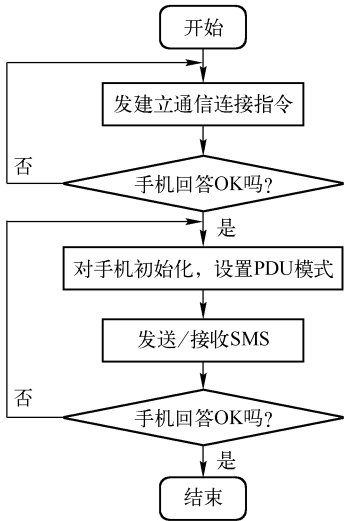


图 4 短消息收发流程

以下是笔者设计的物流数据采集系统中用到的接收 SMS 的一个实例，它说明了 PDU 模式的应用。单片机发送和接收（GSM 模块回答）均为 ASCII 码。所用 GSM 模块为 SIEMENS S3508i。

操作过程如下（{} 内为注释）：

发送：ATE

GSM 模块回答: OK {已建立连接}
发送: AT+CMGF=0 {选用 PDU 格式}
GSM 模块回答: OK {允许选择 PDU 格式}
发送: AT+CMGL=2 {列出已有的短信息}
GSM 模块回答: +CMGL: 1, 2, ..., 24{1 表示信息个数, 2 表示未发信息, 24 表示信息总容量}
0D71683108370105F004000D81683179133208F10000026080410033802632184CF682D
95E0DC2B36D3D170A0243106933D97A0243106933D97A02451068B1983492608 OK
以上这组 PDU 格式的十六进制字符串, 不但包含了短消息的内容, 同时包含了发送者的手机号码、短信息中心号码、短消息发送时间等^[5]。

下面对信息内容进行分析:
0D: 短信息中心地址(号码)长度。
91: 短信息中心号码类型, 91 是 TON/NPI。TON/NPI 遵守 International/E.164 标准, 指在号码前需加“+”号; 此外还可直有其他数值, 但 91 最常用。
683108370105F0: SMSC 短信息所使用的服务中心号码 13807310500。它经过十六进制以字节为单位的高低半字节换位处理, 号码是奇数的添 F, 构成一个 HEX 字节。
04: PDU 类型, 文件头字节。0B: 主叫号码长度。81: 主叫号码类型。
3179133208F1: 0A 主叫号码, 也经过了处理, 实际号码为 13973123801。
00: PID, 为协议标识。
00: DCS 短信息编码类型是 GSM Default Alphabet, 即由 7 位 ASCII 码移位组成 8 位十六进制码(octet), 其方法见表 2。

表 2 7 位 ASCII 码移位组成 8 位十六进制码

1sthex	B0	A6 A5		A4 A3 A2 A1 A0			
2ndhex	C1	C0 B6		B5 B4 B3 B2 B1			
3rdhex	D2	D1	D0	C6 C5 C4 C3 C2			
4thhex	E3	E2	E1	E0	D6 D5 D4 D3		
5thhex F4		F3	F2	F1	F0	E6	E5 E4
6thhex	G5	G4 G3		G2 G1 G0 G6			F5
6thhex	H6	H5 H4		H3 H2 H1 H0 G6			

02608041003380: SCTS 短信息发送时间, 02/06/08/14: 00: 33.08。
本系统采用微计算机应用技术与 TC35i 模块的信息交互, 实现了 GSM 基带处理、GSM 射频的传输与控制, 该系统在短信息接收与发送及终端数据处理的应用方面是一种新的尝试。短信息接收与发送单元主要由单片机系统对 GSM 模块的实施信息交互控制, 终端处理单元完成收、发信息的数据融合与处理。实现单片机控制 GSM 模块与移动台(手机)信息流的查询与控制。该系统在高速公路 LED 显示屏的远程文字编辑方面的使用效果良好。

参考文献

[1] 李广弟. 单片机基础[M]. 北京: 航空航天大学出版社, 2001.
[2] 栋梁. 单片机原理与应用[M]. 北京: 水利水电出版社, 2001.
[3] 李 鸿. 用单片机控制手机收发短信息[J]. 电子技术应用, 2003.
[4] 何立民. MCS51 单片机实用接口技术[M]. 北京: 航空航天大学出版社, 2001.
[5] 苏丽萍. 电子技术基础[M]. 西安: 电子科技大学出版社, 2002.

永煤集团安全监测联网系统及其应用

冯少华¹, 田丰², 王月蓉¹ 黄欢欢¹

(1.防空兵指挥学院, 河南 郑州, 450052; 2.永煤集团技术信息中心, 河南 商丘, 476600)

摘要: 永煤集团安全监测联网系统, 是各矿在生产、安全及管理方面的一个实时监测监控系统, 通过该系统, 将永煤集团本部四矿的各种监测数据集成在公司调度室, 并通过集团公司以太网网络共享数据, 各部门及领导可通过公司网络浏览各自所需的信息, 做到对各矿的安全生产状况进行综合性动态分析, 形成领导决策, 实现资源的有机共享。并将各矿的安全监测信息系统与管理信息系统有机结合, 加强企业内部协作与通信, 提高生产和管理效率, 增强企业的市场竞争力, 使煤矿企业的信息化进程实质性的跨上一个新台阶。

关键词: 煤矿; 永煤集团; 安全监测联网系统

中图分类号: TP274 文献标识码: A 文章编号: 1006-7043 (2010) xx-xxxx-x

The Security Detection Networking System of Yongcheng Coal-electricity Group and Its Application

FENG Shaohua¹, TIAN Feng², WANG Yuerong¹ HUANG Huanhuan¹

(1. Air Defense Forces Command Academy, Zhengzhou 450052, Henan, China;
2. Technology and Information Center of Coal-electricity Group, Shangqiu 476600, Henan China)

Abstract: The security detection networking system of Yongcheng coal-electricity group is a monitoring and supervision system which can monitor the production, security and management of all coal mines. This system can integrate all the data of group's four coal mines into operator's room. Each department and its leader can browse shared data on the Ethernet to get the desired information so as to make a comprehensive and dynamic analysis of security production of each coal mine so as to form leadership decision and realize resources common sharing. This system can combine the security monitoring information system together with the management information system, which can enhance cooperation and communication in our group so as to increase production and management efficiency and promote group's market competition and finally to make coal mine company's information process to reach new heights.

Key words: coal mine; Yongcheng coal-electricity group; security detection networking system

煤矿安全隐患和事故时刻威胁着煤矿职工生命和煤炭行业发展, 国家安监总局多次发文要求所有煤矿安装应用煤矿安全监控系统, 并实现各级区域监控联网。安全对煤炭生产起着保证、支撑和推动作用, 煤矿安全生产也是国家安全生产的重要组成部分。现在发展中的煤炭大国, 如印度、南非、波兰等国家, 百万吨煤死亡率是 0.5 左右。先进国家, 如美国、澳大利亚等国家, 百万吨煤死亡率是 0.03、0.05。我国直到 2009 年, 才历史性地降到了 1 以下, 煤炭生产百万吨死亡率下降到 0.892, 安全形势不容乐观。

1 概述

永煤集团安全监测联网系统, 是各矿在生产、安全及管理方面的一个实时监测监控系统。通过集

作者简介: 冯少华 (1983—), 男, 助教, 学士;
田丰 (1984—), 男, 助工, 学士;
王月蓉 (1985—), 女, 助教, 学士;
黄欢欢 (1982—), 女, 助教, 学士。

团公司以太网网络将各矿远程监测监控信息传送到总部，并可以对各矿井监测监控信息进行分析汇总，为领导提供辅助决策功能。

永煤集团本部四个生产矿井已经具备安全监测系统，本部矿井安全监测联网系统不但实现四个生产矿井的监测系统的联网，还可以在不修改软件程序的前提下，增加后续矿井监测系统的相关数据，实现和已联网矿井同样的监测功能。

2 系统实现

系统基于 Web 技术进行开发，采用 TCP/IP 标准协议，提供开放的、标准的接口，支持第三方的系统集成。满足集团公司、生产管理部门、安全监管部門对井下安全环境的有效监督管理。

(1) 整合接口^[1]。包含以下几个基本模块：页面绘制：完成 HMI 的设计，用于实时数据和历史数据的显示；实时数据中心：用于系统内部的数据交换；数据存储：完成对实时数据的定时或者按需存储；内部脚本：完成界面控制、按需访问、指令下达和策略等功能；通信接口：接入各种通信设备。

(2) 数据存储、查询。集团公司集中存储安全监测联网系统的数据，各子公司存储自有安全监测联网系统的数据。集团公司系统平台查询数据速度不大于 15s，如图 1 所示。

传感器瓦斯 状态报警 时间大于 (秒) 时间2010-05-07 00:00 - 2010-05-07 23:59 生成报表 生成Excel

○页次: 1页/1页, 共条记录 [首页] [上一页] [下一页] [尾页]

测点	类型/单位	测点安装位置	状态	持续时间	开始时间	结束时间	最大值	最小值	平均值
2016	瓦斯/%CH4	2602工作面	报警	00:00:40	05-07 09:32:20	05-07 09:33:00	1.01	1.01	1.01
2416	瓦斯/%CH4	北四轨道下山工作面	报警	00:00:35	05-07 09:46:05	05-07 09:46:40	0.98	0.98	0.98
2521	瓦斯/%CH4	北翼胶带机头高压调室	报警	00:02:00	05-07 10:47:21	05-07 10:49:21	2.02	1.24	1.70
2519	瓦斯/%CH4	北翼胶带机头低压调室	报警	00:02:05	05-07 10:50:51	05-07 10:52:56	2.00	1.89	1.95
2517	瓦斯/%CH4	北翼胶带机电调室	报警	00:02:30	05-07 10:57:21	05-07 10:59:51	1.95	0.77	1.36
900	瓦斯/%CH4	北翼回风巷瓦斯	报警	00:01:35	05-07 11:02:56	05-07 11:04:31	2.02	1.17	1.60
900	瓦斯/%CH4	北翼回风巷瓦斯	报警	00:02:35	05-07 11:12:56	05-07 11:15:31	2.07	0.73	1.48
2518	瓦斯/%CH4	北二采区回风	报警	00:04:10	05-07 11:32:21	05-07 11:36:31	1.93	1.36	1.73

图 1

(3) 瓦斯数据的安全。采用两台应用服务器和数据磁盘阵列，保障系统和数据安全。并在该系统接入口处安装硬件防火墙，以阻止外来黑客的非法入侵。采取对服务器操作系统安装网络防病毒软件和防火墙，并设置自动更新升级周期为一周左右，保障操作系统的安全。

(4) 报表打印存档。本系统具有报表打印功能，可以按测点·时间（段）·矿·报警·异常等各种方式查询各种数据，并能打印出来存档备份，如图 2 所示。

传感器瓦斯 状态报警 时间大于 (秒) 时间2010-05-07 00:00 - 2010-05-07 23:59 生成报表 生成Excel

图 2

(5) 独享取数服务器。目前各矿取数据的服务器是矿上提供的，并且和其他应用系统共用，为了安全和高效，该系统在各矿独享取数服务器。

(6) 兼容现有的监测监控系统。采用通用化、标准化的软件技术实现监测系统的联网，提供标准接口，针对不同的监测系统开发不同的接口组件，更换不同系统只需更换不同的接口组件，方便快捷地将 KJ 系列安全监测系统信息集成到中心服务器上，不用改动矿上现有监测监控系统。

(7) 测点定义自动识别，可扩展性好^[2]。联网系统能自动识别联网矿井监测系统传感器的测点定义的变化，如新增测点、改变、删除等，各矿井及地面中心站无须人工重新定义，在远程联网时，自动把测点定义的变化传输到集团公司监测中心，并在各矿井及地面中心站的实时数据报表画面中及时

反映出来。

(8) 数据实时发布和显示。在地面中心站利用互联网技术建立监测系统的 Web 网站，系统可自动不断更新实时数据报表页面；各级领导及有关人员通过局域网，用 IE 浏览器即可及时方便地浏览各矿井的实时监测数据画面，随时了解矿井生产现场的环境及安全工况参数的变化。并具有实时数据的显示、存储及报表预览打印。在实时数据显示中，可显示各重要传感器的安装位置、类型、当前数值、通信状态，如图 3 所示。

煤矿状态 全部测点 模拟量 数字量 其它量 自定义 测点类型: 全部 轮显 < << 1/9 >> >									
序号	煤矿名称	测点编号	测点类型	测点值	测点状态	馈电	安装位置	时间	测点信息
1	新桥煤矿	1016	瓦斯	0.01%CH4	正常	无	2209外切眼工作面	05-07 16:19	...
2	新桥煤矿	1017	瓦斯	0.00%CH4	正常	无	2209外切眼回风	05-07 16:19	...
3	新桥煤矿	1018	瓦斯	0.00%CH4	正常	无	2209皮顺钻场	05-07 16:19	...
4	新桥煤矿	1019	瓦斯	0.02%CH4	正常	无	2209工作面回风	05-07 16:19	...
5	新桥煤矿	1032	电压	28.50V	正常	无	2209皮顺车场电池电压	05-07 16:19	...
6	新桥煤矿	1033	电压	30.10V	正常	无	2209皮顺车场充电电压	05-07 16:19	...
7	新桥煤矿	1034	电压	15.20V	正常	无	2209皮顺车场电源1#电压	05-07 16:19	...
8	新桥煤矿	1035	电流	0.00mA	正常	无	2209皮顺车场电源1#电流	05-07 16:19	...
9	新桥煤矿	1036	电压	18.20V	正常	无	2209皮顺车场电源2#电压	05-07 16:19	...
10	新桥煤矿	1037	电流	155.00mA	正常	无	2209皮顺车场电源2#电流	05-07 16:19	...

图 3

(9) 数据自动汇总。实现监控中心站监测服务器，通过各种联网方式与各矿井的监测系统联网，并自动汇总各矿实时监测数据等。监测联网历史数据压缩存储。查询速度快、存储量大，支持一年以上。

(10) 报警处理。联网系统具有语音报警，当传感器超限时，系统自动闪烁提示，并执行相应的语音报警。联网系统具有手机短信提示，当传感器超限时，系统自动将超限的矿井名称、超限地点等信息发送到有关领导及值班管理人员的手机上。

(11) 趋势分析并生成报表。联网系统具有历史曲线的显示、查询功能，系统可以显示曲线，使用户可以进行对比分析。用户可以方便选择任意测点，任意时间范围，查看实时数据，以及历史曲线与数据。报表统计自动生成、分析、显示、打印等。

(12) B/S 结构方式，易于使用^[3]。支持各种网络联网方式，如 ADSL、GPRS、DDN、光纤等，及时可靠地实现对煤矿安全生产中各种隐患进行实时监控。系统支持多个厂家的安全监测系统，包括 KJ4、KJ10、KJ31、KJ75、KJ90、KJ92、KJ95、KJ101、KJ2000 等。

(13) 数据联网时，各矿数据进行校时：采用集团公司中心服务器对各矿主、各控机进行校时，各矿监控主机目前采用的双机热备，一旦主控计算机发生宕机或故障，热备计算机升级为主控机，校时需要采用双机同步进行。

(14) 各矿主控计算机定时生成的上传文本文件，到各矿数据发布服务器，建议采用主控机上传程序和热备机上传程序同步进行，分别存放在发布服务器不同目录。

(15) 集团公司中心 Web 服务器能提供曲线分析，保留系统采集的峰值曲线，以便分析各矿瓦斯超限断电功能测试、历史报警、历史超限等实际情况。

(16) 在历史数据显示表中增加按时间查询曲线显示按钮，方便查看，简化操作。

(17) 为避免大量并发数据处理的数据延时、页面刷新不及时，及网络病毒造成数据阻塞，各矿服务主控机、数据发布服务器至公司中心数据服务器建议采用专网，在公司中心数据服务器输出端口配置防火墙。

(18) 短信服务平台，管理员可根据报警信息的类型、时间长短、信息源地点、发布对象、发布

格式、留言（信息摘要）等进行选择、增加、修改操作^[4]。

（19）按 AQ1029-2007、AQ6201-2006、安监总煤装[2008]41 号文件要求，联网系统能提供报警信息，报警信息报表可按传感器的类型、矿名进行分类列表，以便公司值班员打印；提供预案处置模块，各矿将非正常处置程序和应急预案上传，生成静态网页方式（41 号文件有要求，即将颁布监测联网标准中也有此内容）。

3 系统取得的成效

永煤集团安全监测联网系统以计算机网络为基础，能及时、可靠、安全地实现安全生产的信息（数字、图形）采集、存储、传输、处理、Web 信息发布，以及煤矿瓦斯隐患监控功能。在本部各矿已形成的安全监测监控系统的基础上，建立了集团公司级数据监测监控中心：

- （1）实现各矿的安全监测系统以标准的软件接口和信息协议交换信息，并进行综合、分类处理。
- （2）实现四矿安全监测监控数据进行集中存储、备份并向省级数据监控中心实时地上传数据。
- （3）实现集团公司领导及各相关部门可以及时、准确地掌握各矿的生产和安全的实时监测监控数据。
- （4）实现快速地对历史数据查询、统计，汇总报表，并根据实时预警信息，对超限矿井发出整改警告等,为领导决策提供依据。
- （5）实现能够定期按测点·时间（段）·矿·报警·异常等各种方式查询各种数据，并能打印报表并存档备查。
- （6）实现历史数据分库存储、备份，当查询历史数据时导入数据即可方便、快速查询历史数据和历史曲线等。

4 结论

永煤集团安全监测联网系统运行以来，整合了各矿安全监测系统的监测数据，提升了公司安全生产问题的管控能力，实现了公司内部的安全生产资源共享，丰富了安全生产信息的发布手段。

- （1）建立安全生产数据中心。建立集团层面的数据库平台，根据实际需求向各矿的数据库中采集数据，并集中存储在集团公司数据库服务器中，客户端以报表或曲线的形式访问数据，存储的历史数据可为事故分析提供可靠的依据。
- （2）实现数据的共享和实时监测。集团总调度室和安监局可以监测子公司矿井的瓦斯实时数据，子公司监管部门和矿井监控中心则只能监测到管理职权范围内矿井的安全生产情况。
- （3）建立安全报警防范机制。系统将提供对生产安全数据的超限报警功能，并结合移动信息平台，及时将安全信息传递给公司领导和相关人员。大大提高了集团公司对安全生产问题的响应速度，有利于集团公司的指挥和调度，提高了管理效率。杜绝了瓦斯超限漏报、瞒报现象。
- （4）提高数据分析能力。通过数据的汇总分析功能，采用多样化的数据表现方式对安全生产实时数据进行智能化分析，为企业领导和相关管理人员进行科学的生产经营决策提供及时可靠的支持。
- （5）强化煤矿安全生产监督管理。各矿井的监测数据受到集团公司的直接监管，集团公司安全生产部门对瓦斯超限齐抓共管，利用信息化手段获取第一手资料，加强安全生产监督管理，已经取得了明显的效果。

煤矿安全监测联网系统不应只能实现监测监控，还应能实现根据监测指标，结合监测地点、环境等客观因素进行危险性判别、分析并提出临时解决方案的功能。同时系统应用软件应进行标准化建设，实现统一格式提供监测数据，确保数据的准确完整，这对促进矿井监控技术发展和系统推广应用均具有重要意义。同时，也要注重强化技术培训，建设高素质的人才梯队，提高现场管理和对监测系

统的维护水平，对系统的技术创新提供源源不断的动力，确保系统的高效运转。

参考文献

[1] 李伟. 基于 ASP.NET 应用程序的安全性设计和实现[J]. 科技情报开发与经济, 2008, (5): 173-174.

[2] 杨艺清. Web Service 及 ASP.NET 应用程序的安全策略和方法[J]. 湖南科技学院学报, 2006, (5): 173-174.

[3] 谢希仁. 计算机网络[M]. 北京: 电子工业出版社, 2008.

[4] 陈冠军. 精通 ASP.NET 2.0 典型模块设计与实现[M]. 北京: 人民邮电出版社, 2007.

基于三层架构的科研管理系统研究与应用

王辉，孟军，秦兴桥，丁彦芳，黄欢欢

(防空兵指挥学院，河南 郑州，450052)

摘 要：本文阐述了三层架构技术在科研管理系统中的运用，分析了三层架构的优点，并通过一个具体模块的设计与实现，描述了三层架构技术的实现过程。

关键字 科研管理；.NET；三层架构

中图分类号：TP393 文献标识码：A 文章编号：1006-7043 (2010) xx-xxxx-x

The Research and Application of Scientific Research Management System Base on Three-Tiers Architecture

WANG Hui, MENG Jun, QIN Xingqiao, DING Yanfang, HUANG Huanhuan

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: This paper introduces the application of 3-tier architecture technology in the systems of scientific research management. It analyses the advantages of the 3-tier architecture. And it takes the designation and application of one block for example, depicts the process of application of 3-tier architecture.

Keywords: scientific research management; .NET; Three-Tiers Architecture

科研管理工作是现在很多高校教学管理过程中的一个重要环节，传统的管理方式已不能适应信息化建设的要求，经过院校的积极努力，在科研管理中已经逐步向网络化、信息化、规范化方面发展，但还存在着自动化程度不高，大量的管理工作还需要通过办公软件和简单的管理系统来处理大量的数据，大部分工作还要依赖人工来进行等问题，需要进一步改进完善，以适应新的管理形势。基于此开发一个较完善的科研管理系统就显得尤为重要。

当前，基于三层模型设计研究在软件设计中得到快速发展。开发多层体系结构的应用程序已成为研究的热点。因此，在开发过程中采用了三层模式来开发本系统。

1 三层模型分析

1.1 三层模型的结构

.NET Framework 架构下，基本分层模型是三层模型，分别是：界面显示层（UI）、业务逻辑层（Business Logical Layer）和数据操作层（Data Access Layer），其结构如图 1 所示。

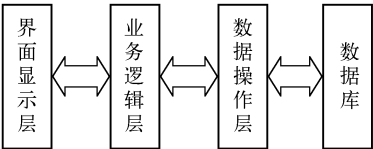


图 1 三层结构示意图

作者简介：王 辉（1974—），男，讲师，硕士；
孟 军（1968—），男，讲师，硕士；
秦兴桥（1976—），男，讲师，硕士；
丁彦芳（1979—），女，讲师，硕士；
黄欢欢（1982—），女，助教，学士。

1.2 各层的任务及功能

界面显示层：用来在显示器中显示的用户界面。该层要以适当的形式显示由业务逻辑层动态传送的数据信息，这个功能要通过使用相应的 **Form** 和相关控件来实现。同时，这一层还要负责获得用户录入的数据，完成对录入数据的校验，并将录入数据传送给业务逻辑层^[1]。

业务逻辑层：是三层模型的中间层，也是整个分层模型中最为重要的层。这一层为界面显示层提供功能调用，同时它又调用数据访问层所提供的功能访问数据库。该层要根据整个系统的设计，构造工程中关键的几个对象，从而实现工程中的大部分逻辑控制功能。业务逻辑层是界面显示层和数据操作层之间的衔接部分。

数据操作层：是整个分层体系的最低层，为应用程序提供统一的数据访问服务，它主要用来实现与数据库的交互，即完成查询、插入、删除和修改数据库中数据的功能。数据操作层为业务逻辑层提供服务，根据业务逻辑层的要求从数据库中提取数据或修改数据库中的数据。由于访问数据库是系统中频繁发生而且最消耗资源的操作，所以在这一层要对数据库访问进行优化，提高系统的性能和可靠性。

1.3 三层模型的优势

使用三层模型进行系统的架构设计具有一些显著的优势。例如：

增加了代码的重用，数据操作层可在多个项目中公用；业务逻辑层可在同一项目的不同子系统和模块中使用。

软件的分层大大提高了系统的模块化程度，增强了系统的可维护性和扩展性。

它方便了软件项目的分工和管理。美工人员可以很方便地进行界面设计，并在其中调用业务逻辑层给出的接口，而程序开发人员则可以专注地进行代码的编写和功能的实现。

2 系统的三层架构设计

使用.NET 平台能够快速方便地设计和部署三层架构，其优势主要体现在：可以混合使用 C#，VB，J#等多种编程语言进行后台代码的编写。.NET 中可以方便地实现组件的装配，代码通过命名空间可以方便地使用自己定义的组件^[2]。界面显示层可以用已有的控件组合实现，数据操作层和业务逻辑层用组件来实现，这样就很方便地实现了三层架构。下面结合科研管理系统的设计介绍如何设计和部署三层架构。

2.1 实体类的构建

实体类是现实实体在计算机中的表示。它贯穿于整个架构，负担着在各层次及模块间传递数据的职责。实体类主要用于存储复杂的数据，简化其他各层的数据操作。下面以一个基本表为例来描述如何构建实体类。

```
public class 计划项目表 // 实体化计划项目表
{ public 计划项目表() {} //构造函数
    //定义属性名
    private int _标识号;
    .....//定义表中所有属性名
    private string _项目名称;
    //声明属性
    public int _标识号
    {
```

```

        set{ _标识号=value;}
        get{return _标识号;}
    }
    .....//声明所有的属性
    public strin _项目名称
    {
        set{ _项目名称=value;}
        get{return _项目名称;}
    }
}

```

2.2 数据操作层的设计

数据操作层主要是对原始数据（数据库或者文本文件等存放数据的形式）进行操作的层，而不是指原始数据，也就是说，是对数据的操作，而不是数据库，具体为业务逻辑层或界面显示层提供数据服务。也可以说是：与数据源操作有关的代码，就应该放在数据操作层中，属于数据操作层的内容。

```

public class 计划项目表                                     //数据操作类
{
    public public 计划项目表() {} //构造函数
    public int Add(Model.计划项目表 model){.....}          //          增加一条数据
    public void Update(Model.计划项目表 model){.....}       //          更新一条数据
    public void Delete(int 标识号){.....}                    //          删除一条数据
    public List<Model.计划项目表> GetModels(string strWhere){.....} //得到数据列表
    .....//添加所有对数据操作的函数
}

```

2.3 业务逻辑层的设计

业务逻辑层主要是针对具体问题的操作，也可以理解成对数据操作层的操作，对数据业务逻辑处理，也就是说把一些数据操作层的操作进行组合。

```

public class 计划项目表                                     //业务逻辑类
{
    private readonly DAL.计划项目表 dal = new DAL.计划项目表();
    public 计划项目表() {}                                   //构造函数
    public int Add( Model.计划项目表 model) {return dal.Add(model);} //增加一条数据
    public void Update(Model.计划项目表 model) {return dal.Update(model);} //更新一条数据
    public void Delete(Model.计划项目表 model) {return dal.Delete(model.ID);} //删除一条数据
    public List< Model.计划项目表> GetModels(string strWhere)
    { return dal.GetModels(strWhere);} // 获得数据列表
}

```

2.4 界面显示层的设计

界面显示层可以以 Web 页面形式展现，也可以以 WINFORM 形式展现，科研管理系统采用 WINFORM 形式进行设计。如果业务逻辑层设计得相当完善，无论界面显示层如何定义和更改，业务逻辑层都能提供完善的服务。

界面显示层只是调用控件提供的方法，具体的逻辑处理完全由业务逻辑层来负责。理论上，界面

显示层的修改不会影响业务逻辑层；反之，业务逻辑层的修改不会影响界面显示层的设计，方便了代码的复用及应用程序的扩展。

3 结束语

本系统通过运行使用，能够较好地完成科研管理工作，效果良好。提高了科研管理工作的效率，保证了相关数据的一致性，对工作需要的大量的表格进行了规范，建立了文档模板库，并可以根据需要自动生成，提高了科研管理的自动化、规范化水平，减轻了工作人员的工作量。因此本系统具有较大的应用推广价值。

参考文献

[1] 张瑞，万建成. 基于.NET 技术的企业国有资产产权登记系统的设计与实现[J]. 计算机应用与软件，2007(2)： 178-180.

[2] 王玲，宋斌，王平立，王克龙. 基于数据仓库三层架构的决策支持系统应用研究[J]. 计算机应用与软件，2007（2）： 69-71.

基于 OSG 的三维虚拟化学实验的建模技术

谭同德, 石奇波, 赵新灿

(郑州大学信息工程学院, 郑州 450001)

摘 要: 虚拟化学实验中, 为了有效支持虚拟化学实验平台, 需要对实验相关内容建立合适的模型, 通过对实验仪器、化学药品、反应现象等的信息及其之间的关系进行的分析、归纳, 对化学药品、实验仪器、整个实验建立了模型, 尤其对化学反应现象及整个实验过程进行了过程性建模。根据化学反应类型, 对初中无机化学所有实验进行了整理, 接着从每类实验中挑选了一个典型示例, 并对其实现了基于图形渲染引擎 OSG 的三维虚拟仿真, 渲染出的实例表明此建模方式真实、灵活的对化学实验进行了模拟, 有效的对虚拟化学实验平台予以了支持。

关键词: 化学实验; OSG; 过程性建模; 粒子系统; 三维虚拟化学实验

中图法分类号: TP391 **文献标识码:** A

Three-Dimensional Virtual Chemistry Experiments Modeling Techniques Based on OSG

TAN Tongde¹, SHI Qibo¹⁺, ZHAO Xincan¹

(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, Henan China)

Abstract: In virtual chemistry experiment, in order to effectively support virtual chemistry experiment platform, the need for relevant experimental content to establish a suitable model, through analysis and summarized the laboratory instruments, chemicals, reaction phenomenon and other information and their relations, establish the models for chemical medicines, laboratory instruments and the whole experiment, especially for the chemical reaction phenomena and the experimental processes of the process of modeling. According to chemical reaction type, right middle Inorganic Chemistry All experiments were consolidated, and then selected from each type of experiment, a typical example, and its realization based on the three-dimensional graphics rendering engine OSG virtual simulation, rendering out examples show that this modeling means real and flexible simulations the chemical experiment, effectively supported to the virtual chemistry experiments platform.

Keywords: chemical experiment; OSG; procedural modeling; particle system; Three-dimensional virtual chemistry experiments

1 引言

教学应用类的虚拟仿真化学实验, 其优点甚多: 解决了现实中药品一次性实验、药品及实验仪器价格昂贵带来的成本问题; 避免了现实中由于不规范操作引起的危险性; 有“存在网络即为实验室”的方便性, 等等。三维虚拟化学实验更是因交互性、逼真性特点引起了热衷的研究^[1]。郑州大学的张艳丽^[2]基于 OSG 的虚拟化学实验平台的设计与实现, 设计与实现了一个智能的、可扩展的虚拟实验平台, 但是对于适合平台需要的关于化学药品、实验仪器以及整个实验的模型的建立问题没有解决和进行深入研究, 因此, 针对化学实验相关内容的建模工作具有很重要的作用。

虚拟现实建模技术是整个虚拟现实系统建立的基础, 模型准确度的高低, 模拟场景的真实与否,

作者简介: 谭同德(1950—), 男, 河南郑州人, 教授, 博士, 主要研究方向为计算机图形学、虚拟现实;
石奇波(1984—), 男, 河南濮阳人, 郑州大学信息工程学院, 硕士研究生, 主要研究方向为虚拟现实。

往往直接关系到应用实例的成败。基于此，在 Visual S tudio 2005 和开源图形渲染引擎 Open Scen e Graph（简称 OSG）环境下，针对初中化学实验进行了三维仿真模拟，并总结出对化学药品、实验仪器及整个实验过程建立的模型，最后，着重对反应现象及其实验过程进行的过程性建模进行了详细的阐述。

2 初中无机化学实验的分类

我们对初中所有的无机化学实验进行了分析，以反应方程式作为分类原则，按反应类型进行了分类，划分成五类，如下所示：

化合反应：是指由两种或两种以上的物质生成一种新物质的反应。例如，碳在氧气中的燃烧。

分解反应：是指化合反应的逆反应。它是指一种化合物在特定条件下（如加热、通直流电、催化剂等）分解成两种或两种以上较简单的单质或化合物。例如，碱式碳酸铜受热分解。

置换反应：指一种单质和一种化合物，生成另一种单质和另一种化合物的反应，可表示为： $A+BC\rightarrow B+AC$ 。例如，氢气还原氧化铜。

复分解反应：由两种化合物互相交换成分，生成另外两种化合物的反应，可简记为 $AB+CD\rightarrow AD+CB$ 。例如，碳酸钙与稀盐酸反应。

其他反应类型：除上面四种反应类型之外的无机化学反应。

3 建模

为了逼真模拟化学实验，要认真研究真实实验需要些什么、是如何做的。其一般步骤是：挑选药品、仪器；装载药品，连接仪器；然后根据需要，加入加热、通电等反应条件；如果反应能够发生，实验开始，同时伴随一系列物理化学现象，比如溶液变色、气体产生、沉淀生成等；另外基于不同的实验，要用到收集装置收集气体，或检验装置检验生成物；最后实验完成，按照实验本身的不同要求，对实验装置进行有序分离、清洁、放回。

仪器包括试管、铁架台、酒精灯、导管等，因不同的实验目的，被组装成为反应装置、收集装置、检验装置。药品则分固体、液体、气体，固体还可细分为颗粒、粉末、固块，这些药品因各自不同的颜色加以区分，做实验时还需要对它们的量进行控制。

这些仪器和药品，具有不同的几何形状、纹理、颜色、形态、质量和重量。每个实验仪器都有具体的尺寸大小，比如半径、长度、宽度等，还有纹理、颜色、材质、光照等属性，都可用几何信息表述，形态、质量、重量等物理属性，则可用物理信息表示，因此要为其建立几何模型和物理模型。药品同样要先对其进行几何建模、物理建模。由于做实验要用到不同的量，所以对于不同的量要对已建成的药品模型进行缩放操作。

试管可以装载药品，铁架台上放置酒精灯，导管需要进行连接，酒精灯经过点燃可以燃烧，药品需要有分子式，判断其参与的反应是否发生，气体需要有密度，与空气比较大小决定其向下或向上排空气的收集方法，随着仪器的摆动，药品溶液具有在重力作用下的流动行为，等等，这些信息都需要加入。经过分析、归纳，对仪器、药品建模的结构如下：

```
仪器{
    物理模型{几何构造；物理属性;}
    装配信息;//此模型与另外某模型可以连接
    行为信息;//比如酒精灯可以点燃，试管装载药品等
};
```

```

药品{
    物理模型{几何构造; 物理属性;}
    化学分子式;
    摩尔数量;
    颜色;//实验中好多药品可以通过气味辨别, 仿真中此处重点突出其颜色
    形态;密度; 溶解度;.....
};

```

接下来, 还需要分析以下内容:

(1) 反应条件: 比如加热、通电等;

(2) 反应方程式: 判断反应是否发生, 连接反应物、生成物、反应现象等之间的关系;

(3) 反应现象: 比如气体的生成、沉淀的产生、反应物的燃烧、溶液反应剧烈时出现沸腾等;

反应物到生成物的过程, 比如氢气还原氧化铜实验, 黑色氧化铜逐渐变为红色的铜, 外在体现的颜色变化;

(4) 反应过程: 实验过程中, 如不同摩尔量的反应物、加没加催化剂、是否中止反应条件和违规操作等不同情况将会导致出现的不同现象。

上面四类内容已经不是几何建模、物理建模等建模方法所能解决的, 这是由于化学实验其本身的特殊性, 动态性、连续性, 不是静态的、离散的, 它要根据实验环境变化而变化; 还有其物理化学现象, 如火焰、气体等, 这些模糊景物, 它们有生存时间、不固定的形态等, 我们都将归入为下面的化学建模, 其结构如下:

```

实验类{
    实验仪器节点{仪器;}
    反应物节点{药品;}
    反应条件节点{催化剂;加热、通电等反应条件;}
    方程式库节点;
    反应现象节点{
        质变节点{产生气体; 生成沉淀; 溶液变色等;}
        量变节点{反应物到生成物;}
    }
    操作节点{中止反应条件; 违规操作等;}
    .....
};

```

实验仪器节点和反应物节点分别对应前面的结构, 仪器和药品。

反应条件节点: 催化剂本身是药品, 同样对应药品模型、摩尔数量、分子式, 另外的作用为加剧实验反应, 所以同一实验加催化剂与否将对应不同的初始化常量数值; 加热、通电等反应条件则通常为必要条件, 有无就能决定反应是否发生, 它们的外在表现模型将和后面的反应现象火焰等所建模型对应的一样。

方程式库节点: 判断反应是否发生, 连接反应物节点、反应条件节点、反应现象节点等之间的关系。

反应现象节点: 如结构中所述, 有反应物到生成物的量变外在表现, 也有突然冒出的气体等给人质变感的现象。

操作节点: 检测实验过程中的操作, 有中止反应条件或违规操作时, 进行相应的处理, 比如重新给控制参数赋值, 以控制对应的现象。

4 过程性建模

过程性建模技术有如下三个重要的特性^[3]:

抽象：它们通常是动态的，是有多个因素作用的复杂过程，比如火焰，有生存时间，形态不定，受风力扰动等，因而将其生成细节往往抽象成一个算法或一套过程。例如基于动态随机性生长算法的粒子系统，基于语法规则算法的 L 系统，利用整体与局部自相似特性的分型迭代算法等。

控制参数：通过定义和调整参数，可在模型生成过程中直接对应于一个具体的行为改变。

灵活性：它可以捕获一个实体的本质，不用把它明确地界定在一个真实的世界里。参数可以改变产生多种多样的结果，而不一定局限对应于原始模型。

总之，过程建模技术是根据一系列的指令来描述实体，包括它的几何造型、纹理或作用等，而不是作为一个数据的静态块。用户直接操作的不是目标模型，而是采用命令行或者其他描述的方式对于所需要的模型或操作给出的描述。然后系统将用户给定的描述转换为内部指令，并做出相应的操作。

4.1 反应方程式库

反应方程式库是根据化学实验的分类来建的，这样就能根据反应物的数量、反应条件去匹配不同的反应方程式库，快速决定出是否发生反应，从而确定出对应的反应生成物、伴随的物理化学现象等。例如分解反应，反应物的数量是一，所以反应物数量是一的，就直接去匹配分解反应方程式库，再根据反应条件进一步缩小判断范围，从而快速判断出是否发生反应。

分解反应方程式的结构如下：

```
Struct 反应方程式{
    方程式 string[];
    反应条件  enum;
    反应物      string[];
    生成物      string[];
    是否生成气体 bool;
    是否发生沉淀 bool;
    各个反应物对应的方程式配平系数 int;
    各个生成物对应的方程式配平系数 int[];
    .....
};
```

另外，方程式所起的作用为：根据参加反应的各个反应物的摩尔数量，比较配平的反应系数计算出控制参数，用来控制反应持续的时间，生成气体及沉淀的多少，模型变换的时间等。

4.2 反应物到生成物的量变

实验中反应物在发生反应的过程中逐渐变为生成物，是利用过程性建模来实现的。反应物到生成物，有个量变过程，碳在氧气中燃烧时的变化，氢气还原氧化铜反应中氧化铜到铜的过程，都有反应物其外在颜色、形体的变化。不同于给人质变感的那些现象，比如突然冒出的气体、突然生成的沉淀等，根本不能看出怎么由反应物到这些现象的。

这个量变过程，本文采取了变换模型的方式予以实现，根据化学药品的数量、反应条件等得来的参数控制变换模型的时间。如图 1 所示，上下分别为碳和氧化铜在变化过程中用到的两组模型，上面一组球状模型为在盛有氧气的集气瓶中燃烧的碳，从左到右，经过燃烧，赤热发红的碳，逐渐由红色变淡变暗以至呈现黑色。下面为黑色氧化铜渐变为红色铜的一组更换模型。此外，另外有两种实现方式：根据参数变换模型纹理，或基于原始模型进行图形图像的混合变化。

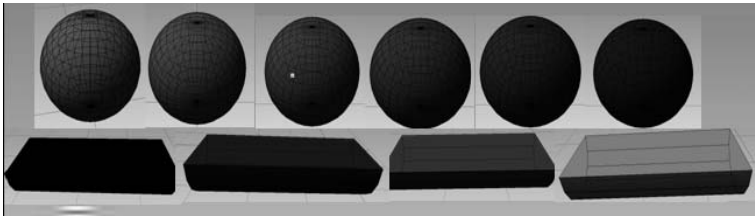


图 1 反应物到生成物的逐渐变化

4.3 整个实验的过程性建模

上面提到的只是实验过程中反应物到生成物部分的变化，其实每一个化学实验的整个过程也需要过程性建模，不同于动画，不是设定好一成不变的。反应是否发生，它要根据反应物、反应条件去判断，如果反应发生，就通过方程式确定生成物是什么，反应伴随的物理化学现象有哪些，反应物的摩尔数量决定了反应发生持续的时间、反应物的变化程度、气体或沉淀生成多少等，反应条件则去判断反应的剧烈程度。

控制参数计算方法如下：
设方程式配平时等式左边为： $aA + bB + cC + \cdots$
参加反应的反应物为： m 摩尔 A， n 摩尔 B， l 摩尔 C...
则控制参数为： $\min\{m/a, n/b, l/c, \cdots\}$
仍以氢气还原氧化铜为例进行说明： $H_2 + CuO = Cu + H_2O$ 。设参加反应的有 10 摩尔 H_2 和 3 摩尔 CuO ， H_2 是 10 比 1 得 10， CuO 是 3 比 1 得 3，控制参数则为最小比值 3，因此等式右边生成物对应系数乘以 3 就是各个生成物的数量，模型变换持续时间乘以 3 就是这个数量的模型变换时间，即反应发生持续时间，控制反应现象用到的也是这个参数乘以它们的初始化常量数值（见表 1）。不论反应持续时间、模型变更时间等都有各自的初始化常量数值，它乘以控制参数才作为最终的值。

表 1 初始化常量数值

实 验	催 化 剂	反应物到生成物（mol）	气体（ml/s）	沉 淀
氢气还原氧化铜	无	氧化铜到铜 15s		
氯酸钾制氧气	无	氯酸钾到氯化钾 20s	1	
	二氧化锰	氯酸钾到氯化钾 10s	2	

另外，反应条件的不同也使得同摩尔量的反应物发生反应的持续时间、反应现象的剧烈程度等不同。例如，在加热氯酸钾制取氧气的实验中，加入催化剂二氧化锰，反应会很快发生，平均时间段内氧气也产生得多，那么必须对有没有催化剂的不同情况设置不同的初始化常量数值。有催化剂时，相同摩尔量的反应物反应快，持续时间短，反应物变成生成物的时间也短，但是产生气体的化学现象要剧烈得多。因此设置的初始常量数值如表 1 所示：有催化剂时，1mol 氯酸钾变成氯化钾为 10s，1s 生成的氧气为 2mL；没有时则分别为 20s、1mL。需要补充一点，此处常量设置并不是实际测量得来的，而是灵活于实际实验的原始模型，考虑了人的感知因素，假设真实实验发生反应需要 20min，那么仿真实验就不会盯着屏幕同样久，可设 5min。

正常操作下如此，若有违规操作，就要有所变化。仍以氢气还原氧化铜为例，若加热时间过短就熄灭酒精灯，那么就要根据已反应的持续时间，与先前计算的反应持续时间的比值作为参数，乘以先前计算得来的控制参数作为实际控制参数。虽然化学药品的量没少，但是由于反应条件不再具备了，就要对此做出符合实际的变动，反应理应发生一部分。如果酒精灯熄灭后，就直接撤去通氢气导管，那么反应也要做出相应的变化，被还原的氧化铜一部分重新被氧化。

5 化学实验反应现象的实现

所做虚拟化学实验，反应条件及反应现象中出现了火焰、气体、气泡、沉淀等，它们形态飘忽不定，受到风力扰动，随时间有消失有生成，这些模糊景物都有个变化过程，我们对其进行了过程性建模，具体实现用的是 OSG 中的粒子系统。

OSG 全称 Open Scene Graph，直译为开源场景图。OSG 包含了一系列的开源图形库，主要为图形图像应用程序的开发提供场景管理和图形渲染优化的功能。它使用可移植的 ANSI C++ 编写，并使用已成为工业标准的 OpenGL 底层渲染 API，OSG 是一种中间件，它提供了高性能 3D 程序所需的空间数据组织能力及其他特性^[4]。

粒子系统基本思想是用许多形状简单且赋予生命的微小粒子作为基本元素来表示物体的。侧重于物体的总体形状和特征的动态变化，把物体定义为许多不规则的、随机分布的粒子，而每个粒子均有一定的生命周期。随时间推移，旧粒子不断消失，新粒子不断加入，借助于粒子的出生、成长、衰老和死亡过程，较好地反应模糊物体的动态特性。同时，与粒子有关的每一个参数均受到一个随机过程的控制，以规定粒子在系统中的形状、特征及运动。

在 OSG 中提供有专门的粒子系统工具，名字空间为 `osgParticle`（本文本小节后面所提的类，为书写简便，名字空间都省去了），OSG 对经常使用的粒子模拟都做了专门的类，如 `ExplosionEffect` 用于爆炸的模拟，`FireEffect` 用于火的模拟，`ExplosionDebrisEffect` 用于爆炸后四散的颗粒模拟等。这些类使用起来极其方便，如图 2 所示酒精灯火焰，就是用 `FireEffect` 类进行的模拟，首先定义个 `FireEffect` 类，然后调用它的成员函数进行相关的设置，即可达到所需效果，比如 `setScale`、`setStartTime`、`setWind`、`setTextureFileName` 等成员函数就是对火焰大小、开始时间、风力作用、火焰纹理等进行设置的。



图 2 OSG 自带粒子系统模拟的酒精灯火焰

OSG 自带的一些特效有限，再者，它们有时不能很好地模拟化学实验中某些反应现象，此时，就需要自己定义粒子系统。

在加热的条件下，浅绿色的碱式碳酸铜分解成黑色的氧化铜，水和二氧化碳，通入后面的检验装置后，试管中澄清石灰水变浑浊。生成沉淀的过程就用自定义粒子系统实现的。沉淀生成过程有点类似于天空中下雪的场景，首先一些雪花，然后慢慢多起来，随后出现模糊浑浊，最后试管底部有些许堆积，所以定义的粒子系统是仿照 OSG 自带的下雪粒子系统建立的。

- OSG 中定义粒子系统一般有以下几个步骤：
- 第一步：确定意图（包括粒子的运动方式等诸多方面）。
 - 第二步：建立粒子模板，按所需要的类型确定粒子的角度（该角度一经确定，由于粒子默认使用有 `Billboard` 所以站在任何角度看都是一样的），形状（圆形，多边形等），生命周期等。实现类为：`Particle` 粒子模板，决定粒子的大小，颜色，生命周期，等等。

第三步：建立粒子系统，设置总的属性。实现类为：`ParticleSystem` 粒子系统的总体属性，粒子总数，纹理等。

第四步：设置发射器（发射器形状，发射粒子的数目变化等）。实现类分别为：Counter 粒子产生的数目范围。Placer 粒子出生点的形状状态，如环形，圆形，点形。Shooter 粒子发射器，决定粒子的初速。Emitter 上述模板及操作都为这个类服务，类名：发射器。

第五步：设置操作（旋转度，风力等因素）。实现类分别为：Program 可接受对粒子的操作，如轨迹的定义，矩阵的变换等。Operator 粒子操作或用户自定义粒子操作。

第六步：加入节点，更新。

编程实现效果如图 3 所示，左面部分是整个实验的装置，右面部分是试管中生成沉淀的放大视图。

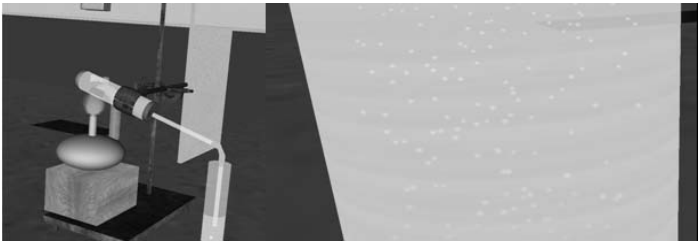


图 3 自定义粒子系统模拟的沉淀

6 实验结果

图 4 是碳酸钙与稀盐酸的复分解反应，生成氯化钙、水和二氧化碳，左面部分集气瓶上面的小火焰，是用来检验里面的二氧化碳气体的，放入集气瓶后，小火焰自动熄灭。

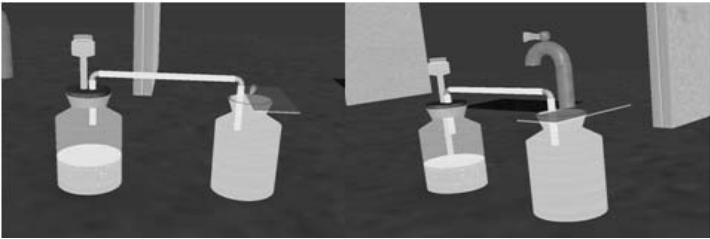


图 4 碳酸钙与稀盐酸的复分解反应

图 5 是氢气还原氧化铜的置换反应，第一部分中左端是氢气的生成装置，先通入氢气，把导管中的空气清干净，这样能达到更好的还原效果；第二部分酒精灯点燃，氢气与氧化铜的置换反应开始，逐渐由黑色的氧化铜变成红色的铜；第三部分是酒精灯熄灭，现在氧化铜已经被完全还原成铜，氢气继续通入直到被还原的铜冷却；第四部分停止通入氢气，实验结束。

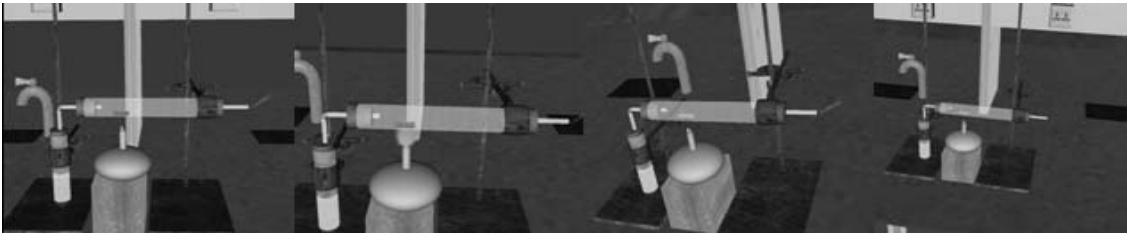


图 5 氢气还原氧化铜的置换反应

7 总结

首先把实验进行了分类，并对每类一个典型示例进行了模拟，起到了应有作用，它使得初中化学

课本中的所有实验都有了类的归属，轻松便可仿照此类所建示例进行模拟；接下来分析了实验内容，分别对仪器和药品进行了建模，考虑了其几何信息、物理信息及化学信息，进而对建立的化学实验模型进行了阐述，内容节点式的划分，很好地理清了它们之间的对应关系；由于化学实验本身的特殊性，动态和连续，我们重点对气体、沉淀等化学反应现象及些许反应条件，比如点燃、加热等和整个实验的控制过程进行了过程性建模，使得模拟的实验更加灵活、生动。存在的不足为：化学仪器、实验药品、反应现象等在逼真度上稍许欠缺；过程性建模过程中，对影响控制参数变化的各个因素，考虑的不太全面。

参考文献

[1] 朱福全, 杨丽平, 李昌国, 谭良, 杨春. 基于 OpenGL 的虚拟化学实验系统的研究[J]. 计算机工程与设计 20 08, 29(3): 723-724.

[2] 张艳丽, 谭同德, 赵新灿, 郭志敏, 基于 OSG 的虚拟化学实验平台的设计与实现[C]. 中国江苏淮安, 第三届全国教育游戏与虚拟现实学术会议论文集 2009 , 75-81.

[3] 郭武, 过程建模技术中若干问题的研究[D]. 吉林大学博士学位论文, 2008010.

[4] [美]Paul Martz 著, 王锐, 钱学雷, 译. OpenSceneGraph 快速入门指导.

[5] 王乘, 周均清, 李利军编著, Creator 可视化仿真建模技术[M]. 华中科技大学出版社.

[6] 李荣辉, 三维建模技术在虚拟现实中的应用研究[D]. 大庆石油学院硕士学位论文, 20070312.

[7] 周伟光, 利用 CAD 数据的虚拟现实视景高效建模技术研究[D], 南京航空航天大学研究生院民航学院硕士学位论文, 200703.

[8] 罗如为, 三维场景漫游与全景拼接的关键技术研究[D]. 贵州大学硕士学位论文, 20070301.

[9] 王红梅, 虚拟实验室的研究与实现[D]. 中国海洋大学硕士学位论文, 20070602.

[10] Olga De Troyer, Frederic Kleinermann, Bram Pellens, and Wesley Bille, Conceptual Modeling for Virtual Reality, Tutorials, Posters, Panels and Industrial Contribution at ER 2007.

[11] Pellens B, De Troyer O, Bille W, Kleinermann F, Romero R, An ontology-driven approach for modeling behavior in virtual environments, On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshop . OTM Confederated International Workshops and Posters AMeSOMe, CAMS, GADA, MIOS+INTEROP, ORM, PhDS, SeBGIS, SWWS, and WOSE 2005. Proceedings (Lecture Notes in Computer Science Vol.3762), 2005 : 1215-24.

知识合作在计算机类课程教学中的应用研究

赵妍, 李玲玲

(郑州航空工业管理学院计算机科学与技术系, 河南 郑州, 450015)

摘要: 本文提出了知识合作教学模式的内涵, 并总结了应用该教学模式过程中应该注重的几个环节, 探讨了知识合作教学模式对培养学生的创新能力、自学能力、独立意识和协作精神的促进作用。

关键词: 知识合作; 团队精神; 教学模式

中图分类号: G642.0

文献标识码: A

Application and Research of A Teaching Model Based on Knowledge Cooperation

ZHAO Yan, LI Lingling

(Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou 450015, Henan China)

Abstract: The meaning of teaching model based on knowledge cooperation is proposed in this paper, and “Network Technology” course is used teaching case, research promoter action of teaching model based on knowledge cooperation about training innovation ability、self-study ability、independent consciousness and team spirit of student.

Keywords: knowledge cooperation;team spirit ;teaching model

1 引言

随着高等教育最近十年跨越式的发展, 我国高等教育在增加“量”的同时一直在强调“质”的提高。教学质量的提高首先要求教学模式的变化, 不但要适应不同学生群体, 还要适应社会科技的发展带来的就业压力, 对于与计算机相关专业的学生这个问题显得尤为突出^[1]。首先是计算机技术所涉及的知识领域很广, 很多学科之间都有相互交叉的部分, 因此计算机公共基础课与公共选修课逐渐增多; 其次计算机技术的发展很快, 几乎每半年就有新的技术与产品的产生, 这就要求我们的学生必须得上市场的发展。本文作者从分析专业培养目标与社会需求变化等相关问题入手, 结合应用性本科教育的系统性和应用性特点, 通过对计算机类课程教学目标的分析和几年来的教学经验的总结, 提出一套基于知识合作的教学模式并应用于实际教学中。

2 知识合作教学模式的含义

知识合作 (Knowledge Collaboration , KC) 是一种提升项目团队工作效率的重要手段, 已经成为知识管理领域研究的热点之一。它是指企业通过整合组织的内外知识资源, 使组织学习、利用和创造知识的整体效益大于各独立组成部分总和的效应^[2]。

计算机类课程重于实践应用^[3], 我们可以把该方法应用于教学当中。一方面教师教授学生学习的过程其实就是一个团队合作的过程。在学校大力倡导培养学生团队合作精神的同时, 我们为何不将其引入教师教学过程之中呢? 另一方面, 知识合作要求教师把知识整合成一个体系, 该体系不是一些孤立的知识点, 它的理论部分包括前导和后继课程的知识, 这是一个纵向知识层面; 它的实践部分包

作者简介: 赵妍 (1979—), 辽宁鞍山人, 硕士, 讲师, 主要研究方向为计算机网络、数据库。

括引导式实践和自主式实践，这是一个横向知识层面。在纵横结合的知识结构体系中，学生是以团队合作方式进行学习。

基于知识合作的教学模式的根本是以“学生为本”，知识需要合作，教师需要合作，学生需要合作，通过合作以保证教师的教学质量，提高学生的学习效率。因此该教学模式从三个方面来理解。

1) 内容合作

我们提出的知识合作是指教师教授的知识内容应该是一个系统的、全面的知识体系，要从学生的整个学习阶段用发展的眼光来组织知识。

2) 教师合作

作为教师在培养学生团队合作精神的同时，自身的团队精神也是重要的，因为教师是教师的同时也是一名学生，也需要不断地学习。教师也是从不同的学校毕业为了自己的梦想走到一起，每个人在自己的知识领域、认知方式、思维理念都有着很大的不同，通过相互沟通和学习提高自身教学盲点^[4]。教师合作体现在多个方面，例如互相听课、科研小组、教研讨论等，最近比较流行的拓展训练也是一种培养团队精神的活动。通过教师之间的合作，教师首先对自我要有一个清楚的认识，认识自身潜能、增强自信心、改善自身形象、克服心理惰性，同时启发想象力与创造力，提高解决问题的能力；其次要认识群体的作用，增进对集体的参与意识与责任心，并且改善人际关系，更为融洽地与群体合作。

3) 学生合作

学生合作是该教学模式的核心思想也是最终目标，如果说内容合作和教师合作是该教学模式的重要组成部分和必要前提，那么最终的目标就是提高学生的知识技能的同时增强学生的团队合作精神。作为一名高校教师必须清楚地认识到所面对的学生是“80 后”或“90 后”的一代，他们有他们这一代青年的鲜明特点，可能与教师自身学习与成长过程截然不同，时代赋予了他们活跃的思想和独立的个性^[4]，同时由于独生子女的增多，相互帮助与合作的思想受到一定的限制，因此培养“团队精神”一直是教育界所关注的热点问题。

在教学过程中教师应该逐渐把“团队精神”融入其中，通过课堂讨论、课后作业、课程实验等教学内容形式，把学生分成“团队”，进行合作学习。这里需要强调“团队”与“部门和小组”的不同^[3]。部门和小组的一个共同特点是：存在明确内部分工的同时，缺乏成员之间的紧密协作。团队则不同，队员之间没有明确的分工，彼此之间的工作内容交叉程度高，相互间的协作性强。

3 教学过程中几个重要环节的总结

为了更好地实现知识合作的教学目标，在教学组织过程中教师应该重点考虑以下几个环节的工作：

(1) 差异分配，相互合作。知识合作教学强调差异分配，实践过程中首先由学生自由组合，然后根据学生的性别、学习能力、学习兴趣等各方面差异进行调整，将具有不同优势的学生分配到不同的团队中。差异分配在一定程度上避免随意分配造成学习能力相差不多的学生扎堆，使得一部分学习能力和学习兴趣较弱的学生失去相互学习、相互督促的氛围。

团队中个体的相互合作是通过相互依赖完成的。以往的团队合作过程中，由于分工、考核及成员间协调等原因，很可能出现本应由成员共同承担的任务，往往变成少部分成员的工作，而另一部分成员则并未完全参与其中，达不到共同进步的效果。该教学过程强调构建团队个体之间的相互依赖，每个成员都要为自己所在团队的其他成员的学习负责^[5]。为达到这一目的，在实践中，我们从学习目标、奖励方式、角色互换等几个不同角度训练和鼓励学生培养积极的相互依赖。

(2) 以引导为主的师生互动。虽然知识合作教学强调学生主动学习，但教师仍然是教学过程的组织者、指导者和促进者。在实践教学过程中，当出现问题时，教师首先引导学生在团队内进行讨

论，当问题超出团队成员能力范围时，教师也不急于给出答案，而是鼓励学生向其他团队求助。当该问题具有一定普遍性或者一定难度时，再由教师组织集体答疑。这种以引导为主的师生互动，能够更有效地激发学生学习兴趣，引导学生自己去完善知识，纠正错误，有利于学生真正成为学习的主体。

（3）合理安排团队合作的频率和时间。在教学实践中，在出现重、难点，或者某些学生出现疑惑或者学生意见不统一时，都可能是教师引入团队合作方式的时机^[6]，但由于实际课时等因素的限制，若太过频繁地组织讨论，则不但可能拖延教学进度，也有可能因为学生学习能力不一致，影响部分学生的学习效果，使得学习较好的学生觉得内容太浅，学习较差的学生觉得处处是难题。我们可以考虑增加团队合作的课外准备工作量，把一部分工作让学生利用课余时间完成，而教师只在课堂上对完成结果进行点评或指导。这样可以更合理地组织和安排教学内容，避免产生为了合作而合作的机械学习。

（4）团队精神的培养。知识合作教学除了完善学生对知识的认识外，团队合作精神的培养、合作技能的训练也是教学目标之一。在实践教学过程中，教师注意引导学生彼此的认可和相互的信任；学会准确的交流；学会彼此理解和支持；学会建设性地解决问题等。

4 结论

知识合作是计算机类课程改革所追求的一种崭新的教学模式，它不仅重视教师与学生的交流，而且强调学生之间的影响力，是对任务驱动、案例法等其他教学模式的有力补充。在下一步的教学改革任务中，我们要进一步明确知识合作的重要性，更好地发挥知识合作在实际教学中的主观能动性。

参考文献

[1] 闫丽.大学计算机基础教育改革探究[J]. 中国成人教育, 2007(8): 134-135.

[2] 白杨, 蔡杰, 张素娟. 一种具有最优知识合作效应的项目团队组建方法[J]. 大连轻工业学院学报, 2007, 26(2): 185-188.

[3] 杨安庆. 基于实验教学的大学生实践能力培养探究[J]. 光学技术, 2007, 33(11): 423-425.

[4] 张宝歌. 高等学校学生质量评价体系的研究[J]. 黑龙江高教研究, 2007, 164(12): 69-71.

[5] 盛群力. 五星教学模式对课程教学改革的启示[J]. 教育发展研究, 2007, 12: 33-35.

[6] DENG Han hui, L IU Fan, ZHAO Wen wen. Rising developments and challenges of american entrepreneurship education[J]. China Youth Study, 2007 (09): 74-76 .

协议分析软件在计算机网络教学中的应用

张巧荣, 郑娅峰

(河南财经学院 计算机与信息工程学院, 河南 郑州, 450002)

摘要: 通过分析计算机专业网络课程的课程特点及教学现状, 提出将协议分析软件贯穿于整个网络课程教学中的思路, 并给出了协议分析软件在计算机网络课程教学中的具体应用, 实践证明协议分析软件对改进计算机网络课程的教学方法具有非常重要的参考价值。

关键词: 计算机网络教学; 协议分析软件; 教学方法

中图分类号: G642

文献标识码: B

The Application of Protocol Analysis Software in the Computer Network Teaching

ZHANG Qiaorong, ZHENG Yafeng

(Henan University of Finance and Economics, Zhengzhou 450002, Henan China)

(College of Computer and Information Engineering, Henna University of Finance and Economics, Zhengzhou 450002, China)

Abstract: The characteristic and status of the teaching of computer network are analyzed in this paper. The idea and the process of using protocol analysis software in the teaching of computer network are proposed. It has been proved that the protocol analysis software has great valuation on improving the teaching method of computer network.

Keywords: computer network teaching; protocol analysis software; teaching method

计算机网络课程是计算机软件、应用、网络实验技术、信息管理、电子商务等专业的重要基础课程。而计算机专业网络的实验教学具有概念理解抽象的特点, 强调对网络理论知识的理解, 尤其是贯穿整个课程的层次协议的原理与实现更是重中之重^[1~4]。

网络实验协议分析软件是一种计算机网络实验调试和数据包嗅探软件, 通过对网络实验数据包的分析确定问题, 常应用于网络实验管理中的故障修复^[5,6]。

由于协议分析软件能够捕获各种网络实验协议的数据包并进行解析, 如果在计算机网络实验课程的网络实验协议教学中应用协议分析软件, 就能够剖析网络实验通信的整个过程, 明确 TCP/IP 网络实验协议在整个通信过程中所起的作用, 能使枯燥的理论教学变成生动的实例解析, 从而让学生更加容易认识掌握网络实验协议在计算机网络实验通信过程中的重要角色。

1 计算机网络教学的现状^[1]

在整个网络教学过程中, 由于网络实验知识的特殊性, 造成学生对课堂讲授的知识一知半解, 不能深刻理解其核心理论。许多高职院校及面向非计算机专业的教师在不断的教学实践中也关注了这个问题, 并进行了有效的教学探索 and 改革。整体而言, 在非计算机专业的网络实验课程教学中淡化基础原理的讲解, 因材施教, 强化与应用联系较紧密的部分, 例如, 如何架设一个 Web 服务器; 如何托管 Web 服务; 如何组建一个局域网; 如何在局域网中实现共享等, 进行了有效的教学改革和探索^[7~9]。

然而对于计算机专业的学生来说, 以开放互连参考模型或 TCP/IP 参考模型为主导的网络实验原

作者简介: 张巧荣 (1978—), 女, 讲师, 硕士;
郑娅峰 (1979—), 女, 讲师, 硕士。

理论知识是不可回避的一个问题。纵观国内外教材，对整个网络实验协议主要是围绕协议报文结构的讲解，协议工作机理的讲解展开的。这部分内容由于涉及网络实验底层，定义比较复杂和抽象，而学生接触较多的是网络实验应用软件，因此学生对这部分知识没有直观的认识，造成对理论学习比较排斥，在做实验的阶段也无法有效地将已学理论和实验结合起来。

协议分析软件虽然是在网络实验故障排除中常用的管理工具，但是由于这种工具能够动态捕获网络实验流量，因此，如果将这种软件与计算机网络实验的理论教学结合起来，让学生能够真正看到网络实验中的数据包，把抽象的东西以直观的形式表现起来，能够加强学生对基础理论的理解，促进学生的学习兴趣。

2 协议分析软件的作用

协议分析软件具有重要作用：跟踪网络实验状况、识别并解决故障^[6]。它通过捕捉流经本主机及局域网网络实验环境中的所有数据包，然后对这些捕捉的数据包进行上层分析，进而得出关于网络实验流量等信息，为网络实验管理员做出某些决策提供证据。一般网络实验管理委员会会在服务器上运行一个协议分析软件，通过捕捉数据包，然后进行详尽的分析，就可以得到一些信息，例如，哪些主机在和服务器网络实验通信，这些主机分别在服务器上做了哪些操作。得到这些信息后，系统管理员就可以做一些决策，使得服务器更加安全。协议分析软件和 TCP/IP 协议栈的关系非常紧密，大多数协议分析软件的实现是严格按照 TCP/IP 协议栈的层次关系实现的。一般来说，协议分析工具可以解析的各种协议在 TCP/IP 协议栈中都有定义和实现，现在最常用的 Wireshark（原名 ethereal）工具支持可达 500 种协议。

目前常用的协议分析软件有 sniffer pro，wireshark，tcpdump 等几种。Sniffer 软件是 NAI 公司推出的功能强大的协议分析软件，具有捕获网络实验流量进行详细分析、实时监控网络实验活动、利用专家分析系统诊断问题、收集网络实验利用率和错误等强大功能。但其功能复杂，对初学者使用有困难。Wireshark 是一个开放源码的网络实验分析系统，也是是目前最好的开放源码的网络实验协议分析器，支持 Linux 和 Windows 平台。借助这个程序，可以直接从网络上抓取数据进行分析，也可以对由其他协议分析工具抓取后保存在硬盘上的数据进行分析。并且能够交互式地浏览抓取到的数据包，查看每一个数据包的摘要和详细信息。Tcpdump 是 Linux 平台下一个很重要的抓包工具，Tcpdump 提供了源代码，公开了接口，因此具备很强的可扩展性，对于网络实验维护和入侵者都是非常有用的工具。这些软件都可以自由选择，在实际的教学过程中，我们选择 Wireshark 进行网络实验课程的辅助教学。

3 利用 Wireshark 进行网络课程教学的案例

3.1 案例一：http 协议传输分析

使用协议分析软件进行协议学习，可以使理论和实践更好地结合。下面介绍一下利用 Wireshark 进行 http 协议分析的过程。

3.1.1 针对 HTTP 协议的 TCP/IP 的分层结构

HTTP 协议是应用层协议，处于最高层，其通过下层传输层的 TCP 进行可靠连接，网络实验层 IP 选路，链路层 Ethernet II，最后在物理上以位（Bit）进行传输。其分层结构如下所示。

应用层——HTTP。

传输层——TCP。

网络实验层——IP。

链路层——Ethernet II。

3.1.2 HTTP 协议工作原理

HTTP 协议是基于请求/响应模式的。一个客户机与服务器建立连接后，发送一个请求给服务器，请求方式的格式为：统一资源标识符（URL）、协议版本号，后边是 MIME 信息包括请求修饰符、客户机信息和可能的内容。服务器接到请求后，给予相应的响应信息，其格式为一个状态行，包括信息的协议版本号、一个成功或错误的代码，后边是 MIME 信息包括服务器信息、实体信息和可能的内容。

3.1.3 数据包分析

进行捕获数据包的分析，通过分析理解 HTTP 协议如何进行建立连接，发送请求信息，发送响应信息，关闭连接。

1) 抓取 http 协议数据包

启动 ethereal 后，选择菜单 Capture → Start。然后通过浏览器访问一台服务器地址（如 www.Baidu.com，抓包前注意把系统缓存清空），当页面显示出来时，单击 Stop 按钮，抓的包就会显示在面板中，如图 1 所示。

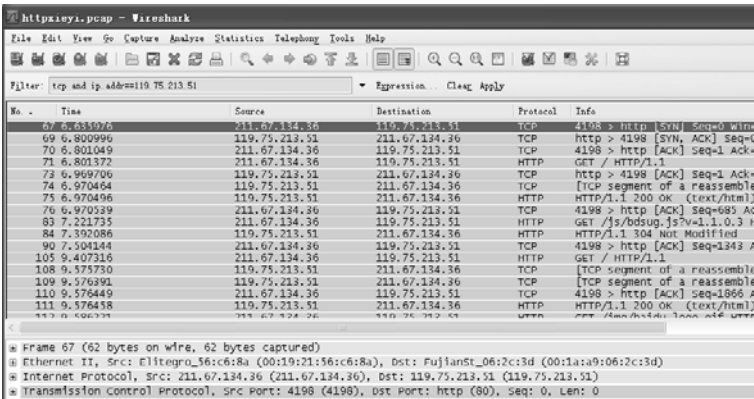


图 1 利用 Wireshark 抓取 HTTP 数据包

2) 建立连接过程

编号为 67、69、70 的数据包是 HTTP 协议使用下层 TCP 协议通过三次握手原则建立连接的过程。通过这几个数据包的分析（如图 2 所示），可以给学生讲解 HTTP 通信是发生在 TCP 协议之上，Web 服务使用的默认端口是 80 端口，所以 HTTP 是一个可靠的协议。还可以给学生讲解 TCP 建立连接的过程中标志位的作用，使学生了解到三次握手过程。在图中可以明确看到客户端向 Web 服务器发送一个 SYN 同步连接请求，syn 标志位被置为 1。Web 服务器收到请求后向客户端发送一个 SYN/ACK 数据包，同意客户端的连接并向客户端发起同步，客户端收到该数据包后再次确认，从而成功建立 TCP 连接。

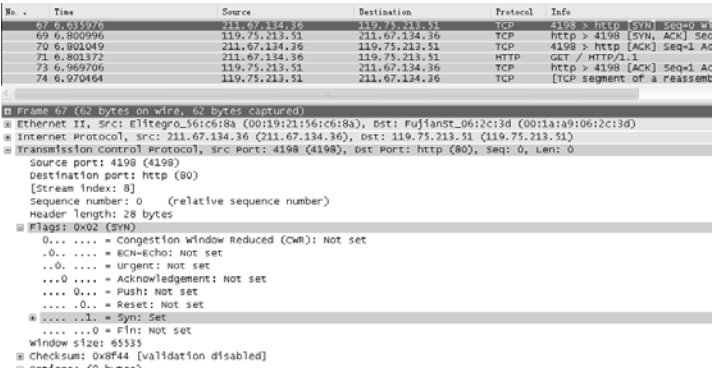


图 2 建立 TCP 连接的数据包

3) 发送请求信息

第 71 号数据包是建立连接后向服务器发出 http 请求的数据包，如图 3 所示，从图中可以看到数据包解析中 HTTP 协议发送请求信息的一些特征。

No. .	Time	Source	Destination	Protocol	Info
67	6.635976	211.67.134.36	119.75.213.51	TCP	4198 > http [SYN] Seq=0 win=6
69	6.800996	119.75.213.51	211.67.134.36	TCP	http > 4198 [SYN, ACK] Seq=0
70	6.801049	211.67.134.36	119.75.213.51	TCP	4198 > http [ACK] Seq=1 Ack=1
71	6.801372	211.67.134.36	119.75.213.51	HTTP	GET / HTTP/1.1
73	6.969706	119.75.213.51	211.67.134.36	TCP	http > 4198 [ACK] Seq=1 Ack=6
74	6.970464	119.75.213.51	211.67.134.36	TCP	[TCP segment of a reassembled

Frame 71 (738 bytes on wire, 738 bytes captured)

Ethernet II, Src: Elitegro_56:c6:8a (00:19:21:56:c6:8a), Dst: FujianSt_06:2c:3d (00:1a:a9:06:2c:3d)

Internet Protocol, Src: 211.67.134.36 (211.67.134.36), Dst: 119.75.213.51 (119.75.213.51)

Transmission Control Protocol, Src Port: 4198 (4198), Dst Port: http (80), Seq: 1, Ack: 1, Len: 684

Hypertext Transfer Protocol

GET / HTTP/1.1\r\n

Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, application/x-shockwave-flash, application/vnd.ms

Accept-Language: zh-cn\r\n

UA-CPU: x86\r\n

Accept-Encoding: gzip, deflate\r\n

[truncated] User-Agent: Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Mozilla/4.0 (compatible Mozilla/4.0

Host: www.baidu.com\r\n

Connection: Keep-Alive\r\n

Cookie: BAIDUID=BC61500204E09F93EA9E9F50E7AFDF:FG=1; BDLFONT=0\r\n\r\n

图 3 HTTP 发送请求的数据包

从图 3 中，可以看到 http 的报文格式。根据图示讲解 http 请求报文，指出如下报文的作用。

HTTP Command: //方法字段，说明其使用的是 GET 方法

URI: / //URL 字段，发送请求至保存该网站的服务器

HTTP Version: //http 协议版本字段，用的是 http/1.1 版本

Accept: //指示可被接受的请求回应的介质类型范围列表

Accept-Language: //限制了请求回应中首选的语言为简体中文，否则使用默认值

Accept-Encoding: //限制了回应中可接受的内容编码值，指示附加内容解码方式为 gzip,deflate.

User-Agent: //定义用户代理，即发送请求的浏览器类型为 Mozilla/4.0

Host: www.baidu.cn\r\n // 定义了目标所在的主机

Connection: Keep-Alive\r\n // 告诉服务器使用持久连接

Cookie: //指明生成的 Cookie 码

4) 响应信息

第 75 个数据包是服务器的响应包信息，从图 4 可以看出，服务器在处理完客户的请求之后，要向客户机发送响应消息。在服务器给的回应请求中，可以从状态码中看到访问的相关信息。

No. .	Time	Source	Destination	Protocol	Info
67	6.635976	211.67.134.36	119.75.213.51	TCP	4198 > http [SYN] Seq=0 win=6
69	6.800996	119.75.213.51	211.67.134.36	TCP	http > 4198 [SYN, ACK] Seq=0
70	6.801049	211.67.134.36	119.75.213.51	TCP	4198 > http [ACK] Seq=1 Ack=1
71	6.801372	211.67.134.36	119.75.213.51	HTTP	GET / HTTP/1.1
73	6.969706	119.75.213.51	211.67.134.36	TCP	http > 4198 [ACK] Seq=1 Ack=6
74	6.970464	119.75.213.51	211.67.134.36	TCP	[TCP segment of a reassembled
75	6.970496	119.75.213.51	211.67.134.36	HTTP	HTTP/1.1 200 OK (text/html)
76	6.970339	211.67.134.36	119.75.213.51	TCP	4198 > http [ACK] Seq=685 Ack=1

Frame 75 (587 bytes on wire, 587 bytes captured)

Ethernet II, Src: FujianSt_06:2c:3d (00:1a:a9:06:2c:3d), Dst: Elitegro_56:c6:8a (00:19:21:56:c6:8a)

Internet Protocol, Src: 119.75.213.51 (119.75.213.51), Dst: 211.67.134.36 (211.67.134.36)

Transmission Control Protocol, Src Port: http (80), Dst Port: 4198 (4198), Seq: 1421, Ack: 685, Len: 533

[Reassembled TCP segments (1953 bytes): #74(1420), #75(533)]

Hypertext Transfer Protocol

HTTP/1.1 200 OK\r\n

Date: Fri, 18 Sep 2009 06:00:16 GMT\r\n

Server: BWS/1.0\r\n

Content-Length: 1745\r\n

Content-Type: text/html\r\n

Cache-Control: private\r\n

Expires: Fri, 18 Sep 2009 06:00:16 GMT\r\n

Content-Encoding: gzip\r\n\r\n

Content-encoded entity body (gzip): 1745 bytes -> 3509 bytes

Line-based text data: text/html

[truncated] <html><head><meta http-equiv=Content-Type content="text/html; charset=gb2312"><title>\260\331\266\310\

[truncated] <body><div id=u>a href=http://passport.baidu.com/?login&tpl=mn>\265\307\302\274</div><center><fr

图 4 HTTP 请求响应数据包

3.2 案例二：网络实验安全原理分析

在讲述计算机网络故障检测时，协议分析软件是一个有力的工具。教师也应该在平时注意收集病毒感染环境的样本，以便在课堂中进行有效的案例教学。下面就以 arp 病毒攻击为例，进行教学示范。首先教师要讲清楚 arp 协议的基本原理及 arp 病毒的攻击特点。

在网络出现符合 arp 病毒攻击的一些特征时，可开启 Wireshark 进行捕获数据包操作。选择 Capture→start 进行抓包，大约 10s 后单击 Stop 按钮停止。下面图示为抓到的数据包。在下面的数据包中可以看到大量的广播（broadcast）数据包，协议类型为 ARP，如图 5 所示。

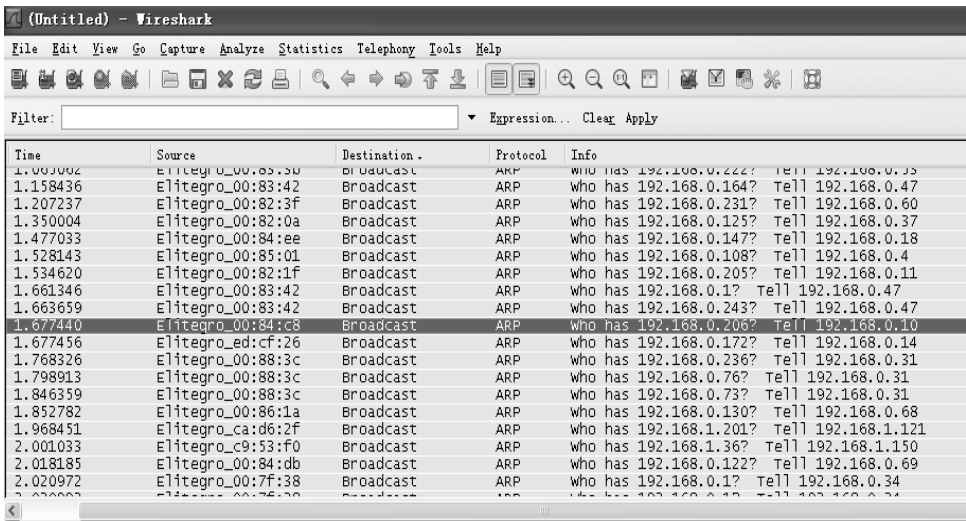


图 5 ARP 攻击环境中的数据包

为进一步进行分析，可要求学生对所捕获数据包进行统计分析。例如，可使用 Statistics→protocol Hierarchy 进行数据包按协议分类，如图 5 所示。依据图中数据引导学生进行分析。在图 6 所示的样本数据中可看到，总抓包量为 4292 个。其中，ARP 数据包有 3921 个，占总抓包量的 91.36%。在正常的网络实验环境中是不会出现的，因此，定位异常情况为过量的 ARP 数据包占用网络实验带宽，由此可判断网络实验内部及其感染了 arp 病毒的可能性，如图 6 所示。

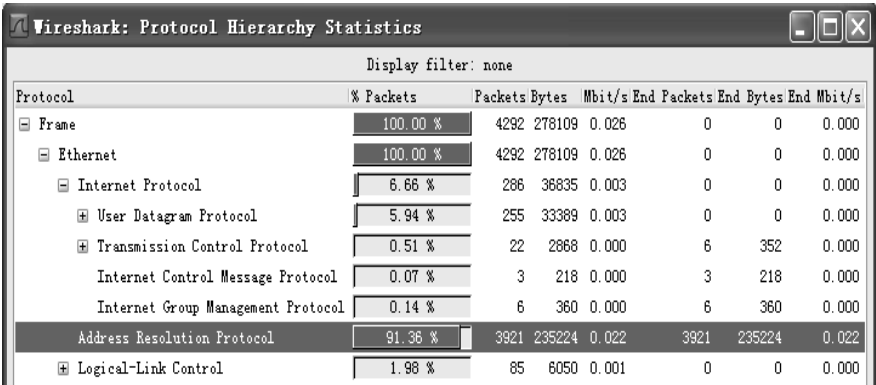


图 6 数据包按协议分类的统计分析

那么寻找感染机器源头成为下一步的重要工作。使用 filter 过滤所有 arp 数据包，然后继续使用 statistics→coversations 可对各个源地址发出的 arp 数据包进行分类统计。如图 7 所示，图中显示多数机器都向外发出了广播信息，尤其以 192.168.0.71 为多。

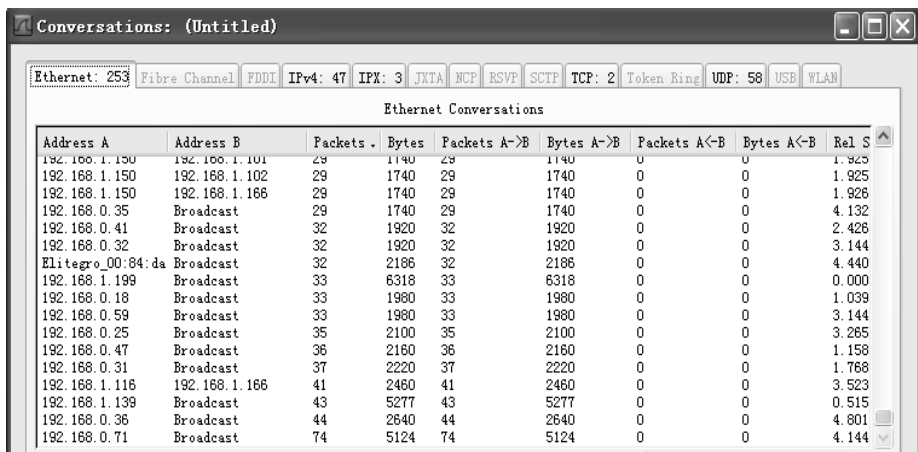


图 7 数据包基于源地址和协议类型的统计分析

4 结束语

本文提出将协议分析软件应用与计算机专业的网络实验课程教学活动中，学生通过对网络实验数据包的跟踪和分析，了解计算机网络实验技术的基本概念、原理。通过协议分析软件，改善计算机网络实验课程的教学工作，提高学生对本学科的理论水平和学习兴趣。同时，这些工具和教学方法可以促进学生在课内外时间对计算机网络实验进行更深入的学习。

参考文献

- [1] 畅卫功, 张爱华. 计算机网络实验教学的研究与探讨[J]. 实验室科学, 2009(4): 108-110.
- [2] 陈旭日, 尹向东. 计算机网络实验类实验教学改革与实践[J]. 实验科学与技术, 2005, 12(4): 77-79.
- [3] 沈军. 对本科“计算机网络实验”课程教学的一点认识和思考[J]. 计算机教育, 2008, (22): 105-107.
- [4] 刘利强, 陈凯文, 周细义. 计算机网络实验教学的改革与实践[J]. 实验技术与管理, 2007, 24(12): 118-120.
- [5] 徐佩锋, 赵中营. 用 packet tracer 模拟软件改进高职计算机网络实验教学[J]. 计算机教育, 2008, (18): 35-39.
- [6] 黄浩. 将协议分析软件引入计算机网络教学[J]. 攀枝花学院学报, 2009, 26(6): 109-112.
- [7] 黄艳琼, 梁俊. 计算机网络课程实验教学改革探索[J]. 计算机教育, 2009, (2): 62-63.
- [8] 刘艳芳, 张力军, 曹庆华等. 在计算机网络实验教学中的体会和思考[J]. 计算机教育, 2009, (03): 51-53.
- [9] 赵靖如, 王宜政. 关于计算机专业“计算机网络”课程教学改革的建议[J]. 计算机教育, 2006, (5): 33-35.

操作系统进程同步的教学实践

李志民, 赵一丁, 底恒

(中原工学院计算机学院, 河南 郑州, 450007)

摘要: 针对操作系统课程中, 存在进程同步内容抽象、难以理解的教学实际, 本文在分析进程相互制约关系的基础上, 归纳出利用信号量机制实现进程同步、互斥的一般性解题思路, 以“生产者—消费者问题”为例阐述了具体的解题步骤, 在教学过程中结合基于 Java 多线程的可编程操作代码, 通过真实的运行结果把抽象知识变为易理解的知识, 取得在理论教学方面提高学生学习兴趣、在实践教学方面提高学生实际动手能力的效果。

关键词: 进程; 同步; 互斥; 信号量; 多线程

中图分类号: G642 **文献标识码:** A

Teaching Practice of Operating System Process Synchronization

LI Zhimin¹, ZHAO Yiding, DI Heng

(Zhongyuan University of Technology, Zhengzhou 450007, Henan China)

Abstract: The content of the abstract process of synchronization in is difficult to understand in the Operating System teaching practice. The article start with analyzing the relationship between the process of mutual restraint and summarize general problem-solving ideas of using semaphore s mechanism to achieve synchronization and mutual exclusion.It Speci fics general problem-solving ideas as the “Producer - Consumer problem” example. In the teaching process, Actual operating results can be observed by a programmable multi-threaded java based operations and make the abstract knowl dge easy to understand and obtain In the theoretical teaching to enhance students in terest in learning, teaching practice to improve the effect of the practical ability of students.

Keywords: process; synchronization; mutex; semaphore; multi-threading

操作系统课程的进程同步问题是该课程的核心知识点, 也是学习难点。解决同步和互斥问题最常用的方法就是信号量机制, 通过在程序算法中使用 P、V 操作达到同步和互斥的目的^[1], 信号量与 P、V 操作是低级的同步机构, 用它们很难表示复杂的并发性问题, 在并发程序中的出现 P、V 操作, 使得程序正确性证明更加困难^[2]。在教学过程中, 发现很多学生对上课所讲的例子大部分都清楚它的算法, 可是当遇到一个新的同步与互斥问题, 很多学生觉得还是无从下手, 感到困惑。下面以“生产者—消费者问题”为案例, 从进程的相互制约关系入手, 归纳出进程同步、互斥的解题思路和解题步骤, 利用 Java 代码实现多线程同步, 把抽象的知识变为具体可理解知识, 提高教学效果。

1 进程同步、互斥的一般性解题思路

对于进程的同步与互斥问题, 学会问题解决的分析过程, 理清解决此类问题的解题思路, 是很重要的。《操作系统》课程中利用信号量机制实现进程的互斥和同步, 以记录型信号量为例, 利用信号量的 P、V 操作解决进程同步问题的解题步骤, 归纳如下:

作者简介: 李志民 (1969—), 男, 汉族, 山东东营人, 中原工学院计算机学院副教授, 硕士, 研究方向: 软件工程。
Biography: LI Zhimin (1969—), Male, Han Nationanlity, Born in Shandong Dongying, Associate profes sor in Zhongyuan University o f Technology, Research Fields: Software Engineering.

1.1 确定进程数和进程间的制约关系

首先，分析待解决的实际问题，提炼出实际问题进程的数量；
然后，根据各进程的活动描述，判断进程间的相互制约关系：
(1) 如果共享某种临界资源，进程间属于间接相互制约关系，称为进程互斥；
(2) 如果进程间的合作带有前趋、后继关系，进程间属于直接相互制约关系，称为进程同步；
(3) 同时存在互斥和同步两种制约关系。
根据这三种制约关系，分别按 1.2 节、1.3 节、1.4 节进行处理。

1.2 进程互斥的解题步骤

(1) 找出临界资源、临界区；
(2) 为每个临界资源设一个互斥信号量，如 `mutex`，初始值为 n (n 是临界资源的个数)；
(3) 在临界区前加入 `wait(mutex)`原语，作为进入区；
在临界区后加入 `signal(mutex)`原语，作为退出区。
注意：`wait(mutex)`和 `signal(mutex)`成对出现在同一个进程中。

1.3 进程同步的解题步骤

(1) 确定进程之间的前趋关系，画出前趋线；
(2) 为每条前趋线设一个同步信号量，如 `empty`；
(3) 在前趋线的箭头前插入 `wait(empty)`原语，作为进入区；
箭尾后插入 `signal(empty)`原语，作为退出区；
(4) 同步信号量 `empty` 的初始值为 n 或 0 (n 为正整数，是申请资源的初始单位数)；
同步信号量的初值则要根据进程的初始状态确定，具体设置相应的值。
一般情况下：不能率先执行的 `wait()`原语，其信号量初值为 0 ，否则为 n ；
注意：`wait(empty)`和 `signal(empty)`成对出现在不同进程中。

1.4 进程互斥、同步同时存在时的解题步骤

(1) 如果进程间存在互斥和同步两种制约关系，分别按照 2.2 节和 2.3 节的解题步骤执行；
(2) 在同一个进程的进入区中可能会出现两个 `wait()`原语，一定要先写同步信号量，后写互斥信号量；否则，可能出现死锁现象；
(3) 在同一个进程的退出区中可能会出现两个 `signal()`原语，书写顺序没有严格要求；规范的写法是：先写互斥信号量，后写同步信号量；
注意：(1) 一定要检查两个 `wait()`原语的顺序，不能颠倒；
(2) `wait()`和 `signal()`必须成对出现，否则会出现死锁现象^[3]，影响系统性能。

1.5 用类 Pacal 语言描述进程同步或互斥算法

根据上面几个步骤分析的结果，就可以类 Pacal 语言或其他语言实现同步与互斥的算法。初学者开始时可按上述步骤解决同步和互斥问题，以上讲述的只是一般的求解规则，有一定的可操作性，但在实际应用中还是要针对具体情况，多做多想，领悟出其中的原理和窍门。

2 基于信号量机制的“生产者—消费者问题”的同步算法

以“多生产者—多消费者—多缓冲”问题为例，给出进程同步与互斥的解题步骤。

2.1 确定进程数和进程间的制约关系

通过分析就会发现，该问题中生产者、消费者各是一类进程。并发执行过程如图 1 所示。

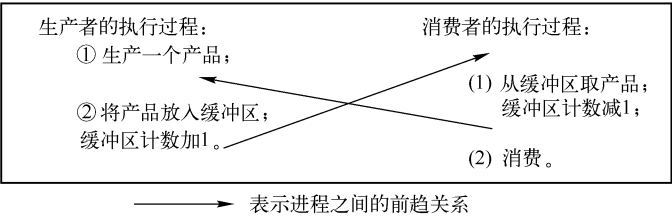


图 1 生产者—消费者的制约关系

两类进程既共享缓冲区计数这一临界资源，又具有“先放再取②→(1)、取后再放(2)→①”的两条前趋关系，进程间同时存在互斥和同步两种制约关系。

2.2 解题步骤

2.2.1 进程同步的解题步骤

- (1) 找出临界资源：共享缓冲区的计数；
- (2) 为临界资源设一个互斥信号量，如 `mutex`，初始值为 1；
- (3) 在临界区前加入 `wait(mutex)`原语、`signal(mutex)`原语；

2.2.2 进程同步的解题步骤

- (1) 确定进程之间“先放再取、取后再放”的两条前趋关系，画出前趋线；
- (2) 为每条前趋线设一个同步信号量，分别是 `empty`、`full`；
- (3) 在前趋线的箭头前分别插入 `wait(empty)`、`wait(full)`，箭尾后插入 `signal(empty)`、`signal(full)`；
- (4) 同步信号量 `empty` 的初始值为 n （缓冲区的大小），`full` 的初始值为 0；

2.2.3 用类 Pacal 语言描述进程同步或互斥算法

生产者进程：	消费者进程：
生产一个产品；	<code>wait(full)</code> ；
<code>wait(empty)</code> ；	<code>wait(mutex)</code> ；
<code>wait(mutex)</code> ；	从缓冲区取产品；
将产品放入缓冲区；	缓冲区计数减 1；
缓冲区计数加 1；	<code>signal(mutex)</code> ；
<code>signal(mutex)</code> ；	<code>signal(empty)</code>
<code>signal(full)</code> ；	消费；

注意：(1) `wait(mutex)`和 `signal(mutex)`成对出现在同一个进程中。
(2) `wait(mutex)`和 `signal(mutex)`成对出现在不同进程中。
(3) 一定要检查两个 `wait()`原语的顺序，不能颠倒；
否则会出现死锁现象，影响系统性能。

3 基于 java 多线程的“生产者—消费者问题”的同步实现

操作系统的课程实验旨在加深学生对理论的理解。对于进程同步问题，线程是轻型进程, Java 为同步线程提供了两个方法:object 类的 `wait()`方法和 `notify()`方法。当线程请求资源而未能满足时，调用

wait()方法使线程等待，并将它排在等待队列上；当线程对资源访问完后，通过 notify()方法唤醒等待队列上的线程^[4]。下面是实现生产者—消费者同步的完整的 java 可运行代码:

```
public class Store {
    private final int MAX;
    private int count;
    public Store(int n){
        MAX=n; count=0; }
    public synchronized void addData() throws Exception{
        while(count==MAX){this.wait();}
        count++;
        System.out.println(Thread.currentThread().getName()
+" add a data:"+count);
        this.notifyAll();}
    public synchronized void removeData() throws Exception{
        while(count==0){this.wait(); }
        System.out.println(Thread.currentThread().getName()
+" remove a data:"+count);
        count--;
        this.notifyAll(); }
}

class Producer extends Thread{
    private Store s;
    public Producer(Store s){this.s=s;}
    public void run() {
        while(true){
            try {
                s.addData();
                Thread.sleep(100);
            } catch (Exception e) {}
        }
    }
}

class Consumer extends Thread{
    private Store s;
    public Consumer(Store s){ this.s=s;}
    public void run(){
        while(true){
            try {
                s.removeData();
                Thread.sleep(100);
            } catch (Exception e) {}
        }
    }
}
```

编写测试类，同时调用多个生产者线程和消费者线程，由于多线程并发执行，每次的执行结果可

能不尽相同，下面分别是 4 个生产者线程和 4 个消费者线程两次并发执行的结果，如图 2 所示。

P01 add a data:1	P01 add a data:1
p02 add a data:2	P03 add a data:2
P03 add a data:3	p02 add a data:3
p04 add a data:4	p04 add a data:4
c02 remove a data:4	c01 remove a data:4
c01 remove a data:3	c02 remove a data:3
c03 remove a data:2	c03 remove a data:2
c04 remove a data:1	c04 remove a data:1
P01 add a data:1	P01 add a data:1
p04 add a data:2	P03 add a data:2
P03 add a data:3	p02 add a data:3
c01 remove a data:3	p04 add a data:4
p02 add a data:3	c02 remove a data:4
c02 remove a data:3	c01 remove a data:3
c03 remove a data:2	c03 remove a data:2
c04 remove a data:1	c04 remove a data:1
P01 add a data:1	P01 add a data:1
P03 add a data:2	P03 add a data:2
p04 add a data:3	p04 add a data:3
p02 add a data:4	p02 add a data:4
c02 remove a data:4	c02 remove a data:4
c01 remove a data:3	

图 2 P-C 同步的两次运行结果

通过多次执行结果的比较，可以直观地理解进程并发的实质，通过案例教学，可以清楚地观察进程互斥和同步的并发执行过程。

4 总结

如何让实验教学配合好理论教学，使枯燥无味的原理变成趣味十足的课程，成为教学改革的主要目标^[3]。在讲解进程的同步与互斥这一比较复杂、抽象的内容时，需要先对问题进行分解，由浅入深，给出进程同步问题的一般性的解题步骤；结合 Java 多线程案例，通过实际的运行结果，把抽象的知识变为具体可理解的知识；在理论教学中调动了学生的积极性，在实践教学中提高了学生的实践能力，培养学生分析问题、解决问题的综合能力，对于提高教学质量、全面提高学生素质有着重要的意义。

参考文献

[1] 汤小丹, 梁红兵,等. 计算机操作系统(修订版)[M]. 西安: 西安电子科技大学出版社, 2003.
[2] 陈向群, 杨芙清. 操作系统教程[M]. 北京: 北京大学出版社, 2006.
[3] 李瑛达, 谢双杰. “操作系统”实例化教学的改革探讨[J]. 计算机教育, 2009(7)27-30.
[4] 郑莉, 王行言, 马素霞. Java 语言程序设计[M]. 北京: 清华大学出版社, 2008.

“数据库系统原理”教学实践与改革

职为梅

(郑州大学信息工程学院, 河南 郑州, 450052)

摘要: “数据库系统原理”课程作为计算机专业的基础课程, 具有理论知识抽象、实践应用性强的特点, 本文分析了“数据库系统原理”课程教学中实际存在的问题, 针对这些问题提出一系列的改革方案, 取得了良好效果。

关键字: 数据库系统原理; 教学改革; 课程设计; 数据库实验

中图分类号: TP3-05 **文献标识码:** A **文章编号:** 1006-7043 (2004) xx-xxxx-x

Practice of Teaching and Reformation of Database System Principle

ZHI Weimei

(Department of Computer Science, Zhengzhou University ZhengZhou 450052, Henan China)

Abstract: As the basic courses of Computer Science, the course of Database System Principle has the feature of abstract theory and very strong practicality. This paper analyzes the actual problem in the teaching process of the course of Database System Principle, and bring up a series of reformation measures. We obtain better effect.

Keywords: database system principle; practice of teaching; design of courses; experiment of database

1 引言

数据库技术产生于 20 世纪 60 年代中期, 经历了四十余年的发展, 已经成为计算机学科的重要分支。数据库技术是信息系统的核心和基础, 它的出现极大地促进了计算机应用向各行各业的渗透。现在, 数据库的建设规模、数据库信息量的大小和使用频度已经成为衡量一个国家信息化程度的重要标志^[1]。

“数据库系统原理”课程是计算机本科专业的必修课。课程的教学以目前流行的关系数据库系统为主要内容。通过对“数据库系统原理”课程的学习, 使学生能够正确地理解数据库技术的基本原理、掌握数据库设计的基本方法, 并能够熟练地使用学到的知识设计基于特定应用背景下的数据库。因此, 本课程对学生提出了更高的要求, 不仅要准确地掌握理论知识, 还要能够应用理论知识。这就要求授课老师在讲解本课程时选择适当的教学方法, 使得学生能更好地接受, 而不仅是“填鸭式”地灌输知识。但是, 在实际的教学过程中, 存在很多问题, 本文通过分析现有“数据库系统原理”课程存在的一些问题, 提出关于“数据库系统原理”课程改革的新思路。

2 传统教学现状分析

结合自己的教学体会, 以及当前数据库的发展现状, 通过分析传统“数据库系统原理”课程的授课模式, 我认为在“数据库系统原理”课程的教学中存在以下问题: 过分强调课程的理论知识, 忽略实践的重要性; 重视学生对理论知识的掌握程度, 忽略学生的解决问题能力的培养; “以教师讲解为中心”的授课模式; 教学方法和方式不够灵活、教学资源匮乏、考核制度不合理等问题。

作者简介: 职为梅 (1977—), 女, 讲师, 硕士, 长期从事数据挖掘的研究, 发表论文六篇, 译著一本, 合著一本。

下面, 详细分析“数据库系统原理”课程存在的每个问题。

1) 过分强调课程的理论知识, 忽略实践的重要性

“数据库系统原理”课程和“数据库原理与应用”课程不同, 前者以理论教学为主, 后者以应用为主。课程本身的特点使得授课老师在教学上过分强调理论知识的学习, 对实践不屑一顾。这是一种错误的认识。虽然“数据库系统原理”课程理论性强, 但很多内容抽象, 不辅助实践教学, 很难理解这些抽象的理论知识。相反, 如果实践环节跟得上, 不仅能促进学生对理论知识得理解, 还会提高学生学习的积极性。

现在的数据库教学中也安排了相应的实践环节和理论教学相配套的上机实践。由于传统的考核制度中很少依赖上机实践的分數, 学生从心理上没有重视上机实践。而上机使用的数据库管理系统(即 DBMS)不是教学内容, 加上学时紧张, 授课老师往往没有时间去讲解一个具体的 DBMS, 这要求学生课下学习, 但长期以来的“填鸭式教学”使得学生不主动学习, 导致虽然上机实验内容布置下去, 但学生完成情况却很差。上机时学生往往是在熟悉实验环境, 并没有完成老师布置的任务, 达不到教学实践要求。使得教学和实践脱节。

2) 重视学生对理论知识的掌握, 忽略学生的解决问题能力的培养

“数据库系统原理”课程内容抽象、琐碎、庞杂, 涉及很多概念和技术, 且均各成体系, 相互之间的衔接线索很少, 总体感觉内容零散, 没有一个整体的知识框架体系。同时, 因教学内容多、知识量大, 很难取舍, 所以不免在教学过程中变得面面俱到, 重点、难点不突出, 学生觉得课程抽象、理解困难。课程的特点有时导致授课老师精力放在理论教学上, 忽略了学生解决问题能力的培养。比如, 在传统的教学方式中, 遇到问题的解答时, 授课老师往往直接告诉学生答案。这种做法没有起到对学生解决问题能力的培养。

3) “以教师讲解为中心”的授课模式

教学中, 授课老师多采用“填鸭式”教学, 把课堂变成自己表演的舞台。认为我把自己知道的知识全给学生讲出来, 这样学生就会学得好, 这样做往往适得其反。“数据库系统原理”课程的教学在很大程度上沿袭着“以教师讲解为中心”的传统教学观念: 教师是知识的传递者, 教学以传授知识为主等。这样的教学理念直接影响和制约了教学模式、教学方法、教学手段的改革和创新, 也不利于发挥学生学习的主动性和积极性。

4) 教学方法和方式不够灵活

目前“数据库系统原理”课程采用多媒体教学, 减少了授课老师的板书, 增加了授课教学信息量。但同时多媒体教学也存在一些问题: 只是使用课件演示, 有时并不能使学生完全理解, 使用板书详细讲解实现的步骤, 往往取得更佳的效果; 信息量多, 会使得课堂变成纯粹的演示, 学生只是走马观花, 但不知所云; 教师过分依赖课件, 一旦多媒体设备出现问题, 竟然不知道该如何上课。

5) 教学资源匮乏

除了讲课教材和课件外, 学生在学习“数据库系统原理”课程时, 往往没有其他的学习资料, 这对开阔学生的学习视野很不利。

6) 考核制度不合理

传统的考核方式以笔试为主, 甚至有的学习只以笔试作为学生的最终成绩。对“数据库系统原理”课程来说这样的考核方式很不合理。该课程不仅是一门基础理论课程, 同时它还有很强的实践性, 传统的考核方式使得学生疲于应付考试, 往往只是死记硬背理论知识, 考高分学生的动手能力往往还没有一些低分学生的动手能力强。

3 改革的具体内容和实施方法

上小节中详细地分析了“数据库系统原理”课程在教学中存在的问题。针对上述问题, 提出了对

该课程教学进行改革的具体内容和实施方法。

1) 理论知识与实践并重

充分调动学生的学习积极性，变被动学习为主动学习，学生在学习理论知识的同时也掌握一个具体的 DBMS 系统。能够很好地完成上机任务，上机任务的良好完成又可以带动学生学习的积极性，形成良性互动。而不是学生在上机时疲于熟悉上机环境完不成上机任务。

自 2005 年以来，针对数据库系统原理课程实践少这一问题，我们在上机实践外额外增加了“数据库原理课程设计”。将学生分组，以组为单位，给每个组一个应用背景，进行需求分析，设计相应的数据库系统，在此基础上开发应用系统。开设四年以来，课程设计取得良好的效果。通过课程设计学生不仅锻炼了动手能力，亲身体会了设计一个系统的全过程，还加深了对理论知识的理解。

2) 重视学生对理论知识的掌握，更重视学生解决问题能力的培养

教师在理论教学过程中，采用问题驱动学习模式，提出与所学理论相关问题，要求学生思考解答，并逐步引导学生自己提出问题、分析问题、解决问题，着重培养学生创造性思维的能力，充分调动学习的积极性，使学生进入积极思考的主动状态，在解决问题的过程中获取知识，提高能力。

3) “以学生为主体，以能力为中心”的授课模式

如果教师讲授课程时把自己也作为初学者，充分理解学生的学习需求，转换学生与教师作为提问者和解答者的角色，教师要发现问题、设置问题，启发、引导学生思考解答，充分调动学生的主观能动性，使教学成为教师和学生共同的事业。

4) 丰富教学资源

充分利用本专业的优势，将所有资源全部上网。提供教学网站，在网站上提供课程的全部信息，比如课程的授课范围、教学大纲、教学课件、参考书目、数据库知识的参考网站、数据库新技术的一些超链接等。学生可以从网站上获得和教学相关的所有资源。

可以建立在线答疑和留言板块，使学生随时可以和老师交流，有问题及时向老师请教。

提供网上试题库：根据每章的教学内容，给出适当的练习题目，学生可以及时了解自己对所学内容的掌握情况。同时提供综合试题卷，有利于学生对全部内容的掌握。

5) 建立合理的考核制度

传统的考核制度往往造就“高分低能”的学生。必须对这种考核制度进行改革。“数据库系统原理”课程是实践性很强的一门课程。在考核方法中，既应该看重学生对理论知识的掌握，也应该重视学生的动手能力。因此，考核方法应该采用笔试与实践成绩相结合的方法，比如笔试成绩占 70%，实践成绩占 30%，其中平时上机占实践成绩的 40%，课程设计占实践成绩的 70%。通过这种考核制度，使学生充分认识到实践的重要性，从而达到既掌握理论知识又提高了动手能力的目的。

4 总结

本文分析传统的“数据库系统原理”课程教学中存在的若干问题，一一提出了改革措施，在我们的教学中付诸实施。取得了良好的教学效果，学生对理论知识掌握的同时动手能力也得到了很大的锻炼。在以后的教学中我们将根据数据库技术的发展，继续探讨新的改革思路，以适应时代对数据库技术人才的要求，从而培养符合信息化社会需要的数据库人才。

参考文献：

[1] 范明，叶阳东，邱保志，职为梅. 数据库原理教程[M]. 北京：科学出版社，2008.

- [2] 萨师煊, 王珊. 数据库系统概论 (第四版) [M]. 北京: 高等教育出版社, 2006.
- [3] 徐保民, 孙丽君, 李爱萍. 数据库原理与应用[M], 北京: 人民邮电出版社, 2008.
- [4] 吴达胜, 刘丽娟, 孙圣力. 数据库原理与技术的理论与实践教学的整体优化研究[J]. 计算机时代, 2005 (11): 31-32.
- [5] 任建军. 数据库原理及应用教学改革探讨[J]. 四川师范学院学报 (自然科学版), 2003 24(1):22-24.
- [6] 丁宝康, 董健全. 数据库实用教程[M]. 北京: 清华大学出版社, 2001.
- [7] 翟中. 数据库教学方法改革的探索与实践[J]. 黑龙江高教研究, 2006 (2): 113-114.

计算机人才培养研究

曲宏山¹，崔清民

(¹河南工程学院计算机科学与工程系，河南 郑州，450007)

摘要：计算机相关专业毕业生已经产生了严重的过剩，这种过剩的原因无外乎两个，一是数量，二是质量，本文探讨如何提高毕业生的质量，并探讨计算机人才培养的相关问题。

关键词：计算机；人才培养；实践教学；校企合作；创新意识

The Research of Computer Person's Training

QU Hongshan¹，CUI Qingmin²

(Henan Institute of Engineering,Zhengzhou 450007, Henan China)

Abstract: Coumputer-related professional graduates has result in serious surplus. There are two main reasons of this phenomena: one is the quantity of graduates, an other is the quality. This paper investigat ed the method how to improve the quality o f graduates and other relative questions on computer-related personnel training.

Keywords: computer; personnel training; practice teaching; school-enterprise combine ; innovation awareness

近几年，计算机专业的毕业生也由学生挑工作单位转变为工作一时难求的局面。原因是多方面的，但其中之一是学生的质量不高，学生普遍缺乏实际动手能力，用人单位需要花费大量的时间和精力培训才能够上岗。这就形成了一方面计算机企业招不到软件人才，另一方面毕业生又为工作难找而苦恼，而一些宣称几个月成就软件大师培训就业的机构又异常热闹的尴尬局面。

学校不能只作为旁观者，而应该正确面对，认真思考。计算机专业的定位和课程设置要适应社会需求，不断地调整以应对日趋激烈的市场竞争。探索计算机人才培养模式。本文结合我系的情况做探讨。

1 明确人才培养的学科定位

学校培养人才要以市场为导向，以服务经济和社会发展为主要目标，培养应用型人才。培养的学生要适应信息现代化建设需要，牢固掌握计算机基本理论与知识，能够熟练进行计算机程序设计，熟悉本专业的工作流程，真正做到毕业就能上岗，上岗就能工作。

2 抓住人才培养的关键环节

2.1 以培养学生创新意识和实践能力为目标

我校作为新升本院校，新生层次不是很高，入学前接触计算机的机会少，许多同学抽象思维能力薄弱，缺乏自主学习的能力。然而我们面临的却是一个多元化的信息社会，要求我们掌握多方面的计算机知识和能力，这样如何针对绝大多数同学制定一个合适的培养目标就显得特别重要。根据以上实际情况，决定以培养学生的创新意识和实践能力为目标，对教学过程的各个要素和环节进行整体优化，努力形成多元的、以学生为本的、注重个性发展的人才培养过程。以此来适应现在市场的需求，为学生提供一个良好的发展空间。

教学计划在保证计算机主干课不变的同时，设立了数据库与网络、游戏程序开发、网页制作等子

作者简介：曲宏山，（1959—），教授，研究方向：网络安全、数据库、人才培养。

方向的选修课程。让不同类型和不同层次的学生都能够发展自己的兴趣爱好。这个教学计划把“高级科学技术专业人才”的培养定位改变为“多类型、多层次的初步具有实际工作能力和研究能力的应用型人才”，根据这个培养目标制订新的教学计划。

在教学过程中，要求学生较、扎实地掌握计算机软、硬件的基础理论，系统地学习利用计算机解决实际问题的基本方法和完整过程，实际地掌握数字媒体构件加工、软件编程、作品制作和网络建设的基本方法。通过课堂教学、上机实习、课程设计多个环节，培养出初步具有实际工作能力和研究能力的应用型人才。

2.2 合理设计课程

(1) 开设计算机导论、数字媒体基础、计算机原理、操作系统、计算机网络、数据库、计算机图形学等专业基础课和专业课，加强学生对基础理论课程的学习；强化 C、C++、Java 等基础编程能力和软件开发能力；设置动画课程设计、网站课程设计、数据库课程设计、游戏课程设计等子方向的实践课程，供不同发展方向的学生选择，要求学生独立完成一个较大型的课程设计作品，以提高学生的实践动手能力。

(2) 在教学计划中充分发挥学分制的优点，按照教学次序和难易程度分开，将要开展的课程分别安排到各类必修课和选修课中，这样学生可以根据自己的情况选择不同层次的课程，达到培养不同层次人才的目标。强化主干课程的同时，允许学生根据兴趣选修课程，这就需要学校在选修课中设置相应的序列课程，供学生根据自身的优势和市场需要进行个性化的自我设计，努力向选定的方向去发展，以便深入掌握相关领域的专业技术和技能，同时保证所学知识体系的整体性和连贯性。

2.3 重视实践教学

作为对理论教学的补充，实践教学环节应该包括上机实习、课程设计、专业社会实践、作品制作、开放实验、项目开发、毕业实习和毕业设计等多方面的内容。我们进行了有益的尝试，例如，鼓励学生参加天津市乃至全国的各种竞赛；鼓励学生组成团队进行专业社会实践活动；组织一些大规模的练习性项目，在项目中实践软件开发和软件工程的知识等。

为了提升整体的教学质量，我系不仅聘用外校的师资，而且聘请了社会上一些知名企业的技术人员，以及一些具有创业经验的企业家承担教学任务，这种方式极大地提升了教学的多样性，同时也对学生的应用能力培养起到了至关重要的作用。

2.4 帮助学生树立正确的择业目标

从计算机专业毕业生所从事的工作性质来划分，大致上可以分为以下 3 类：从事研究型工作、从事工程型工作和从事应用型工作。目前高校计算机专业在本科阶段对前两类人才的培养已有一定基础，对于第三类人才的专门培养则几乎是空白。

根据市场的这些特点，我们不是只将就业指导停留在毕业生，而是贯穿整个大学 4 年的教育，针对不同的年级，做不同的指导。一年级新生应该开展新生入学专业介绍，培养学生的专业认识能力，教育学生开始建立职业规划的意识，确立职业目标，培养学生愿意学习、主动学习、善于学习的能力。二年级注重相关能力的培养，比如学生有意向数据库、网络、游戏软件研发方向发展，就应教育学生在强化程序设计课程的同时，加强专业社会实践活动，积累计算机相关项目的经验；如果学生有意向网页设计、数字媒体作品设计方向发展，应指导学生多参与各类设计创作大赛，并指导学生选修相关的艺术类课程，提高对艺术的关注度和感悟能力。三年级要加强学生的价值观念和技能技术的训练；提高学生的学习能力、社会实践能力、组织管理能力、团队协作能力。四年级要着重就业的服务性工作，多渠道、多手段给学生提供就业信息和咨询指导。四个年级的不同阶段要相互贯通，与平时

的日常事务管理和思想政治教育结合。

3 突出人才培养的特色模式

3.1 鼓励课外开展兴趣小组，培养团队意识

丰富的课外活动，将有专业兴趣的同学集合起来，邀请老师和有专业特长的同学讲座，激发学生对专业的学习兴趣。鼓励学生结成兴趣小组，共同学习探讨，带动周围的同学，营造良好的学习氛围。如数字媒体、游戏软件、网站建设等兴趣活动小组，每一个学生都可以亲身体验到系、专业对应用型人才培养的力度。

3.2 组织参与各种专业技能竞赛

每年的计算机技能大赛都会产生优秀的具有社会价值的作品，正是这些作品冲击着每一个同学，在对作品惊叹之余每个人都亲身体验着“实践”带来的乐趣，而这种“乐趣”恰恰是培养应用型和创业型人才必不可少的根基，也是学生提高学习主动性的动力。

参加全国和本地举行的各类计算机能力竞赛，使学生可以在更大的范围内进行竞争，在取得优异成绩的同时，更能了解不同学校不同学生的差异，清楚自己的优势和弱点，明确学习方向。

3.3 校企结合，实现共赢

加强高校与各类培养机构及企业的主动联系，在高校、企业培训机构之间组建多渠道、符合社会需求的软件业人才教育体系，促进高校与专业培训机构的良好合作与互动，形成良好软件人才培养链，发挥各自的优势，加强实践环节，加大继续教育和培训的力度、使人才培养工作充分地与市场需求相结合。

3.4 开展相关的专业资格认证工作

普通院校计算机专业毕业生在就业难，社会认可度不高的事实面前，也要不断发挥优势，突出特点，提高质量，最大限度地帮助毕业生顺利就业。

提高社会的认可度，取得相关的专业资格认证不失为一种很好的解决方法。去过大大小小的招聘会的毕业生经常看到 Java 工程师、DSP 工程师、网络设计师等的招聘，我们应该指导学生根据自身情况，选报相应的认证。在学校开设考点和相关的培训课程，方便学生通过考试。

4 师资问题

现代教育观中，教师的角色已由传统意义上的知识传授者转变为对学生探索实践活动的引导者、协作者、帮助者、监督者和评价者。这就要求教师既要具备课堂教学的基本教育素质，还应进一步拓展知识结构，提高计算机专业理论素养，钻研实践教学中特别需要的沟通能力、关注能力、管理能力和协调能力。

学校是为社会服务的，学校的根本价值是为社会培养合格的人才，而人才的培养是通过教师实现的，这就要求教师必须要经得起社会和企事业单位的考验，但大多数教师都是从学校到学校，实践经验极度匮乏，直接影响到学生的质量，所以，教师如何与社会挂钩，也是我们必须研究的问题。

学校要树立人才使用的新观念，建立计算机应用技术人才的多层次结构，建立教师资源的研究、实施和保障部门。重视计算机学科师资培养与提高，采取多种有效手段为中青年优秀人才的脱颖而出创造必要条件，可以制度化地选送教师到比较好的高校进修学习。由于现在计算机学科的师

资力量中青年教师比例相当大，大多数都有硕士研究生或是博士生学历，其中一部分表现出教学经验不足或是专业素质欠缺等问题，因此高校应该考虑在一个时间段和有针对性地进行青年教师业务培训。另外就是请求“外援”来提高专业水平，不仅可以带来先进技术，而且还能带来较高的学术和教学标准。

总之，为了培养合格的计算机专业人才，必须转变教学理念、深入进行教学的改革研究，加强实验实训基地建设，加强校企合作，坚持产学研相结合之路。计算机的发展日新月异，我们要不断地研究计算机人才培养的方案，努力为社会培养更多合格的计算机人才。

参考文献

[1] 王金如. 计算机教学创新能力培养[J]. 《中小企业管理与科学》，2009.10

[2] 王鑫. 加强计算机实验教学，培养高素质创新人才. 《湘潭师范学院学报》，2009.3.

[3] 于金玲. 地方院校计算机专业人才培养方案的探索与研究. 《科技信息》，2009.6.

[4] 徐蕾. 计算机专业产学研结合的学生能力培养方式探索与实践. 《沈阳航院学报》，2009.8.

[5] 欧阳成. 以应用能力为核心的计算机基础课教学改革. 贵阳学院学报，2009.3.

高等学校教学资源共享运行机制研究

郑娅峰, 张巧荣

(河南财经学院信息学院, 郑州 河南, 450002)

摘要: 分析了高等学校实验教学资源共享的现状, 分别从管理层面和技术层面阐述了高等学校实验教学资源共享运行机制的建立。并提出了将网格技术应用于实验教学资源共享平台建立的具体实施办法, 对提高高等学校实验教学资源的共享有参考价值。

关键词: 网格; 高等教育; 资源共享

中图分类号: (G642) **文献标识码:** B

Research on the Sharing Mechanism of Teaching Resource in College

ZHENG Yafeng, ZHANG Qiaorong

(College of Computer and Information Engineering, Henan University of Finance and Economics Zhengzhou 450002, Henan China)

Abstract: This paper analyzes the present situation of teaching resource sharing in college, and presents the effective method to establish operational mechanism from the management and the technical way. This paper applies the grid technology in the experiment teaching resource sharing platform and proposes that concrete implementation. It has the reference value to enhance the College teaching resources sharing.

Keywords: teaching resource; sharing mechanism; grid

1 引言

实验教学资源作为高校教学资源的重要组成部分, 在高等学校的学科建设、专业发展中具有举足轻重的地位。实验教学资源的共享是指在一定的区域内, 教育部门对其所有的实验资源打破现有界限, 实行共同享用。实行实验教学资源共享、优势互补、服务社会, 产学研的发展模式等, 有利于高等教学的稳定发展, 也是优化实验教学资源合理配置的有效途径。实验资源的共享不仅包括专业实验师资共享, 而且包括教学仪器、设备共享和实验教学信息共享等多种形式和内容^[1]。实验资源的共享与共建问题是我国高等教学信息化进程中一个重要的研究内容。由于实验教学资源通常在地理上分布较广, 各种资源在底层数据格式、实现方法、运行环境等方面存在不同程度的差异, 再加上各部门间协调、管理, 以及版权保护等方面的问题, 使得在各高校间要实现真正意义上的实验资源共享还存在一定距离^[2]。

目前, 随着技术水平的发展, 一些实验演示文件, 实验操作文件开始出现在一些院校进行内部共享。同时, 各高校都在大力建设基于本校精品课程的校内在线学习环境, 各种学习资源日益丰富。但受网络传输速度影响, 外部人员很难通过远程访问获得这些资源进行在线学习。将流媒体技术引入后, 一定程度上解决了传输受限的问题。使得校内局域网环境下的资源共享得到了极大的发展。目前, 很多高校都已建立了自己的实验网络平台, 供本校学生进行在线的实验及课程学习。但校际之间的共享仍然受到很大的制约。近几年来, 如何进行校际之间的实验资源发现与共享成为研究的热点。

基金项目: 河南财经学院校级重点教学改革项目——高等学校教育教学资源共享运行机制研究。

作者简介: 郑娅峰 (1979—), 河南洛阳人, 讲师, 硕士, 研究方向为分布式系统, 计算机网络。

2 高等学校实验教学资源共享存在的问题

目前教育资源共享中存在诸多问题。由于多年信息化建设投入，教育信息化已经在各方面有了不同程度的发展，但其使用效率、应用效果及投资效益与人们的预期相比仍存在着较大的差距。主要是资源建设与共享及管理问题。一方面，由于实验设备基本由各部门自行维护管理，因此设备资源普遍存在开放程度低、协同使用少、利用率低的现状，造成资源的浪费；另一方面，许多教师科研手段单一，并不知道已购置设备的类型，无专业仪器可用，直接影响了科研水平，许多仪器设备和实验教学资源在教学科研中的作用还没有发挥出来。主要的问题概括为以下几个方面：

（1）实验室网络基础设施薄弱。目前在大多数学校中，很多没有专用的实验资源网站。实验教学资源缺乏全面性与鲜活性，成为信息技术应用中的一个主要障碍；即使建设了这样的网站，也多是以文字信息为主。对于有些实验食品录像等信息量较大的应用，用户在看到页面显示之前又往往要等待较长的时间。实验教学网的建设及应用没有和学校理论教学目标有机地结合起来，有效整合和资源建设仍然有待加强。

（2）设备重复投资。各高校目前都在加大软、硬件投资的力度。有数据显示，目前高校中，新添置的单价超过5万元的贵重设备已大量进入各实验室和教学科研单位，如果把这些投资集中在一个学校，这个学校的发展速度是难以想象的。但现在分散的，造成了实际上的重复投资。而且，这些设备只满足了各部门自己的小型服务要求。如果能将一些巨型机资源共享，让各个高校有环境进行并行计算的教学和试验，能够节约投资成本。

（3）缺乏相关的仪器设备共享制度。各学科之间的交叉可能涉及多部门的实验环境，由于缺乏相关的仪器设备共享制度，如手续，共享成本费用标准，使得共享的复杂度变高，无法调动设备管理人员的积极性，导致对仪器设备的跟踪管理不到位，最终影响了设备的利用率。现在就高校内部而言各学院、实验室间互借设备情况已经较为普遍，但由于缺乏相关的仪器设备共享制度。因此应该建立起共享的平台，提高设备资源的利用率，做好资源共享，同时能够在高校建立共享实验资源，扩展和挖掘仪器设备的利用价值，优化学校的实验设备的资源配置，推动教学资源的共享。

如何消除实验室资源配置的孤岛现象，做到资源共享、优势互补，避免重复开发，节省人力、财力，是大家都在努力解决的重要课题。实验教育资源共享将为教育信息化的改革和发展提供广阔的前景，并促成新的教育应用。

3 高等学校实验教学资源共享的机制建立

实验教学资源的共享能够得到有效的发展主要需要解决管理和技术两个层面的问题。管理层面主要包括协调机制和激励机制的建设。技术层面主要指共享平台的建立，管理和资源内容形式的多样化建设。

3.1 管理层面机制的建立

3.1.1 协调机制

由于各高校在规模、类型、层次方面的差别，对实验教学资源共享的需求及由此获得的利益各不相同：资源相对充足的院校，共享的积极性不高，而资源相对匮乏的院校，则强烈希望共享。因此，仅依赖各高校自发地去沟通和交流，显然是不够的。在区域高等实验教学资源共享系统中，应该由各级高等教育主管部门出面组织一个常设的专门机构，其主要职能是在宏观层面上协调各高校之间的利益关系，推动公共信息平台和资源标准化建设，以保证和推动区域资源共享的实现。

3.1.2 激励机制

各高校希望占有设备教学资源，强化自身的地位和优势，在此基础上实现共享。这样就可能出现以各自利益为重，故步自封，画地为牢，从而影响到校际间合作及资源共享的进程。所以应该制定相应的策略，对在校际实验资源共享过程中具有突出贡献的单位与个人予以奖励或补贴，同时对阻碍推进资源共享或置之事外的单位与个人予以告诫或惩罚，来影响高校主体需要的实现，提高他们参与资源共享的积极性，引导他们的理性行为。

3.2 技术层面机制的建立

3.2.1 实验教学资源共享平台的建立

推进实验资源和设备的共享，最主要的是共享平台的建设。共享平台具备以下功能。

1) 资源共享的网格支撑平台

资源共享的网格支撑平台采用网格理论与典型的应用技术，采用分布式资源存储和统一目录管理的模式，目标是将高校中自治的、分布的、异构的海量实验信息进行整合和集成，实现教育资源的有效共享，为高校的教学与科研提供高效的服务。各高校拥有自己的资源共享节点，并自行管理资源的共享，各高校的共享资源采用自动发现的机制，通过统一的平台界面入口进行发现和使用。各高校节点之间采用认证的方式使用对方共享资源，能有效地保证安全性。

2) 仪器仪表的共享管理系统

仪器仪表的共享管理系统部署在网格支撑平台上，包含仪器仪表的登记、管理，贵重仪器设备管理运作，共享机制等。

3) 实验资源共享系统

实验资源共享系统包含优质实验教学资源共享利用、远程实验和虚拟实验、贵重仪器设备共享利用等方面。

4) 网上服务系统

提交实验仪器的共享申请，管理员进行申请的审核和确认。给出具体的共享开放时间，登录口令等。

3.2.2 资源共享的内容形式

平台中支持的共享资源的内容形式较传统的方式要有大的改变。传统多以文档，基础网页内容为主要资源，现有平台建立起来后，应能够有效支持各种格式的实验录像视频文件，动画演示，实验题库，虚拟实验等多种形式资源。其中具体分为：

1) 计算资源的共享

这里的计算资源包括设备资源和专家资源。随着网络技术的发展，国际互联网已遍及各单位，地方高校都不同程度引进了一些大型精密仪器设备。但是有些设备利用率不高，多数处于闲置状态。而有的学校由于资金问题没有能力购置，这样在一个区域内就出现了有的设备闲置，有的急需此设备。为了有效利用资源，以提高仪器的利用率，避免大型仪器设备资源浪费，同时也为了避免不必要的重复投资，优化本地区高校的资源配置，改变各校原有资源分割闲置浪费的状况又可以在一个区域避免不必要的仪器重复购置，实现大型仪器的资源共享。例如，在欧盟资助的“数据网格”项目中，世界各地的 3000 名高能物理学家在网络上共同开发并共享数据和程序，因而提高了研究效率，同时加强了研究活动的竞争。

2) 数据资源的共享

各高校根据各自不同的应用需求、信息结构和计算机软、硬件环境等分别建立了各种各样的管理信息系统，这些系统依托与底层不同的数据库系统。这些数据库系统采用不同的数据结构、不同的数据类型、不同的编码方式、不同的表示形式和不同的检索方法等。如有 Foxpro，Access 等桌面型的数

据库管理系统，也有 Oracle，SQL Server 之类的大中型数据库管理系统。各类数据库分散在各个区域的分支机构中。现在，越来越多的用户需要同时访问和处理网络节点的多个异构数据库的数据，希望屏蔽各个层次的异构特性，他们不必知道各物理数据库系统的分布，也不必知道各物理数据库的结构组成，不必自己去进行数据转换和结果汇总，只需通过简单的全局查询便可以得到一个综合结果。

3) 网络教育资源的共享

教育资源是信息化教学的基础，随着教育信息化的深层次推进，互联网中的信息资源以指数方式增长，这些资源不仅在内容上多种多样，在表现形式上更是丰富多彩。这些资源主要包括素材类教育资源建设，如试题库、试卷素材、媒体素材、文献素材、课件素材、案例素材、常见问题素材和教育资源索引；网络课程建设，它是按照学科知识体系及网络教学的要求，对各种教育资源的综合集成，如各地各校的名师讲座，精品课程等。

4) 图书资源共享

把整个区域高校的图书馆及信息资源联结在一起，提高电子图书的共享信息，使得各高校不需要耗费大量的资金建设传统的图书馆。所有的信息资源均来自网络，图书馆建设可以实现跨越式发展，只要配备个人计算机，接入网格，就可以获得与全区域一样的信息处理网格服务，不需要知道哪台计算机在处理，不需要知道数据存储在哪里，网络总是以最快的速度、最经济的方式返回处理结果。

4 利用网格技术实现教育资源共享

4.1 网格技术的特点

网格技术是网格使多个联网的服务器节点充分利用各自的资源共同向用户提供服务能力的模式。它是伴随着互联网技术而迅速发展起来的一种新技术，这不仅能够为信息资源的获取、分布、传输和有效利用带来革命性和结构性的巨大变化，而且将根本改变我们的研究方式、教育方式、生活和生产方式。通过网格，可以在多个动态的虚拟组织之间共享资源，协同解决问题。网格计算建立在网格基础设施之上，它使人们可以以更自由、更方便的方式使用计算资源，解决更复杂的问题。首先，网格计算机扩展了以前十分有限的计算能力。在网格计算的支持下，人们可以方便地使用大型机，计算机机群和并行计算机来完成许多以前无法想象和无法完成的工作。其次，网格计算机突破了地理位置的限制，资源的提供者和使用者完全位置无关。再次，重要的一点就是网格打破了传统的共享和协作的限制。过去对资源的共享往往停留在数据文件传输的层次，而网格资源的共享允许对资源进行直接的控制。

目前，网络上的教育资源分布在不同的资源孤岛上。网格计算技术与网络教育资源的结合，可以消除这些资源孤岛，使用户一次登录就可以使用整个网格中的教育资源。

4.2 网格技术应用于高等学校教育资源共享平台的优势

网格技术最直接的应用是能促进教育资源的优化整合、全面共享。网格技术^[3]的实质就是实现 Internet 上汇集了成千上万的计算资源、数据资源、软件资源、各种数字化设备和控制系统的全面共享。基于网格技术将分布在各所高校的名师讲座、精品课程、特色课程、多媒体课件、数字图书馆、数字博物馆等各种海量信息资源集成起来，建立一个多媒体教育资源网格平台，提供统一、高效的信息服务。该平台将真正实现优质教育资源共享、实现高校的优势互补，大大节省高等教育投资成本，有效地避免了重复开发，充分发挥各所高校的优秀教育科研资源的效用。

(1) 各所高校的教师可以将典型案例在教师论坛上进行交流。特别，通过建设教师备课应用交流平台，提高教师的信息意识，促进各所高校的教师的交流。同时，加强了师生的互动性，培养学生学习的主动性和积极性，扩展教师教学思维的广度与深度。

(2) 通过建立网格化的资源部署平台和网格化的资源共享平台, 并利用数据网格技术规划资源部署实现资源共享管理体制。利用先进的个性化用户管理门户技术来实现用户界面。新的资源平台实现对多媒体教育资源的生命周期管理, 通过建立资源索引信息、描述信息、定位信息来实现各所高校的资源共享与访问。

(3) 基于该平台, 各高校都可部署一个网格节点, 存储和管理自身的多媒体教育资源, 实现各所高校的优质资源的共建共享环境, 促进各所高校教育科研的可持续发展。

(4) 各级各类题库, 仿真试验, 大规模数据计算等多种形式的共享内容均可以在该共享平台上进行有效访问。

4.3 网格技术应用于高等学校教育资源共享平台技术实现

(1) 在基于网格技术的教育资源组织方面, 将各种多媒体教育资源进行分类管理, 针对每种资源的特点采用不同的表示和组织方式。对所有的分布式教育资源建立统一的管理策略, 支持资源的共享服务。多媒体教育资源服务由多个节点互相协作来提供。利用面向对象思想对教育资源进行分类管理, 采用 XML 语言对资源统一描述, 支持资源的高效管理。

(2) 在构建教育资源网格平台方面, 开发教育资源管理、任务调度、用户管理和服务管理模块, 开发统一的系统门户。关注教育资源网格平台环境的硬件、软件配置。充分利用已有的硬件资源来搭建教育资源共享平台。利用平台将多台服务器互连, 每个物理节点都可以充当服务提供者和服务接受者两种角色。新加入的服务器只需部署上多媒体教育资源平台, 就可以融入网格环境中。

(3) 在资源的的服务发现和统一检索方面, 传统的共享系统支持一对多的服务相比, 应充分利用各服务器分布式部署视频资源, 利用网格服务的概念, 采用多服务器对多用户的服务模式, 为用户提供更好的可扩展的紫云共享服务。利用网格技术中MDS (Monitoring and Discovery Service) 的信息监控和发现的功能, 采用动态可扩展的框架来管理网格环境中各种多媒体教育资源的静态和动态信息, 支持各种服务的灵活集成。将分布在高等教育领域中的精品课程、名师讲座、特色课程、多媒体课件等各种海量资源集成起来, 提供统一分类整理, 为无缝式的资源共享提供资源基础。

5 结束语

本文将网格技术和教育信息技术结合在一起, 运用到教育资源的大规模共享。同时提出利用网格技术构建了即自主管理又互相协作的网络共享模式, 并给出了有效的数据管理和服务方法的具体实现, 较好地解决了高等教育资源共享问题的网络速度, 资源管理, 资源搜索等问题, 可以作为一个很好的参考方案。但各校之间的实际情况比较复杂, 应根据实际情况进行不断的完善和改进, 以便能更好地实现资源共享。下一步在实现基本共享的同时, 应该引入了由授权、认证、审核三整合机制决定的权限对资源管理分配使用的等级, 以确保最终使用的公平合理和安全性。

参考文献

[1] 朱瑾, 李捍无.从大学城模式谈高校资源共享[J]. 西安建筑科技大学学报: 社会科学版, 2006, 25(1): 94-95.
[2] 陈乙雄, 吴中福等. 利用网格技术促进高校教育资源的共享与共建研究[J]. 现代远距离教育, 2009, 4: 60-62.
[3] I Foster , C kesselman , 金海, 袁平鹏, 石柯译. 网格计算 (第二版) The Grid2: Blue print for a new computing Infrastructure [M]. 北京: 电子工业出版社, 2004. 149-162.

任职教育中计算机教学方法浅析

王月蓉, 冯少华, 黄欢欢, 金玉

(防空兵指挥学院, 河南 郑州, 450052)

摘 要: 本文针对任职教育培养对象特点, 结合多媒体技术在日常教育中的应用, 以《计算机科学文化基础》为例, 讲解任职教育中如何将多媒体技术与传统教育方法结合, 达到让学生高效率、高质量掌握计算机基础知识, 达到学以致用的教学目的。

关键词: 信息化; 任职教育; 多媒体技术; 教学方法

中图分类号: TP3-0 **文献标识码:** A **文章编号:** 1006-7043 (2010) xx-xxxx-x

On The Teaching Methods of Computer in Professional Education

WANG Yuerong, FENG Shaohua, HUANG Huanhuan, JIN Yu

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: Aiming at the characteristics of the objects of professional education, combining with the application of multimedia technology in their daily lives, taking the computer science and culture foundation for example, this paper explains how to combine the multimedia technology with the traditional education methods to make students master computer elementary knowledge with high efficiency and high quality and achieve the teaching goal of practicing.

Keywords: Informatization; Professional Education; Multimedia Technology; Teaching Method

目前, 为满足人才知识、技术、能力、素质和管理的新要求, 院校人才培养提高了对任职教育工作的重视。按照任职教育培养新要求, 任职教育应坚持能力本位, 重在提高任职需要的实用能力和素质, 突出实践应用, 重在提高分析解决岗位实际问题的能力水平^[1]。在使学生掌握扎实的专业基础理论的基础上, 强调实践应用, 突出培养学生利用所学知识解决实际工作中实际问题的能力。

任职教育与以往的学历教育相比, 培训周期明显缩短, 而且在参加任职教育培训的学生中, 绝大多数学生都是非计算机专业学生, 如何在短期内让学生掌握好计算机基础知识, 并能熟练操作, 达到在现实工作岗位上利用所学知识解决实际问题的目的, 成为任职教育培养教员不断思考、尝试和创新的方向^[2]。

在信息化要求下, 掌握相关的专业知识显得十分重要, 在任职教育中, 《计算机科学文化基础》是计算机相关知识学习的前导课程, 这门课程既有理论知识讲解, 也有操作实践学习, 是一门理论与实践相结合的课程。以笔者在这门课程教授过程中的体会, 结合多媒体教学方法, 帮助学生熟练掌握所学知识, 提高学习质量的心得, 对任职教育中计算机教学方法进行阐述。

1 深入了解教学对象, 针对不同教学对象, 同课不同上

在任职教育中, 学生在现实工作中担负的具体工作不同, 不同的工作对计算机基础操作要求也不尽相同, 针对不同学生, 深入了解学生的学习目的和学习方向, 对相同的课程内容, 根据教学对象的不同, 侧重点也不相同。

作者简介: 王月蓉 (1985—), 女, 助教, 学士;
冯少华 (1983—), 男, 助教, 学士;
黄欢欢 (1982—), 女, 助教, 学士;
金玉 (1984—), 女, 助教, 学士。

对于《计算机科学文化基础》这门课，在普通大学里，无论是计算机专业学生还是非专业学生都是作为基础课程来学习，主要引导学生了解计算机的硬件、软件组成，了解常用应用软件基本操作，了解网络组成和网络安全知识，是作为一个计算机入门课程来学习的。对于任职教育，教学对象虽然不是计算机专业学生，但是对计算机也有相当的认识，如果还是把这门课程当做入门课程来讲解就失去了任职教育的教学目的，针对教学对象，需要深入了解学生对计算机的掌握情况，了解他们平时在工作中需要用到哪些知识，针对了解情况，教学内容上就有了不同的侧重。比如，计算机硬件这部分内容，主要侧重讲解计算机硬件维护保养和兼容机箱内部构造和接线方法，帮助学生在平时工作中对计算机进行日常养护和简单故障排除；计算机网络部分，网络组成部分理论简单讲解，针对日常工作中关于安全保密的相关需要，将重点放在计算机安全网络安全上，教给学生一些实用的计算机防护技术和方法。做到同课不同上，针对问题有所侧重，达到学以致用教学目的。

2 发挥计算机多媒体的优势，激发学生学习兴趣

“知之者不如好知者，好知者不如乐知者”，学生是教学的本体。学生真正学到知识才是我们教育的最终目的。激发学生学习兴趣就是在现有的教学条件下利用一切教学手段，达到提高学习效率和质量的目的。在所有能提高学习效果的方法中，最高效的都一定是自主学习。通过调研：学生感觉到他们正在做有意义的事情，自己对学习内容充满兴趣，在教学过程中参与到学习中，并且知道自己所学的知识可以解决怎样的实际问题，并且在实践过程中利用所学知识完成某项内容，得到成就感和挑战感，在这样的情况下，学习效果才会是非常有效的学习。要促进学生的自主发展，达到高效学习的目的，就必须最大可能地创设让学生参与到自主学习中来的情境与氛围。“兴趣是最好的老师”，在课堂教学过程中，使用多媒体课件，将图像、音频、视频加入到课堂教学，吸引学生注意力，调动学生学习兴趣，这样学习才会有主动性和积极性，只有产生了兴趣，才会有动机，学生有兴趣时，注意力高度集中，思维异常活跃，求知欲异常强烈，能够发挥出潜在的学习积极性、主动性和自觉性，这样创造思维活动得以启动运行。

在计算机课堂教学中，首先应唤起学生的学习兴趣。比如在讲解计算机网络安全时，我们可以把好莱坞电影中有关信息窃取的片段剪辑后在讲课之前给学生放映，在学生看短片时，吸引他们的注意力，当他们对影片当中信息窃取技术和手段的高明折服时，对信息的安全保密问题有所思考时，他们的好奇心和学习的兴趣就被调动起来了，我们可以告诉学生一些相关的信息安全技术和保密措施。通过这样的方法，充分调动他们学习的学习兴趣，提高教学效果。

3 讲解要形象生动，多用生动的语言和形象的比喻

计算机学科中有些内容比较抽象，不容易被学生接受和理解，作为老师要多钻研教材教法，结合实际生活，将抽象的概念用现实的例子体现出来，化繁为简，将深奥的理论讲得通俗易懂。在教学中巧用生活中的实例，形象的比喻方法最容易被学生接受。例如，在讲操作系统分类时，批处理系统可以用会计账目管理的实例进行讲解，会计每收到十份账目，抽出一个小时时间集中将这些账目处理好，或者是每两天对收到的所有账目一次性全部结算一遍，说明批处理系统对用户请求的处理方式，当用户请求到达一定的数量或者到一定的时间时，操作系统集中在一个时间段内一次性处理这些请求。用饭店给同一个时间段内各个餐桌上顾客上菜的方式，说明分时处理系统的处理方式。这样学生就很容易懂得操作系统的工作方式和各种工作方式间的优缺点。把抽象复杂的原理简化为我们生活中常见的事物，这样既有利于学生的接受，又活跃课堂气氛。所以比喻法是我们教师的一把利刃，需要灵活利用。

4 真正把课堂的重心让给学生，要注意精讲多练

为了充分调动学生的积极性，尝试把学生分成若干小组，将学习能力强，掌握知识快的学生平均分到各组，分给每个小组特定的学习任务，学生有明确的任务还可以互助性学习。在课下各小组自主学习，课上利用练习时间抽取学生现场演示所学内容，各小组之间相互竞争，相互比较，有利地促进了学生动手的能力。合作学习还有助于学生整体的提高，每个小组的小教员帮助教员给本组成员辅导，督促学习，从而真正实现使每个学生都得到发展的目标。

在课堂教学中利用有限的时间，精讲重点理论知识和实际操作应用，在教学中主要讲清课程的要点和学生容易混淆的知识，节省下来的时间就交给学生，给他们布置一些实际工作中常用的一些任务，用课堂所学知识，给他们练习时间，注重学生学习方法的练习，把简单易懂的内容交给学生独立完成。多练就是让学生多上机操作，其目的是从培养学生的操作技能和实践能力入手，让学生多动手、多动脑，提高操作的准确性、迅速性、灵活性和协调性，尽可能调动学生积极性，训练他们动手能力，以便熟练掌握计算机应用技术。我们这里说的精讲就是，让我们的学生不但学会现有知识，还要培养他们独立自主的学习能力，才能跟得上计算机技术的迅猛发展，最终达到最好的教学效果。

参考文献

[1] 李贤温, 张术环. 论高职教育目标对师资队伍建设的要求[J]. 前沿, 2005, (8).
[2] 潘懋元, 谈松华, 吴岩, 陈智. 高等职业教育: 体系定位、发展与模式(笔谈)[J]. 教育研究, 2005, (5).

二叉树的四种遍历的非递归算法

王正辉，姜鹏飞，张锋

(防空兵指挥学院，河南 郑州，450052)

摘 要：二叉树是数据结构中典型的、也是非常重要的非线性结构，它在实际生活中有着广泛的应用。本文主要介绍数据结构中二叉树的先序、中序、后序和层序的非递归算法。

关键词：二叉树先序遍历；二叉树中序遍历；二叉树后序遍历；二叉树层序遍历

Binary Tree Traversal Non-recursive Algorithm for the Four

WANG Zhenghui, JIANG Pengfei, ZHANG Feng

(Air Defense Forces Command Academy, Zhengzhou 450052, Henan China)

Abstract: Binary tree is an important and typical nonlinear structure in data sturacture.It is widely used in our life. This paper describes the binary tree preorder, inorder,post-order and layerorder non-recursive algorithm in data structure, .

Keywords: preorder binary tree traversal ; binary tree inorder traversal ; post-order binary tree traversal ; layerorder binary tree traversal

1 引言

二叉树是一种重要的树形结构，其结构规整。许多实际问题抽象出来的数据结构往往是二叉树的形式，而且其存储结构及运算都较为简练，因此，二叉树在数据结构课程中显得特别重要，这里先了解一下二叉树。

二叉树是由节点的有限集合构成，这个有限集合或者为空集，或者是由一个根节点及两棵互不相交的分别称为这个根的左子树和右子树的二叉树组成。从定义来看，二叉树定义是个递归定义，但由于学生对递归算法的理解存在一些误区，同时递归算法效率较低，并且在实际应用中有些问题也不允许调用递归算法，故有关二叉树的试题通常要求采用非递归算法，这就使得掌握二叉树的生成及遍历的非递归算法成为必要。

2 二叉树的先序、中序、后序和层序的非递归遍历

二叉树的遍历就是对二叉树的每一个节点访问一次且仅访问一次。教材上讲述的都是递归的算法，但是递归算法效率比较低，时间复杂度和空间复杂度都比较大。换一种角度思考，采用非递归算法进行遍历。下面对每种遍历都采用非递归算法进行遍历。二叉树非递归遍历是用显示栈来存储二叉树的节点指针。

2.1 先序遍历

从递归说起

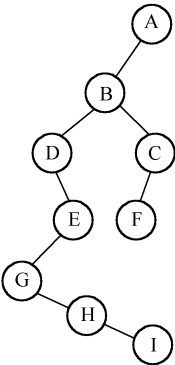
```
void preOrder(TNode* root)
{
    if (root != NULL)
    {
        Visit(root);
        preOrder(root->left);
        preOrder(root->right);
    }
}
```

递归算法是先访问根节点，然后访问左节点，再访问右节点。如果不用递归，那该怎么做呢？仔细看一下递归程序，就会发现，其实每次都是走树的左分支（left），直到左子树为空，然后开始从递归的最深处返回，再开始恢复递归现场，访问右子树。

其实过程很简单：一直往左走 $root \rightarrow left \rightarrow left \rightarrow left \cdots \rightarrow null$ ，由于是先序遍历，因此一遇到节点，便需要立即访问；由于一直走到最左边后，需要逐步返回到父节点访问右节点，因此必须有一个措施能够对节点序列回溯。有两个办法如下：

（1）用栈记忆：在访问途中将依次遇到的节点保存下来。由于节点出现次序与恢复次序是反序的，因此是一个先进后出结构，需要用栈。

（2）节点增加指向父节点的指针：通过指向父节点的指针来回溯。



```
void preOrderTNode* root)
{
    Stack S;
    while ((root != NULL) || !S.empty())
    {
        if (root != NULL)
        {
            Visit(root);
            S.push(root);                //先序就体现在这里了，先访问，再入栈
            root = root->left;          //依次访问左子树
        }
        else
        {
            root = S.pop();             //回溯至父亲节点
            root = root->right;
        }
    }
}
```

采用此算法对上树的先序遍历结果是 ABDEGHICF。

2.2 中序遍历

中序遍历是先尝试向左走，一直到左边不通后访问当前节点，然后尝试向右走，右边不通，则回溯（这里不通的意思是：节点不为空，且没有被访问过）。

```
void InOrder (TNode* root)
{
```

```

while ( root != NULL )           // 回溯到根节点时为 NULL，退出
{
while    ( root->left != NULL && !root->left->bVisited )
    {
        // 沿左子树向下搜索当前子树尚未访问的节点
        root = root->left;
    }
    if ( !root->bVisited )
    {
        // 访问尚未访问的最左节点
        Visit(root);
        root->bVisited=true;
    }
    if    ( root->right != NULL && !root->right->bVisited )
    {
        // 遍历当前节点的右子树
        root = root->right;
    }
    else
    {
        // 回溯至父节点
        root = root->parent;
    }
}
}

```

采用此算法对上树的中序遍历序列为 DGHIEBFCA。

2.3 后序遍历

从直觉上来说，后序遍历对比中序遍历难度要增大很多。因为中序遍历节点序列有一点的连续性，而后续遍历则感觉有一定的跳跃性。先左，再右，最后才中间节点；访问左子树后，需要跳转到右子树，右子树访问完毕，再回溯至根节点并访问。这种序列的不连续造成实现前面先序与中序类似的两个程序比较困难。但是按照第 2 个思想，直接来模拟递归还是非常容易的。

```

void PostOrder(TNode* root)
{
    Stack S;
    if(    root != NULL )
    {
        S.push(root);
    }
    while  ( !S.empty() )
    {
        TNode*    node = S.pop();
        if    ( node->bPushed )
        {
            // 如果标识位为 true，则表示其左右子树都已经入栈，那么现在就需要访问该节点
            Visit(node);
        }
        else
        {
            // 左右子树尚未入栈，则依次将右节点，左节点，根节点入栈
            if    ( node->right != NULL )

```

```

        {
            node->right->bPushed      = false;           // 左右子树均设置为 false
            S.push(node->right);
        }
        if      ( node->left != NULL )
        {
            node->left->bPushed      = false;
            S.push(node->left);
        }
        node->bPushed      = true;           // 根节点标志位为 true
        S.push(node);
    }
}

```

采用此算法对上树后序遍历序列为 IHGEDFCBA。

2.4 层序遍历

层序遍历的非递归算法，用队列完成

```

void LevelOrder(TNode *root)
{
    Queue Q;
    Q.push(root);

    while  (!Q.empty())
    {
        node      = Q.front();           // 取出队首值并访问
        Visit(node);

        if      (NULL != node->left)      // 左孩子入队
        {
            Q.push(node->left);
        }
        if      (NULL != node->right)     // 右孩子入队
        {
            Q.push(node->right);
        }
    }
}

```

采用此算法对上树层序遍历序列为 ABCDEFGHI。

3 总结

二叉树在计算机科学中有着重要的作用，现实生活中有好多问题都是用树这种数据结构描述的，而二叉树在计算机中操作和实现都非常方便，因此二叉树的建立及其遍历是非常重要的。同时二叉树

的四种遍历算法是二叉树运算的基础，大多数二叉树上的复杂运算都是建立在这四种遍历之上的。因此，要深刻理解这四种遍历算法是十分必要的。

参考文献

[1] 严蔚敏，吴伟民．数据结构（C 语言版）[M]．北京：清华大学出版社，2007.

[2] 王昆仑，李红．数据结构与算法[M]．北京：中国铁道出版社，2007.

[3] 降春葆．数据结构习题与解析（第 2 版）[M]．北京：清华大学出版社，2004.

[4] 刘大有，唐海鹰．数据结构[M]．北京：高等教育出版社，2001.

[5] 朱战立编著，数据结构——使用 C 语言（第 3 版）．西安：西安交通大学出版社，2003.

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：(010) 88254396；(010) 88258888

传 真：(010) 88254397

E-mail: dbqq@phei.com.cn

通信地址：北京市万寿路 173 信箱

电子工业出版社总编办公室

邮 编：100036

征 稿 启 事

电子工业出版社是工业和信息化部直属的科技与教育出版社，享有“全国优秀出版社”、“‘讲信誉、重服务’的优秀出版社”、“全国版权贸易先进单位”、“首届中国出版政府奖”、“先进出版单位”等荣誉称号，在全国拥有良好的品牌声誉和市场占有率。

我社已在河南地区出版教材、学术类著作多本，在研发和编写过程中积累了丰富的经验。我们希望，有志于打造优秀计算机学科教材的学界同仁们，能够奉献出更多优秀的作品。

如果您对本套教材有什么好的意见和建议，也可以随时向我们反馈。

投稿咨询：

何况：010-88254596 E-mail : hekuang@phei.com.cn

地址：北京市万寿路南口金家村 288 号院华信大厦

邮编：100036