文章编号:2095-476X(2012)06-0001-04

基于语义的KNN短文本分类算法研究

张素智, 刘婧姣

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450002)

摘要:针对短文本分类关键词特征稀疏和样本数量多,难以处理的技术难点,提出一种基于语义的KNN短文本分类算法.该算法采用基于字的分词策略提取出短文本的特征词,结合中国知网对关键词进行概念映射以提高短文本的语义表达,并针对短文本特点,通过使用LSA降维处理,对KNN分类算法加以改进.实验结果表明,该算法能够有效提高短文本的分类性能.

关键词:短文本;文本分类;语义扩展;KNN分类算法

中图分类号:TP391 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.001

A short text KNN classification algorithm based on semantic

ZHANG Su-zhi, LIU Jing-jiao

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450002, *China*)

Abstract: Aiming at the problems of key words sparse features, sample quantity of the short text classification and differcult dealing with, a method based on semantic KNN short text classification algorithm was presented. The algorithm extracts short text feature words based on the word segmentation strategy, combining CNKI to key for concept mapping to improve the short text semantic expression, KNN classification algorithm was improved according to the characteristics of short text through application of LSA dimensionality reduction. The experiment results showed that the algorithm can effectively improve the short text classification performance.

Key words: short text; text classification; semantic expansion; KNN classification algorithm

0 引言

随着我国互联网的发展,微博、在线聊天记录、BBS标题、博客观点等各种形式的短文本迅速增多,并逐渐成为人们沟通交流和信息获取的一种重要方式.短文本数量多、信息量大,包含了人们对社会各种现象的评价反映,话题涉及政治、经济、娱乐、军事、生活等众多领域.如果能够对这些短文本进行有效分类,将会有助于舆情预警、流行语分析、话题跟踪与发现等.因此如何对海量的网络短文本数据进行分类,逐渐成为近年来相关研究领域的热点.

短文本即包含字符数量较少(通常不超过200个)的文本,其内容简短,语言多不规范[1],所以在常规文本分类中成熟运用的分类技术并不能较好地适用于短文本分类.短文本分类面临的难点主要有:1)关键词特征稀疏.与一般的长文本相比,每个

收稿日期:2012-02-26

基金项目:郑州市科技攻关计划项目(0910SGYG23259—3)

作者简介:张素智(1965—),男,河南省孟州市人,郑州轻工业学院教授,博士,主要研究方向为Web数据库、分布式计算和异构系统集成.

短文本中一般只含有数十个甚至几个关键词,描述信号弱,难以充分挖掘出特征之间的关联性. 因此,要对短文本分类最重要的工作就是提高特征词的表达能力. 针对中文短文本分类,最常用的方法是结合知网进行特征词语义扩展. 2)样本数量多,处理复杂. 单条短文本长度太短,难以挖掘出有效特征和有价值的信息,因此对短文本的研究和处理一般针对整个短文本语料库,数量庞大. 目前专门针对短文本分类的研究工作还较少. 在方法上,主要还是直接采用长文本的分类算法,常用的有 KNN, Bayes,SVM 和决策树等. 向量空间模型中最好的分类算法之一[2],KNN 最直接地利用了样本与样本之间的关系,减少了类别特征选择不当对分类造成的不利影响,可以最大程度地减少分类过程中的误差项;但是为了找出待分类样本的 k 个邻居,需要与样本空间中的每个样本向量做比较,当训练样本较多时会产生巨大维数的文本特征向量,计算开销很大,导致分类速度下降,因而 KNN 算法并不能直接应用于短文本分类.

针对短文本分类的特点,本文提出一种基于语义的 KNN 短文本分类算法,以期提高短文本分类的精确率、查全率和 $F1$ 测试值.

1 基于语义的 KNN 短文本分类算法

1.1 算法思想

KNN 算法的基本思想[3]是:如果一个样本在特征空间中的 K 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别,则该样本也属于这个类别. KNN 算法是一种基于实例的学习方法,它不需要建立模型,只需要逐个计算待分类 K 个样本的类别作为待分类文本 D 的候选类别,然后将 D 与这个邻居的相似度作为 K 近邻文本所在类别的权重,将各类里邻居文本的类权重之和作为该类别和测试文本的相似度,再把待分类文本 D 分到权重最大的类别中去.

KNN 算法可以最大程度地减少分类过程中的误差项. 但是其缺点主要是高维文本向量规模较大时,算法的时间和空间复杂度较高,并且有相当多的维数对于文本分类意义不大甚至成为噪音数据,影响分类的准确性. 因此本文为了降低维数,针对短文本的特点,首先从特征处理开始进行精简,提取价值量高的特征,从源头降低噪音;然后对文本特征向量运用 LSA 理论进行降维处理,以提高 KNN 算法的运行效率和分类精度. 算法整体框架如图 1 所示.

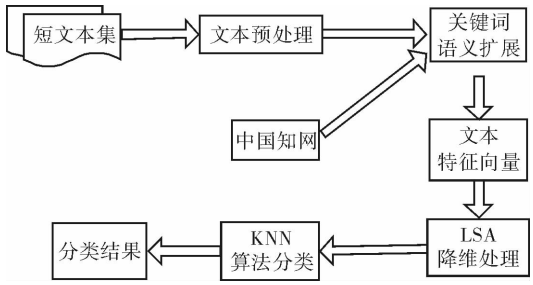


图1 基于语义的 KNN 短文本分类算法框架

1.2 文本预处理

对输入的短文本集,首先根据特殊字符,如句号、逗号、顿号等标点符号,还有空格、换行、回车、制表符等间隔符作为字符串划分的依据进行划分,将一个短文本划分成若干以字为单位的字符串,以便下一步处理. 本文从含有包括符号和数字在内的 1 028个常见停用词表中筛选得到 362 个停用字,包含数字、符号、连词等,作为停用字表. 通过对短文本中特征字与停用字表的匹配,删除相匹配的停用词.

1.3 分词算法

为了得到便于计算机识别和处理的语义单位,如“开发”、“和谐”等词,需要经过一定的分词策略将中文文本切割成由词组成的词序列,把独立的词从原文中区分出来. 由于短文本往往只有一句或者几句话,词汇量较少,采用基于词的预处理方式往往会忽略掉许多文本的源信息. 近几年来,基于字的分词方法渐渐引起关注[4]. 如果能充分挖掘出短文本中字与字之间的关系,则使文本主题信息能够表达得更为精确,因此本文参照文献[5]中基于字的分类方法对短文本进行分词.

文献[5]的思想是以互信息的方法把汉字分类,那么分词的过程就变成字分类的过程,即将文本与建立好的字库进行匹配,然后输出 4 个类别:跟它前面结合的字、跟它后面结合的字、跟它前后结合的字、独立的字. 本文将其应用到短文本的分词处理上,先建立字库,然后选取出关键字集,只对关键字集里面的字进行字库匹配,输出关键字相关的字串,用于后期结合中国知网进行的语义扩展.

互信息来源于信息论,是一个基于熵的信息度

量概念,它用来度量 2 个随机变量间的统计相关性. 形式化可以表示为

$$MI(W_i,W_j) = \log\frac{P(i,j)}{P(i)P(j)}$$

式中,W_i 和 W_j 分别代表字 i 和字 j,以一个短文本作为一个窗口单元,其中

$$P(i,j) = \frac{f_{ij}}{f_i + f_j - f_{ij}} \qquad P(i) = \frac{f_i}{\sum f}$$

其中,f_i,f_j 分别表示 W_i 和 W_j 在短文本中出现的次数;$\sum f$ 为一个短文本中的总字数. MI 值越大,说明这 2 个字之间的结合程度越高,关联程度越强.

互信息能够充分考虑文本中低频词的重要性,使得低频词可能具有较大的信息值. 这对于字数少、低频词大量存在的短文本运用非常有优势. 本文中互信息有两方面的用途,一是根据互信息进行字库的建立[5],二是根据互信息进行字的共现权重的计算. 具体方法是:针对每一个短文本句子,计算出每个字对的互信息值 $MI(i,j)$,分别对每个相关的特征字 i 和 j 进行赋值,作为它们各自的共现信息量;然后扫描文本中的所有特征字,分别统计出每个特征字的共现信息之和,其中特征字 i 的共现信息之和为 $\sum_{j=1}^{n} MI(i,j)$,并计算出文本中所有特征共现信息总和 $\sum_{i=1}^{n}\sum_{j=1}^{n} MI(i,j)$.

每个特征项的共现度量

$$M = \sum_{j=1}^{n} MI(i,j) \Big/ \sum_{i=1}^{n}\sum_{j=1}^{n} MI(i,j)$$

采用向量空间模型作为短文本的特征表示模型,特征项权重计算采用 TFIDF 归一化公式.

$$w(t,d) = \frac{tf(t,d) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t\in d}[tf(t,d) \times \log(N/n_t + 0.01)]^2}}$$

其中,$w(t,d)$ 为特征项 t 在文本 d 中的权重,$tf(t,d)$ 为特征项 t 在文本 d 中的词频,N 为训练集文本总数,n_t 为训练集中出现特征项 t 的文本数,分母为归一化因子.

综合考虑词频权重和特征字的共现度量信息,给出特征字的共现权重计算公式

$$W(t,d) = \lambda \cdot w(t,d) + (1-\lambda)M_t$$

其中,M_t 为特征项 t 的共现度量;λ 为共现度量参数,取 $\lambda = 0.4$. 由此可以依据共现权重的大小进行排列,取前 m 个权值最大的特征字作为短文本的关键字集.

1.4　基于中国知网的概念映射

短文本的样本特征非常稀疏,并且表达简洁,用语不规范,造成短文本中语义特征稀疏,同义词、多义词现象颇多. 因此本文结合中国知网中类层次结构及属性约束等特点,将关键字对映射到概念,以此进行关键字对的语义扩展,以提高短文本的语义表达能力,使得主题相同、包含不同同义词和近义词的文档能更好地分为同类. 具体步骤为:

1)预处理后得到的关键词集 $C = \{c_1, c_2, \cdots, c_m\}$ 中第 i 个关键词 c_i 送入到中国知网中进行语义扩展(因为中国知网中的概念是通过属性集合进行定义的,所以通过关键词与属性的匹配就可以得到相应的概念).

2)若匹配成功,则用概念 k_m 来代替关键词 c_i.

3)若匹配不成功,则表明关键词 c_i 不能满足中国知网中的任何概念,直接保留,作为未登录词对待.

4)关键词全部匹配后,要处理所有的概念特征和未登录词. 前者需要合并相同概念的词,后者需要事先设定一个阈值,如果该词的权重大于事先设定的阈值,则对其予以保留,否则就将其删除.

1.5　改进的 KNN 短文本分类算法

基于语义的 KNN 短文本分类算法具体步骤如下:

1)采用基于字的共现分析对文本进行预处理,根据文本特征词形成测试文本特征向量矩阵;

2)运用 LSA 理论对文本特征矩阵做降维处理;

3)利用余弦定理计算测试文本与训练集中每个文本的文本相似度,根据相似度,在训练文本集中选出与新文本最相似的几个文本作为邻居;

4)在测试文本的几个邻居中,依次计算每类的权重;

5)比较类的权重,将文本分到权重最大的那个类别中.

2　实验结果与分析

2.1　性能评估指标

笔者采用如下 3 个指标对分类实验结果的性能进行评估.

1)查准率 $P = A/B \times 100\%$,其中 A 代表正确分

为某类的文本数,B 代表实际分为此类的文本数.

2)查全率 $R = A/C \times 100\%$,其中 A 代表正确分为某类的文本数,C 代表属于该类的文本数.

3)查准率和查全率反映了分类质量的 2 个不同方面,两者必须综合考虑. 因此综合考查查全率和查准率得到一个新的评估指标,即 $F1$ 测试值(*F-measure*):

$$F1 = \frac{2PR}{P+R} \times 100\%$$

2.2 实验评估及分析

本文选取新浪、搜狐等各大网站的新闻标题或者专题网友评论为语料集中的2 916个短文本语料,包含娱乐、经济、体育、军事、计算机、汽车等类别. 从中随机抽取1 853篇作为训练样本集,其中娱乐 423 篇、经济 362 篇、体育 268 篇、军事 190 篇、计算机 315 篇、汽车 295 篇,其余 1 063 篇作为测试样本集. 实验结果见表 1.

表 1 各个实验的查准率、查全率及 $F1$ 测试值对照表 %

	类别	P	R	$F1$
实验 1	娱乐	70.81	73.59	72.17
	经济	56.83	79.42	66.25
	体育	53.25	60.87	56.81
	军事	62.48	75.83	68.51
	计算机	50.22	76.94	60.77
	汽车	60.45	65.41	62.83
实验 2	娱乐	70.94	83.16	76.57
	经济	65.29	80.71	72.19
	体育	56.54	86.59	68.41
	军事	67.98	84.37	75.29
	计算机	69.66	72.47	71.04
	汽车	62.51	73.15	67.41
实验 3	娱乐	71.46	87.64	78.73
	经济	70.89	81.23	75.71
	体育	60.78	90.12	72.6
	军事	74.13	87.35	80.2
	计算机	79.68	80.19	79.93
	汽车	69.88	78.24	73.82

实验 1 是未经中国知网语义扩展的短文本分类情况;实验 2 是经中国知网扩展但没有进行 LSA 降维处理的短文本分类情况;实验 3 是结合中国知网,并且经过 LSA 降维处理的短文本分类情况. 从表 1 可看出,经过中国知网的语义扩展,并经 LSA 维数处理的 KNN 短文本分类算法在查准率、查全率和 $F1$ 值方面都有较大的提高.

3 结论

面对短文本分类难的问题,本文提出了一种基于语义的 KNN 短文本分类算法,通过中国知网对提取出的文本关键字进行概念映射,扩展语义表达能力;针对 KNN 分类算法在短文本分类处理方面的弊端,使用 LSA 降维处理实现改进. 实验结果表明,该方法能够有效提高查准率、查全率和 $F1$ 测试值.

参考文献:

[1] 王盛,樊兴华,陈现麟. 利用上下位关系的中文短文本分类[J]. 计算机应用,2010,30(3):603.

[2] Yang Y M,Liu X. A re-examination of text categorization methods[C]//Proceedings 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'99),Berkeley:ACM Press,1999:42-49.

[3] 江涛,陈小莉,张玉芳,等. 基于聚类算法的文本分类算法研究[J]. 计算机工程与应用,2009,45(7):153.

[4] Xue N W,Shen L B. Chinese word segmentation as LMR tagging[C]//Proceedings of the Second SIGHAN Workshop on Chinese Language Processing,Stroudsburg:ACL,2003:176-179.

[5] 韩月阳,邓世昆. 基于字分类的中文分词的研究[J]. 计算机技术与发展,2011,21(7):29.

文章编号:2095-476X(2012)06-0005-04

基于对话的多 Agent 协作模型研究

邓璐娟, 陈培, 潘凯洁

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450000)

摘要:针对传统协作设计过程中的盲目性和随意性问题,提出一种基于对话的多 Agent 协作模型. 该模型首先利用"与/或"结构任务树的形式来简化任务,再通过划分通信区域建立区域 Agent 服务器,以提高通信效率;再依据对话模型定义 Agent 对话交互语义,并将其应用于多 Agent 的协作过程. 仿真结果表明,该模型能够为多 Agent 协作提供一种灵活、有效的交互手段,并能够显著提高系统的运行效率.

关键词:任务分解;Agent 通信;对话语义;协作模型;交互手段

中图分类号:TP311.4 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.002

Study on the cooperation model of the multi-Agent based on dialogue

DENG Lu-juan, CHEN Pei, PAN Kai-jie

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450002, *China*)

Abstract: Aiming at the blindness and randomness of the traditional cooperation in design process, a dialogue-based multi-Agent cooperation model was proposed. The task was simplified by the task tree form of "and /or" structure. Communication efficiency has been raised through dividing the communications area and establishing Agent server. Agent dialogue interaction semantics were defined according to the dialogue model and used in multi-Agent cooperation process. The simulation results showed that the model which is Agent cooperation design provides a flexible and effective means of interaction, and the efficiency of the system can be significantly improved.

Key words: task decomposition; Agent communication; dialogue semantics; cooperation model; interactive tools

0 引言

多 Agent 系统提供了一种解决复杂问题的分而治之的方法. 传统的多 Agent 协作设计过程中,每个 Agent 对于所要完成的任务拥有不全面的信息和能力,无法根据协同任务中其他 Agent 的行为需求来调整优化自身行为,从而影响协作效率和系统的灵活性[1]. 本文针对传统的多 Agent 协作设计过程中的盲目性和随意性,提出一种基于对话的多 Agent 协作模型,以期在一定程度上实现多 Agent 之间的理性交互,为多 Agent 协作提供一种灵活、有效的交互手段.

1 任务分解

把问题不断地分解和细化是解决复杂问题的重要方法. 对于任务耦合度较大的 Agent,建立一种

收稿日期:2012-09-14

基金项目:国家自然科学基金项目(61040025)

作者简介:邓璐娟(1964—),女,湖南省浏阳市人,郑州轻工业学院教授,博士,主要研究方向为控制理论和控制工程.

"与/或"结构来描述任务分解是一种有效的方法. 根据这种层次的树状结构,规定树的根节点为总请求任务,树中同层节点间具有"与/或"关系. 若节点 T 有 n 条边分别链接子节点 $T_1,T_2,\cdots,T_n$,且这 n 条边为逻辑"与"的关系,则需要多个 Agent 协作完成任务 T,即 $T = T_1 \wedge T_2 \wedge \cdots \wedge T_n$;若 n 条边为逻辑"或"的关系,则只需要其中的某个 Agent 就可以独立完成任务 T,即 $T = T_1 \vee T_2 \vee \cdots \vee T_n$. 采用这种方式将任务逐层地划分下去,一直划分到满足最小子任务的条件为止,子任务之间也存在"与/或"关系. 根据这种逻辑关系,对每个分支通过目标函数 M 进行综合开销运算,即

$$M = \sum_i \sum_j Z_{ij} ExecFun(A_i,E_j) + \sum_i \sum_j W_{ij} Commfun(E_i,E_j)$$

式中,$ExecFun(A_i,E_j)$ 为 Agent j 操作 i 的执行开销,$Commfun(E_i,E_j)$ 为 2 个 Agent 间的通信开销,Z_{ij} 和 W_{ij} 分别代表 $ExecFun(A_i,E_j)$ 和 $Commfun(E_i,E_j)$ 在综合开销运算中所占权重. 根据 M 算出每个分支的综合开销,保留最小消耗分支去除其他分支并删除该"或"分支的根节点,可完成对该任务树的修剪,达到简化任务的目的.

2 多 Agent 通信模式

多 Agent 通信模式是:建立一个 Agent 通信区,将交互程度紧密的 Agent 通过总的区域 Agent 置于同一个通信区内,共同信息以黑板系统的方式经过共享通道传递给区域内成员.

符号名是各个 Agent 识别的标识. 为了保证一致性,采用统一的 Agent 命名机制,将区域管理 Agent 命名为管理 Agent(MA),MA 的职责是管理区内其他 Agent 的名字,维护相关的信息(如 Agent 的专业领域、角色类型、完成任务的能力大小和当前活动的状态等). 新加入系统的 Agent,必须先向 MA 进行名称注册,注册成功后的 Agent 具有同域中其他 Agent 进行通信的权利[2]. MA 也会及时记录各个 Agent 的能力或其他参数的变化情况. 当任务发生时,请求 Agent 通过询问 MA 来选择相应的通信对象 Agent 来承担任务,此时可以将 MA 理解为通信服务器. 在原始的通信模式中,每个 Agent 需要维护一张各个 Agent 的地址和功能表,不仅耗费通信资源,而且效率也不高. 相比较而言,通过 MA 的方式来完成 Agent 之间的寻址和定位的方式更加灵活有效.

3 基于对话的多 Agent 协作模型

3.1 基于对话的多 Agent 协作模型设计

对话模型分为信息搜索对话、查询对话、劝说对话、协商对话、慎思对话 5 种语言[3]. 一次对话至少发生在 2 个 Agent 之间,说话者 Agent 需要选择正确的语言和语法有计划地表达自己的意图,听话者 Agent 通过分析说话者的言语行为来准确理解其意图,它不仅关心自身的状态和行为,还要关注其他 Agent,最终在服务于总体任务的条件下,根据说话者 Agent 的建议修改自身的愿望和意图.

在对话过程中,用 T 表示协作请求任务,Agent A 为任务的发起者,通常 A 即为通信模型中的 MA,$B = \{b_1,b_2,\cdots,b_n\}$ 表示目前整个协作系统中的 Agent 集合.

基于对话的多 Agent 协作过程可分为以下 3 步.

第 1 步:任务发起者 A 依次与任务承担着 b_i 建立信息搜索对话,询问其专业领域及参与完成任务 T 的能力、意愿、机会,以确定 b_i 能否成功参与协作并完成任务. 查询过程如下:

A 查询 b_i 专业领域的征询式为

$$\mathrm{req}_{A,b_i}(\text{specific of } b_i,t)$$

b_i 根据自身状态知识库并结合对话策略做出合理的应答,即

$$\mathrm{statement}_{b_i,A}(\text{"Vehicle Enigeering"},t+1)$$

根据 b_i 的回答,A 能够初步判断 b_i 能否承担起任务 T. 若 b_i 具有这样的能力,则 A 进行对话;否则结束对话,继续寻找下一个合适的服务 Agent.

A 询问 b_i 承担任务 T 并提供服务的愿望征询式为

$$\mathrm{req}_{A,b_i}(\mathrm{Accept}(b_i,T),t+2)$$

b_i 对于承担 T 的意愿有 2 种可能,即 Accept (b_i,T)(b_i 愿意承担任务 T)和 Refuse (b_i,T)(b_i 不愿意承担任务 T). 因此,b_i 可做出如下 2 种应答:

1) $\mathrm{prom}_{b_i,A}(\mathrm{Accept}(b_i,T),t+3)$

2) $\mathrm{prom}_{b_i,A}(\mathrm{Refuse}(b_i,T),t+3)$

基于 b_i 的应答,任务发起者 A 就会根据"if trust,then believe"的规则更新其社会模型知识库,最终形成一个愿意共同完成协同设计任务的潜在 Agent 集合 G.

第 2 步:根据信息搜索对话的结果,A 选用劝说对话,依次劝说 G 把实现请求任务 T 作为其意图,也就是说 G 中每个 Agent 均把实现任务 T 作为它们各

自的意图. 劝说过程如下:

当 b_i 做出的应答为 Refuse (b_i,T) 时,往往还会给出拒绝的原因,即

$$\inf \mathrm{or}_{b_i,A}(\text{"reason"},t+4)$$

A 对 b_i 的拒绝原因给予考虑,并再次提出改进建议后的请求,即

$$\mathrm{prom}_{A,b_i}(\mathrm{Accept}(b_iT'),t+5)$$

若 b_i 能够接受此建议,则向 A 做出回应

$$\mathrm{prom}_{b_i,A}(\mathrm{Accept}(b_i,T'),t+6)$$

或者 b_i 也可根据自身的情况对 A 的建议提出反建议

$$\mathrm{req}_{b_i,A}(\text{Counter-Advice}(A,T''),t+7)$$

最后 A 接受 b_i 的反建议,协商成功,做出回应

$$\mathrm{prom}_{A,b_i}(\mathrm{Accept}(A,T''),t+8)$$

第 3 步:当 G 中的所有成员都拥有了上述意图,A 将以联合信念的形式予以确认,并广播到每个成员,最终形成联合意图,即各个 Agent 间能够成功进行合作. 到此,此次协作过程结束.

3.2　基于对话的多 Agent 协作过程效用分析

不确定性和不可靠性一直是多 Agent 间协作的重要障碍. 例如完成某项总任务需要 5 个 Agent 共同协作,然而潜在的协作 Agent 集合 $H=\{20\}$,根据随机事件规律,恰好选中的这 5 个 Agent 的概率低于 0.01,这样选择的盲目性很可能导致执行任务结果异常. 针对这一问题,采用对话协作模型的优势显而易见.

根据每个预分解设计子任务,A 采取广播方式对 b_i 的专业领域进行并行查询,进而了解到该查询对象 Agent 的相关参数,包括完成子任务的能力、兴趣和机会. 在满足专业领域匹配的条件下,再比较兴趣和机会的概率值[4],若两者也满足某给定值,那么能力值高的 Agent 优先被选择. 由于对话带有时间属性,因此若对象 Agent 在规定时间内没有做出响应则会被认为是拒绝请求,放弃任务. 一旦确定选择 A,它将会从后续任务的备选集合中被剔除. 在这种对话模式下,任务分解分配的盲目性和随机性可以得到有效控制.

由于 Agent 的专业领域、能力、兴趣和机会具有差异性,如何选择并确定最合适的 Agent 是衡量协作效率的另一个重要指标[5]. 然而任何的协作交互都需要占用系统资源,造成通信开销,能否正确地对 Agent 间的协作效用与开销进行合理的计算分析,就显得非常重要. 多 Agent 协作效用与开销的计算式对话协商效用评价模型为

$$(A,B,P_B,P_{A,B})$$

其中,A 表示 Agent A,即对话协商发起者;B 表示 Agent B,即对话应答者;P_B 表示对话中 Agent B 所需占用的系统资源,即所需要付出的代价;$P_{A,B}$ 表示对话协商中 Agent A 认为 Agent B 所需要付出的代价,即理论上系统认为可以为此次对话分配的系统资源. 将代价 P 简化成维持一个 Agent 的代价 C_1,建立对话的代价 C_2,进行对话协商的代价 C_3. C_1 的权重为 ξ_1,C_2 和 C_3 的权重均为 ξ_2,则得到

$$P_B=\xi_1C_1+\xi_2(C_2+C_3)$$

$$P_{A,B}=\xi_1C'_1+\xi_2(C'_2+C'_3)$$

其中,$C_1=\sum_{i=1}^{\lambda}\theta_iP_it$,$\theta_i$ 表示第 i 个 Agent 的状态,活动状态 $\theta_i=1$,挂起状态 $\theta_i=0$;t 是活动的时间. λ 个 Agent 的协作效用 $F(\lambda)=\sum_{i=1}^{\lambda}P_{A,i}$,该协作效用等价于协作中所有 Agent 的代价之和.

对话协商的效用分析可以定义为

$$\begin{cases}F(\lambda)=\sum_{i=1}^{\lambda}P_{A,i}\\ F(\lambda)\leqslant\sum_{i=1}^{\lambda}P_i & \text{拒绝 Agent } i \text{ 加入协作集合}\\ F(\lambda)>\sum_{i=1}^{\lambda}P_i & \text{接受 Agent } i \text{ 加入协作集合}\end{cases}$$

综上所述,管理 Agent 需要对新加入的 Agent 进行效用和代价的评估:希望协作效用大于协作代价,否则协作的效用全部用来维持集合中各个Agent 的代价开销.

4　仿真试验

下面以室内温度调节控制为例,进行仿真验证. 在不采取任何措施的条件下,若房间中的热负荷变化,则温度也会相应地发生改变. 当房间实际温度与设定温度差超过人体舒适度要求时,房间 Agent就会向管理 Agent(MA)发出室温调整任务请求,并将温差参数传递给 MA. MA 接收到任务请求后,根据知识库中的知识,通过调节风机转速和送风温度这 2 个子任务即可完成温度调节任务,而各子任务又需要若干个子任务以并发或是串行的方式来完成[6]. 因此,可根据“与/或”结构任务树的划分规则进行任务分解.

图 1 为温度调节任务的一个协调过程,在芝加哥大学社会科学计算研究中心研制的 multi-Agent 建模工具 Repast 软件平台上实现了此过程中各个

Agent的协作交互.经过系统仿真,试验结果如图2所示.

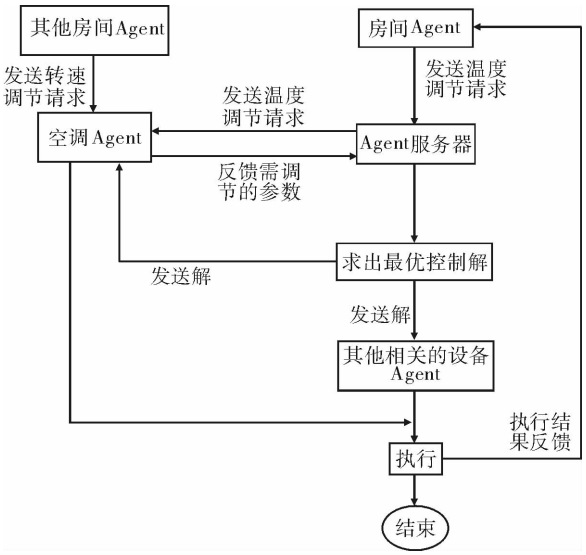


图1 温度调节协调过程

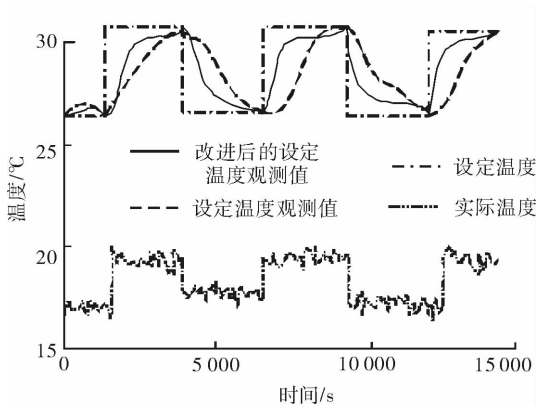


图2 温度变化参数曲线

从图2可以看出,采用对话语义的Agent协作交互能够有效提高MAS控制的时效性和准确性,送风温度可以更快地达到设定值,性能明显优于一般的Agent协作模型.图3为在Repast软件平台上,当任务书由10个增加至80个时,通信代价测试结果.试验结果证明,随着任务数的增多,基于对话语义的Agent协作模型的通信代价大大低于普通的协作方式,节省了时间,提高了运行效率.

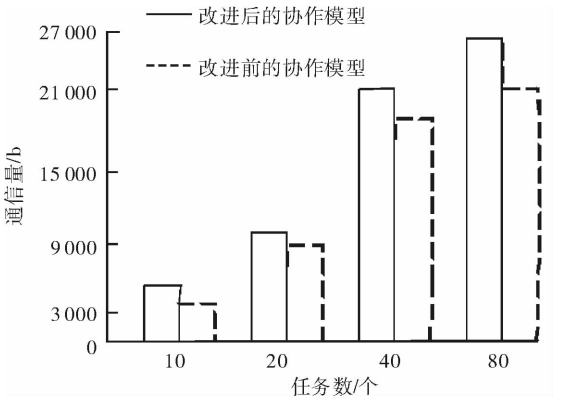


图3 通信代价测试结果

5 结论

本文提出了一种基于对话的多Agent协作模型:基于Agent协作模型中"与/或"结构任务树任务分解模式,采用划分通信区域、建立区域Agent的通信模式,给出了基于对话交互语义的对话交互过程.通过该模型,不仅将完成一个独立复杂的任务转变成多个Agent的联合意图,使多个Agent各自的行为具有相关性、一致性,而且使得任务的分配充分考虑各辅助Agent的专业领域、能力、兴趣和机会.试验结果表明,通过对话协作模式大大提高了多Agent间相互协作的有效性和灵活性.

参考文献:

[1] 陈志.基于Agent的无线传感器网络若干问题研究[J].南京邮电大学学报,2007,27(3):216.

[2] 谢学科.多Agent交互协作研究及系统模拟[D].西安:西北工业大学,2005.

[3] 安毅生,李人厚.基于对话的多Agent协作交互模型[J].西安交通大学学报,2005,40(12):1344.

[4] 孟建良,孔维莉,庞春江,等.基于系统体系结构的多Agent协作[J].微机发展,2005,15(12):73.

[5] 杨爱琴,朱玲玲,程学云.基于流演算的多Agent请求/服务协作模型的研究[J].计算机工程与设计,2011,32(2):681.

[6] 马琴.基于多Agent的智能建筑集成管理平台的研究[D].上海:同济大学,2009.

文章编号:2095-476X(2012)06-0009-05

基于可调整邻域阈值的DBSCAN算法在应急预案分类管理中的应用

金保华, 林青, 赵家明

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450002)

摘要:针对庞大的预案文本资源分类难的问题,将可调整的邻域阈值*Eps*取代原有的全局*Eps*,得到了改进的DBSCAN密度聚类算法.以预案文本间的相似度作为聚类基本的度量属性,将改进的DBSCAN算法应用于应急预案分类管理,以去除边界.仿真结果证明该方法不仅不影响预案本来的基础分类方式,而且更易于实现,在一定程度上能够缓解噪音点误识别问题,对提高预案文本的重用性和分类的准确率有一定的参考意义.

关键词:DBSCAN算法;文本相似度;应急预案文本管理;可调整邻域阈值

中图分类号:TP391 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.003

Application of DBSCAN algorithm based on adjustable threshold in the emergency plan classification management

JIN Bao-hua, LIN Qing, ZHAO Jia-ming

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450002, *China*)

Abstract: Aiming at large plan texts resource classification problems, adjustable threshold Eps replaced the original global threshold Eps. An improved DBSCAN clustering algorithm based on density was put forward. The similarity between plan texts was taken as measurement attribute. Improved DBSCAN was applied in the field of plan classification to remove the boundary identification error. The simulation results showed that this method not only does not affect the result in basis classification way, but also have certain reference significance to improve accuracy and reusability of classification.

Key words: DBSCAN algorithm; text similarity; emergency plan text management; adjustable threshold

0 引言

应急预案需明确规定应急管理的责任主体、工作范围、运行机制等,因此,预案管理的工作效率是应急事件处置过程中的关键.但现存的预案信息系统只能单纯地实现文本上传与查看功能,检索效率低下.因此,与现代信息技术、文本挖掘技术相结合的数字化应急预案,是应急预案管理发展的方向.本文将改进的DBSCAN算法应用在应急预案分类管理方面,力求在面对突发情况之时,提供分类整合原有预案文本资源的功能,为应急指挥人员提供高层次有侧重的决策方案,以提高预案重用性.

收稿日期:2012-06-29

作者简介:金保华(1966—),男,河南省郑州市人,郑州轻工业学院副教授,主要研究方向为人工智能、计算机辅助决策系统.

1 预案文本间的相似度

应急预案作为一种规范性与结构性较好的文本,有一定的灵活性. 使用效果较好的向量空间模型[1]能对半结构化的预案文本信息做结构化处理,整个预案文本空间由经过范化的预案文本特征向量构成. 该特征向量可以表示为

$$\boldsymbol{V}(d) = \sum_{i=0}^{n} (t_i, w_i(d))$$

其中,t_i 为预案文本经过预处理后,利用分词工具提取出来的文本 d 中的关键词 i; $w_i(d)$ 为关键词 i 的权重,表示该词反映预案文本 d 主题的重要程度. 应用矢量空间模型的前提为:每篇预案文本是由一组相互独立的词条所构成的集合,表示为 $d = \{t_1, t_2, \cdots, t_n\}$.

设预案 $plan_x$ 可以表示为$(w_{x1}, w_{x2}, \cdots, w_{xn})$,预案 $plan_y$ 可以表示为$(w_{y1}, w_{y2}, \cdots, w_{yn})$,则两者之间的相似度为

$$\mathrm{Sim}(plan_x, plan_y) = \cos(plan_x, plan_y) = \frac{plan_x \times plan_y}{\| plan_x \| \times \| plan_y \|} = \frac{\sum_{i=1}^{n} (w_{xi} \times w_{yi})}{\sqrt{\sum_{i=1}^{n} w_{xi}^2} \cdot \sqrt{\sum_{i=1}^{n} w_{yi}^2}}$$

由此可以推断,设输入的预案文本集为$(plan_1, plan_2, \cdots, plan_i)$,则可以得到文本相似度倒排表

$$\mathrm{SimList}(plan_i, plan_j) = \sum_{\substack{i=1, j=2 \\ i<j}}^{n} [\mathrm{Sim}(plan_i, plan_j), plan_i - plan_j]$$

以及编号为 i 的文本与其他文本的相似度倒排表

$$\mathrm{SimList\_}i(plan_i, plan_j) = \sum_{j=1, i \neq j}^{n} [\mathrm{Sim}(plan_i, plan_j), plan_i - plan_j]$$

倒排表均按从大到小的相似度进行排列,输出形式均为

$$< \mathrm{Sim}(plan_i, plan_j), plan_i - plan_j >$$

其中,$\mathrm{Sim}(plan_i, plan_j)$ 为预案 $plan_i$ 与 $plan_j$ 之间的相似度,$plan_i - plan_j$ 为预案文本标识码.

2 DBSCAN 的改进及其在预案分类管理中的应用

2.1 DBSCAN 密度聚类算法原理

DBSCAN(density-based spatial clustering of application with noise) 是一种经典的基于密度的聚类算法,应用广泛. 其计算的时间复杂度为 $O(n^2)$,空间复杂度为 $O(n\log n)$,相关概念定义如下[2].

定义 1 类半径:数据集空间中的类簇半径,表示为 *Eps*.

定义 2 密度:空间中任意一点的密度是以该点为圆心、以 *Eps* 为半径的圆区域内包含的点数目.

定义 3 邻域:空间中任意一点的邻域是以该点为圆心、以 *Eps* 为半径的圆区域内包含的点集合.

定义 4 核心点:空间中某一点的密度如果大于某一给定阈值 *Minpts*,则称该点为核心点.

定义 5 直接密度可达:已知 *Eps* 和 *Minpts*,对于点 x 和点 y,如果 y 是核心点,而且 x 属于 y 的 *Eps* 邻域,则点 x 从点 y 直接密度可达.

定义 6 噪音:事先给定 *Eps* 和 *Minpts*, 基于密度聚类中的一个聚类就是可以密度连接所能包含的最多数据点的集合,不属于任何聚类的数据点的集合称为噪音.

假定输入参数为 *Eps* 和 *Minpts*,DBSCAN 算法的流程图如图 1 所示. 具体描述如下:

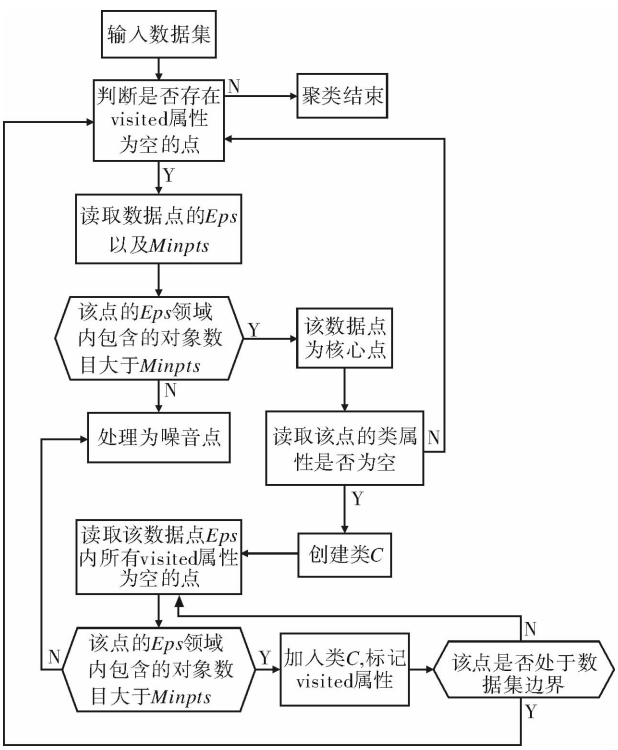


图 1 DBSCAN 算法的流程图

1)输入聚类数据,然后任意选取 1 个数据点 x,检查数据点 x 的 *Eps* 邻域.

2)如果 x 是核心点,而且没有被划分到某一个类,则创建一个类 C,找出所有从 x 密度可达的点并

加入 C.

3)依次检查 C 中未处理过的对象,如果该对象的 Eps 邻域内包含的对象数目大于 $Minpts$,就把该邻域中未包含于类 C 的对象加入 C 中.

4)如果 x 不是核心点,则被当作噪声处理.

5)转到第 1 步,重复执行算法;如果数据集合中所有的点都被处理过,则算法结束.

2.2　DBSCAN 算法的改进

DBSCAN 算法最大的优点是可以发现任意形状的类簇,而不受到噪音的影响. 聚类的结果会随着用户设定的全局 Eps 邻域值的改变而改变,为了确定 Eps 的初始值,DBSCAN 需要计算所有数据对象与其第 k 个(此处 $k = 6$)[3] 最邻近的对象之间的距离,并将结果按距离排序,由此得到 k-dist 图. 同时,DBSCAN 算法需要搜索从某个核心点出发到所有密度可达的点,这一步是经过反复进行区域查询实现的,因此需要建立 R^*-tree 来查询返回给定查询区域中的所有对象. 这 2 个点相应的代价都是系统的大量内存与 IO 开销[4].

原始 DBSCAN 算法建立 k-dist 图的时间复杂度为 $O(n^2)$, 建立 R^*-tree 的时间复杂度为 $O(n\log n)$, 基于聚类过程的大部分时间用在区域查询操作上,因此该算法平均时间复杂度为 $O(n\log n)$. 在明确了 SimList 表的本质即是 k-dist 图的基础上,笔者对该算法进行改进:DBSCAN 原始算法的第 2 步 R^*-tree 的建立过程被搜索 SimList 表种子队列 SeedList 中满足要求的记录所取代(在后面一节详细展开). 仿真试验时效图见图 2,经过改进的算法减少了原来搜索查询区域的时间,降低了系统的内耗,提高了效率.

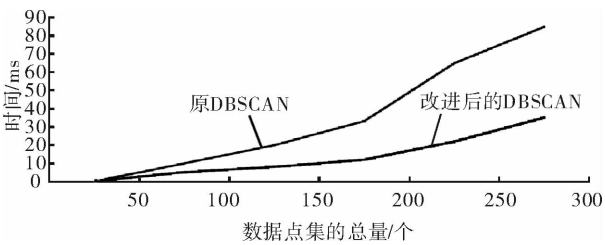


图 2　仿真试验时效图

原 DBSCAN 中,在预案文本密度与类之间的间距不均匀的情况下,较小的 Eps 值会将边界点处理成噪音点[5];反之,则可能将离得较近而密度较大的那些类合并为同一个类,使得聚类结果不够准确. 在改进的算法中,使用可调整的邻域阈值 Eps,可调整的 Eps 能够有效地针对密度不均的现状,与软聚类结合后使因数据过于稀疏而造成的错误识别边界点的情况得到一定的改善.

2.3　改进的 DBSCAN 算法在预案分类管理中的应用

随着各类突发公共事件的发生日益频繁,相关应急预案的框架体系逐步完善,最基本的分类已经无法满足应急决策者在进行应急决策时搜索与分类整合的需求. 利用文本聚类中 DBSCAN 密度聚类的方法,不仅能够对预案文本进行快速分类,方便决策者获取相关的预案信息,而且能够去除预案文本的边缘性信息,提高预案重用性.

一个预案要具备重用性,往往基于以下 2 个方面的原因:

1)从预案本身的功能来说,预案文本资源属于政府公文,而且都是依据相应的法律以及行政法规制定的,虽然地区间的实际情况不相同,但是既然法律存在通用性,预案的制定者就可以借鉴同级预案的经验进行编制,决策者也可以参考以往同类预案的方案进行应急指挥工作.

2)从聚类的目的来说,以文本相似度为属性度量依据,对预案文本使用软聚类后,即一个预案文本可以归入多个簇中,以达到预案文本可以在不同侧面体现参考价值的目的.

因此,结合预案文本的特性,可以对以预案文本相似度为度量依据形成的文本空间使用改进的 DBSCAN 算法. 改进后的算法的伪代码如下所示:

1)首先,定义预案文本对应的数据结构.

```
// 定义数据集数据点的类,并设定其属性 Minpts 为 int 类型,Eps 为 float 类型,visited 为枚举类型(0 代表未访问,1 代表已访问)
    public class MetaSet{
        private int Minpts,visited;
        private float Eps;
        public int getMinpts(){
        return Minpts;
        }
        public void setMinpts(){
        this.Minpts = Minpts;
        }
        public int getvisited(){
        return visited;
```

```
        }
        public void setvisited(){
        this. visited = visited;
        }
        public float getEps(){
        return Eps;
        }
        public void setEps(){
        this. Eps = Eps;
        }
    }
```

2)按序搜索 SimList 中未访问过的数据点形成种子队列,并对种子调用聚类方法.

```
public void SeedList(){
        < MetaSet > ls = new < ListArray > ();
        ls. add(数据点);
        for(MetaSet ms:ls){
        if(ms. visted = = 0&& ms. Eps 内 Minpts > a){
        Cluster(ms);
        ms. visited = 1;
        }
        else("类库生成完毕");
        }
        else("类库生成失败");
    }
```

3)生成以 MetaSet 类中的种子数据为核心点的类 C_0.

```
public void Cluster(MetaSet sm){
    CreatClusterClass(C0).
    < MetaSet > ls0 = new < ListArray > ();
    //以 Eps0 启发式的计算出当前需要扩展搜索的区域邻域阈值 Epsi
    for(int i = 0; ;i ++){
    Epsi = (Epsi - 1 + Min(Sim))/2 - 0.1
    for(MetaSet ms0:ls0){
    if(ms0. visted = = 0&& ms0. Epsi 内 Minpts > a){
    C0. add(ms0);
    }
    else{
    将 ms0 处理成为噪音点;
    }
  }
 }
}
```

3 仿真结果与分析

3.1 试验环境

使用 Java 编写仿真软件,在分词上采用的是中国科学院提供的 imdict-chinese-analyzer 开源分词工具. 本实验所使用的语料为郑州轻工业学院应急研发中心的预案语料库,其中预案文本严格按照预案的分类方法,即按预案类型主要分为自然灾害、事故灾难、公共卫生与社会安全 4 类. 从各类中选择一定数量的文档共 100 篇作为训练集,并随机选择自然灾害类 80 篇,事故灾难类 70 篇,公共卫生类 50 篇,社会安全类100篇作为测试的文本数据集合.

3.2 试验结果

取数据集合的 1/25 作为 *Minpts* 的值是一种行之有效的方法[6],即此处训练集与测试集的 *Minpts* 均取 4. 在此基础上,对固定 *Minpts* 参数的训练预案文本集选取不同的 *Eps* 值,并使用原始的 DBSCAN 对其聚类,精度(P) 与召回率(R) 分布情况分别如图 3,图 4 所示.

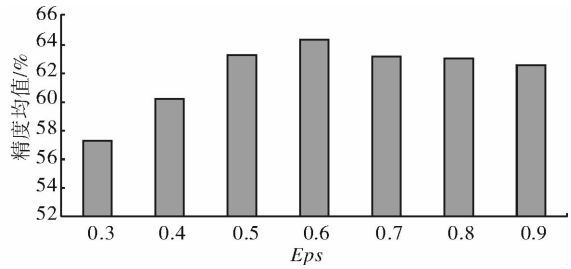


图 3 精度分布图

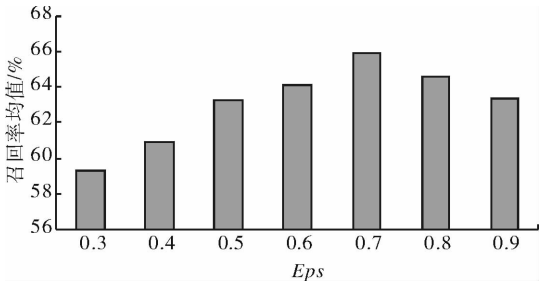


图 4 召回率分布图

由此可以得知,训练集的精度均值在 $Eps = 0.6$ 处取得峰值,召回率均值在 $Eps = 0.7$ 处取得峰值. 因此根据训练集的训练情况,设置测试集 *Minpts*, *Eps* 的初始值为 $Minpts = 4$, $Eps = 0.8$,并设定 *Eps* 的变化区间为[0.5,0.8]. 对测试集进行聚类试验,结果见表 1 和表 2.

表 1　原 DBSCAN 测试结果　　篇

大类	小类	随机抽取	系统分类	正确分类
自然灾害	2	80	99	62
事故灾害	2	70	80	46
公共卫生	2	50	39	26
社会安全	4	100	82	63

表 2　改进后的 DBSCAN 测试结果　　篇

大类	小类	随机抽取	系统分类	正确分类
自然灾害	11	80	96	68
事故灾害	6	70	60	50
公共卫生	4	50	54	32
社会安全	12	100	90	78

3.3　结果分析

模糊矩阵、熵、整体相似度、分类正确率、精度和召回率都是文本聚类的质量评价方法. 本文采用精度和召回率作为改进的 DBSCAN 算法性能的评价标准. 原始的 DBSCAN 算法与改进 DBSCAN 算法对 150 篇随机抽取的预案文本进行聚类的精度均值分别为 64.55% 与 75.02%,召回率均值为 65.75% 与 74.60%. 试验数据表明以下 4 点:

1)虽然测试集的文本被划分为 33 个子类,但是划分出来的子类之间均属于原来的大类,因此划分出小类对原来的数据文本的基本分类并无影响.

2)小类中预案文本之间在文本结构、内容上表示出一定的相似之处,较之原算法精度与召回率均得到了提高,符合为预案制定者提供参考的功能要求.

3)聚类效果明显,同一文本会被划分到不同的子类之中,预案管理系统用户可以根据需求将对决策者有利的子类保存,并进行个性化的调整. 而且可以看到,使用改进的 DBSCAN 算法划分出来的小类的数量明显比原始算法大,倘若从类中随机抽取预案文本来生成参考组,采用可调整的 *Eps* 能够更细致地得到更佳的参考组.

4)改进后的 DBSCAN 算法能够有效降低系统内存与 IO 的消耗,在预案文本的数据点个数持续增加的过程中,时间消耗仅是原算法的 1/3 左右.

4　结论

本文对原有的 DBSCAN 密度聚类算法进行改进,将可调整的 *Eps* 邻域阈值应用到预案文本聚类中,以取代原来的全局 *Eps*,提高了 DBSCAN 算法的准确性. 仿真结果表明,该方法不仅不影响预案本来的基础分类方式,而且更易于实现,在一定程度上能够缓解噪音点误识别问题,对提高预案文本的重用性和分类的准确率有一定的参考意义.

参考文献:

[1] 刘志勇,耿新青. 基于模糊聚类的文本挖掘算法[J]. 计算机工程,2009,35(5):44.

[2] Han J, Kamber M. 数据挖掘概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社, 2008.

[3] 夏鲁宁,荆继武. SA-DBSCAN:一种自适应基于密度聚类算法[J]. 中国科学院研究生院学报,2009,26(4):530.

[4] Das S, Abraham A, Konar A. Automatic clustering an improved differential[J]. IEEE Transactions on Systems Man and Cybernetics (Part A): Systems and Humans, 2008,38(1):218.

[5] Sanjay C, Sun Pei. SLOM:a new measure for local spatial outliers[J]. Knowledge and Information Systems, 2006, 9(4):412.

[6] 于亚飞,周爱武. 一种改进的 DBSCAN 密度算法[J]. 计算机技术与发展,2011,21(2):30.

文章编号:2095-476X(2012)06-0014-03

AOP技术在数据交换与共享系统中的应用

钱慎一, 付中举, 林青

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:针对数据交换与共享系统通用模块存在代码冗余的问题,引入AOP技术,对系统通用服务进行代码植入操作,实现了AOP框架在数据交换与共享系统的应用.试验结果证明,AOP技术能够有效减轻系统负担,提高系统的时效.

关键词:数据交换;数据共享;AOP技术

中图分类号:TP39 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.004

Application of AOP technology in data exchange and sharing system

QIAN Shen-yi, FU Zhong-ju, LIN Qing

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*)

Abstract: Aming at the probelm of code redundancy in data exchange and data sharing system, AOP technology was introduced in general service of system by code woven operation, the application of AOP frame was realized in data exchange and data sharing system. The experiment results showed that AOP technology can effectively reduce the burden and improve security of system.

Key words: data exchange; data sharing; AOP technology

0 引言

数据交换与共享系统的核心功能是用户可以通过系统的服务器端统一进入、统一访问、统一管理,用户登录时由系统自动进入相应的角色界面进行管理,系统服务器端管理员负责登记前置机的相应信息等.这要求该系统访问控制机制必须严谨高效.

目前,在用户的访问与系统基础管理、维护等方面,大多数数据交换与共享系统采用的是基于角色的访问控制(RBAC)[1-2].但是,单纯地使用RBAC机制,并不能很好地解决在访问控制、部门管理、共享管理与日志管理等模块中代码大量重复、分散以及效率低下等问题.

AOP(aspect-oriented programming)是面向方面(切面)的编程OOP(object-oriented programming)的补充与延续[3].AOP作为一种新的软件开发思想,是为了更好地解决对象中的方法具有通用性而代码冗余的问题[4-5].鉴于此,本文在数据交换与共享系统中引入AOP技术,以解决系统通用模块代码的冗余问题[6],并解除代码在数据交换与共享系统中的强耦合性,从而提高系统效率.

1 AOP在数据交换与共享系统中的应用设计方案

1.1 数据交换与共享系统中的AOP技术路线

数据交换与共享系统应用面较宽,现已广泛应用于银行机构、金融债券公司、数字认证中心等领

收稿日期:2012-10-10

基金项目:河南省科技攻关项目(122102210124)

作者简介:钱慎一(1975—),男,河南省郑州市人,郑州轻工业学院副教授,主要研究方向为数据库应用和数据挖掘.

域. 其核心功能是对用户访问各前置机的数据资源进行统一控制,通过登录与权限机制来访问获取资源列表,保证数据的安全性与及时性. 值得注意的是,系统前置机上装载的是统一的数据交换标准,为引入 AOP 技术提供了便利. 采用 AOP 后,系统能够调用方面编程中的代码,识别存储在前置机中的特殊字段中的数据结构,并抽取日志、安全、事务等非业务代码为一个独立模块,插入到各执行业务中. 这样不仅解决了非核心模块业务逻辑与核心模块业务逻辑的代码需解耦合的问题,也降低了系统维护的难度. 数据交换与共享系统的整体系统结构设计如图 1 所示.

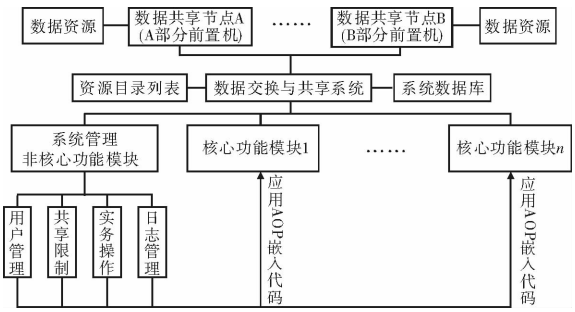


图 1　基于 AOP 的数据交换与共享系统结构图

1.2　AOP 技术在数据交换与共享系统中的设计理念

AOP 是在 OOP 的基础之上发展起来的,这并不能说明 AOP 的发展对 OOP 有取代性的作用[7]. 作为对 OOP 的补充,AOP 针对 OOP 中代码强耦合的问题,提出将应用程序中的商业逻辑和对其进行支持的系统通用服务进行分离的思想[8].

AOP 把软件系统分成 2 部分,即核心关注点和横切关注点. 核心关注点是业务处理的主要流程,也就是说这个解决方案要做的事;横切关注点是与核心关注点无关的部分. 笔者结合数据交换与共享系统的总体设计,明确系统非核心模块(如共享限制、事务操作与日志管理等)都可以很方便地使用 AOP 来实现(此处用户管理子模块的安全交由 RBAC 机制保障). 这种将影响多个类的公共行为封装到一个可重用方面的编程思想,具体表现在:在涉及用户调用核心模块时,使用 AOP 技术对非核心模块子模块的核心关注点进行操作,紧接着以模块化在横切关注点处采用与应用程序无关的方式处理相应的操作[9]. 图 2 给出了用户在操作之前、之中、之后捕获目标的应用程序并进行相应操作的过程.

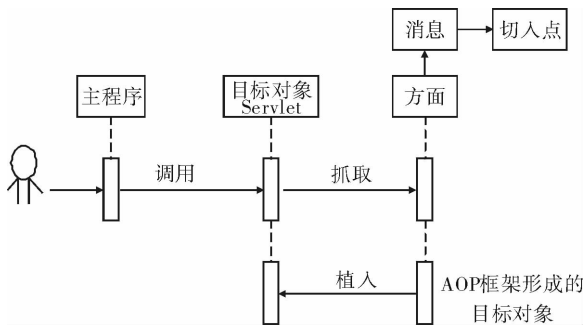


图 2　系统操作顺序图

当用户调用需要响应的 Servlet 程序后,执行响应的操作,aspect 将会通过 advice 捕捉到切入点,并进行植入操作,处理方面编程的相关操作.

2　AOP 技术在数据交换与共享系统中应用的实现

目前,宣称能够支持 AOP 的架构平台已达近百种,实现了基于 Java 语言的 AOP 框架也有 20 多种,其中较为完善的是 Spring AOP 和 Aspect J. 综合考虑系统的构架,本文使用 Aspect J 作为面向方面编程的一个框架,使用 myeclipse 作为编程工具,在 2.5 GHz主频的实验环境下进行数据交换与共享系统的软件开发.

数据交换与共享系统被分为 2 部分:核心功能模块所处理的业务流程将被设计为核心关注点;非核心功能模块(不包括用户权限的设计与用户管理部分)被设计成横向关注点. 以日志管理为例,部门管理员角色的用户在管理资源时存在对文件与数据的增加、删除与更新操作,这时系统需要将这些操作过程记录在案,因此需要引入 AOP 编程技术. 通过用户的增删改操作,可以插入一个横切面,形成一个切入点,从而实现跟踪操作.

下面给出的是在项目中部门管理员对相应的文件或者数据进行增加、删除与更新操作时如何插入切入点进行日志记录的部分代码:

```
public class DepManServlet
{
    public void insert(HttpServletRequest request, Ht-
tpServletResponse response)
    {
    /* insert the corresponding data or filestreams into
database */
    }
    public void update(HttpServletRequest request, Ht-
```

```
tpServletResponse response)
    {
    /* update the corresponding data or filestreams from database */
    }
    public void delete(HttpServletRequest request, HttpServletResponse response)
    {
    /* delete the corresponding data or filestreams from database */
    }
}
```

使用 Aspect J 定义一个名为 Logging 的方面,以实现用户对日志方面的管理,部分代码如下:

```
pulbic aspect Loggin
{
    pointcut Logcap( HttpServletRequest request, HttpServletResponse response):
    execution ( * DepManServlet. insert ( HttpServletReque, HttpServletResponse))
    &&args(request,response);
    after ( HttpServletRequest request, HttpServletResponse response);
    returning:Logcap(request,response)
    {
    /* do logging */
    }
}
```

以代码执行相同行数的时间为度量,分别测试原系统与应用 AOP 框架后的系统性能,具体的系统时效图如图 3 所示.

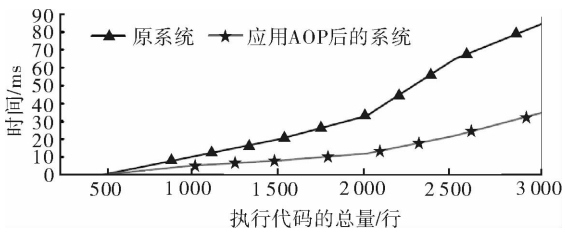


图 3 使用 AOP 技术的系统时效图

从图 3 可知,未采用 AOP 方面编程策略的数据交换与共享系统在代码执行量为 500 行、1 000 行、1 500 行、2 000行、2 500 行与 3 000 行的耗时分别是8 ms,15 ms,28 ms,52 ms,74 ms 与 88 ms. 但在原系统应用 AOP 后,系统耗时降低 50%,时效性得到增强. 当在 Java 代码中添加类或者添加新的方法后,程序员不必加入新的横切面来处理相应的操作,只需要改动类中的连接点标识,日志管理等模块就会自动应用到相应的位置上. 从图 3 可知,AOP 的方面编程以其独有的优势不仅解决了用户在非核心模块上代码冗余的问题,而且提高了系统的效率.

3 结论

本文研究表明,原始开发模式的通用业务中非核心部分的代码强耦合性、冗余性可以通过引入 AOP 技术得到很好的处理. 试验结果证明,将 AOP 应用于数据交换与共享系统中,不仅能够很好地协助 RBAC 访问机制控制对用户的身份认证,而且能够高效分离核心业务逻辑与非核心业务逻辑,使得访问控制过程更加严密清晰. 通过对相应权限用户安全记录、事务处理、共享限制与日志的管理,可以跟踪后台数据库的存储情况,进一步提高系统安全分析员对系统安全的审计情况,符合数据交换与共享系统的总体设计思想.

参考文献:

[1] Gradecki J D, Lesieckin N. Mastering Aspect J: Aspect-Oriented Programming in Java[M]. Indianapolis: Wiley Publishing Inc,2003.

[2] 伍建晖. 基于角色的访问控制在 J2EE 中的研究及扩展[D]. 南京:南京理工大学, 2006.

[3] 张英捷,刘万军. SpringAOP 技术在 J2EE 系统安全性验证中的应用研究[J]. 计算机工程与科学, 2008,30(8):137.

[4] 钟秀琴,符红光,余莉,等. 基于本体的几何学知识获取及知识表示[J]. 计算机学报, 2009,33(1):167.

[5] 李淼,白勇,张波. 基于领域特征的 AOP 编织实现方法[J]. 计算机科学, 2009,36(2):299.

[6] 郑汉雄,郑汉英,周晓聪. 在 AOP 中使用标注改进日志功能的实现[J]. 计算机工程, 2009,34(17):71.

[7] Havinga W, Nagy I, Bergmans L, et al. Detecting and resolving ambiguities caused by inter-dependent introductions[C] //Proc of the 5th International Conference on Aspect-oriented Software Development, New York: ACM Press,2006:214 – 225.

[8] Kiczales. An overview of aspectJ[C]//ECOOP, Object-Oriented Programming, Budapest: Springer-Verlag, 2001: 327 – 353.

[9] Laddad R. Aspect J in Action: Practical Aspect Oriented Programming [M]. Greenwich: Manning Publications Co,2003.

文章编号:2095-476X(2012)06-0017-04

基于 Android 平台的人员定位与监控系统的设计与实现

刘玉玮, 刘爱莲, 谢涛, 宋耀莲

(昆明理工大学 信息工程与自动化学院, 云南 昆明 650500)

摘要:针对传统定位与监控系统成本较高和不够便捷的缺点,采用C/S架构,结合Oracle数据库和Google Maps技术,在Android手机平台下设计并实现了一个方便、实时的人员定位与监控系统.测试结果表明,该系统能够在合法的前提下用Android手机实时准确地进行多人监控,同时又具备记录轨迹、移动距离计算等功能.

关键词:Android;Google Maps;C/S构架;人员定位与监控系统

中图分类号:TN966 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.005

Design and implementation of a personnel positioning and monitoring system based on the Android platform

LIU Yu-wei, LIU Ai-lian, XIE Tao, SONG Yao-lian

(*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500, *China*)

Abstract: Because of the disadvantage of the traditional positioning and monitoring system with high cost and inconvenience, using the C/S structure, combined with Oracle database and Google Maps, a convenient and real-time personnel positioning and monitoring system was designed and implemented in the Android platform. The test results showed that under the premise of leagl, the system completes some functions on the Android mobile phones: monitoring multiplayer real-time and accurately, track record and distance calculation.

Key words: Android; Google Maps; C/S structure; personnel positioning and monitoring system

0 引言

随着现代生活节奏不断加快,实时的地理位置信息正在成为人们最渴求的信息之一.传统的GPS定位与监控系统研究大多集中在交通运输领域,主要应用于车辆船舶等交通工具的定位、导航与监控,系统开发成本较高,且便捷性和经济实用性不强,不适用于人员的监控管理.因此,设计一个既节约成本、方便实用,又能对人员进行科学实时的监控和管理的定位与监控系统非常必要.Android为普通开发者提供了非常灵活的Google Maps的展示与控制功能[1],在Android手机平台上可以方便地实

收稿日期:2012-11-06

基金项目:云南省科技厅基金项目(2011FB035)

作者简介:刘玉玮(1987—),男,河南省安阳市人,昆明理工大学硕士研究生,主要研究方向为无线通信与Android应用开发.

现这一系统. 针对在 Android 平台下的人员定位与监控系统的研究,已取得了一定的成果[2-3]. 代敏[2]对自己的位置进行了定位,无法达到定位和监控他人的目的;李武钲[3]虽然在 Android 平台下实现了定位与监控的功能,但其监控端为 PC 机,而不是便于携带的移动终端,系统便捷性不强.

本文针对已有系统的不足,拟设计并实现一种人员定位与监控系统,用户无需购买专业的 GPS 设备和 PC 机,利用日常使用的 Android 智能手机即可方便快捷地对人员进行监控管理.

1 系统设计

Android 是一个基于 Linux 平台的开源手机操作系统,由底层的 Linux 操作系统、中间件和核心应用程序组合而成[4]. Google 重新设计了 Java 虚拟机和系统,使 Android 具有以下 3 个特点:Java 应用更接近于底层系统,效率更高;应用在被监控情况下运行,安全性更高;第三方软件完全开放的平台,代码完全开源免费[5]. Android 平台的这些优势提高了程序开发的便捷性、兼容性和可扩展性. 本系统的构架为 C/S 构架,服务器开发环境采用 Windows + Apache + Oracle + Servlet + JSP 的配置方案,客户端采用 Java 语言在 Android SDK 环境下实现. 此外,监控者客户端整合了 Google Maps,可方便地显示出被监控人员的位置,显著提高系统的开发效率和实用性.

人员定位与监控系统分为 3 部分,即被监控客户端、监控客户端、服务器端. 系统的体系结构如图 1 所示. 被监控端将自己的实时经纬度信息上传至服务器中,经过服务器端程序处理后将数据保存至数据库;监控端获取到被监控人员列表后向服务器数据库提取被监控端经纬度信息,显示在监控端的监控地图界面.

2 系统各功能模块设计

2.1 被监控端设计

此模块功能是被监控端向服务器提交自身的实时地理位置信息,即实时经纬度. 应用程序提供给用户一个操作界面,点击“开始服务”按钮,程序每隔 5 s 向服务器提交用户所在的经纬度信息,点击“停止服务”按钮,停止上传. 此界面采用 Android 平台下的 Service 技术,使得用户提供的服务可以在后台隐藏运行,增加了程序的实用性.

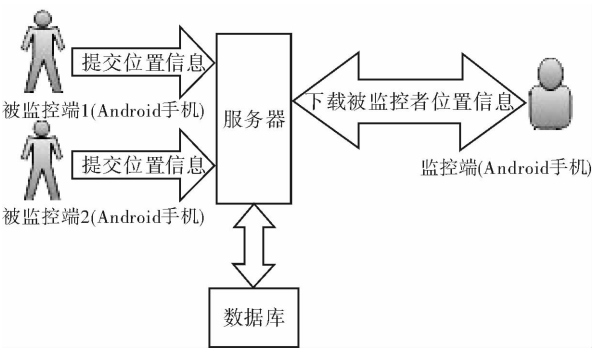


图 1 系统体系结构图

2.2 监控端设计

此模块功能主要是定时从服务器下载被监控人员的实时经纬度信息,并在监控地图界面上显示,此外还提供记录轨迹、计算距离等功能. 该模块主要包括 2 个子功能模块,即被监控者菜单和监控地图.

2.2.1 被监控者菜单 此模块功能主要是向监控者提供可供监控的人员列表以方便在多个被监控人员间快速切换. 被监控者向服务器注册时指定自身受某一监控人员监控,所以每名被监控者对应着各自的监控人员. 另外,被监控者许可认证后才能受到监控,增加了系统的安全性. 被监控者菜单界面如图 2 所示.

2.2.2 监控地图 此模块为监控端的主模块,向用户提供一个地图监控界面. 当选中一被监控者后进入此界面,显示被监控者的当前地理位置. 此外该界面还提供了记录轨迹和计算距离等功能,点击“开始”按钮开始记录被监控者的行动轨迹,并开始计算行走距离,按下“停止”按钮停止记录. 监控地图界面如图 3 所示.

2.3 服务器设计

本系统服务器采用 Apache + Oracle + Servlet + JSP 的设计方案,主要功能是接收被监控端上传的经纬度信息和向监控端提供经纬度信息,以及向用户提供注册登录验证和监控者与被监控者之间的数据匹配. 服务器端的功能模块主要包括被监控端注册、登录、经纬度读信息接收模块,监控端提取经纬度信息、注册、登录模块等. 监控端提取经纬度信息模块采用 JSP 技术,其优势是从数据库中提取出

图 2　监控端被监控者菜单界面

图 3　监控地图界面

来的经纬度信息可以直接在 JSP 中生成 XML 文档,便于采用 SAX(simple API for XML)技术进行数据解析.

本系统所有数据信息存储在名为 db_monitor 的数据库中,其中包含的表主要涉及监控者用户注册、登录验证,被监控者用户注册、登录验证,被监控者实时经纬度信息存储、提取等.

2.4　系统试验运行效果分析

经过系统各个模块设计,结合程序开发工具 MyEclipse 进行系统开发,最终实现了基于 Android 的人员定位与监控系统. 配置好相应的网络环境并将系统安装在 Android 手机上,进行系统运行试验.

从被监控者菜单界面选择被监控人员后,进入监控地图界面如图 3 所示. 当被监控者不断移动时,监控端定时提取该人员的实时位置信息并在地图上显示人员的行走轨迹,同时显示该人员的移动距离. 经过试验验证,本系统定位准确,轨迹记录和距离计算实时精准,达到了系统设计要求.

3　系统实现中的关键技术

3.1　Google Maps API 技术

Google Maps 是 Google 公司提供的电子地图服务,包括局部详细的卫星照片[6]. Google Maps API 是 Google 为开发者提供的 Maps 编程接口,允许开发者在不建立自己的地图服务器的情况下,将 Google Maps 地图数据嵌入网站或程序之中,从而方便实现 Google Maps 的地图服务应用.

Google Maps API 除了帮助开发者将地图嵌入自己的应用中之外,还允许开发者对地图进行应用开发拓展. Android 系统很好地兼容了 Google Maps 服务,为基于位置服务的应用程序开发提供了极大的方便.

在本系统的开发中,监控端应用程序使用 MapView 对象,将 Google Maps 嵌入到应用程序中. 在使用 MapView 开发应用程序之前,需先向 Google 申请一组经过验证的 Android Maps API Key[7],才能正常地在手机上使用 Google Maps 服务.

3.2　解析 XML 文件 SAX 技术

XML 现在已经成为一种通用的数据交换格式,平台的无关性使得很多场合都需要用到 XML. 在本系统设计中,使用 XML 存储交换被监控者经纬度信息等数据. 解析 XML 文件有 DOM(document object model)和 SAX 2 种基本的方法. SAX 是一个用于处理 XML 事件驱动的“推”模型,它不像 DOM 那样建立一个完整的文档树,而是在读取文档时激活一系列事件,这些事件被推给事件处理器,然后由事件

处理器提供对文档内容的访问. 相比于 DOM,SAX 可以在解析文档的任意时刻停止解析,速度更快,但操作复杂. 本系统采用 SAX 方法把被监控人的经纬度信息、人员编号、注册信息等从服务器端的 XML 文件中解析出来.

4 系统的安全性保障

4.1 身份鉴别和加密技术

本文研究的系统用于人员的定位和监控,因此安全性尤为重要. 为防止未授权用户绕过用户登录页面进入系统主页面,需要进行用户身份验证. 用户需要正确输入用户名和密码后才能进入本系统,凡验证失败都将停留在登录页面. 同时,为了保证用户的口令在系统数据库中存放的安全性,口令字采用单向加密的方式进行保护. 由于监控系统涉及他人隐私问题,为防止非法监控他人,监控人必须输入由被监控人提供的验证码才能正常监控.

4.2 数据有效性的验证

经过笔者在智能 Android 手机上的测试,该系统能够准确验证用户输入信息的有效性. 当用户注册或登录信息填写不合法时,系统会弹出相应的错误提示,让用户重新输入. 此外,测试证明系统可以快捷实时准确地显示被监控人的位置、轨迹和移动距离,并能够方便地在不同被监控者之间切换,为用户提供了良好的用户体验.

5 结论

本文采用 C/S 架构、结构 Oracle 数据库和 Coogle Maps 技术,在 Android 手机平台下设计实现了人员定位与监控系统. 系统经过真机测试,运行稳定可靠、定位准确无误、轨迹记录和距离计算实时准确,达到了人员定位与监控的目的. 用户可以在合法前提下,在 Android 手机上方便快捷地对被监控者进行监控和管理,加强了系统的安全性和便捷性,满足了用户的使用需求. 此外,在开发过程中预留了扩展系统功能的程序接口,可以方便地在监控地图界面增加各种附加功能,便于系统升级. 当然,系统还有一些待改进的地方,可以围绕地图界面开发更为丰富的功能,增加系统实用性. 例如增加发送信息功能,快捷地向被监控人员发送信息,以提高人员管理效率等.

参考文献:

[1] 姜文周,王彦超,李先毅. 基于 Android 的个性化校园地图服务设计[J]. 实验技术与管理,2012,29(3):109.

[2] 代敏. 基于 Android 平台下手机定位程序的设计与实现[J]. 计算机与数字工程,2012,40(4):143.

[3] 李武钲. 通过 Android 平台装置的人员追踪系统[P]. CN:200920163222. X,2010 - 05 - 26.

[4] 郭宏志. Android 应用开发详解[M]. 北京:电子工业出版社,2010:3 - 4.

[5] 刘胜前,陈立定. 基于 Android 平台的车辆导航系统设计与实现[J]. 自动化与仪表,2012,27(4):1.

[6] 杨丰盛. Android 应用开发揭秘[M]. 北京: 机械工业出版社,2010:283 - 284.

[7] 王世江,余志龙,陈昱勋,等. Android SDK 开发范例大全[M]. 北京:人民邮电出版社,2009:557 - 558.

文章编号:2095-476X(2012)06-0021-03

基于马尔柯夫过程的交叉路口车流量预测模型研究

蒋亚平, 郭俊亮

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:为了预测城市交叉路口交通控制系统中每个相位的车辆流量,进而在一个信号周期内合理分配每个相位的时间,建立了交叉路口车流量预测模型.该模型运用马尔柯夫分析方法,把各相位定义为当前状态,经片段时候后,系统只要掌握转化为另一状态的可能性,即可制订出相应的控制策略.试验结果表明该算法预测的车流量与实测车流量之间的误差比较小,在短时预测车流量方面是可行的.

关键词:马尔柯夫过程;交叉路口短时交通预测;车流量预测

中图分类号:U491.1 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.006

Study on vehicle flowrate prediction model of crossroads based on Markov process

JIANG Ya-ping, GUO Jun-liang

(*College of Computer and Communication Engineering, Zhengzhou Univrsity of Light Industry, Zhengzhou* 450001, *China*)

Abstract: In order to predict the flow of vehicles of the each phase in the crossroads traffic controlling system, which can reasonable distribute the time of the each phase in one signal period, the vehicle flowrate prediction model of crossroads was built, this model uses the Markov analysis method and define the each phase as the current state, after the fragment of the time, as long as the system master the possibility of the transform the phase to another phase, the system can work out the corresponding control strategy. Experimental results showed that the errors between prediction of the flow of vehicles and the actual flow were small, and the method was feasible in the short-term traffic prediction.

Key words: Markov process; crossroads short-term traffic prediction; vehicle flowrate prediction

0 引言

基于混沌理论、粗糙集理论、神经网络理论等的预测方法是目前常用的交通流量预测方法.徐启华[1]提出了一种实时的基于动态递归神经网络的交通流量预测方法,董春娇等[2]提出基于Elman神经网络的道路网设计方法,史其信等提出基于BP神经网络的路径形成时间预测方法[3-6].

一般情况下,高度复杂性、随机性和不确定性是交通流所表现出来的特性,在每个周期各个相位时间内车辆的达到数量是随机分布的,具有高度的随机特性.在预测的状态转移方面,马尔柯夫方法可以进行较准确的判断.故本文拟利用马尔柯夫方法对到达交叉路口车流量预测模型进行研究,把各

收稿日期:2012-09-03

作者简介:蒋亚平(1970—),男,河南省商丘市人,郑州轻工业学院副教授,博士,主要研究方向为网络安全与智能交通.

相位定义为当前状态,经片段时候后,系统只要掌握转化为另一状态的可能性,即可制订出相应的控制策略.

1 马尔柯夫过程的定义及分析

1.1 马尔柯夫过程的定义

马尔柯夫过程定义为:在一些随机现象中,所表现出来的事物特性是,其结果不依赖于前几次实验的状态,只与前 1 次结果相关. 连续的马尔柯夫过程,就形成了马尔柯夫链,马尔柯夫分析法主要用于观测或预测随机事件未来状态的变化趋势.

在城市单交叉口多相位控制系统中,在一个交通信号周期内,各相位的分布是随机的,而且在各相位内交通流所呈现的性质也是随机的,这符合马尔柯夫的特性. 运用马尔柯夫分析法,根据随机变量的现在动向和状态预测变量未来的状态和动向,采取的交通管制策略更加妥当.

1.2 马尔柯夫过程分析

定义 1 任意一个向量 $\boldsymbol{V}=(v_1,v_2,\cdots,v_n)$, 向量中每个元素都为≥0 的整数,所有元素的总和为 1,此向量就叫做概率向量.

定义 2 在矩阵 $\boldsymbol{P}=(\boldsymbol{P}_{ij})_{n\times n}$ 中,每个列向量都是概率向量,那么此方阵就叫做随机矩阵.

定义 3 对任意一个概率矩阵 $\boldsymbol{P}$ 而言,若 $\boldsymbol{P}$ 中的所有元素都是正数,则称此矩阵为正规则概率矩阵.

系统取 n 个状态 $z_1,z_2,\cdots,z_n$,则状态空间为 $\{z_1,z_2,\cdots,z_n\}$ 的有限集,实验的第 i 次结果如果是 z_i,则称系统在第 i 步的状态是 z_i. 转移概率 p_{ij} 的定义是系统在 z_i 转移到下一时刻状态 z_j 的概率.

状态 z_i 发生后,紧接着状态 z_j 发生的转移概率矩阵为 $\boldsymbol{p}$,$i=j$ 时为保留同一状态的概率.

$$\boldsymbol{p}=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \quad ①$$

其中,p_{ij}是从状态 z_i 转移到 z_j 的概率;$i,j=1,2,\cdots,n$.

如果系统从状态 z_i 经过 k 步转移到状态 z_j($z_i\to z_1\to z_2\to\cdots\to z_{k-1}\to z_j$)的概率为 $p_{ij}^{(k)}$,把 $p_{ij}^{(k)}$ 排成一个矩阵 $\boldsymbol{P}^{(k)}$,则称之为 k 步转移矩阵,表示为

$$\boldsymbol{P}^2=\boldsymbol{P}^{(2)}=\boldsymbol{PP}$$

$$\boldsymbol{P}^{(n)}=\boldsymbol{P}^n=\boldsymbol{P}^{n-1}\boldsymbol{P}=\boldsymbol{PP}^{n-1}=\boldsymbol{P}^{(n)} \quad ②$$

假定 $X^{(0)}=(x_1^{(0)},x_2^{(0)},\cdots,x_n^{(0)})$,第 k 步的概率分布表示为 $X^{(k)}=(x_1^{(k)},x_2^{(k)},\cdots,x_n^{(k)})$,$\boldsymbol{P}^{(k)}$ 表示 k 步转移概率矩阵,则

$$X^{(k)}=X^{(0)}\boldsymbol{P}^{(k)} \quad ③$$

由式① ② ③可得预测模型

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}^k \begin{pmatrix} x_1(k) \\ x_2(k) \\ \cdots \\ x_n(k) \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} x_1(k+1) \\ x_2(k+1) \\ \cdots \\ x_n(k+1) \end{pmatrix}^{\mathrm{T}} \quad ④$$

其中 $k=1,2,\cdots,n$.

2 交叉路口车流量预测模型

2.1 交叉路口的相位组成

以一个十字型的单交叉口为例(见图 1),有东西南北 4 个方向的入口,在不同的入口又分为 2 个车道,分别是左转与直行,直行车道包括右转车道,该入口还包括有导流导线车道,可以使右转车辆随时通过路口,不会影响其他方向的车辆. 在一个周期内共有 8 个相位,这些相位交替进行切换,各自放行各相位内的交通流.

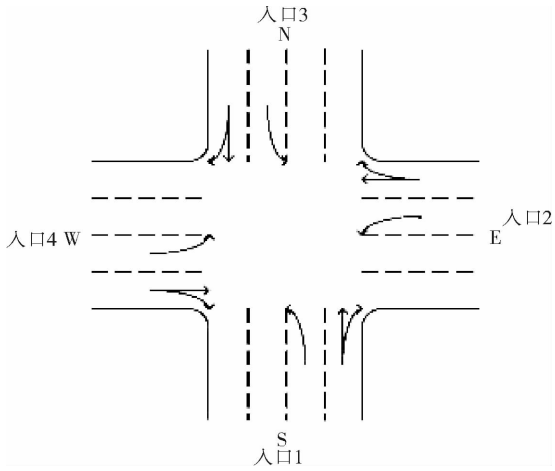


图 1 单交叉路口

2.2 单交叉路口的车流量变化

表 1 的数据是由入口 1 的检测器检测到的 2 个周期的入口车流量. 本文仅讨论交叉路口的入口 1 的车流量,在后续研究中再讨论上游交叉口车流量对下游交叉口的影响.

表 1 入口 1 在 3 个方向的车流量 veh

行驶方向	前一周期	当前周期
直行	60	40
左转	30	34
右转	20	26

表 2 的数据是在一个相位内,左转、直行、右转方向之间的车辆转移. 利用表 2 中的数据就可以得到转移概率矩阵 $\boldsymbol{p}$.

表 2　入口 1 每个行驶方向的车辆转移　veh

行驶方向	直行	左转	右转
直行	25	6	8
左转	2	20	3
右转	2	2	30

由表 2 可以得到转移概率矩阵为

$$\boldsymbol{p}=\begin{bmatrix}0.64 & 0.15 & 0.21\\ 0.08 & 0.80 & 0.12\\ 0.03 & 0.03 & 0.88\end{bmatrix}$$

根据式④,可计算得出进入入口 1 的下一周期的交通流占有率. 在左转车道方向上将有 31% 的车辆进入,直行车道方向上将有 20.7% 的车辆进入,右转方向上将有 48.3% 的车辆进入. 假定转移概率矩阵是不变的,想要预测 k 步以后的交通流量,可以采用 k 步转移概率的方法.

$$\boldsymbol{P}^{k-1}\begin{bmatrix}q_{s1}\\ q_{l1}\\ q_{r1}\end{bmatrix}^{\mathrm{T}}=\begin{bmatrix}q_{sk}\\ q_{lk}\\ q_{rk}\end{bmatrix}^{\mathrm{T}}$$

式中,q_{s1},q_{l1},q_{r1}分别为第一信号周期内直行、左转、右转的占有率;q_{sk},q_{lk},q_{rk}为 k 信号周期内直行、左转、右转的占有率.

3　预测结果与分析

交通探测器可以得到实际的交通流量,表 3 给出了未来 9 个周期内实测交通流量;表 4 给出了由马尔柯夫分析法预测的未来 9 个周期内交通流量;表 5 给出了由 Elman 神经网络法预测的未来9 个周期内交通流量;表 6 中的数据是运用马尔柯夫法预测的误差;表 7 中的数据是运用 Elman 神经网络法预测的误差. 其中,q_{sf},q_{lf},q_{rf}为信号周期内的预测值;q_{so},q_{lo},q_{ro}为信号周期内的观测值.

表 3　未来 9 个周期内实测交通流量　veh

周期	q_{sf}	q_{lf}	q_{rf}
1	0.320 1	0.280 1	0.399 8
2	0.218 2	0.321 7	0.460 1
3	0.160 6	0.342 6	0.496 8
4	0.128 3	0.352 2	0.519 5
5	0.110 1	0.355 7	0.534 2
6	0.099 7	0.355 3	0.544 0
7	0.094 0	0.354 7	0.511 3
8	0.090 6	0.352 4	0.556 0
9	0.088 7	0.350 3	0.564 0

$$E_x=q_{xo}-q_{xf}/q_{xo}\times 100\%\qquad x=s,l,r$$

表 4　由马尔柯夫法预测的未来 9 个周期内的交通流量　veh

周期	q_{so}	q_{lo}	q_{ro}
1	0.300 3	0.268 5	0.431 2
2	0.203 2	0.327 8	0.468 0
3	0.165 1	0.363 5	0.471 4
4	0.135 7	0.324 2	0.540 1
5	0.101 8	0.399 6	0.498 6
6	0.101 6	0.344 8	0.553 6
7	0.088 5	0.410 7	0.500 5
8	0.101 4	0.380 1	0.518 5
9	0.101 3	0.362 6	0.536 1

表 5　由 Elman 神经网络法预测的未来 9 个周期内的交通流量　veh

周期	q_{so}	q_{lo}	q_{ro}
1	0.302 9	0.268 5	0.430 6
2	0.203 4	0.327 8	0.469 8
3	0.167 6	0.363 5	0.479 9
4	0.135 7	0.324 2	0.540 1
5	0.103 4	0.399 6	0.498 0
6	0.103 8	0.344 8	0.553 8
7	0.086 5	0.410 7	0.503 8
8	0.103 5	0.389 1	0.518 4
9	0.101 4	0.362 6	0.536 0

表 6　运用马尔柯夫法预测的误差　%

周期	E_s	E_l	E_r
1	-3.21	-1.14	-2.14
2	-4.26	-2.18	3.48
3	2.63	-3.58	5.41
4	3.42	-2.45	2.19
5	3.11	3.24	-5.17
6	2.12	3.26	4.13
7	-1.25	2.48	4.15
8	-1.32	3.68	2.14
9	-2.15	5.14	-3.26

由表 6 和表 7 可以看出,运用马尔柯夫法预测的误差比运用 Elman 神经网络法预测的误差要小.

在现实中,准确预测交叉口车辆是一个复杂的问题. 从预测值与实测值对比来看,运用马尔柯夫

(下转第 31 页)

文章编号:2095-476X(2012)06-0024-04

停车泊位动态分配算法的研究

窦亚星[1], 张明明[2], 张杰[1], 樊霄艳[1]

(1. 郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001;
2. 河南中烟工业有限责任公司 信息中心, 河南 郑州 450000)

摘要:针对大型智能停车场车位众多,如何对车位进行合理分配从而让用户快速找到适合自己的车位的问题,设计了基于停车场内车位动态分配算法的车位信息采集与发布系统. 其核心是建立的table表与车位形成对应关系,为停车场空闲车位的均匀分布和行车路径选择一个平衡点. 仿真结果证明了该算法的合理性.

关键词:智能停车;车位预订;车位动态分配
中图分类号:TU248.3 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.007

Study on dynamic distribution algorithm in parking

DOU Ya-xing[1], ZHANG Ming-ming[2], ZHANG Jie[1], FAN Xiao-yan[1]

(1. *College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*;
2. *Information Center, China Tobacco He'nan Industry Limited Company, Zhengzhou* 450000, *China*)

Abstract: Large-scale parking lot owns numerous spaces, how to distribute parking spaces and make users get suitable parking spaces is problem which should be resolved instantly. So a parking spaces information acquisition and release system was designed based on dynamic distribution algorithm. The research distribution strategy of the spaces in parking lot was to establish the correspondence of data table and parking spaces and it makes a balance point for the free spaces uniform distribution and the path weight. The simulation results verified that the algorithm is reasonable.

Key words: intelligent parking; parking reservation; parking dynamic distribution

0 引言

随着机动车数量的飞速增长,停车位短缺成为影响城市交通不可忽视的问题. 停车泊位管理是智能交通管理的重要部分,它是动态交通管理的起点和延续[1]. 据统计,城市交通中30%的车辆因为找不到停车位而在道路上缓行,既浪费了宝贵的时间,又加剧了道路拥堵. 因此能够让用户快速找到适合自己的车位是解决交通拥堵的一项重要措施[2].

停车场的建设不断向大型化、智能化的方向发展. 在大型购物中心、火车站等车辆停放密集场所,停车场一般拥有上千个车位,如何合理分配车位是一个难题. 多数停车场采用的规则是随机停车,这种停车泊位方式不利于停车场的管理. 国内外学者对如何运用优化算法在停车场内寻找最短路径的车位做了大量的研究. 王一军等[3]采用蚁群算法求得入口到达停车位的最短路径,其所对应的车位定

收稿日期:2012-05-15

作者简介:窦亚星(1988—),男,河南省新乡市人,郑州轻工业学院硕士研究生,主要研究方向为计算机网络与智能控制.
通信作者:张杰(1972—),男,河南省郑州市人,郑州轻工业学院副教授,主要研究方向为计算机网络与智能控制.

义为最佳车位;刘子文等[4]采用新型的粒子群算法解决车位的诱导问题,即车辆如何到达距离最近的停车位. 现有用户预订车位方法的研究集中在基于模糊逻辑的预约实时决策方法上[5-6],在一定程度上可以解决用户预订车位的问题. 然而在实际的停车泊位过程中,需要考虑的因素很多,如果仅仅按照最短路径停车,容易造成区域性拥挤,特别是入口处,因此现有最短路径导向的停车位管理不一定是最佳的.

鉴于此,本文拟提出一种新的停车泊位动态分配算法,考虑多方面因素,在保证停车场空闲车位均匀分布的基础上,为用户寻找最短停车路径,实时为无预定用户和预订用户选择最合适的停车区域. 另外设计 2 种车位预订方式,同时制定对应的车位分配策略.

1　车位状态信息的数字化

1.1　采集内容

车位状态采集终端所采集的信息包括停车场内已有车辆数量及停放车位信息、已预订车位信息、空闲车位数量与位置信息、车辆进入停车场后是否驶入指定车位信息等.

假设停车场内共有 n 个车位,每个车位的空闲用 P_{i0} 表示,已停车用 P_{i1} 表示,已预订用 P_{i2} 表示,不开放用 P_{i3} 表示.

1.2　传输流程

车位状态信息采集与发布系统结构如图 1 所示. 车位信息采集部分由传感器监控节点和汇聚节点 2 部分组成. 车位传感器监控节点检查每个车位的状态,以多跳的方式将信息传递到汇聚节点;汇聚节点收集并且处理信息,通过网关传输到数据处理中心.

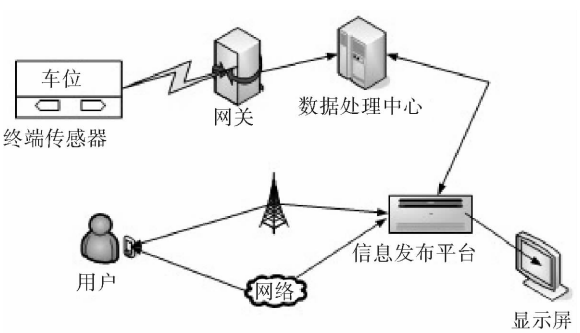


图 1　车位状态信息采集与发布结构图

数据处理中心支持多种网络接入方式,提供标准化的接口使得数据传输易于实现,处于系统的核心地位,其主要作用是综合各种因素进行车位的分配,包括计算最佳停车位置 P_{i0},预订车位以及对所有车位进行综合管理.

通过信息发布平台,用户可以查询所有车位的状态,进行车位预订,无预定用户到达停车场时可以通过显示屏得到最佳停车位的信息.

2　车位的分配算法

车位分配算法由在没有预订情况下用户进入停车场时的车位分配方法,以及用户预订车位后进入停车场时的车位分配方法两部分构成,两方法相结合更便于用户停车.

2.1　无预定车位的用户进入停车场

首先,将所有车位按照位置划分为 w 个区域,为了保证区域分割鲜明,每个区域总的车位数是不固定的. 建立停车区域选择表 table,表内包含 m 行,每行代表 1 个区域. m 与 w 的关系应满足

$$c \cdot m = w \qquad c \in R$$

式中,R 为非负整数集合. 为了确保进入停车场的车辆分布均匀且具有一定的规律,区域与 table 之间的对应方式应满足

$$w_j = c \cdot m + i \qquad c \in R, i \in [0, m-1]$$

将多个区域有效地结合起来,table 表格对应指定的区域,简单方便. 由于区域内车位数目都不多,一般为几十个,可以采用用户自主选择的方式选择空闲车位,一方面可以减少算法的复杂度,另一方面可以增加用户的车位选择权.

每一次只能有一个区域对应到 table 的相应行内,其替换策略是比较能够进入 table 对应行的所有区域的车位空闲率 η 的大小,空闲率最大的进入到 table 的对应行. 每当一辆车驶入停车场,空闲率 η 就发生变化,重新比较 η,进行替换.

$$\eta_{\max} = \begin{cases} \max\{\eta_i, \eta_{m+i}, \eta_{2m+i}, \cdots, \eta_{(c-1)m+i}\} & i \geqslant 1 \\ \max\{\eta_0, \eta_m, \eta_{2m}, \cdots, \eta_{cm}\} & i = 0 \end{cases}$$

此时,在 table 表格中的每一行存放的都是最大空闲率的区域. 若直接比较每一行 $\eta_{\max}$ 的大小来确定停车区域,则陷入了完全按照空闲率停车的思路,在实际的应用中,停车时间也是一项重要的因素. 一方面要保证停车场管理方便及空闲车位均匀

分布,另一方面也应使用户在最短的时间内停车. 考虑到停车场内影响行驶车辆停车时间的主要因素为路径的长短,因此可以将求最短时间的问题等效转化为求最短路径的问题.

大型停车场一般有多个出入口,各个出入口到达每个区域的最短路径都不相同,因此在每一个出入口处的数据是不一样的. 考虑车辆在停车场内行驶的总路径,车辆进入时的入口是一定的,而驶出往往是寻找最近的出口. 设在某一个入口处到达第 j 个区域的最短路径为 d_j,此区域到最近的出口的距离为 l_j. 为了寻求车位空闲率最大、路径最短的区域,引入公式

$$\gamma_i = \frac{\eta_{\max}}{\rho \cdot d_j + (1-\rho) \cdot l_j}\theta \qquad \rho \in [0,1]$$

式中,θ 为比例参数,取常数;ρ 为出入路径比重参数. table 表中每一行对应一个权值 γ_i,比较各行 γ_i 的大小,其中最大 γ_i 对应的区域就是下一辆车应该停放的区域;若最大 γ_i 值相同,则选择距离入口最近的区域.

2.2 用户预订车位的分配

车位预订可以避免用户在路上盲目地寻找停车位,使用户放心地出行. 本文提出 2 种预订方式,分别为实时性预订和排队性预订,如图 2 所示.

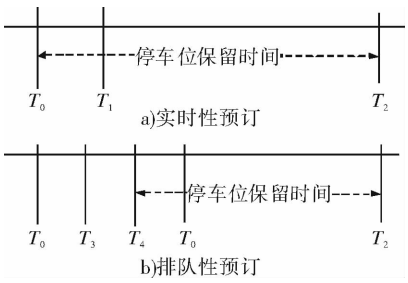


图 2 2 种预订方式比较图

2.2.1 实时性预订 由图 2a) 可知,T_0 表示用户预订车位的时刻,T_1 表示用户计划进入停车场的时刻,T_2 表示用户离开停车场的时刻. 用户在 T_0 时刻查询、预订车位,估计到达停车场的时间 T_1,预定时用户需要根据电子地图选择停车场入口,数据处理中心按照上述寻找最大权值 γ_i 的方法为用户分配停车区域.

若预订成功,对应区域空闲车位数目减少 1 位,停车场内一直为用户保留车位,直到用户到达,用户所付的费用为 T_0 到 T_2 时间段的费用. 若在 T_1 时刻用户未到达,则取消车位预订,只需扣除用户 T_0 到 T_1 时间段的费用. 实时性预订适合离停车场路程较近,或者有急事停车的用户,此种预订方式可以保证用户到达停车场就有车位. 为了防止车位长时间在预订状态,实时性预订需要规定 T_0 到 T_1 的时间上限.

2.2.2 排队性预订 此方法比较复杂,为了提高停车场的车位利用率,用户发送预订请求后,数据处理中心不需立即为用户安排车位,而是根据用户到达的时间为用户安排车位. 如图2b) 所示,T_0 表示用户开始预订的时刻,T_1 表示用户计划进入停车场的时刻,T_2 表示用户离开停车场的时刻,T_3 表示开始为用户分配车位的时刻,T_4 表示为用户分配好车位的时刻. 具体方法如下:将一天平均分为若干个时段,每个时段的长度为 t.

1)用户查询停车场空位,在 T_0 时刻发送预订请求,预计车辆到达时间是 T_1 时刻.

2)数据处理中心接到用户发送的预订请求后,判断 T_1 时刻所处第 f 时间段是否会有空闲车位. 可以根据以往统计的此时间段内车辆的流量、此时间段起点和终点空闲车位数量的均值以及现有多少用户在第 f 时段预订进行判断,然后短信通知用户预订成功或者无法预订. 简单的判断方法是若下式成立,则允许下一位用户进行预订.

$$f(b) - f(a) - x > 0$$

其中,$f(a)$,$f(b)$ 分别为第 f 时间段起点、终点的车辆均值;x 为此时间段已预订车位数目.

3)若用户预订成功,为了确保用户在预订时间内能够拥有车位,在 T_1 时刻到来前,即 T_3 时刻开始为用户分配车位. T_3 与 T_1 的间隔根据历史车辆数据进行确定. 若有空闲车位,按照上述寻找最大权值 γ_i 的方法,在 T_4 时刻为用户分配好停车区域,并发送短信通知用户具体的停车区域.

4)用户到达时,直接进入短信通知的区域即可. 若用户在 T_1 时刻未到达,则取消车位预订.

排队性预订适合距离停车场较远或者不急于停车的用户. 为了保证车位的利用率,虽然用户到达前会为用户分配车位,但是根据以往的情况进行车位预测,在停车场繁忙时仍有可能需要用户等待车位. 此预订方式中用户只需支付 T_1 到 T_2 时间段的费用,不包括预订的费用.

2 种预订方式预订时间段不同,收费方法不同,可以满足不同用户的要求. 同已有的预订方式相比,停车场采用 2 种方法相结合的预订方式,既可满足用户随时停车的要求,又可提高停车场的车位利用率.

3　仿真分析

鉴于 Matlab 具有强大的数据处理和函数功能,因此采用 Matlab 7.0 进行仿真实验. 设图 3 是某停车场示意图,该停车场共有 A,B,C 3 个出入口,其中 A 为主出入口,B,C 为次出入口,停车场内共 290 个车位. 将停车场分为 9 个区域,各个区域的总车位数目为 n. 到达 A,B,C 出入口的路径权值以及在某一时刻的空闲车位 P_{i0} 数目见表 1.

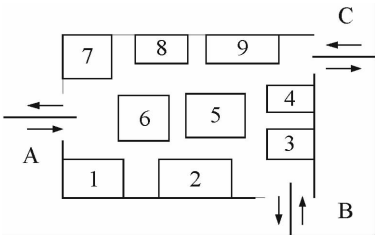


图 3　某停车场平面示意图

建立 table 表,表格包含 3 行. 第 1 行对应 1,4,7 区域,第 2 行对应 2,5,8 区域,第 3 行对应 3,6,9 区域. 参数 $\theta = 100$,$\rho = 0.7$. 在接下来的时间段内,出入口 A 有 40 辆车,B,C 分别有 20 辆车预订车位或者直接进入停车场.

表 1　各个区域状态表　　个

区域	n	P_{i0}	出入口 A		出入口 B		出入口 C	
			d_j	l_j	d_j	l_j	d_j	l_j
1	30	6	6	6	12	6	20	6
2	45	20	11	7	7	7	15	7
3	25	13	18	6	6	6	9	6
4	24	3	17	5	8	5	5	5
5	38	15	13	7	7	7	7	7
6	34	15	7	7	10	7	11	7
7	30	9	6	6	21	6	14	6
8	28	15	10	10	16	10	11	10
9	36	20	6	6	14	6	12	6

图 4 是随机分配和按照车位动态分配算法分配车位的均方差变化情况,可以看出,在车辆进入停车场后,随机分配的各个区域空闲率的均方差变化没有规律,时而增大,时而减小;而按照动态分配算法分配停车,各个区域空闲率的均方差一直呈减小趋势. 这表明,动态分配算法能够很好地平均各个区域的车位空闲率,使它们不断趋于平衡.

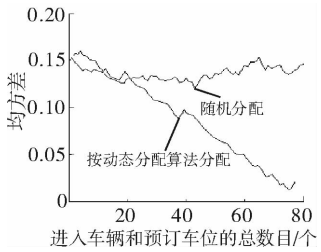


图 4　不同车位分配方式的均方差变化

在图 4 中,按动态分配算法分配的方差曲线在个别的点处会出现增长的情况,这是由于在考虑车位使用率的同时考虑区域到出入口路径的权值因素,以保证用户停车时间最短.

4　结论

针对如何合理分配停车场内众多车位的问题,提出了车位动态分配算法,同时设计了 2 种预订车位的方法. 根据动态分配算法在分配区域的时候,以区域车位空闲率为主,兼顾路径的权值,在各个区域的空闲率趋于均匀的基础上,保证用户停车时间最短. 仿真结果表明,本文提出的算法能够有效地解决停车场内的车位分布和用户行车路径的问题,使用户能够更加方便地停车泊位.

参考文献:

[1] Gallo M,D'Acierno L,Montella B. A multilayer model to simulate cruising for parking in urban areas[J]. Transport Policy,2011,18(5):735.

[2] 颉靖. 基于物联网技术的城市停车诱导系统研究[D]. 北京:北京邮电大学,2011.

[3] 王一军,陶杰. 现代大型停车场车位诱导优化算法及仿真[J]. 计算机仿真,2007,24(11):176.

[4] 刘子文,杨恢先,许翔,等. 新型 PSO 算法在停车场车位诱导问题中的研究[J]. 计算机工程与应用,2010,46(30):233.

[5] Leephakpreeda T. Car-parking guidance with fuzzy knowledge-based decision making[J]. Building and Environment,2007,42(2):803.

[6] 张富,严丽,马宗民,等. 基于模糊描述逻辑的模糊 XML 模型的表示与推理[J]. 计算机学报,2011,34(8):1437.

文章编号:2095-476X(2012)06-0028-04

基于综合负载动态分组的负载均衡算法研究

李永明[1], 李冬[2]

(1.平顶山学院 计算机科学与技术学院, 河南 平顶山 467000;
2.新乡学院 计算机与信息工程学院, 河南 新乡 453003)

摘要:针对服务器集群负载多变和动态算法系统开销大的问题,结合轮转法和动态反馈法的优点,提出了综合负载动态分组的负载均衡算法.该算法兼顾了集群系统中服务器异构和请求类型不同的问题,并配置了综合负载阈值和强制刷新的最小时间间隔.试验表明该算法系统开销小,负载均衡效果显著.

关键词:服务器集群;负载均衡算法;综合负载;动态分组
中图分类号:TP301.6 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.008

Study on load balancing algorithm based on integrated load and dynamic group

LI Yong-ming[1], LI Dong[2]

(1. *Computer Science and Technical Academy*, *Pingdingshan University*, *Pingdingshan* 467000, *China*;
2. *College of Computer and Information Engineering*, *Xinxiang University*, *Xinxiang* 453003, *China*)

Abstract: In order to resolve the problems that server cluster load is variable and dynamic algorithms spends too much system overhead, a load balancing algorithm based on integrated load and dynamic group was presented. This algorithm considers problem of server isomerism and different requirement categories in computer cluster and is equipped with comprehensive load threshold and the minimum interval of compulsory reload. The experimental results showed that the overhead of the algorithm is smaller and its load-balancing effect is remarkable.

Key words: server cluster; load balancing algorithm; integrated load; dynamic group

0 引言

目前网络通信的信息量迅速增大,服务器不堪重荷,但是服务器软硬件升级又面临众多障碍.这种现状下,服务器集群技术应运而生[1].服务器集群就是将一组服务器作为一个整体,代替单个服务器为用户提供透明的服务[2].负载均衡算法对服务器集群至关重要,它决定着集群系统的性能[3].负载均衡的前提是找到一种能准确反映服务器负载情况的表示方法.常用的方法是通过服务器的维持连接数来表示,这种方法虽然简单,但是存在弊端:第一,如果集群中服务器是异构的,那么它们的性能差别会很大;第二,不同类型的服务消耗的资源不同.因此,单纯使用连接数表示不能准确地反映负载的状态.

目前国内外提出的负载均衡算法主要有LIAC,

收稿日期:2012-08-17
基金项目:河南省科技攻关计划项目(122102310082)
作者简介:李永明(1979—),男,河南省焦作市人,平顶山学院讲师,主要研究方向为网络安全、物联网系统设计.

LCB 和 RSLB[4-6]. LIAC 算法需要先获取服务器的负载上限和当前信息,然后挑选出负载最小的服务器来调度,其缺点是服务器计算量大,维护成本高. LCB 算法使用负载容率(LC)来表示节点负载状况,LC 能够自适应地调整. 其优点是 LC 能够准确地反映节点信息,缺点是在网络状态不稳定的情况下,数据传输会受到很大的干扰. 针对上述问题,本文提出一种基于综合负载动态分组的负载均衡算法 ILDGB(integrated load and dynamic group based algorithm),将简单高效的轮转算法和动态反馈算法结合起来,以达到既能很好地均衡服务器集群负载,又避免算法大量开销的效果.

1　ILDGB

在 ILDGB 中,有 6 个需要考虑的输入信息,分别是服务器新连接比例、处理器负载、磁盘情况、内存情况、进程数和响应时间. 把它们存储在输入矩阵 $\boldsymbol{M}$ 中,表示为 $\boldsymbol{M}=[I,C,D,F,P,R]$. 引入参数矩阵 $\boldsymbol{K}=[K_1,K_2,K_3,K_4,K_5,K_6]$,负载计算公式为

$$L_j=[K_1,K_2,K_3,K_4,K_5,K_6]\begin{bmatrix}I\\C\\D\\F\\P\\R\end{bmatrix}$$

其中,K_i 表示 $\boldsymbol{M}$ 中第 i 项的权值,$\sum K_i=1$,$i=1,2,\cdots,6$.

该算法的基本思想是,一个调度周期 T 结束时,更新综合负载表的信息并按照负载的大小排序;然后,按照负载的大小将服务器集群分成 2 组,负载较小的一组服务能力强,为调度组;负载较大的一组服务能力弱,为非调度组. 下个调度周期内的服务请求按轮转法分配给调度组中的服务器. 这样,既考虑了服务器异构的性能差异和服务请求差异,又避免了在高密度请求时出现服务器负载倾斜的现象,而且体现了轮转调度简单高效的优点. 算法的程序流程图如图 1 所示.

算法:综合负载动态分组调度(ILDGB)

输入:未分配服务请求的集群服务器

输出:已分配服务请求的集群服务器

步骤:

1)初始化综合负载表;

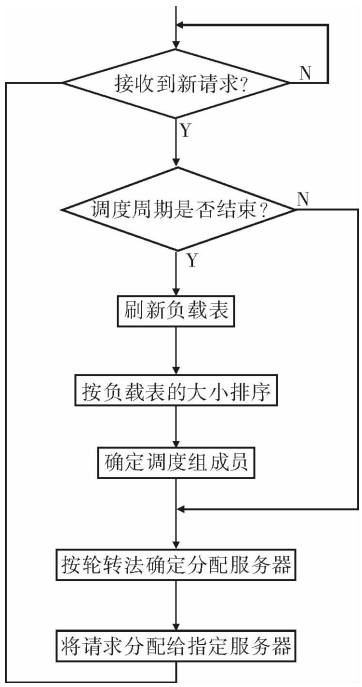


图 1　ILDGB 算法流程图

2)接收一个客户机的新请求,如果调度周期 T 结束,则刷新综合负载表,按照综合负载的值大小排序,并将其分为 2 组,即调度组和非调度组;

3)将下个调度周期 T 内到达的请求按照轮转调度原则分配到调度组的服务器上;

4)若调度周期内没有收到某台服务器的信息,就将该服务器置于非调度组中;

5)转到 2)循环执行.

调度周期 T 的设置会影响算法的性能. T 太大,则当请求密集时,调度组中的服务器负载急剧上升,出现负载倾斜的现象;T 过短,虽能更准确地反映服务器的负载状况,但会增加调度算法本身的系统开销. 一般情况下,更新周期设置为 1 ~ 11 s[7]. 但是在 1 个调度周期内,若服务请求过于密集,就会出现调度组中的服务器超负荷工作而非调度组闲置的现象. 为了解决这个问题,可以为服务器设置一个综合负载阈值. 当服务器的综合负载超过这个阈值时,就向负载调度器发出警告. 若在一个调度周期 T 未结束时,调度组内有超过 1/2 的服务器发出警告,则调度器强制刷新负载表,重新分组. 试验表明,恰当地设置阈值,在短时间请求超常密集时,能得到很好的效果;但是当服务器集群的整体负载很大时,设置阈值会导致频繁地分组,使集群系统的性能雪上加霜. 因此,系统需要设置刷新负载表进

行重新分组的最小时间间隔,这样就可以使集群系统长期处于高效稳定的服务状态.

2 仿真结果与分析

试验设备:计算机 10 台,软件压力测试工具采用 WAS.

试验方法:10 台计算机中,1 台负责调度,8 台作为服务器,另外 1 台通过模拟测试工具来模拟外界用户.

试验重点:比较轮转法(RR)、综合负载动态分组法(ILDGB)和最小负载优先法(LLF)3 种算法的平均响应时间和吞吐率. 平均响应时间用参数 *TTFB* 表示,$TTFB = \overline{s_t - s_0}$,其中 s_t 是客户端接收到服务器信息的时刻,s_0 是客户端发送请求的时刻. 客户端单位时间收到的字节数用 *BRR* 表示,反映系统的吞吐率[8].

试验结果:请求量分别设置为 100,200,300,依次增加到 1 000 时,记录下 3 种算法的 *TTFB* 和 *BRR*,图 2 和图 3 分别是 3 种算法在不同请求数时 *TTFB* 和 *BRR* 的对比情况.

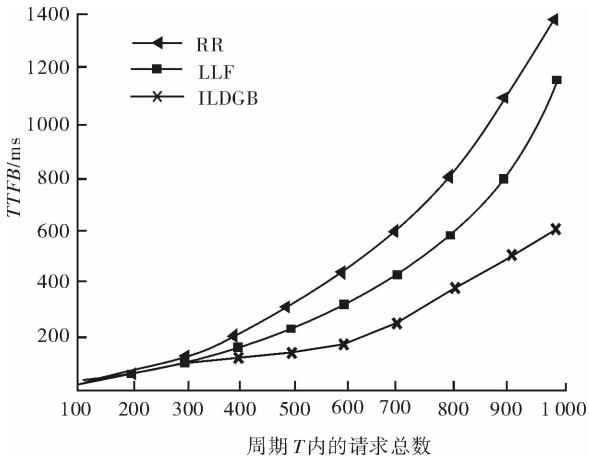


图 2 3 种算法的 *TTFB* 比较

从图 2 和图 3 可知,当请求数较小时,RR 法的响应时间相比其他 2 种使用动态反馈机制的算法要略小一些,这是因为使用动态反馈机制收集负载信息会有一定的系统开销;随着请求数逐渐增加,使用动态反馈机制的 LLF 算法和 ILDGB 算法就显示出明显的优势,吞吐率比 RR 法高,响应时间也更小. 此外,在请求数增加的过程中,ILDGB 算法表现出来的性能要优于 LLF 算法的性能,这是因为在请求数较多时,LLF 算法将一个周期内到达的全部请求都分配到同一台服务器上,造成了这台服务器的负载突然大幅度增加,从而出现负载倾斜的现象,而 ILDGB 算法则是将一个周期内到达的所有请求平均到调度组中的服务器上,因此在响应时间和吞吐率上都好于 LLF 算法.

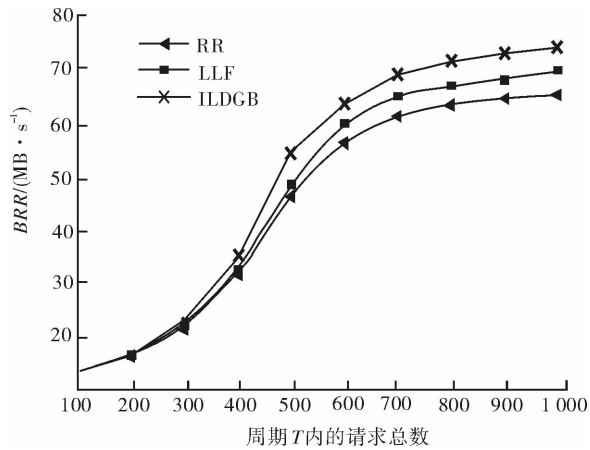


图 3 3 种算法的 *BRR* 比较

当一个调度周期内服务请求过于密集时,就会出现调度组中的服务器超负荷工作而非调度组闲置的现象. 通过使用阈值和设置负载表刷新的最小时间间隔能很好地解决这个问题. 图 4 是 ILDGB 算法在未用阈值、使用阈值和使用阈值且配置最小刷新间隔 3 种情况下,在不同请求数时的 *BRR* 对比图. 从图 4 可知,在周期 T 内请求数 < 500 时,这 3 种情况基本是相同的,因为请求不够密集,在调度周期 T 内,很少有服务器负载超过阈值,不会导致强制刷新. 当调度周期 T 内的请求数为 500 ~ 800 时,

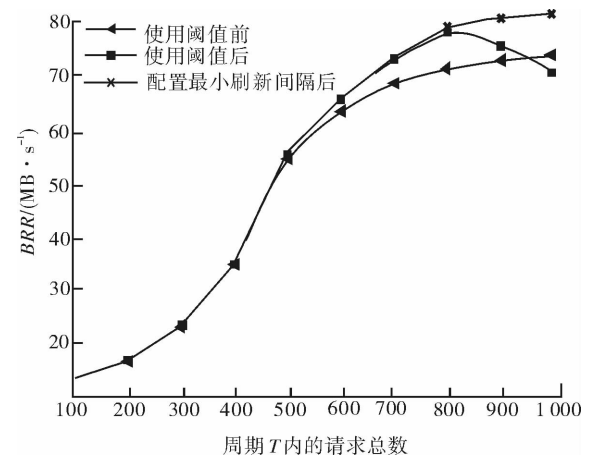


图 4 ILDGB 算法在未用阈值、使用阈值和使用阈值且配置最小刷新间隔 3 种情况下,在不同请求数时的 *BRR* 对比图

调度组内会有过半数的服务器负载超过阈值,导致强制刷新,及时避免负载倾斜现象,提高了系统的吞吐率. 但当调度周期内的请求数继续增加至 800 ~1 000 时,由于集群中服务器的负载都很大,会频繁产生强制刷新,急剧增加调度算法的系统开销,从而降低系统吞吐率. 因此,该算法设置了强制刷新的最小时间间隔(0.5 s),避免请求过于密集时频繁刷新. 从试验数据看,设置最小刷新间隔可以在周期内请求数 > 800 的情况下,使系统的吞吐率持续增长.

3　结论

本文提出了综合负载动态分组调度负载均衡算法,该算法有效地结合了轮转法和最小负载优先法 2 种算法的优点,在配置了综合负载阈值和最小刷新间隔后,达到了很好的实用效果. 与现有的负载均衡算法相比,ILDGB 算法不仅考虑了集群系统中服务器异构和请求类型不同的问题,还具有算法系统开销小和负载均衡效果显著的特点. 仿真结果表明,在分布式综合服务的应用环境中,ILDGB 算法能使服务器集群系统长期处于负载均衡、服务高效的状态.

参考文献:

[1] 张前进,齐美彬,李莉. 基于应用层负载均衡策略的分析与研究[J]. 计算机工程与应用,2007,43(32):138.

[2] 黄光球,刘兆明. 基于随机高级 Petri 网模型的服务器均衡集群[J]. 微计算机信息,2006,22(15):134.

[3] 李永喜,陈小平,杨兴良. 一种基于内容的 Web 服务器集群调度算法[J]. 计算机应用与软件,2008,25(3):215.

[4] Raman B, Katz R H. Load balancing and stability issues in algorithms for service composition[C]//Twenty-Second Annual Joint Conference of the IEEE Computer and Communications, San Francisco: IEEE Infocom, 2003: 1477 - 1487.

[5] 李文中,郭胜,许平. 服务组合中一种自适应的负载均衡算法[J]. 软件学报,2006,17(5):101.

[6] 陈亮,王加阳. 基于粗糙集的负载均衡算法研究[J]. 计算机工程与科学,2010,32(1):101.

[7] 蒋澜,朱明. 综合负载变化和分发代价的负载均衡方法研究[J]. 计算机工程与应用,2009,45(19):110.

[8] Cardellini V, Colajanni M, Yu P S. Dynamic load balancing on Web-server systems[J]. IEEE Internet Computing,1999,3(3):28.

(上接第 23 页)

法预测最大误差不超过 5.41%,说明在短时交通流预测方面,该方法是可行的.

表 7　运用 Elman 神经网络法预测的误差 %

周期	E_s	E_l	E_r
1	-3.23	-1.17	-2.15
2	-4.30	-2.19	3.49
3	2.66	-3.60	5.43
4	3.45	-2.47	2.20
5	3.20	3.26	-5.18
6	2.14	3.27	4.14
7	-1.27	2.49	4.16
8	-1.34	3.70	2.15
9	-2.17	5.15	-3.27

4　结语

本文基于马尔柯夫过程建立了道路交叉口车流量预测模型. 该模型把各相位定义为当前状态,经片段时候后,系统只要掌握转化为另一状态的可能性,即可制订出相应的控制策略. 该模型可用于预测短时间内交叉口每个行驶方向的交通占有率,在较长时间内的预测还需进一步的研究.

参考文献:

[1] 徐启华. 一种基于动态递归神经网络的交通流量实时预测方法[J]. 淮海工学院学报:自然科学版,2010,12(4):14.

[2] 董春娇,邵春福. 基于 Elman 神经网络的道路网短时交通流预测方法[J]. 交通运输系统工程与信息,2010,10(1):145.

[3] 史其信,郑为中. 道路网短期交通流预测方法比较[J]. 交通运输工程学报,2004,4(4):68.

[4] 杜长海,黄席樾,杨祖元,等. 基于神经网络和 Markov 链的交通流实时滚动预测[J]. 系统仿真学报,2008,20(9):2464.

[5] Sun S, Zhang C. A Bayesian network approach to traffic flow forecasting[J]. IEEE Transactions on Intelligent Transportation Systems,2010,7(1):124.

[6] Moortlty C K, Ratcliffe B G. Short term traffic forecasting using time series methods[J]. Transportation Planning and Technology,1988,12(1):45.

文章编号:2095-476X(2012)06-0032-03

一种基于网页划分的Web应用程序测试新方法

刘小园

(罗定职业技术学院 电子信息系,广东 罗定 527200)

摘要:在模型测试技术的基础上,通过对网页进行分类,提出了一种针对不同类别的网页特点的Web应用程序测试方法.静态网页采用黑盒测试,数据库网页采用白盒测试,动态网页采用灰盒测试.通过试验验证了该方法的有效性.

关键词:软件测试;Web应用程序;网页划分

中图分类号:TP393.1 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.009

A new Web application program test method based on Webpage classification

LIU Xiao-yuan

(*Department of Electronic Information, Institute of Luoding Polytechnic, Luoding* 527200, *China*)

Abstract: Based on the model test technology, a testig method for Web application program was put forward according to Website feature through Webpage classification: the static Webpage with black-box testing methods, database Webpage with white-box testing methods, dynamic Webpage with gray-box testing methods. The experiments verified the effectiveness of this method.

Key words: software testing; Web applications program; Webpage classification

0 引言

基于B/S结构的Web应用程序具有方便、快捷、操作简单等特点,是目前软件开发的主流模式.但是,由于Web应用程序具有规模大、复杂度高、开发周期长等特点,将其部署到实际运行环境后有可能出现不同程度的错误,因此人们开始重视Web应用程序的质量问题.与传统软件相比,Web应用程序具有异构、并发、跨组织和跨平台等特征,传统的软件测试方法基本无法实现对Web应用程序的充分测试,这也给软件测试领域提出了新的挑战[1].

目前,针对Web应用程序大多采用白盒测试技术,且大部分测试都需要人工手动生成测试用例,难度大,而且用例生成不充分.目前,国内外关于Web应用程序测试方面的研究工作较多,研究重点也各有不同.例如J. Kong等[2]提出用状态来表示Web应用程序中网页和网页中的各种元素,规定用状态的迁移来表示网页的超级链接或各页面间的跳转,该方法虽然理解起来很直观,但是最终形成的状态图错综复杂,表达起来很麻烦;基于用户会话的测试方法利用Web应用程序的域数据进行软件测试,但该方法没有考虑到Web应用程序具有多用户交互的特性,测试不够全面;在该方法的基础上武晋南等[3]提出了基于用户行为和用户会话的Web应用测试新方法,该方法注重了用户的行为,在一定程度上满足了软件的功能测试需求;W.

收稿日期:2012-10-11

基金项目:罗定职业技术学院科学研究基金资助项目(KY10030)

作者简介:刘小园(1978—),男,江西省樟树市人,罗定职业技术学院讲师,硕士,主要研究方向为网络与数据库系统、软件工程.

Alfond等[4]提出结构测试方法,该方法把与结构相关的所有实体如超级链接、表单和窗体等网页间的跳转关系都表示出来,强调链接和动态交互等带导航特性的内容,由于忽略了 Web 应用程序的网页中包含脚本、组件、接口对象以及服务端网页间的重定向关系,测试显然不够全面;D. C. Kung 等[5]提出应用于 Web 应用程序的图形测试模型,并进行了一些 Web 应用程序测试的基础定义,但该方法基本忽略了服务器端网页的动态行为.

本文在借鉴以往 Web 应用程序测试方法优点的基础上,提出一种基于 Web 应用程序的网页分类测试方法,以期达到不仅提高测试效率,而且使测试更加充分的目的,并通过对新生报到管理系统测试进行验证.

1　基于网页划分的 Web 应用程序测试

Web 应用系统是一系列 Web 网页和部件组成的系统,其模型如图 1 所示. 用户在客户机上使用浏览器向中间服务器发送 Web 应用请求,中间服务器将收到的请求转发给应用服务器,并将从应用服务器收到的所馈结果以静态网页的形式转发到客户机上,客户机浏览器显示给用户.

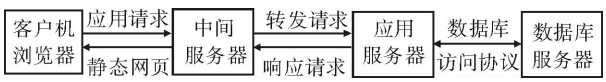


图 1　Web 应用系统模型

Web 应用程序由大量的 Web 网页和网页间的连接组件组成. Web 网页是能在客户机浏览器上显示的信息体,分为用 HTML 编写的静态网页和用 ASP(active server pages)或 JSP(java server pages)动态技术生成的网页. 连接组件则通过超级链接、表单和窗体等与 Web 页面相关联,用户通过 Web 网页实现对系统的访问及使用. 网页可分为如下 3 种不同的类别.

静态网页:用 HTML 创建的网页. 每个静态网页都有固定的 URL,且都存储在 Web 服务器上,都是一个独立的 HTML 文档.

数据库网页:是动态网页的一种,与数据库关联,其内容是数据库中的数据元素,会因为数据库内容的变化而发生变化.

动态网页:以 HTML 为基础,结合 ASP 或 JSP 创建的网页. 动态网页不独立存在于服务器上,只是在用户发送请求后,才在服务器上运行并动态生成一个 HTML 网页返回给用户,其内容会因为系统状态和用户实时操作的不同而发生变化.

基于网页划分的测试技术有以下 3 种:

1)静态网页采用黑盒测试(HHCS). 黑盒测试也称为功能测试,是一种穷举测试,需要把所有可能作为测试用例,适合用来测试功能确定的内容. 由于静态网页的内容和功能一般是确定的,且 URL 是固定的,因此非常适合采用黑盒测试. 测试内容包括浏览静态网页的每个页面,检查页面中的文字内容是否都能正确显示;检查网页中的每个超级链接,查看是否能够跳转到正确的页面;检查网页中的每张图片是不是都能及时正确地显示;检查网页中的所有表单对象是不是拒绝错误数据和接收正确数据. 特别需要强调的是,由于 Web 应用程序用户平台具有不确定性,所以需要针对 Web 应用程序的网页界面和软件功能使用多种平台和浏览器进行搭配测试,以检测软件的跨平台错误.

2)数据库网页采用白盒测试(BHCS). 白盒测试深入到网页代码一级进行测试,优点是发现问题早、效果好;缺点是需要开发人员在 Web 应用程序编码阶段,根据自己对代码的理解进行软件测试,测试的工作量大,只适合检测少量网页. 在 Web 应用程序中,数据库起着重要的作用,为整个软件系统数据查询和存储提供空间. 数据库网页可能发生的错误主要有 2 种:一是由于用户提交的表单信息不正确而造成数据不一致;另一种是由于程序设计本身引起的错误. 第一种错误主要靠黑盒测试时的表单测试来检查,第二种错误则需要对数据库驱动网页进行基于网页划分的白盒测试.

3)动态网页采用灰盒测试(FHCS). 灰盒测试以黑盒测试为主,白盒测试为辅,非常适合动态网页动态生成的特点. 其工作原理是通过简单查看动态网页的内部代码(不是像白盒测试那样完整地查看),了解动态网页的运行状况,有助于把测试用例设计得更加合理. 动态网页比程序更容易被查看,利用这个特点,在黑盒测试基础上对动态网页进行检查即可实现网页的灰盒测试.

Web 应用程序的测试用例往往是靠手工生成,测试成本很高,测试效率却很低. 为了提高测试的效率和覆盖率,基于网页划分的白盒测试的测试用

例可以借助如下算法来自动生成.

1)选择数据库网页 $H(h_1,h_2,\cdots,h_n)$ 的 URL 地址 $M_i(1\leqslant i\leqslant n)$ 作为测试用例;

2)从网页 h_i 中选择一个操作 J_i;

3)将 J_i 跳转到的新网页地址登记为 N_i;

4)若 N_i 没有包含在 H 中,则将 N_i 作为新的测试用例添加到 H 中,若 N_i 已经包含在 H 中,直接进入下一步;

5)将 J_i 登记为"已登记",重复第 2 步直到 H 的所有网页的所有操作都已登记;

6)对最终的 H 进行测试.

2 试验验证

笔者通过试验验证了基于网页划分的 Web 应用程序测试方法的有效性. 随着校园信息化建设的深入开展,新生报到管理系统是目前校园内较为普遍的 Web 应用程序,故选用该系统进行试验测试. 网站面向学生和学校职能部门管理员,学生可以网预报名、核对本人信息、申请错误信息更改、申请转专业等;部门管理员可以实现收费管理、宿舍分配、审批转专业申请、审批错误信息、更改申请等共 50 个主要功能. 为了达到测试效果,分别插入了 3 种不同类型共 45 个错误,其中与变量和控制流有关的脚本错误 15 个,与表单对象有关的表单错误 15 个,与数据库操作有关的数据库查询错误 15 个. 测试过程包括如下 4 个步骤:

1)将 Web 应用程序的网页进行分类;

2)对不同类别的网页分别生成测试用例;

3)单个激活错误,进行测试,并记录测试结果;

4)根据测试记录,主要从功能覆盖率和错误检测率两方面对测试方法进行有效性评估.

为了评估基于网页划分的 Web 应用程序测试方法的有效性,每次只激活一个错误,但执行所有测试用例,并采用结构测试方法进行了对比测试. 测试过程如图 2 所示,所得到的功能覆盖率和错误检测率见表 1. 从表 1 可知,基于网页划分的 Web 应用程序测试方法发现的错误更多,覆盖范围更大,是 Web 应用程序测试技术中较为有效的测试方法.

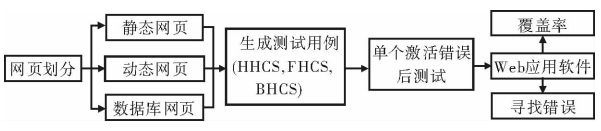


图 2 测试框架

表 1 功能覆盖率和错误检测率 %

测试方法	功能覆盖率	错误检测率
HHCS	97	73
BHCS	99	81
FHCS	97	76
基于网页划分测试方法	98	77
结构测试方法	92	71

3 结论

本文提出一种基于网页划分的测试方法,有效地解决了 Web 应用程序测试技术中测试用例代价高和生成不充分的问题. 试验结果表明该方法在功能覆盖率和错误检测率等方面都有更好的表现. 未来将重点研究 Web 应用程序测试框架、Web 应用程序测试的对象模型与应用、Web 应用程序测试中测试用例及复用等[6].

参考文献:

[1] 路晓丽,董云卫. Web 应用软件的结构测试研究[J]. 计算机科学,2010,37(12):110.

[2] Kong J,Zou C,Zhou H. Improving software securityvia runtime instruction-level taint checkingp[C]//Proc of the 1st Workshop on Architectural and System Support for Improving Software Dependability,Califorria:ACM Press,2006:18 - 24.

[3] 武晋南,高建华. 基于用户行为和会话的 Web 应用测试方法[J]. 计算机工程,2010,36(8):83.

[4] Alfond W,Orso A,Manolios P. Using positive tainting and syntax-aware evaluation to counter SQL injection attacks [C]//Proc of the 14th ACM SIGSOFT International Symposiumon Foundations of Software Engineering, New York:ACM Press,2006:175 - 185.

[5] Kung D C,Liu Chien-hung,Hsia Pei. An object-oriented Web test model for testing Web applications[C]//Proceedings of the 24th Annual International Computer Software and Applications Conference (COMPSAC. 00),Taipei:[s. l.],2000: 537 - 542.

[6] 尚冬娟,郝克刚,葛玮,等. 软件测试中的测试用例及复用研究[J]. 计算机技术与发展,2006,16(1):69.

文章编号:2095-476X(2012)06-0035-06

基于能量块与峰度特征的联合检测算法研究

杜海明[1], 孙明权[2]

(1.郑州轻工业学院 电气信息工程学院, 河南 郑州 450002;
2.中国人民解放军91286部队 航空管制中心, 山东 青岛 266003)

摘要:为提高信号检测概率,提出了基于归一化峰度的特征检测与能量块检测相结合的联合检测算法:在信号和噪声均服从高斯分布时,采用能量块检测的相关计算公式;对数据块中信号点个数远小于数据块长度时,采用数学方法分析归一化峰度值的变化情况.仿真结果表明,联合检测方法提高了检测概率和检测性能,具有一定的应用价值.

关键词:归一化峰度;能量块检测;特征检测;联合检测
中图分类号:TN911 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.010

Study on the joint detection algorithm based on energy block and kurtosis characteristic

DU Hai-ming[1], SUN Ming-quan[2]

(1. *College of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450002, *China*;
2. *The Air Traffic Control Center of No.* 91286 *Unit of PLA, Qingdao* 266003, *China*)

Abstract: A joint detection algorithm was presented based on characteristic detection of normalized kurtosis and energy block detection to improve the detection probability. When signal and noise are Gaussian distribution, the related calculation formula was applied for the energy block detection. Normalized kurtosis variation was analyzed based on mathematic theory when the number of signals changes from small to maximum in the data block. The simulation results showed that the detection probability and detection performace were improved by using the joint detection algorithm, so it will be useful in the detection of data block.

Key words: normalized kurtosis; energy block detection; characteristic detection; joint detection

0 引言

信号检测是进行信号处理的首要步骤,是现代信息与信号处理的基础,在通信、雷达、声纳等研究领域一直受到重视.信号检测方法的优劣,在很大程度上决定了信号处理的复杂度和系统处理的整体效果.1967年H. Urkowitz[1]第一次提出了能量检测,它属于一种非相干信号检测方法,不需要发射信号的先验知识,适用于任何信号,并且其硬件复杂度低,实现信号检测非常简单,因此它是高斯背景下应用最广泛的信号检测方法之一.目前能量检测技术已经成为认知无线电、移动通信、卫星通信以及超宽带通信[2-6]中常见的检测技术.

峰度是现代数字信号处理中衡量随机信号与高斯信号区别度的一个重要参数,属于四阶统计量,在现代信号处理中的应用很多.胡啸等[7]将归

收稿日期:2012-03-22
基金项目:河南省科技厅科技攻关项目(112300410144)
作者简介:杜海明(1977—),男,河南省灵宝县人,郑州轻工业学院讲师,博士,主要研究方向为信号预处理和UWB通信.

一化峰度用于弱非线性系统的盲辨识,表明归一化峰度能够精确辨识弱非线性系统;赵锡凯等[8]将最大峰度准则与非线性优化中的梯度法相结合,并将其应用到非因果 AR 系统的盲辨识. 在 CFAR 检测算法的研究中,M. E. Smith 等[9]提出的采用统计量 VI 判断背景环境是否均匀的方法,本质上也是利用了归一化峰度的特点. 基于峰度准则,I. Guvenc 等[10]针对脉冲超宽带无线网络系统提出了一种基于动态门限的 TOA 估计算法,主要思想是利用接收信号峰度与接收信号的最小、最大能量值之间的关系来自适应调整归一化门限值.

在阵列信号处理中,利用能量检测的方法检测目标信号[11],由于外辐射源信号辐射的时间长短以及信号辐射发生和结束的时间、辐射位置等都具有很强的随机性,采用能量块检测方法检测是否存在有用信息时,数据段长度的选取对非相干检测的性能有很大影响. 如果该数据段长度选取得比较短,将无法满足 DOA/TDOA 估计时所需要的快拍数要求,并因能量累积太小而影响接收性能;如果数据段的长度选择得太大,信号采集时,尽管接收到了一定长度的有用信号,但是如果有用信号的长度远小于数据段的长度,将可能因为引入了过多的噪声而削弱有用信号,造成有用信号的检测概率降低. 为了降低有用信息的丢失,合理选择数据块长度就显得非常重要. 归一化峰度可以用来描述数据块的波形形状特征,具有可加性的重要性质. 因此,本文拟在数据块的能量检测中引入归一化峰度的概念,将特征检测和能量块检测相结合形成联合检测,以提高有用信号的检测概率.

1 能量块检测与信号模型

基于阵列信号处理原理,在利用多个子站和一个中心站构成的外辐射源定位监测系统中,取其中一路信号并采用能量块检测的方法检测目标信号,能量块检测的原理如图 1 所示.

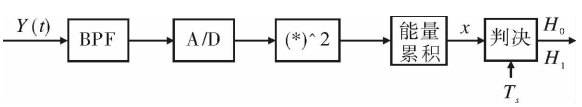


图 1 能量块检测的原理框图

图 1 中 $Y(t)$ 是通过天线或者传感器接收到的待检测信号,待检测信号先通过带通滤波器(BPF)滤除带外噪声或者其他干扰信号;然后经过模拟/数字(A/D)转换器、平方器和能量累积器,得到检测统计量 x(能量值);将 x 与预先设定的门限值 T_s 进行比较,当 $x > T_s$ 时,则判定目标信号存在,输出 H_1;当 $x < T_s$ 时,则判定目标信号不存在,输出 H_0.

由经典信号检测理论可知,信号检测问题相当于一个二元假设问题,即

$$y(t) = \begin{cases} n(t) & H_0 \\ s(t) + n(t) & H_1 \end{cases} \quad ①$$

其中,$n(t)$ 为均值为零、方差为 σ^2 的高斯白噪声. 经过 A/D 转换后,假设得到的信号为 $y(n)$,在 H_1 时,假设信号 $s(n)$ 是零均值的高斯随机信号. 再经过平方运算和累加运算后求平均得到 x,其计算表达式为

$$x(i) = 1/M \cdot \sum_{l=1}^{M} y(i,l)^2 \quad ②$$

其中,i 表示第 i 个数据块,l 表示第 i 个数据块中进行累积量计算的第 l 个数据点. 在式①中,若只存在噪声,由②式可知

$$x_0(i) = 1/M \cdot \sum_{l=1}^{M} n^2(i,l)$$

因 $n(l)$ 是均值为 0 且方差为 σ^2 的独立同分布的高斯白噪声,$x_0(i)$ 是 M 个数据的平方和,因此 $x_0(i)$ 是自由度为 M 的中心 χ^2 分布,它的概率密度函数为

$$f_{H_0}(x) = \frac{x^{\frac{M}{2}-1} e^{-\frac{x}{2\sigma^2}}}{\sigma^M 2^{\frac{M}{2}} \Gamma(\frac{M}{2})}$$

同理可得

$$f_{H_1}(x) = \frac{x^{\frac{M}{2}-1} e^{-\frac{x}{2\sigma_1^2}}}{\sigma_1^M 2^{\frac{M}{2}} \Gamma(\frac{M}{2})} \quad ③$$

式③中,$\sigma_1^2 = \sigma_s^2 + \sigma^2$,信号与噪声相互独立,前面已假设信号属于均值为 0 的高斯信号,因此式③也是中心 χ^2 分布.

假设门限值取为 T,其虚警概率 P_F 和检测概率 P_D 分别为[12]

$$P_F = P_r\{X > T \mid H_0\} \quad ④$$

$$P_D = P_r\{X > T \mid H_1\} \quad ⑤$$

由 Neyman-pearson 准则可知,在一定的虚警概率的要求下,通过公式⑥的计算,可以求得对应的门限值

$$T = F_{H_0}^{-1}(1 - P_F, k_0, \theta_0) \quad ⑥$$

将该门限值代入式 ⑦ 可求出其检测概率

$$P_D = 1 - F_{H_1}(T, k_1, \theta_1) \quad ⑦$$

在式 ⑥ 和 ⑦ 中,$k_0 = k_1 = M/2$,$\theta_0 = \frac{2}{M}$,$\theta_1 = \frac{2}{M}\left(1 + \frac{\sigma_s^2}{\sigma^2}\right)$,则

$$F_{H_1}(x, k_1, \theta_1) = \int_0^x \frac{1}{\theta_1^{k_1}\Gamma(k_1)} t^{k_1 - 1} e^{-\frac{t}{\theta_1}} \mathrm{d}t \quad ⑧$$

$$F_{H_0}(x, k_0, \theta_0) = \int_0^x \frac{1}{\theta_0^{k_0}\Gamma(k_0)} t^{k_0 - 1} e^{-\frac{t}{\theta_0}} \mathrm{d}t \quad ⑨$$

其中 $F_{H_0}^{-1}(\,)$ 为 $F_{H_0}(\,)$ 的逆函数.

因此,由 ⑥ 式可以得到一定虚警概率要求下的归一化门限值 T,然后用噪声的方差值 σ^2 与 T 相乘,得到的值作为图 1 中的门限值 T_s. 由于在实际中噪声的方差未知,因此可以通过求 $E[X_0] = M\sigma^2$ 得到 $\sigma^2 = \frac{E[X_0]}{M}$,从而得到检测所需的门限值 T_s. 由公式 ⑦ 和 ⑧ 可知,在门限值 T 确定后,信噪比越大,则 q_1 的值就越大,因此公式 ⑧ 的计算结果越小,则公式 ⑦ 计算得到的检测概率就越高.

在信号采集并进行检测时,由公式 ② 可知,有用信号点的功率一定,在理想情况下,信号与噪声个数相等时,其信噪比为 SNR,并且期望在该信噪比下检测到有用信息的概率为 p. 而在实际检测过程中,如果信号点的个数 N 远小于数据块的长度 M,则相对于整个数据块来说,实际的信噪比 SNR 大大降低,因此将会引起检测概率降低,漏检概率增大. 因此要求在一定虚警率条件下,最大程度地提高存在辐射信号时的检测概率,但不能降低检测门限. 因此研究在数据块长度一定、适度信噪比的信号点个数较少的情况下,如何提高检测概率就显得很有必要.

2　峰度的定义与性质

若 $x(k)$ 是一个实平稳随机过程且存在 n 阶矩,则其 4 阶累积量与矩之间的关系为

$$C_4 = m_4 - 3m_2^2 - 4m_1m_3 + 12m_1^2m_2 - 6m_1^4 \quad ⑩$$

归一化峰度的定义为

$$Kurt(x) = \frac{m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4}{(m_2 - m_1^2)^2} \quad ⑪$$

若 $m_1 = 0$,则由 ⑩ 和 ⑪ 可分别得到

$$C_4 = m_4 - 3m_2^2 \quad ⑫$$

$$Kurt(x) = m_4/m_2^2 \quad ⑬$$

又因为零均值高斯随机变量的 k 阶矩可用 2 阶矩表示[13],即

$$m_k = E[x^k] = \begin{cases} [1,3,5,\cdots,(k-1)]\sigma^2 & k \text{ 为偶数} \\ 0 & k \text{ 为奇数} \end{cases} \quad ⑭$$

联合 ⑫⑬⑭ 可知,若信号为高斯信号,则计算 ⑬ 式,得到归一化峰度值应当为 3, 基于归一化峰度的可加性,对 2 个统计独立的高斯随机变量 x_1 和 x_2,若 $(x_1 + x_2)$ 仍然属于高斯信号,那么归一化峰度值 $Kurt(x_1 + x_2) = 3$ 仍成立. 由公式 ⑬ 可知,在零均值高斯信号的条件下,归一化峰度值的大小只与 4 阶矩 m_4 和 2 阶矩 m_2 有关,因此利用归一化峰度的定义,讨论当 2 个不同长度的信号加和后,通过计算该数据段的 2 阶矩和 4 阶矩,可以分析对该段数据块的归一化峰度值变化的影响,后面给出其数学分析和仿真分析.

假设数据块长度为 M,其中只含有 N 个目标信号,噪声的长度和数据块的长度相同. 求其方差

$$D(y) = E(y^2) = \frac{1}{M}\sum_{i=1}^{M} y_i^2 = \frac{1}{M}\left[\sum_{i=1}^{N}(s_i + n_i)^2 + \sum_{i=N+1}^{M} n_i^2\right] = \frac{1}{M}\left[\sum_{i=1}^{N} s_i^2 + \sum_{i=1}^{M} n_i^2\right] = \frac{1}{M}\left[\frac{N}{N}\sum_{i=1}^{N} s_i^2 + \sum_{i=1}^{M} n_i^2\right] \quad ⑮$$

由式 ① 和均值的性质可以得到其 $E(y) = 0$. 令 $\sigma_s^2 = \frac{1}{N}\sum_{i=1}^{N} s_i^2$,$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M} n_i^2$,则

$$D(y) = \frac{N}{M}\sigma_s^2 + \sigma^2 \quad ⑯$$

同理

$$E(y^4) = \frac{1}{M}\left(\sum_{i=1}^{N}(s_i + n_i)^4 + \sum_{i=N+1}^{M} n_i^4\right) = \frac{1}{M}\left(\sum_{i=1}^{N}(s_i^4 + n_i^4 + 6s_i^2n_i^2) + \sum_{i=N+1}^{M} n_i^4\right) \quad ⑰$$

因为信号和噪声都是零均值高斯分布,将公式 ⑭ 代入 ⑰ 计算,可以得到

$$E(y^4) = 3\left(\frac{N}{M}\sigma_s^4 + \sigma^4 + \frac{2N}{M}\sigma_s^2\sigma^2\right) \quad ⑱$$

将式 ⑱ 和 ⑯ 代入 ⑬,可得此时归一化峰度的计算公式为

$$Kurt(x) = \frac{3\left(\frac{N}{M}\sigma_s^4 + \sigma^4 + \frac{2N}{M}\sigma_s^2\sigma^2\right)}{\left(\frac{N}{M}\sigma_s^2 + \sigma^2\right)^2} \quad ⑲$$

上式中 $\sigma_s^2 = SNR \times \sigma^2$,因此当 $N \ll M$ 时,运用数学分析的方法对式 ⑲ 进行讨论,可以得到 $Kurt(x) > 3$. 在一段数据中,若只存在高斯噪声,其归一化峰度值就在 3 附近;若高斯噪声和信号的长度基本相同或者相差不大时,计算得到的归一化峰度值也在 3 附近;但是若在一段数据中信号与噪声的长度相差很大且信号的信噪比较高时,其归一化峰度值将会发生变化,相当于该段数据的波形形状特征发生了变化.

因此,通过数学分析可知,数据段中存在有用信号的个数远小于数据段长度时,在不同的信噪比下,对归一化峰度的计算结果有一定的影响,相当于波形形状发生了变化,通过设定归一化峰度的检测门限 T_{VI},可以弥补能量块检测中实际信噪比降低所引起的检测概率降低,以达到提高有用信号检测概率的目的.

3 能量块与峰度特征的联合检测

前面谈到,在利用能量块检测外辐射源的目标信号时,由于辐射信号发生的起始时间、结束时间以及辐射持续时间的长短都是随机的,无论该数据段选得过短或者过长,都会对检测以及信号处理的性能产生很大影响,因此选取数据段长度的参数就显得十分重要. 根据实际系统需要来选择能量块检测的长度 M,假设能量块检测的数据段由 M 个采样点构成,理想情况下,辐射信号时整个数据段都由信号和噪声组成,或者说接收信号都是由噪声和信号组成的,此时将由②式求出的 x 与 T_s 相比较,判断有用信号存在的概率就会很大. 但是如果该数据段中只含有 N 个信号加噪声的采样点,其余 $M-N$ 都是噪声的采样点,若 $N=M$ 时,可能会产生噪声淹没有用信号的现象,相当于降低了信噪比,因此利用能量块检测法检测到有用信号的概率将会降低,即在假设每个信号点功率相同和每个噪声功率相同的情况下,检测概率 $prob(x(N,M) > T \mid H_1)$ 随着 N 的增大而增大,当 $M=N$ 时达到最大. 然而当 $N \ll M$ 时,该段数据的归一化峰度值将会远大于 3,因此可以通过归一化峰度值来衡量波形形状发生变化的情况,完成该段数据的特征检测;且 $prob(k(x(N,M)) > T_{VI} \mid H_1)$ 将会随着 N 的增大先变大然后变小,当 $M \ll N$ 时其归一化峰度计算结果又将在 3 附近,随着信号点个数的增加,其峰度检测概率将会变得很小,然而此时能量检测的概率将达到最大.

因此,为解决能量块检测过程中目标信号点个数小于数据块长度时有用信息丢失,进而造成分析外辐射源的物理特征或者其他性质的信息不完备的问题,本文提出采用能量块检测和峰度检测的联合检测,在一定虚警概率的要求下,使有用信号的检测概率达到最大,即当 $N \ll M$ 时,可以利用峰度值的大小作为波形形状特征进行检测,将特征检测与能量块检测相结合;当 N 值逐渐变大时,归一化峰度值降低,此时能量块检测起主要作用,其实现的原理框图如图 2 所示. 图 2 中,检测判断时,只要峰度值或者能量值其中一个条件满足,就可以判定为目标信号存在;若 2 个条件都不满足,则说明目标信号不存在. 在具体应用时,应该根据虚警率、信噪比以及能量块检测等要求,来具体设定 T_s 和 T_{VI} 的值.

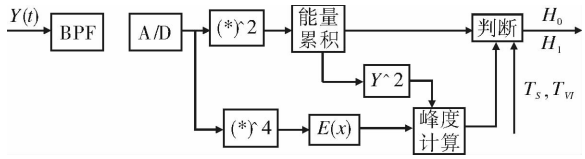


图 2 能量块与峰度特征联合检测原理框图

4 性能仿真与分析

首先分析在数据段长度相同、信噪比相同时,数据段信号个数 N 的变化对归一化峰度计算结果的影响如图 3 所示. 取数据段长度 $M=100$,数据段的数目为 10^5 个. 图 3 中曲线 1,3,5,7 分别表示信噪比为15 dB,归一化峰度值是 4,5,7,9 时大于归一化峰度的概率;曲线 2,4,6,8 分别表示信噪比为 10 dB,归一化峰度值是 4,5,7,9 时大于归一化峰度的概率. 由图 3 可知,其曲线的统计趋势与数学分析的结果相同,即当数据块中信号点的个数逐渐增大时,归一化峰度值也是逐渐增大到一个极大值后又慢慢变小;其次,信噪比不同时,相同数据块所含信号点数相同,大信噪比计算得到的归一化峰度值大,因此信号检测时,若只关心某一信噪比以上的信号时,可提高归一化峰度值的检测门限.

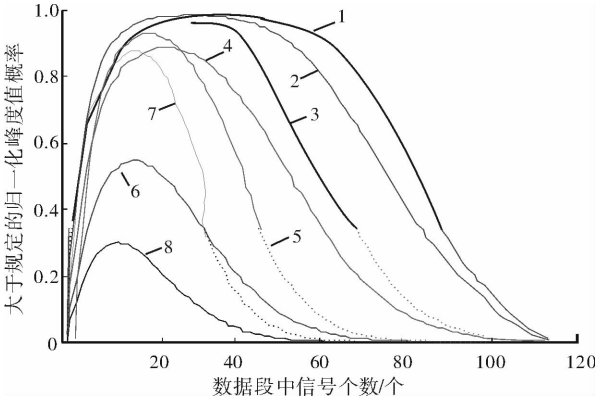


图 3　归一化峰度值的变化规律的统计

在信号检测时,应当在系统要求的虚警概率条件下提高检测概率. 因此对特征检测的门限进行蒙特卡洛实验,其数据块的个数为 5×10^5,数据块长度分别取 50,100 和 200,特征检测的门限取 4—8,表 1 给出了虚警概率的统计分析结果.

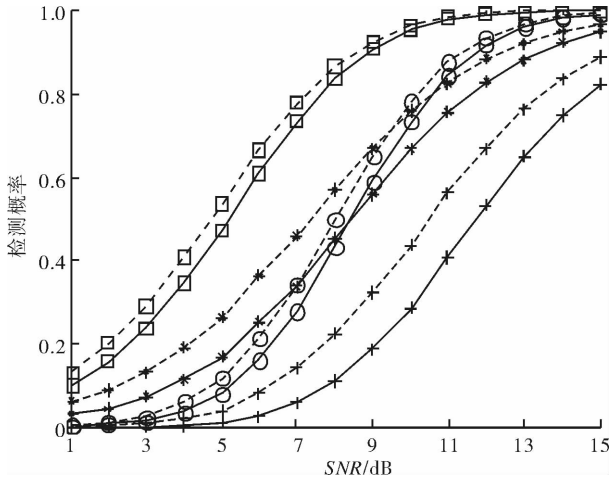
表 1　不同程度下虚警概率的统计结果

数据块长度	$Kurt(x)>4$	$Kurt(x)>5$	$Kurt(x)>6$	$Kurt(x)>7$	$Kurt(x)>8$
50	4.82e -2	4.90e -3	1.30e -3	3.79e -4	1.07e -4
100	2.71e -2	2.30e -3	2.86e -4	5.70e -5	1.45e -5
200	8.90e -3	2.84e -4	2.60e -5	4.00e -6	

由表 1 可见,数据块长度相同,随着归一化峰度检测门限的升高,虚警概率降低,这符合信号检测的基本理论;在相同的检测门限下,随着数据块长度的增加,其虚警概率也降低,其主要原因是,随着数据块长度的增加,基于大数定理可知,其统计特性更符合高斯分布,因此当数据块的长度越大时,其归一化峰度的值越接近 3,数据块长度大到一定程度后,归一化峰度值在 3 附近起伏. 因此利用归一化峰度进行特征检测时,应该根据系统需要选定数据长度,然后进行一定门限下虚警概率的仿真与测试.

为了分析能量块与峰度特征的联合检测的性能,数据块长度分别取 50 和 100,分析对比能量块检测与联合检测检测性能.

噪声方差为标准方差,数据块长度为 50,能量块检测门限 $T_s=1.61$,根据公式⑨其对应的虚警概率为 4.024e -3;数据块长度为 100,能量块检测门限 $T_s=1.65$,对应的虚警概率为 4.654e -5. 根据表 1,在数据块长度为 50 和 100 时,分别选择归一化峰度值 5 和 7 作为特征检测门限值. 图 4 给出了在信号点个数分别为 5 和 10 的情况下,随着信噪比的增大,采用能量块检测与联合检测时检测性能的变化与比较;图 5 给出了信噪比为 10 dB 和 5 dB 时,随着信号点个数的增加,能量块检测与联合检测在检测性能的变化与比较. 图 4 和图 5 中,能量块检测的检测性能为实线表示,联合检测的检测性能为虚线表示,两图其他曲线标志意义相同.



+ 表示(100,5);O 表示(100,10),* 表示(50,5),方块表示(50,10),数字 100 和 50 表示数据块的长度,数字 5 和 10 的单位分别为 dB 和点数

图 4　信号点个数分别为 5 和 10 的情况下在信噪比增大时的检测性能

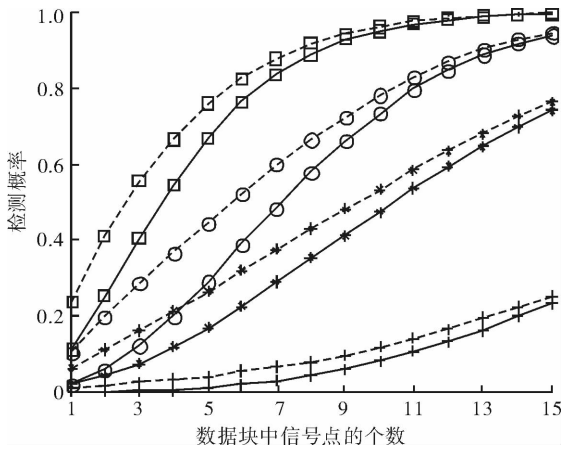


图 5　信噪比相同信号点个数增大时的检测性能

由图 4 和图 5 可知:1)联合检测的检测性能比能量块检测优越;2)在信号点个数相同的情况下,随着信噪比的升高,其检测性能越来越好,特征检测对提高检测性能的影响越来越小;3)在信号点信噪比相同的情况下,随着信号点个数的增加,其检

测性能越来越好,并且特征检测对提高检测性能的影响越来越小; 4)相同的数据块长度和相同的信号点信噪比时,数据块中信号点个数小时,其检测性能提高很大; 5)在信号点个数小且信噪比大时,特征检测对提高检测概率的影响很大,随着信号点个数增加,其影响逐渐变小; 6)相同信号点个数和信噪比下,数据块长度大时,噪声淹没信号的现象严重. 为了更好地说明噪声淹没信号,数据块长度为 100 和 50 时,在 10 dB 时分别取 10 个信号点,计算得到其整个数据段的平均信噪比为 0 dB 和 3.025 dB,代入能量检测的公式⑧,理论计算得到的检测概率分别为 0.898 0 和 0.997 1,联合检测的检测概率分别为 0.781 3 和0.961 4,其检测概率都高于只用能量块检测,说明联合检测在数据块检测中可以有效提高检测性能,其结果与数学分析的结论基本一致.

因此,在数据块中信号点的个数远小于数据块的长度时,在一定的虚警概率下,基于能量块和归一化峰度的联合检测方法与能量块检测相比较,前者能够有效地提高检测概率和检测性能,因此联合检测方法在数据块检测中可以更好地检测有用信号.

5 结论

本文提出了基于归一化峰度的特征检测与能量块检测相结合的联合检测算法:在信号和噪声均服从高斯分布时,采用能量块检测的相关计算公式,在信噪比和数据块长度以及信号点个数等条件下,利用数学方法分析了归一化峰度值的变化特征. 仿真结果与数学分析相吻合,由此可知联合检测算法具有一定的应用价值. 更重要的是,该方法在硬件实现上不增加过多的器件,只需要增加乘法器和累加器即可,在 FPGA 中实现该功能很简单. 该方法还可以用于其他地方,例如大量采集了某一随机信号,并将该信号存储起来,为了分析该随机信号的特征或者用于其他的信号处理,就需要在大容量存储空间查找该有用信息的信号段的存储位置,这时联合检测就是一个不错的选择.

参考文献:

[1] Urkowiz H. Energy detection of unkown deterministic signals[J]. Proceedings of the IEEE,1967,55(4):523.

[2] 虞贵财,罗涛,乐光新. 认知无线电系统中协同能量检测算法的性能研究[J]. 电子与信息学报,2009,31(11):2681.

[3] 隋丹,葛临东,屈丹. 一种新的基于能量检测的突发信号存在性检测算法[J]. 信号处理,2008,24(4):614.

[4] 张震廷,张钦宇,张乃通. 基于能量检测的脉冲超宽带接收机[J]. 吉林大学学报:工学版,2010,40(1):281.

[5] 张震廷,张钦宇,张乃通. 针对 IR-UWB 无线传感器网络的两步能量测距法[J]. 通信学报,2009,30(8):96.

[6] 吴绍华,张乃通. 基于 UWB 的无线传感器网络中的两步 TOA 估计法[J]. 软件学报,2007,18(5):1164.

[7] 胡啸,马洪. 归一化峰度及其在弱非线性系统盲辨识中的应用[J]. 信号处理,2010,26(9):1389.

[8] 赵锡凯,张贤达. 基于最大峰度准则的非因果 AR 系统盲辨识[J]. 电子学报,1999,27(12):126.

[9] Smith M E,Varshney P K. Intelligent CFAR processor based on data variability [J]. IEEE Trans On AES(S0018—9251),2000,36(3):837.

[10] Guvenc I,Sahinoglu Z. Threshold selection for UWB TOA estimation based on kurtosis analysis[J]. IEEE Communications Letters,2005,9(12):1025.

[11] 杨晨辉,马远良,杨益新. 峰值能量检测及其在被动声纳显示中的应用[J]. 应用声学,2003,22(50):31.

[12] Chen Yunfei. Improved energy detector for random signals in Gaussian noise[J]. IEEE Transactions on Wirelesss Communications,2010,9(2):558.

[13] 姚天任, 孙洪. 现代数字信号处理[M]. 武汉:华中科技大学出版社,1999:184 - 190.

文章编号:2095-476X(2012)06-0041-03

一种改进的动态帧时隙ALOHA算法

陈燕, 李娜娜, 张娜

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:针对目前动态帧时隙ALOHA算法所需时隙数较多的问题,基于电子标签数量和发生碰撞概率的关系,提出了一种改进的动态帧时隙ALOHA算法.在改进算法中,帧长的确定不需事先估算电子标签的数量,而只需根据上一帧中电子标签发生碰撞的概率来确定.仿真试验表明,改进算法所需的时隙数和计算量较少.

关键词:射频识别;动态帧时隙ALOHA算法;时隙数

中图分类号:TP391 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.011

An improved dynamic frame slotted ALOHA algorithm

CHEN Yan, LI Na-na, ZHANG Na

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*)

Abstract: Aiming at the problem that the dynamic frame slotted ALOHA algorithm required more timeslots, an improved dynamic frame slotted ALOHA algorithm was proposed through analyzing the relation between the number of electronic tag and probability of collision, which determine the length of frame don't need estimate the number of tags, only need a frame timeslot conflict probability. Simulation results showed that the improved algorithm has less timeslot and computation.

Key words: RFID; dynamic frame slotted ALOHA algorithm(DFSA); timeslot

0 引言

射频识别RFID(radio frequency identification)是一种非接触式的自动识别技术,应用广泛.它以射频信号作为信息和能量传递的媒介,完成与被测物体间的信息交互[1].在很多场合,读写器需要在很短时间内准确地识别出多个电子标签,但是多个电子标签相互独立且共享同一个射频信道,因此若多个电子标签在同一时刻发出信息,就会造成信号混叠,发生碰撞.

基于时隙的ALOHA算法是电子标签防碰撞中常用的一种方法,其中动态帧时隙ALOHA(dynamic frame slotted ALOHA, DFSA)算法由于能够适应电子标签数量不同的场合,具有较高的电子标签识别效率,被广泛应用于RFID系统.帧时隙ALOHA算法(frame slotted ALOHA, FSA)把多个时隙组成一个帧,电子标签在每帧内随机选择一个时隙发送信息.

在FSA算法中,一帧内的时隙数量即帧长是固定不变的,因此当电子标签数量远大于帧长时,读取电子标签的时间会大大增加;当电子标签数量远小于帧长时,则会造成时隙浪费.而在DFSA算法中,帧长能够随阅读区域中电子标签的数量而动态改变,当待识别电子标签数大于帧长时会增加帧长度;反之,则会减小帧长度.因此在DFSA算法中,如

收稿日期:2012-11-01

基金项目:河南省科技攻关项目(112102210321)

作者简介:陈燕(1979—),女,河南省南阳市人,郑州轻工业学院讲师,硕士,主要研究方向为射频识别技术、嵌入式系统.

何确定帧长是提高电子标签识别效率的关键技术.

目前,大多数帧长确定方法是先估算待识别电子标签的数量,然后再确定帧长. 文献[2]假设电子标签参与响应时间按照泊松分布,该算法简单易行,但电子标签数量估算误差较大;文献[3]根据阅读器检测到的某一帧中空闲时隙数、成功发送的时隙数、碰撞的时隙数和概率估计量来寻求最优解,该方法需要的总时隙数较少,对电子标签数量的估计相对精确,但是计算量较大;文献[4]通过定义碰撞比,并利用逼近算法来估算电子标签数量,其迭代过程复杂,耗时较长,识别效率较低;文献[5]通过迭代算法先计算出在不同帧长情况下发生不同次数碰撞时的电子标签数,并把计算结果存放在数组 $C(2,2^8)$ 中,然后直接读取矩阵数据获得电子标签数量,从而提高系统的效率,但该方法中存放表格需要占用一定的内存空间. 另外,虽然理论上当帧长等于待读取的电子标签数量时,算法能达到最高的工作效率,但实际应用中读写器能设定的帧长通常是定值,估算出电子标签数量后还需进一步调整,文献[6-7]给出了估算的电子标签数量和帧长的对应关系. 本文拟提出一种改进的 DFSA 算法,只需根据上一帧中发生碰撞的概率即可直接确定下一帧的帧长,不需事先估算电子标签的数量,减少计算量和所需的时隙数.

1 改进的 DFSA 算法

1.1 帧长的确定

假设电子标签总数为 n,帧长为 L,由于各个电子标签相互独立,因此在某一个时隙中同时出现 k 个电子标签的概率为

$$P_k = C_n^k\left(\frac{1}{L}\right)^k\left(1-\frac{1}{L}\right)^{n-k}$$

有 2 个及以上的电子标签选中某一时隙,即发生碰撞的概率为

$$P_k(L,n) = 1-\left(1-\frac{1}{L}\right)^n-\frac{n}{L}\left(1-\frac{1}{L}\right)^{n-1} \qquad ①$$

由式①可得,帧长分别为 16,32,64,128 和 256 时,电子标签数量和发生碰撞的概率 $P_k(L,n)$ 的关系,如图 1 所示. 当帧长确定时,电子标签数量和发生碰撞的概率呈单调递增关系.

为了便于实现,通常读写器设定的帧长为 2^Q ($Q=0,\cdots,8$),即 1,2,4,8,16,32,64,128,256. 不同电子标签数量对应的帧长见表 1.

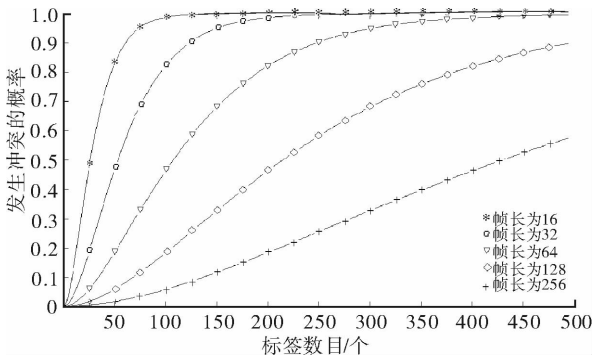


图 1 电子标签数量和发生碰撞概率的关系曲线

表 1 不同电子标签个数对应的帧长度

电子标签数目/个	帧长度	分组数/组
884—1 141	256	4
623—883	256	3
355—622	256	2
177—354	256	1
89—176	128	1
45—88	64	1
23—44	32	1
12—22	16	1
5—11	8	1
1—4	4	1

根据图 1 和表 1,本文改进算法中初始帧长为 16,发生碰撞的概率可由发生碰撞的时隙数/帧长求得. 确定帧长的过程如下.

1)由于 $P_k(16,12)=0.170\ 3$, $P_k(32,23)=0.160\ 7$, $P_k(64,45)=0.156\ 1$, $P_k(128,89)=0.153\ 8$, $P_k(256,177)=0.152\ 6$, $P_k(16,22)=0.043\ 7$, $P_k(32,44)=0.401\ 6$, $P_k(64,88)=0.400\ 5$, $P_k(128,176)=0.400\ 0$, $P_k(256,354)=0.402\ 5$,因此当发生碰撞的概率为 15% ~40% 时,帧长保持不变.

2)由于 $P_k(64,22)=0.045\ 9$, $P_k(128,44)=0.046\ 5$, $P_k(256,88)=0.046\ 8$,因此当发生碰撞的概率为 4% ~15% 时,帧长缩小为原来的 1/2. 若帧长为 16,则帧长保持不变.

3)当发生碰撞的概率小于 4% 时,帧长缩小为原来的 1/4. 其中,若帧长为 16,则帧长保持不变;若帧长为 32,则帧长缩小为原来的 1/2.

4)由于 $P_k(16,45)=0.780\ 8$, $P_k(32,89)=0.770\ 6$, $P_k(64,177)=0.765\ 4$, $P_k(128,355)=0.765\ 6$,因此当发生碰撞的概率为 40% ~76% 时,

帧长扩大为原来的 2 倍. 若帧长为 256,则帧长保持不变.

5)当发生碰撞的概率大于 76% 时,帧长扩大为原来的 4 倍. 其中若帧长为 256,则帧长保持不变;若帧长为 128,则帧长扩大为原来的 2 倍.

1.2　改进算法的步骤

Step1:阅读器发送请求命令,该命令包含初始的时隙数,本文算法中初始时隙数为 16.

Step2:电子标签接收到阅读器的命令后,在(1, N)中随机选择一个时隙,同时将自己的时隙计数器复位为 1.

Step3:如果电子标签的时隙计数器等于电子标签自己所选择的时隙数,则电子标签向阅读器发送信息;否则,电子标签不发送信息,并保留自己的时隙数.

Step4:若阅读器只收到 1 个电子标签的响应,则在阅读器正确读取电子标签信息后,电子标签进入休眠状态;若阅读器收到多个电子标签的响应,统计发生碰撞的时隙数 $Nc = Nc + 1$, 通过(Nc/帧长)确定发生碰撞的概率.

Step5:阅读器根据 1.1 中帧长的确定算法,按照该帧发生碰撞的概率动态地调整下一帧的帧长 N,然后转到 Step1,开始下一帧的识别.

2　仿真结果与分析

为了检验本文所提出的改进算法的识别效率,分别在不同电子标签数量情况下,对 FSA 算法、DFSA算法、文献 6 所提出的算法和本文提出的算法进行仿真分析. 其中文献[6]提出的算法是先根据文献[3]的方法进行电子标签的估算,然后根据估算出的电子标签数量确定帧长.

在每个仿真实验中, 给定电子标签的数量区间是 1 ~400. 为了提高仿真精度,仿真结果是经过 100 次运算的平均值. 仿真结果见图 2.

从图 2 可以看出,本文提出的改进算法和文献[6]提出的算法所需的时隙数较小,即识别效率较高. 文献[6]需要通过寻找使实际测得的成功、空闲和碰撞的时隙数值与理论上的成功、空闲和碰撞的时隙数值的差值最小,从而确定电子标签数,然后再确定帧长. 而本文提出的算法仅根据每一帧发生碰撞的时隙数就可确定下一帧的帧长,其计算量要远小于文献[6]提出的算法. 所以,本文所提出的算法具有计算量较小、所需时隙数较少的优点.

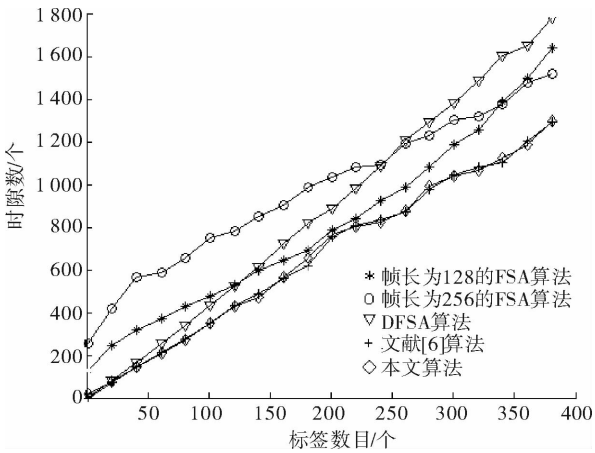


图 2　仿真结果

3　结论

在 DSFA 算法中,帧长的确定算法决定了电子标签的识别效率. 针对目前大多数帧长的确定方法是先估算待识别电子标签的数量,然后再确定帧长,计算过程复杂这一问题,本文基于不同帧长情况下电子标签数量与发生碰撞的概率关系,提出了改进的动态帧时隙 ALOHA 算法,简化了帧长的确定过程,仅仅根据上一帧发生碰撞的概率即可确定下一帧的帧长,从而减小了计算量,提高了DSFA算法的识别效率.

参考文献:

[1]　单承赣,单玉峰,姚磊,等. 射频识别(RFID)原理与应用[M]. 北京:电子工业出版社,2008.

[2]　Schoute F C. Dynamic frame length ALOHA[J]. IEEE Transactions on Communieations,1983,31(4):565.

[3]　Vogt H. Efficient object identification with passive RFID tags[C]//International Conference on Pervasive Computing, Berlin:Springer-Verlag, 2002:98 - 113.

[4]　Cha Jae-Ryong, Kim Jae-Hyun. Novel anti-collision algorithms for fast object identification in RFID system[C]// Proceedings of 11th International Conference on Parallel and Distributed Systems, Washington:IEEE Computer Society,2005:63 - 67.

[5]　黄仁,张静,程平. 一种 ALOHA 算法的帧长度调整方法[J]. 计算机工程与应用,2011,47(9):115.

[6]　尹君,何怡刚,李兵,等. 基于分组动态帧时隙的 RFID 防碰撞算法[J]. 计算机工程,2009,35(20):267.

[7]　李飞高,张贵林. 基于 ALOHA 的分组动态帧时隙 RFID 系统防碰撞算法[J]. 郑州轻工业学院学报:自然科学版,2012,27(3):80.

文章编号:2095-476X(2012)06-0044-04

基于ZigBee的热压机监控系统的设计

李帷笳[1], 梁万用[2]

(1.河南职工医学院 教务处, 河南 郑州 451191;
2.郑州轻工业学院 电气信息工程学院, 河南 郑州 450002)

摘要:针对多台热压机的无线实时监控和管理的问题,基于无线通信模块JN5148,开发了一个基于ZigBee的热压机监控系统.该系统采用基于继电整定的PID控制算法,通过ZigBee无线网络建立了PC机与多台烧结炉的无线通信.运行结果表明,该系统设计方法应用方便,通信可靠性高,温度控制误差<0.3℃,压力控制误差<0.1 kN.

关键词:热压机;JN5148;PID;无线监控
中图分类号:TN80 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.012

Design of sintering machine monitoring system based on the ZigBee

LI Wei-jia[1], LIANG Wan-yong[2]

(1. *Teaching Affairs Office, He'nan Medical College for Staff and Workers, Zhengzhou* 451191, *China*;
2. *College of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450002, *China*)

Abstract: Aiming at the problem of wireless monitoring and management for sintering machines, the wireless communication module JN5148 was used to develop a sintering machine monitoring system based on ZigBee technology. The system builds the wireless communication among the PC and many sets of sintering machines which using PID control algorithm based on relay setting and ZigBee wireless network. The running results showed that the methods is convenient, and its reliability is high, and tempreature control error is less than 0.3 ℃, and pressure control error is less than 0.1 kN.

Key words: sintering machine; JN5148; PID; wireless monitoring

0 引言

热压机是通过加热加压对材料进行加工成型的机器,广泛应用于金刚石制品、板类制品、石墨制品等的生产,产品的质量主要取决于3个方面,即温度压力的控制精度、工艺的设置合理性和整个工艺过程的实际工艺曲线[1-2].目前,在产品生产过程中,为了保证产品质量、工艺协同性和安全性,每台热压机往往由多人实时监控,造成人为失误的可能性较高,对企业的人力资源要求较高,车间环境也对工人的健康不利.同时,由于工人工作量大多是计件统计,这就导致有些工人可能为了更多的计件会去修改工艺时间,或者将不合适的产品混入.针对这些问题,已有学者、热压机制造企业和软件企业对热压机监控系统进行了研究和改进,如基于PLC的热压机PID控制系统研究[1]、烧结炉控制系

收稿日期:2012-06-19
基金项目:国家自然科学基金项目(51205372);郑州市科技攻关项目(112PPTGY249—6)
作者简介:李帷笳(1983—),男,河南省郑州市人,河南职工医学院助教,主要研究方向为计算机应用与电子信息技术.

统设计[2]、热压机控制系统的 PID 改进[3]. 现有研究和应用只能对 1 台热压机进行有线控制,无法对多台和分布式的热压机系统进行控制.

鉴于此,本文设计一种基于 ZigBee 的热压机监控系统,在实现对温度和压力精确控制的同时,解决对多台热压机的无线实时监控和管理问题.

1　系统设计原理

1.1　无线模块选择及无线网络构成

无线通信模块选用的是 Jennic 公司的超低功耗高性能无线通信模块 JN5148,该模块采用增强的 32 位RISC 处理器,集成了 2.4 GHz IEEE802.15.4 兼容的收发器,具有 128 K ROM,128 K RAM 以及各种丰富的模拟和数字接口. JN5148 具有 28 位 AES 安全处理器,MAC 加速器,500 kb/s 或 667 kb/s 数据速率.

该系统中 ZigBee 网络由网络协调器、中继器和终端设备(无线传感器节点)3 部分组成. 无线传感器节点主要实现对温度、压力等数据的采集和处理,无线传输数据,执行主控 PC 命令等功能;网络中只有一个协调器节点,负责与所控制的子节点通信;中继器负责网内信息的路由.

1.2　系统构成及原理

基于 ZigBee 的热压机无线监控系统由多台热压机(从机或终端节点)、协调器、RS232 通信电路及一台主控 PC 机(主机)组成(如图 1 所示). 该系统最多可接入65 536个终端节点,没有专用的路由节点,每个终端节点同时具备路由节点功能,能够入网. 各终端节点有各自的编号(地址),其运行由各自控制系统控制,运行实时数据(温度、压力、报警和故障等)由控制终端的无线模块 JN5148 转变为数字信号并传输给 PC 机,同时接收 PC 机发送来的命令并执行,具有交换控制消息和收发数据的功能[5]. 主控 PC 机对数据处理后通过开发的软件界面显示出来,管理员可根据相应信息通过键盘或鼠标输入相应的控制命令,该命令经由 PC 机端无线模块发送到对应的热压机,实现对热压机的实时无线监控. 同时,主控 PC 机具有报警和警告功能.

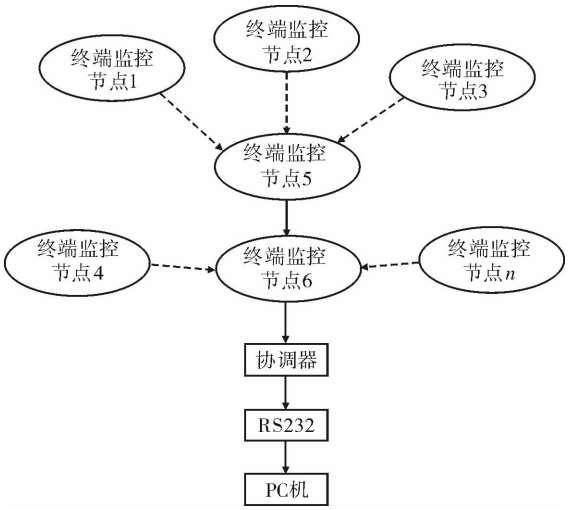


图 1　基于 ZigBee 的热压机无线监控系统结构图

2　系统电路设计

2.1　终端节点总体结构

终端节点主要由热压机控制系统和无线模块 2 部分组成. 该系统为了维持原热压机控制系统的独立性,终端节点设计采用了 2 个独立的微处理器,一个负责热压机自身的控制,另一个归属无线通信模块,两者之间通过 SPI 接口实现数据通信.

2.2　热压机控制系统的设计方法

2.2.1　热压机控制系统的组成原理　热压机控制系统的组成原理如图 2 所示. 该控制系统基于 ARM7,由数据采集模块、微处理模块、LCD 显示模块、键盘模块及控制输出模块组成. 数据采集模块负责信息的采集并将采集的信号(温度、压力等信号)转变为数字信号,传送给微处理模块;微处理模块主要控制整个节点的处理操作、路由协议、功耗管理、任务管理、控制数据显示和 JN5148 通信等[5];输出模块主要实现数模转换、触发和控制电路等功能.

2.2.2　通信电路设计　JN5148 无线模块与热压机控制系统采用 SPI 通信方式,JN5148 采用隔离电源独立供电,数据线间选用 6N137 进行光电隔离,保

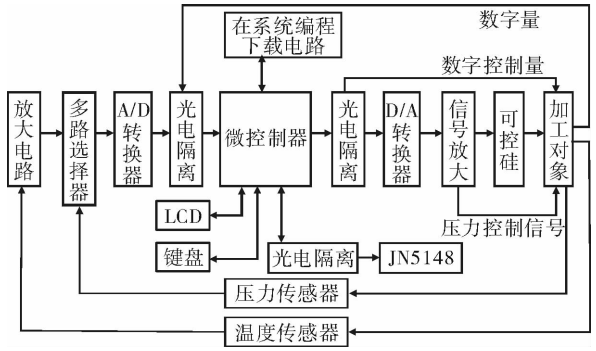


图 2　热压机控制系统结构框图

证了通信系统和控制系统电气上的独立.

2.3 协调器设计

协调器节点负责网络的组织和维护,是主控 PC 机与每个监控节点的数据通道,该节点与 PC 通信采用 RS232 接口实现.

3 系统软件设计

3.1 热压机控制算法设计实现原理及效果

3.1.1 控制算法原理及实现方法 由于热压机的加工对象类型较多,温度变化范围和随机性较大.因此,常用的 PID 控制方法不能满足热压机的这些特殊要求.本文采用基于继电整定的模糊 PID 控制算法来实现系统控制,其系统结构和原理如图 3 所示.首先,可以通过继电整定法得到 PID 参数 K_{P0}, K_{I0}, K_{D0},将这些参数分别作为模糊 PID 的初始比例系数、积分系数和微分系数.然后,根据系统偏差 e 和偏差变化率 e_c,采用高效的模糊推理法,实现系统 PID 参数 K_P, K_I 和 K_D 的实时调整,这种独特的实时操控能力使得 PID 控制器具有较强的自适应能力,从而使系统处于最优状态,并能达到期望的温控效果[3-5].

3.1.2 控制算法实现方法及应用效果 该系统中 PID 控制分温度控制和压力控制 2 个独立的控制算法,温度的继电整定的实现过程是:先满功率加热,当温度达到 600 ℃时,以 600 ℃为设定值,高于设定值加热停止,低于设定值满功率加热,连续进行 3 个周期后,系统可以得到振荡周期和振荡幅度,从而利用 Z-N 公式[5]计算出 PID 的相关参数.由于压力反应速度快,容易造成危险,因此压力继电整定采用限幅整定的方式.

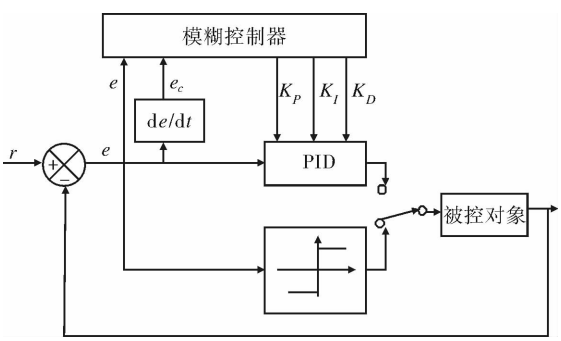


图 3 模糊 PID 控制器算法结构图

模糊 PID 智能控制算法在该系统中的设计实现,解决了成品 PID 仪表操作不便、价格昂贵等缺点,提高了温度和压力的控制精度.在实际运行中,温度控制在对温度要求较高的保温阶段控制误差 <0.3 ℃,升温阶段误差 <0.4 ℃,压力控制在整个工艺过程中误差均 <0.1 kN.

3.2 无线通信系统程序设计

基于 ZigBee 的热压机控制系统的无线通信程序设计主要包括终端节点程序设计和协调器程序设计,这 2 部分的流程图分别如图 4 和图 5 所示.系统的有效运行需要根据通信协议建立有效的通信方式,以保证数据传输的可靠性[6-7].当协调器启动后,进行通信硬件和传感器网络的初始化,形成一个基于 ZigBee 的无线传感器网络.首先,设置各从机为数据接收状态,然后设置协调器为等待状态,等待主控 PC 机发送指令或路由器和节点发送的信息;然后,当协调器节点接收到主控 PC 机的控制指令时,协调器节点需要快速解析这些指令,并发送这些指令到相关的终端节点;当协调器节点接收到节点数据后,迅速对数据包进行解析,并通过数据接口将其发送到主控 PC 机上.

3.3 主控 PC 机的系统程序设计

基于 ZigBee 技术的无线通信数据的发送和接收均由 PC 机控制,由传感器节点中的 ZigBee 通信传输模块对信息进行简单处理后,主控 PC 机请求连接,依次与各个节点建立通信并得到实时数据;建立通信联系后的节点及控制器按照主控 PC 机的命令执行任务[8-9].主控 PC 机得到来自各节点实时数据后,先对这些数据包进行解析,将这些数据保存,每个节点均有历史数据,主控 PC 机也根据这些数据(温度、压力数据)进行及时处理,并对这 2 类数据在坐标的纵横轴上进行描点,可以得到实时的温度和压力曲线.对特定时间内的曲线进行实时分析,对于程序可自动处理的异常信息,则及时发出处理命令;否则,通过弹出框和声音提示,管理员根据实际情况进行决策,判断热压机的实时运行情况,以便设备能正常有效地运行.

4 结论

本文基于 ZigBee 技术对热压机监控系统进行设计,该系统采用基于继电整定的 PID 控制算法,通过 ZigBee 无线网络建立了 PC 机与多台烧结炉的无线通信.利用 JN5148 模块的无线接收发送功能实现了多台热压机的实时无线监控,给生产管理人员带

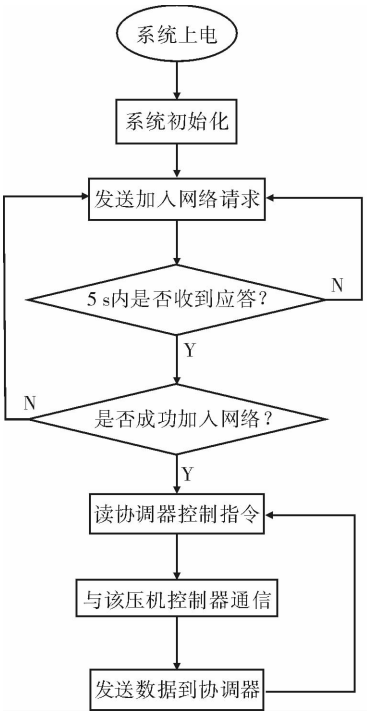


图 4　终端节点程序流程图

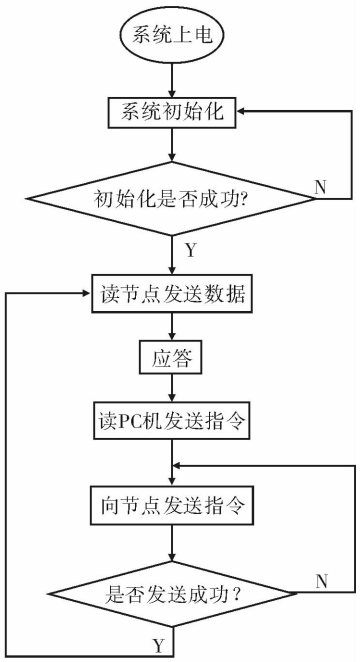


图 5　协调器程序流程图

来很大的方便. 管理者可以通过该系统了解和实时监控每个机器的运行情况,通过该系统同时对每台热压机进行工艺修改和系统设置,保证了产品的合格率;模糊 PID 智能控制算法在该系统中的应用,提高了温度和压力的控制精度,在实际使用中温度控制误差 <0.3 ℃,压力控制误差 <0.1 kN. 实际运行情况表明,该系统运行安全可靠.

参考文献:

[1] 李明,杨承,杨成梧. 基于 PLC 的热压机 PID 控制系统[J]. 林业机械与木工设备,2004,32(12):56.

[2] 张峰. 立式高温真空烧结炉控制系统的设计[J]. 真空,2010,47(2):68.

[3] 李明,杨承,杨成梧. 热压机控制系统的 PID 改进[J]. 控制工程,2006,13(1):45.

[4] 任秀丽. ZigBee 无线通信协议实现技术的研究[J]. 计算机工程与应用,2007,42(6):143.

[5] 王耀南,孙炜. 智能控制理论及应用[M]. 北京:机械工业出版社,2008.

[6] 纪友芳,林美娜. 模糊 PID 复合智能控制参数自整定研究[J]. 微计算机应用,2007,8(28):828.

[7] 昂志敏,金海红,范之国,等. 基于 ZigBee 的无线传感器网络节点的设计与通信实现[J]. 现代电子技术,2006,29(10):47.

[8] 刘瑞强,冯长安,蒋延,等. 基于 ZigBee 的无线传感器网络[J]. 遥测遥控,2006,29(5):57.

[9] 梁万用,张宇翔,胡智宏,等. 基于 NRF401 的烧结炉无线监控系统的设计[J]. 郑州轻工业学院学报:自然科学版,2007,22(4):55.

文章编号:2095-476X(2012)06-0048-03

基于SIP协议的临床呼叫语音网关的设计

陈晓雷, 梁坡, 邓蕾

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450002)

摘要:采用SIP协议,基于ARM9 CPU Mini2440设计了了一个控制临床呼叫信号的小型语音网关.该设计在通话时将模拟信号经过采样、数字化、压缩编码、打包分组、分配路由、存储交换、解压缩等一系列交换处理,在IP网实现语音通信,实现了复杂协议向SIP标准协议的转换,使信号可以在整栋大楼进行传输.系统测试表明,通信语音质量良好,不存在语音延迟.

关键词:SIP协议;临床呼叫语音网关;Mini2440

中图分类号:TP393 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.013

Design of clinical call voice gateway based on SIP protocol

CHEN Xiao-lei, LIANG Po, DENG Lei

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450002, *China*)

Abstract: Based on the ARM9 CPU Mini2440 and using the SIP protocol, a small voice gateway which is controlled the clinical call signaling was presented. The analog signal was processed with sampling, digitization, compression coding, packaging and grouping, distribution route, storage and exchange, unzip etc, the voice communication was realized in the IP network. The conversion from complex protocol to SIP standard protocol was realized in the gateway, the signal can be transmitted in the whole building. The test results showed that the communication voice quality is good and there is no speech delay.

Key words: SIP protocol; clinical call voice gateway; Mini2440

0 引言

随着互联网的快速发展,多媒体网络通信已成为人们关注的热点.通过IP网络进行语音传输是语音网络发展的主导方向,并逐步成为下一代网络(NGN)的主要发展目标之一.随着VoIP(互联网协议电话)技术的发展,与之相关的业务在世界范围内也取得了较快的发展.SIP协议(session initiation protocol)以其简单、灵活的固有优势,正在逐步替代原有的H.232协议,成为IP语音技术的标准协议.SIP的功能扩展性以及网络伸展性较好,为开发各种增值业务和会议呼叫提供了很大的方便.因此,SIP协议近年来得到了很大的关注与发展,而基于SIP协议的语音网关的研究也随之成为热点[1-2].

本文拟提出一种IP电话网关的实现方案以及模块化实现方式,以期更好地利用网络资源,降低语音业务成本,在软交换的控制下实现增值业务,为嵌入式语音网关的发展提供一种思路[3-6].

1 硬件设计及语音信号处理流程

1.1 系统硬件设计

本系统采用基于Samsung S3C2440微处理器的

收稿日期:2012-04-21

基金项目:河南省教育厅自然科学研究计划项目(2009A510015)

作者简介:陈晓雷(1964—),男,河南省郑州市人,郑州轻工业学院副教授,主要研究方向为嵌入式系统及应用、工业控制计算机及其软件开发.

ARM 开发板 Mini2440. 该开发板采用专业稳定的 CPU 内核电源芯片和复位芯片来保证系统运行时的稳定性,具有较低的功耗和高速的处理计算能力,主频为 400 MHz,适合各种控制应用. 其采用的 Samsung S3C2440 微处理器,可在一个芯片上支持通信物理层、协议堆栈、特定设备应用和特定设备的外设软件模块,用户还可对其进行编程和重编程,便于人机之间建立真正的单片式网络应用方案. 基于 Mini2440 的临床呼叫语音网关系统结构如图 1 所示.

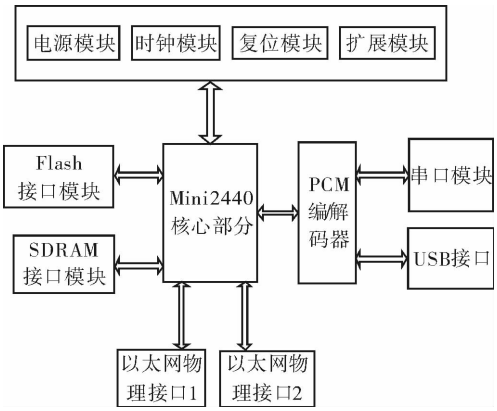


图 1　临床呼叫语音网关系统结构图

1.2　语音信号处理流程

模拟音频信号送入 SLIC 芯片以后被放大,同时进行去干扰的前置处理,然后再送入 CODEC 芯片进行 A/D 转换,并转换为 PCM(脉冲编码调制)编码的音频信号,再进行压缩编码,即转换为 RTP 包格式的有效数据净荷;信号送入 CPU,通过 CPU 上运行的协议栈对有效数据净荷(PayLoad)进行封装、打包;最后通过以太网交换芯片传输到目的网络. 语音网关对以太网收到的语音包则采用与之相反的处理流程. 语音网关内部结构如图 2 所示.

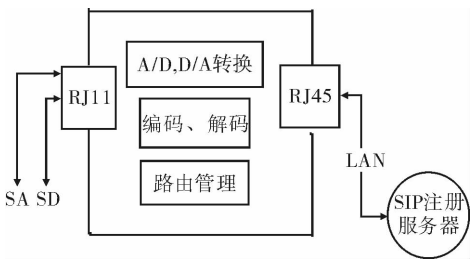


图 2　语音网关内部系统框图

系统使用 SA 线传输控制指令,例如:

1)X 房 X 床呼叫请求时,SA 线上会有类似“01 01 E0 FF FF”的呼叫指令,其中前 2 个十六进制表示床号及房号,第 3 个数字表示控制指令,E0 表示呼叫.

2)护士站或者门口接收到呼叫时,SA 线上会有类似“F0 01 E1 FF FF”的接听指令,其中前 2 个十六进制表示接听的分机号,第 3 个数字表示控制指令,E1 表示接听.

总线上 SD 线的作用是,当 1 房 1 床呼叫,护士站接听时,SD 线上传输模拟语音信号. 当前只能 1 路通话,其他的呼叫为等待状态,系统循环报号. 控制总线上发送呼叫挂断指令时,SD 线上的语音信号停止.

有床头呼叫时,语音网关通过系统总线接收,经过 A/D 转换、编码、解码等,将呼叫转移到 SIP 服务器上的可接收的终端,如 IP 电话、电脑等.

每一个语音网关有一个固定的 IP 地址及编号,此编号实际上为病区编码. 总线上输入的编码是房号 + 床号,语音网关输出的是病区号 + 房号 + 床号. 此外,语音网关还有可配置的界面.

2　语音网关模块软件的设计

语音网关主要完成拨号、DTMF(双频多音信号)传送、呼叫建立、基本会话等功能.

本文采用开放源代码 OSIP 的基本会话功能,作为 SIP 开发库,其允许构建互操作的注册服务器、用户代理(软件电话)和代理服务器. 网关的原设计思想就是实现基本的会话功能,以达到所用器件最少、整机体积最小的目的. OSIP 足够灵活和微小,可以在小的操作系统(如手持设备)上满足其特定要求.

下面以输液呼叫对讲指令为例来说明该语音网关如何实现呼叫指令中的对讲和等待等功能. 具体执行过程见图 3.

2.1　语音网关的呼叫对讲功能

病人按下呼叫按钮时,呼叫转移到语音网关设置的 IP 电话终端上. 具体过程如下:

1)病人在床头或者卫生间按下呼叫按钮;

2)语音网关通过总线收到呼叫信号;

3)语音网关接口通过总线 SD 线解析出当前呼叫的房号及床号,加上语音网关本身的病区号(如 010101 表示 1 病区 1 房 1 床),或者用#号作为分隔符,形成分机号;

4)接口板将总线收到的呼叫指令转换成标准的 O 口,作为语音网关的输入;

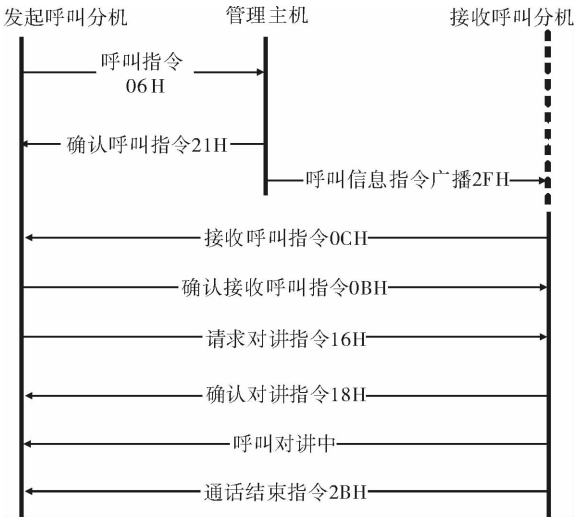


图 3 输液呼叫对讲指令执行过程

5)语音网关根据预先设定的 IP 地址呼叫对应的 IP 电话;

6)IP 电话端听到呼叫铃声后,值班人员接听电话;

7)值班人员与病人进行对讲;

8)对讲结束后,挂断;

9)语音网关向总线发送挂断指令;

10)总线模拟语音线 SA 释放,处于空闲状态.

2.2 语音网关的呼叫等待功能

病人按下呼叫按钮时,如遇 SA 线上正在通话中,该病人的呼叫会处于等待状态. 语音网关完成一次呼叫对讲之后,继续检测 SD 线.

3 系统测试及分析

为验证语音网关的基本呼叫控制及通话功能,构建了基于局域网的测试平台,如图 4 所示. 语音网关登录 WebServer 服务器进行网络配置,再通过语音网关下话机 A 拨打语音网关为话机 B 分配的号码,在拨打过程中使用抓包软件 Sniffer _Por 抓包,分析抓包得出本语音网关 SIP 信令流程符合 RFC3261 规范,双方通话时的 RTP 流正常,通话过程中语音质量良好,也不存在语音延迟.

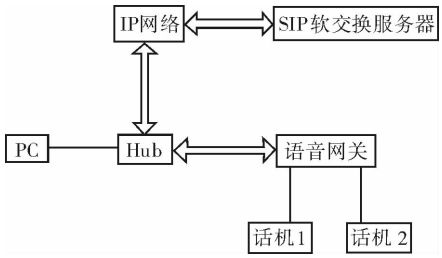


图 4 语音测试环境

4 结论

本文采用 SIP 协议,基于 ARM9 CPU Mini2440,设计了一种嵌入式控制临床呼叫信号的小型语音网关,完成了复杂协议向标准 SIP 协议的转换,实现语音信号的传输和整栋大楼的呼叫功能. 系统能够按照 SIP 协议规范进行通话,并能在以太网上进行语音传输,通话过程中语音质量良好,此外还具有简单、灵活等优点,在功能性和增长潜力方面有明显优势.

参考文献:

[1] 肖峰. 基于 SIP 协议的嵌入式 IP 电话的研究和实现[D]. 北京:北京邮电大学出版社,2006:22 - 34.

[2] RFC3261,SIP:Session Initial Protocal[S].

[3] 肖永军,李海标,杨文,等. 基于 SIP 协议的嵌入式语音网关的设计与实现[J]. 计算机系统应用,2009(8):120.

[4] 曹玖新,张德运. VoIP 实现技术研究[J]. 计算机工程,2000,26(S1):497.

[5] 周海华,边恩炯. 下一代网络——SIP 原理与应用[M]. 北京:机械工业出版社,2006:15 - 25.

[6] 于明,范书瑞,曾祥烨. ARM9 嵌入式系统设计与开发教程[M]. 北京:电子工业出版社,2008:260 - 266.

文章编号:2095-476X(2012)06-0051-04

基于 Android 的手持终端 CoreProcess 系统的设计

韩冰[1], 闫红岩[2]

(1. 北京理工大学 软件学院, 北京 100081;
2. 郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:为了在手持终端系统上实现音/视频的转发、播放、录制等功能,设计了手持终端 CoreProcess 系统. 该系统分为嵌入式客户端和 Windows 服务器端 2 部分,采用三星公司以 S5PC110 为核心芯片的 Android 系统,使用了 V4L2 视频驱动、ALSA 音频驱动等技术. 交付前期的测试验证表明,系统能够稳定工作,在网络传输过程中出现的丢包、码流控制方案不稳等问题也都得到了解决.

关键词:Android;手持终端;CoreProcess
中图分类号:TP393.09 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.014

Design of handheld terminal CoreProcess system based on Android

HAN Bing[1], YAN Hong-yan[2]

(1. *School of Software, Beijing Institute of Technology, Beijing* 100081, *China*;
2. *College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*)

Abstract: To realize the functions of sound/video transmission, playing, recording etc. on the handheld terminal system, the handheld CoreProcess terminal system was designed, which is divided into two parts: embedded client and Windows server. Besides V4L2 video driver, ALSA audio driver, it adopts Android system with the Samsung S5PC110 as the core chip. The test results showed that it can work stably, the problems of losing package, the instability of controlling code stream project and so on are solved in the process of net transmission.

Key words: Android; handheld terminal; CoreProcess

0 引言

随着物联网技术的推广应用,智能视频监控系统越来越多地引起人们的关注. 而智能视频分析技术作为智能视频监控的核心技术,引起了国际上许多著名科研机构以及研究人员的兴趣,尤其在美国、英国等国已经开展了大量相关研究,包括运动检测、基于三维模型的车辆与行人的定位识别和跟踪、基于移动摄像机的视觉监控等. 智能视频技术的研究已经取得初步的成果,但是还处于初级阶段.

音/视频编解码适用于专业的音/视频图像分析和处理,基于音/视频编解码技术的高清高精度采集卡可以应用于工业检测、工业测量、智能交通、显微成像等领域多种图像采集处理分析. 然而目前市场上大部分音/视频编解码板都没有手持终端,且只有一些专用的功能. 鉴于此,本文拟设计基于

收稿日期:2012-09-14
基金项目:国家"十一五"科技支撑计划项目(2006BAK01A38);郑州轻工业学院 2009 年校科研基金项目(No.24)
作者简介:韩冰(1990—),男,河南省新野县人,北京理工大学本科生,主要研究方向为软件工程;闫红岩(1975—),男,河南省三门峡市人,郑州轻工业学院讲师,硕士,主要研究方向为数据库安全、嵌入式系统.

Android 的手持终端 CoreProcess 系统,以实现音/视频的转发、播放、录制等功能,并解决网络传输过程中的丢包、码流控制方案不稳等问题.

1 CoreProcess 系统主要技术

在 CoreProcess 系统中,主要用到了 Android 和 S5PC110 处理器、串口驱动、V4L2 视频驱动、ALSA 音频驱动等技术.

1.1 Android

Android 是以 Linux 为核心,由软件堆迭的架构延伸发展而来的一套软件平台与操作系统,主要用于便携设备. 相比其他手机操作系统,Android 更具开放性,有丰富的硬件选择、开发商不受任何限制、无缝结合的 Google 应用等优势[1].

1.2 S5PC110 处理器

三星公司的 S5PC110 处理器采用主频为 1 GHz 的 45 nm 制程的数据处理芯片,是目前功能最强的移动单核处理器. S5PC110 采用的图形显示芯片是 Power VR SGX540,可以达到9 000万个/s 多边形输出和 10 亿/s 的像素填充率,S5PC110 处理器上集成了多个 Power VR SGX540 芯片.

S5PC110 内部集成了 4 Gb MuxOneNAND,2 Gb OneDRAM 和 1 Gb mDDR,不但减小了硬件电路设计的复杂度,而且低功率技术可确保电池有更长的使用寿命. S5PC110 配备功能超强的内建 3D 图形引擎,能支持复杂的 3D UI 及高显示能力的游戏,整合了1 080 p高画质编码引擎,能支持 30 f/s 的高画质的影片播放与录像. S5PC110 支持以太网口,VGA 接口,液晶触摸屏,TV OUT,TV IN,I2S 和 AC97 接口,4 个 USB HOST,1 个 USB OTG,4 个串口和 3 路 SD/SDIO/MMC 接口,硬件配备完全能够满足音/视频编解码的需要[2].

1.3 串口驱动

在 Linux 系统中,终端是一种字符设备,通常使用 Tty(Teletype 的缩写)来简称各种类型终端设备. 串行端口终端是其中的一类,它是使用计算机串行端口连接的终端设备. 串行端口所对应的设备名称是/dev/ttyS0 (或/dev/tts/0),/dev/ttyS1 (或/dev/tts/1)等.

1.4 V4L2 视频驱动

V4L(Video4Linux 或 Video for Linux)是 Linux 内核中标准的视频驱动程序,目前其版本是 Video4Linux2,简称 V4L2. 在 Linux 系统中,摄像头视频一般规范到了使用 V4L2 驱动程序. V4L2 可以支持多种设备,可以有多种接口. V4L2 驱动的 Video 设备可以支持捕获及视频输出方式,通常使用它作为摄像头的驱动程序. V4L2 驱动的 Video 设备在用户空间通过各种 ioctl 调用进行控制,并且可以使用 mmap 进行内存映射[3].

1.5 ALSA 音频驱动

高级 Linux 声音体系 ALSA(advanced linux sound architecture)是为音频系统提供驱动的 Linux 内核组件,以替代原先的开放声音系统 OSS(open sound system).

ALSA 是一个完全开放源代码的音频驱动程序集,除了像 OSS 那样提供一组内核驱动程序模块之处,还专门为简化应用程序的编写提供了相应的函数库,与 OSS 提供的基于 ioctl 等原始编程接口相比,使用 ALSA 函数库要更加方便一些. 利用该函数库,开发人员可以方便、快捷地开发出自己的应用程序,细节则留给函数库进行内部处理.

ASOC(ALSA system on chip)是 ALSA 在 SOC 方面的发展,对 CPU 和 Codec 的相关代码进行了分离. 本文对于嵌入式系统的声卡驱动开发,采用 ASOC 框架,主要由 Codec 驱动、平台驱动、板驱动 3 部分组成. 前 2 部分是通用的驱动,只有板驱动是不通用的,它由特定电路板上具体的 CPU 和 Codec 确定[4].

2 系统架构设计

CoreProcess 系统分为嵌入式客户端和 Windows 服务器端 2 部分. 客户端是 S5PC110 为核心芯片的 Android 系统,可将模拟视频转化为数字信号,经过 H.264 main profile 编码,再使用 RTP/RTCP 协议传输到网络中;客户端还可以接收别的客户端或者 Windows 端传送过来的 RTP/RTCP 码率,对音/视频码流进行解码显示. 在服务器端,系统提供了登录一个 Android 终端和登录多个 Android 终端、音/视频转发、音/视频的本地播放、录制文件、播放文件等大部分 API 的 demo 实例,用户可以参考这些 demo 实例,研发出符合自己需求的 Windows 端产品. CoreProcess 系统硬件使用 S5PC110 CPU,具有硬编和硬解的功能,系统硬件架构设计如图 1 所示.

CoreProcess 系统的软件设计架构如图 2 所示,CoreProcess 底层是一个 Linux 系统,应用程序通过 JNI 对硬件进行控制.

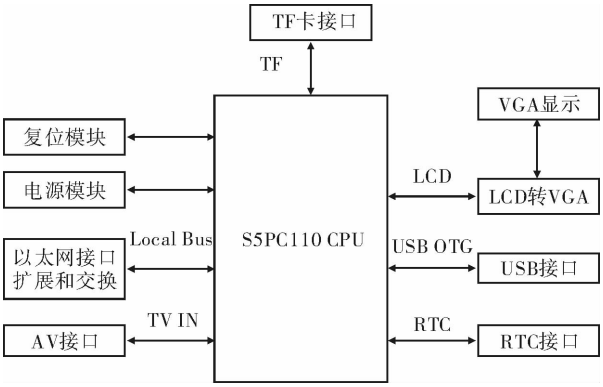


图 1　CoreProcess 系统硬件架构图

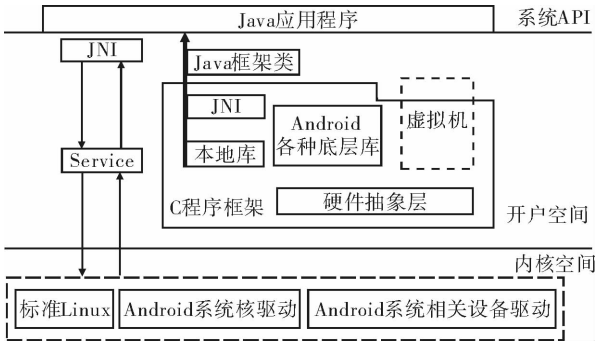


图 2　CoreProcess 系统软件架构图

3　系统的移植编程设计

CoreProcess 系统的移植编程设计主要包括串口移植、VGA 驱动、板移植和平台移植.

3.1　串口移植

在. config 文件中添加 CONFIG_CMDLINE = " console = ttySAC0,115200 init = /linuxrc"[5].

3.2　VGA 驱动

在文件 drivers/video/samsung/s3cfb_lte480wv. c 中修改 s3cfb_lcd lte480wv 结构[6].

```
static struct s3cfb_lcd lte480wv =
{. width = 800, . height = 600, . bpp = 24, . freq = 60,
    . timing = {. h_fp = 40, . h_bp = 470, . h_sw = 110, . v_fp = 2, . v_fpe = 1, . v_bp = 18, . v_bpe = 1,
    . v_sw = 3 },
    . polarity = {. rise_vclk = 1, . inv_hsync = 0, . inv_vsync = 0, . inv_vden = 0}
};
```

3.3　板移植

针对板卡 saa7113 的移植,修改 saa7113_init 函数:

```
static int saa7113_init(struct v4l2_subdev * sd, u32 val)
{ struct i2c_client * client = v4l2_get_subdevdata(sd);
    unsigned char data8;
    unsigned char ID = 9;
    int i = 1,k;
    u32 pup;
    reg_con = ioremap(T_GPD1_CON,24);
    t_gpio_init(reg_con);
    v4l_info(client, "% s: camera initialization start \n", __func__);
    for(k = 0;k < SAA7113_INIT_REGS;k + + )
    { I2C_Start();
    I2C_reg_Write(0x4a, saa7113_init_reg[k][0], saa7113_init_reg[k][1]);
    i = I2C_reg_Read(0x4a,saa7113_init_reg[k][0], &ID);
    }
    i2c_deinit();
    return 0;
}
```

3.4　平台移植

对于音/频驱动要有 Codec 和 CPU 的平台驱动,本系统的平台板卡选用 uda1341,对平台驱动做如下修改.

1)在文件 sound/soc/s3c24xx_wm8580slv. c 中修改函数 smdk64xx_hw_params.

```
static int smdk64xx_hw_params(struct snd_pcm_substream * substream,struct snd_pcm_hw_params * params)
{...
epll_out_rate = rclk * psr; ret = set_epll_rate(epll_out_rate);
if (ret < 0)return ret;
ret = snd_soc_dai_set_sysclk(cpu_dai, S3C64XX_CLKSRC_CDCLK, 0, SND_SOC_CLOCK_OUT);
if (ret < 0)return ret;
ret = snd_soc_dai_set_sysclk(cpu_dai, S3C64XX_CLKSRC_MUX, 0, SND_SOC_CLOCK_IN);
if (ret < 0)return ret;
ret = snd_soc_dai_set_fmt(&s3c64xx_i2s_dai[I2S_NUM], SND_SOC_DAIFMT_I2S
|SND_SOC_DAIFMT_NB_NF | SND_SOC_DAIFMT_CBS_CFS);
if (ret < 0)return ret;
ret = snd_soc_dai_set_clkdiv(cpu_dai, S3C_I2SV2_DIV_PRESCALER,psr - 1);
```

```
if (ret < 0)return ret;
ret = snd_soc_dai_set_clkdiv(cpu_dai, S3C_I2SV2_DIV_BCLK, bfs);
if (ret < 0)return ret;
ret = snd_soc_dai_set_clkdiv(cpu_dai, S3C_I2SV2_DIV_RCLK, rfs);
if (ret < 0)return ret;
ret = snd_soc_dai_set_sysclk(codec_dai, 0, rclk, SND_SOC_CLOCK_OUT);
if (ret < 0) return ret;
return 0;
}
```

2)在 sound/soc/codecs/l3.c 中添加.

```
static unsigned int read_l (u32 ptr)
{return *((volatile unsigned int *)ptr); }
static void write_l(u32 value, u32 ptr)
{ *((volatile unsigned int *)ptr) = value; }
void setclk(u8 value)
{u32 pup; pup = read_l(reg_dat); pup &= 0xfffffffd;if (value) pup |=2; write_l(pup,reg_dat); }
static void sendbytes(struct l3_pins *adap, const u8 *buf, int len)
{int i;
for (i = 0; i < len; i++)
{if (i) {udelay(adap->mode_hold);setmode(0);udelay(adap->mode);}
setmode(1); udelay(adap->mode_setup); t_sendbyte(buf[i]);}
}
int l3_write(struct l3_pins *adap, u8 addr, u8 *data, int len)
{setclk(1);setdat(1);setmode(1); udelay(adap->mode);setmode(0);
udelay(adap->mode_setup); t_sendbyte(addr); udelay(adap->mode_hold);
sendbytes(adap, data, len); setclk(1); setdat(1); setmode(0); return len;
}
```

4 测试验证

在测试阶段,对动态壁纸、SD 卡、图片查看器、媒体播放器、音乐播放器、有线网卡/无线网卡上网来检测音/视频的连接、解码、播放、录像等功能进行了验证. 通过对 VC 客户端进行的模块调试用例测试,测试结果如下:VC 端播放音/视频 5 路;输入码流为 800 000 b/s;CPU 占用率 8%;内存使用 153 208 K;带宽 402 kb·s^{-1}/1 Gb·s^{-1};实际收到的码流为 805 000 b/s;音/视频质量良好;延迟时间为音频 500 ms,视频 2 s;稳定运行时间 3 h;循环切换视频稳定,为 6 次/s.

在 30 min 内把环境温度提升到 70 ℃,对 2 个 Android 终端在 2 h 内进行压力测试. 结果表明:在测试环境中,音频能够正常播放和使用,视频也能够正常进行编解码和播放. 后期,又进一步把 Recorder 做成服务、RecorderSeting(完成 Recorder 部分参数设置)、更正 Recorder bug、增加 OSD 参数设置功能等;同时,分别解压 app-jni.rar 到 NDK/samples 目录、解压 SerialPort.rar 到 NDK/samples 目录,并进行了编译,实现了 JNI 层的底层操作.

5 结论

本文设计了一个基于 Android 平台的嵌入式手持终端 CoreProcess 系统,它集成了音/视频编解码卡,采用 S5PC110 CPU,进行了音/视频驱动的移植设计. 音频驱动使用了 ALSA 架构,视频驱动使用了 V4L2 架构;此外,还做了 LED 灯、VGA 等驱动,设计了一个 Android 应用程序,对 Android 底层类库进行了修改. 该系统已经应用在了某部队的夜视镜上,到目前为止,CoreProcess 系统能够进行长时间稳定的音/视频编解码,同时,在网络传输过程中出现的丢包、码流控制方案不稳等问题也都得到了解决.

参考文献:

[1] 郭宏志. Android 应用开发详解[M]. 北京:电子工业出版社,2010:40-69.

[2] 韦东山. 嵌入式 Linux 应用开发完全手册[M]. 2 版. 北京:人民邮电出版社,2008:50-111.

[3] 宋宝华. Linux 设备驱动开发详解[M]. 北京:人民邮电出版社,2008:21-92.

[4] 张石锐,郑文刚,申长军,等. 嵌入式手持无线农产品价格信息采集终端[J]. 计算机工程与设计,2012,33(2):514.

[5] 孙弘羿. 基于智能移动终端的数据监控系统的解决方案[J]. 软件,2012,33(2):61.

[6] 宋立波. 嵌入式手持智能导航终端摄像头模块设计与实现[D]. 武汉:华中科技大学,2011.

文章编号:2095-476X(2012)06-0055-04

欠费信息语音自动呼叫和自助查询系统的设计与实现

姚妮, 张林林, 朱付保

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:针对部分高校学生欠缴学费信息管理耗时费力、效率低下的问题,设计了一套欠费信息自动呼叫和自助查询系统. 系统底层硬件采用SHT-4B/USB语音卡,以SQL Server为数据库服务器,VS2010为开发环境,C#为开发语言. 试验结果证明,该系统具有良好的灵活性和扩展性,满足了学校对欠费信息管理的需求.

关键词:欠费信息;自动呼叫;自助查询;呼叫策略

中图分类号:TP29 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.015

Design and implementation of debt information speech automatic calling and self-help querying system

YAO Ni, ZHANG Lin-lin, ZHU Fu-bao

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*)

Abstract: Aiming at the information management problem of the college students' debt of tuition with time-consuming and bad performing, a speech automatic calling and self-help querying system was designed. The underlying hardware of the system uses the SHT-4B/USB voice card, the database server uses the SQL Server, and the development environment uses the VS2010, and the development language use the C#. The experiment results showed that the proposed system has good flexibility and expansibility to meet the needs of the management of the arrears in tuition.

Key words: debt information; automatic calling; self-help querying; calling strategy

0 引言

由于各种原因,各高校每年都会出现部分学生欠缴学费的现象. 为了减轻学校财务人员统计欠缴学费信息、催缴学费的工作量,及时通知学生补缴学费,以便学生能够及时了解自己欠缴学费的情况,利用语音合成技术开发学费欠缴信息语音自动呼叫与自助查询系统非常必要.

目前,用于语音处理的一种重要技术手段就是电话语音卡[1],它是一种用于计算机并能够实现语音处理的插件,简称语音卡. 语音卡通过计算机与电信网相连,提供录音、放音、收码(DTMF码、PULSE码)、自动拨号、振铃检测与控制摘挂机、信令检测、转接内线、监控录音、传真、数据传输、主叫

收稿日期:2012-09-03

基金项目:河南省科技厅科技攻关项目(122102210492)

作者简介:姚妮(1978—),女,土家族,湖南省张家界市人,郑州轻工业学院助理实验师,硕士,主要研究方向为智能信息处理和地理信息系统.

号侦测等服务功能[2-3].语音卡近几年的发展很快,其应用领域从最初的证券委托,逐步拓展到邮电通信、办公自动化、金融、公安、医疗、商业等领域[4-5].

目前大多数高等学校的学费管理仍处于纸质材料或计算机电子文档阶段[6],通过人工打电话或发短信通知欠费对象及时缴费,这种管理方式耗时费力、效率低下.本文拟针对高校学费管理的问题,应用成熟的语音板卡技术和电话网,设计学生欠费信息语音系统,实现学生欠费信息管理的智能化与自动化,减轻催缴学费的工作量,提高学费管理的效率.

1 系统分析

1.1 系统结构

本系统的欠费对象可以是学生或者家长,欠费对象可以向系统发出呼叫,查询欠费情况.语音系统通过与驱动程序的交互,完成语音卡与用户之间的信息互动,从而实现语音自动呼叫和语音自助查询功能,系统逻辑结构如图1所示.

图1 系统业务流程

为了使用户登录之后能够正常地使用其他功能,在用户成功登录之后,语音自动呼叫功能和语音自助查询功能将分别以2个线程启动运行,这样,用户在管理数据的时候不影响系统的其他呼叫和查询处理.图2所示为系统的体系结构,其中,财务人员把欠费人员的名单通过系统提供的接口导入系统,教辅人员把欠费对象的联系方式也导入系统,语音系统将按照呼叫策略对欠费对象进行自动语音呼叫,并自动应答欠费对象拨打的欠费信息查询请求.

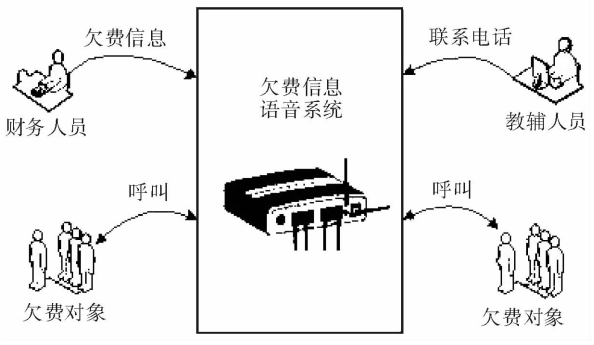


图2 系统体系结构

1.2 语音自动呼叫

系统向一个被叫人发起语音呼叫时,使用的通道必须是空闲的.在通道空闲的情况下,向被呼叫对象发起语音呼叫的过程如图3所示.其中,最理想的情况是摘机→拨号→对方摘机→通话→挂机.但是,数据库中存储的数据有可能是错误的,如下几种呼叫情况也必须考虑到,并能够进行相应的处理.

1)对方空号:在用户输入数据的时候可能会出现手误,也可能提交的电话号码数据本身就是错误的,在这些情况下,存在数据库中的电话号码可能就是空号,因此必须对号码是空号的情况进行处理.

2)呼叫超时:在号码不是空号的情况下,对方电话能够接到此次电话呼叫,但是,对方可能没有及时接听,就会出现呼叫超时的情况,此时应挂断电话,开始呼叫下个学生.

3)挂断和停机:在对方接起电话后,系统开始播放自我介绍和学生欠缴学费情况的语音.如果此时被呼叫对象先于语音卡挂断电话,或者被呼叫对象在振铃后直接挂机,或者被呼叫对象电话停机,这种情况必须做相应的处理.

1.3 语音自动接听

为了让学生能够知道自己是否欠缴学费,了解自己欠缴学费的情况,系统需实现语音自助查询功能.系统检测到有外部电话打进来时,进行摘机操作,然后播放欢迎语音,提示输入要查询的学号,系统通过DTMF得到对方输入的学号后,再从数据库查找,此时可能会出现以下3种情况:

1)对方输入的学号不存在,系统提示对方学号不存在,重新输入学号或者挂断电话;

2)对方输入的学号存在,但是没有欠费信息,系统提示其没有欠费情况,重新输入学号或者挂断电话;

3)对方输入的学号存在,并且有欠费信息,系统将其欠费情况语音播放给对方之后,提示其重新输入学号或者挂断电话.

在外部打电话进来时,需要等待对方输入学号,这时系统会进入不断检测对方是否输入学号完

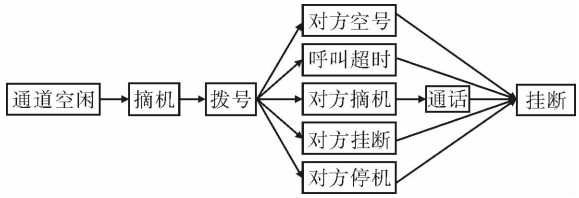


图3 语音自动呼叫过程

毕的死循环,同时对方可能会挂断电话,如果不注意检测对方挂断电话的情况,系统将会陷入死循环.

2　系统设计

2.1　系统功能结构

欠费信息语音系统主要由数据信息管理功能和语音功能 2 部分组成. 数据信息管理功能主要完成部门、班级、欠费信息等管理,语音功能主要实现呼叫时段设置、语音自动呼叫、语音自助查询以及语音呼叫自动记录. 系统功能结构如图 4 所示.

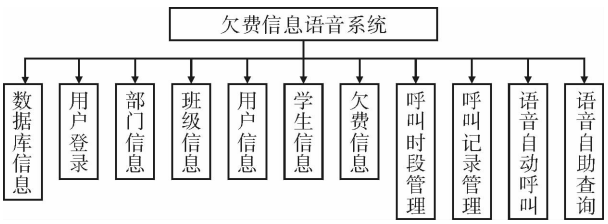


图 4　系统的功能结构图

系统中设计了超级管理员、财务处用户和部门用户 3 种用户类别. 超级管理员能够对部门、部门管理员、班级、学生、欠费信息和呼叫信息进行管理;财务处用户能够管理除用户信息之外的数据信息;部门管理员通过 Web 系统对本部门的学生信息进行采集与处理.

语音自动呼叫模块根据欠费对象的联系电话等相关信息和管理员设置的呼叫策略,自动地周期性地提醒欠费对象,并根据欠费对象联系电话是否可用,是否存在占线、停机、直接挂断等情况,做出相应的处理.

语音自助查询模块根据事先录制的音频播放语音提示,并根据呼叫者通过话机输入的身份识别信息(如学号)和查询选项(如欠费),自动从系统中检索该呼叫者的相关信息,并通过合成的语音文件播放给呼叫者.

2.2　系统业务流程

1)语音自动呼叫流程. 该流程是通过一个线程启动的,只要系统已启动,该功能将一直处于工作状态. 管理员要预先设置呼叫的时段和频率,系统则根据当前时间判断是否发起呼叫. 在不能发起呼叫的时段,线程将处于短时间的休眠状态,唤醒之后再次判断能否发起呼叫. 如此循环,直到能够发起呼叫,整个呼叫过程如图 5 所示.

2)语音自动接听流程. 该流程用于实现语音自助查询. 与语音自动呼叫一样,自助查询也将通过

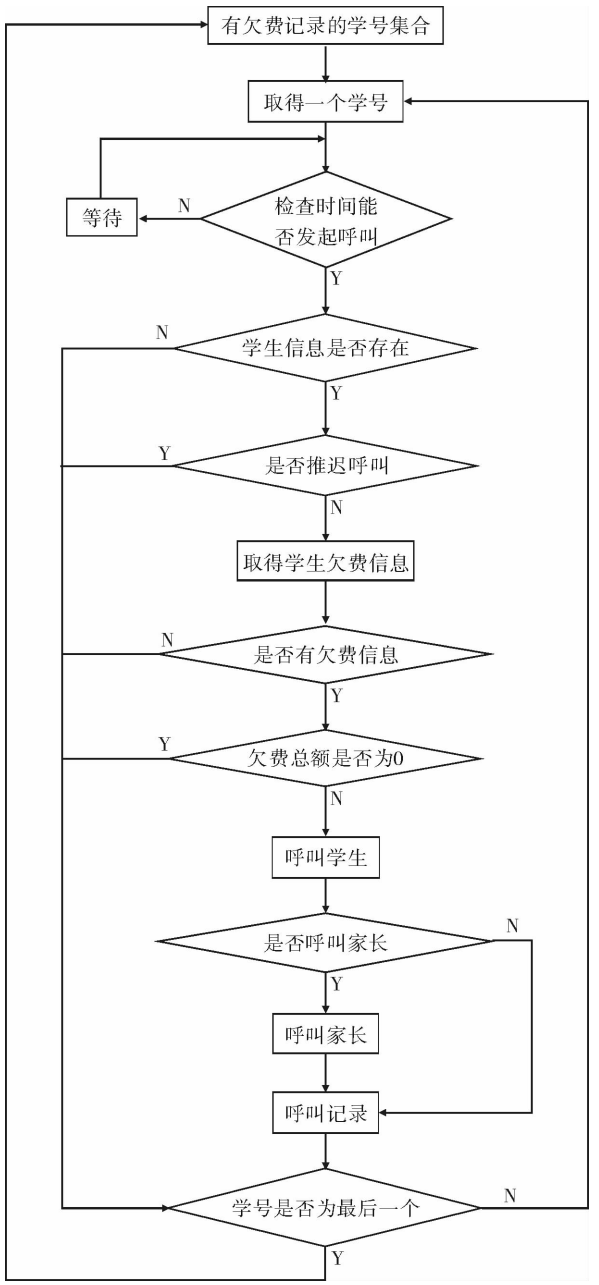


图 5　自动呼叫流程

一个线程的方式启动. 该线程是一个守护线程,它不断检测是否有电话呼叫进来,检测到有电话打进来,进行摘机操作,提示输入要查询的学号,通知呼叫人查询结果,具体流程如图 6 所示.

3)呼叫设置. 呼叫设置用于实现呼叫频率和呼叫时段的管理. 为了减少呼叫频率,不影响学生上课和休息,用户可以设置语音自动呼叫的频率和呼叫时段. 呼叫频率确定语音自动呼叫每隔多少天才能发起,设置信息保存在系统的配置文件中;呼叫时段确定星期几的哪几个时间段可以用来呼叫,时段数据保存在系统数据库中.

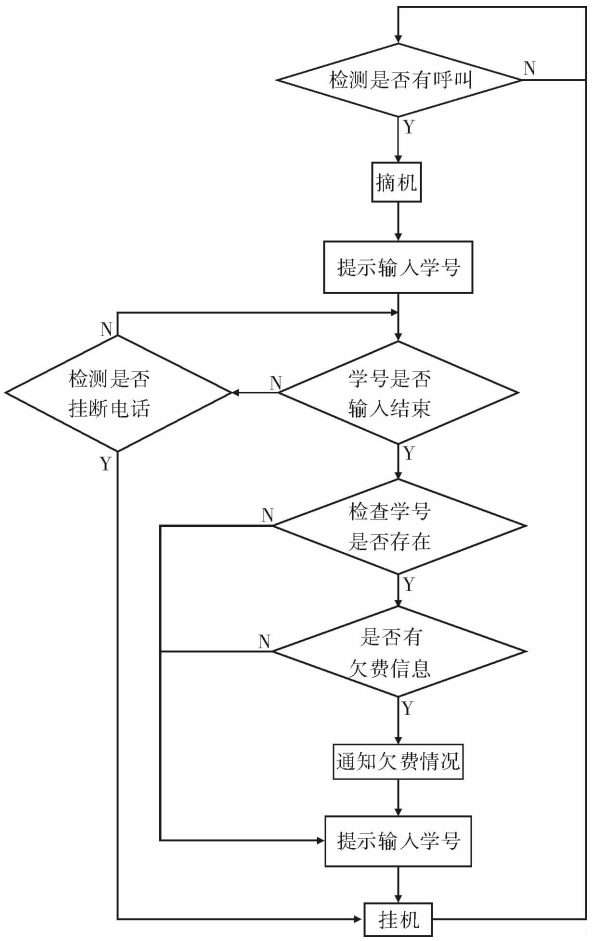


图 6 自动接听流程

3 系统实现

3.1 平台环境

本系统对硬件环境要求不高,只需 1 台普通 PC 和 1 部语音卡. 系统采用的是三汇语音卡 SHT-4B/USB,包括 LED 指示灯、馈电和铃流电源接口、LINE 通道接口、语音输入和输出接口、电话机手柄接口以及与 PC 连接的 USB 接口. 在软件环境方面,则需要安装 Windows 操作系统、. NET 框架、SQL Server 数据库管理系统,以及语音卡驱动程序 SYNWAY_PCI(USB)_5310_CN.

3.2 自动呼叫与自助查询的实现

语音卡的外线通道有挂机、等待、放音、接受 DTMF 码、检索数据、拨号 6 种状态. 初始默认值为挂机状态,当用户摘机拨打电话后,主程序会检测到语音卡的振铃信号,将该通道的信号量设置为放音状态,在下一个时间片中程序转入执行放音事件代码,向用户播放事先准备好或者程序组合生成的语音. 系统运行时若用户启用电话自动通知功能,那么通道的信号量直接设置为拨号状态,待接收方摘机后,信号量便设置为放音状态,将要通知的信息通过语音片告知接收方.

3.3 系统性能分析

本文实现的系统中,语音自动呼叫和语音自助查询功能是 2 个独立的线程,在测试系统性能的过程中发现,系统占用的内存不多,但由于 2 个线程一直在循环检测欠费对象的信息,因此占用 CPU 的资源较多.

系统中的语音自动呼叫功能,经常会读取数据库中的数据,为了减少系统读数据的时间,建议将系统和数据库安装在同一个 PC 机上. SQL Server 数据库服务器占用内存较多,但是占用 CPU 较少,因此两者在一台 PC 机上运行不会对系统的性能造成影响.

4 结论

本文设计了基于语音卡的欠费信息自动呼叫和自助查询系统,给出了系统各模块的功能及主要业务流程. 该系统是在 Microsoft Visual Studio 2010 环境下开发,数据库服务器为 SQL Server 2005,采用 C/S 架构完成开发. 为了方便使用多条电话线部署系统,将电话线的编号信息写入配置文件,只需修改配置文件即可实现多路电话线分别完成语音自动呼叫和语音自助查询的功能. 试验表明本系统具有良好的容错性和可扩展性,节约了人工催缴欠费的时间,提高了高校学费管理工作的效率.

参考文献:

[1] 程铃钫. 基于语音卡的话费自动催缴与查询系统[J]. 机电技术,2010,33(1):32.

[2] 陈超. 基于语音卡的费用催缴呼叫中心系统的设计与实现[D]. 成都:电子科技大学,2011.

[3] 来洪孝,崔颖安,崔杜武. 基于语音卡的呼叫中心通用架构[J]. 计算机工程, 2007, 33(22):283.

[4] 刘卫涛. 基于 CTI 板卡的电信语音增值业务平台的设计与实现[D]. 南宁:广西师范大学, 2009.

[5] 雷国平,谭泽富. 基于语音卡的呼叫中心在乡镇电子政务中的应用[J]. 西安邮电学院学报,2009, 14(1):12.

[6] 汪家常,徐昶,季小明,等. 基于工作流的高校学费管理系统研究[J]. 计算机应用与软件,2012,29(6):294.

文章编号:2095-476X(2012)06-0059-03

基于SIP的嵌入式手持终端的设计与实现

邓蕾, 陈晓雷, 梁坡

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:将嵌入式技术与SIP技术相结合,采用模块化的设计思想,实现了基于S3C2440的嵌入式手持终端的设计.仿真结果表明,该设计在Linux系统中实现了SIP协议栈的移植,手持终端运行良好,符合SIP通话流程,满足设计需求.

关键词:SIP;嵌入式手持终端;Linux

中图分类号:TP368.1 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.016

Design and implementation of embedded handheld terminal based on SIP

DENG Lei, CHEN Xiao-lei, LIANG Po

(*School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*)

Abstract: Combining embedded technology with SIP technology, using the modularizing design thought, the design of embedded handheld terminal based on S3C2440 was realized. The simulation results showed that the design realized transplantation of SIP protocol stack in the Linux system, the handheld terminal operated well, conformed to the SIP call process and meet the design demand.

Key words: SIP; embedded handheld terminal; Linux

0 引言

由于全球互联网技术的迅速发展,Internet技术在多媒体通信业中的竞争日趋激烈,SIP协议的提出与发展使其逐渐替代了传统的H.323协议,以满足人们对各种新业务的需求.SIP协议是目前VOIP系统中运用最广泛的信令控制协议,它能够保证通话的正常实现及语音质量,占据了VOIP系统的核心地位.

随着嵌入式技术的发展[1],集移动通信、嵌入式系统、网络技术为一体的智能化通信终端使通信服务在任何时间、任何地点均成为可能,这是未来嵌入式系统的重要应用,因此嵌入式VOIP终端的研究具有广阔的发展前景.鉴于此,本文拟基于SIP进行嵌入式手持终端的设计与实现.

1 SIP协议及系统构架

1.1 SIP协议的信令机制

SIP[2]协议用于建立、修改和终结一方或者多方会话,会话可以是Internet多媒体会议、IP电话呼叫、多媒体发布、即时消息、在线游戏等.SIP协议因其简单易用且对于新型的应用具有较好的适应性,目前被广泛采纳用作各种VOIP系统的控制信令协议.

SIP协议基于文本格式,采用ISO10646字符集的UTF-8字符集格式进行编码.SIP消息被分为SIP请求和响应,由1个起始行,1个或者多个消息头和可选的消息体组成.2种消息的不同之处在于SIP请求的开始行是1个请求行(request-line),SIP响应的开始行是1个状态行(status-line)[3].

收稿日期:2012-10-14

作者简介:邓蕾(1987—),女,河南省洛阳市人,郑州轻工业学院硕士研究生,主要研究方向为嵌入式系统及应用.

1.2 SIP 的网络元素

SIP 采用客户机/服务器(C/S)的工作方式,包含 2 类组件,即用户代理(user Agent)和网络服务器(network server)[4].

1)用户代理(UA). 用户代理是发起或响应 SIP 事务处理的逻辑功能,它包括 2 部分,即用户代理客户端(UAC)和用户代理服务器(UAS),前者产生请求,后者产生对应的响应.

2)网络服务器. 网络服务器主要实现用户定位与域名解析,主要包括代理服务器、重定向服务器和注册服务器.

2 系统总体结构设计

嵌入式手持终端是运行在嵌入式操作系统的用户代理,主要完成 SIP 电话的注册、呼叫、接听、挂断等功能,与用户直接交互,接收输入信息,检验消息语法,根据用户的操作初始化 SIP 消息,与此同时获得本地的语音接口信息,完成信息的封装与传输[5].

本手持终端基于 SIP 协议,直接或间接地连接到网络上进行通信,且各个终端需要注册到相应的服务器上,以使终端之间进行相互识别. 本终端采用 Mini2440 开发板;CPU 处理器为三星公司的 ARM S3C2440A,主频 400 MHz ,最高可达533 MHz;Flash 采用三星公司的 NAND Flash 芯片. 移动终端通话过程中所传输的数据是通过网络传输的,需要配置以太网接口,为了提高移动通信的方便性,也可以通过无线 WiFi 网络进行传输. 该终端的系统总体结构设计如图 1 所示.

3 软件系统设计

本系统采用嵌入式 Linux[6] 作为操作系统平台,实现的主要功能为:SIP 协议栈功能、作为 SIP 网络的用户代理、音频数据的采集与回放及其编解码处理、音频数据的有线传输. 根据这些功能,将软件设计进行模块化划分,其组成如图 2 所示.

1)主控程序模块主要提取用户输入的信息进行会话控制,提取 RTP 相关信息及会话地址,配置通话相关参数,并将 SIP 协议模块、媒体处理模块反馈的信息传递给用户界面.

2)用户界面模块用来设置相关的配置信息,并提供键盘输入、信息查看等功能,是用户与系统交互的平台. 本模块采用 MiniGUI 作为用户图形界面支持系统.

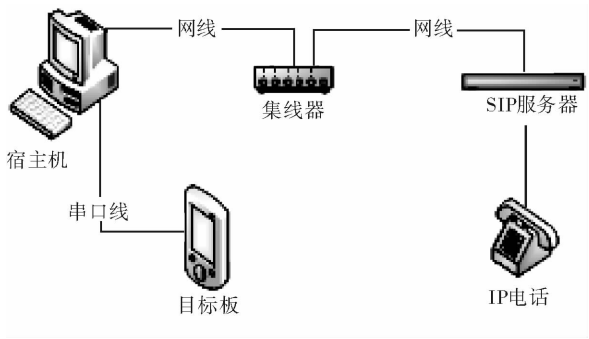


图 1 终端系统结构图

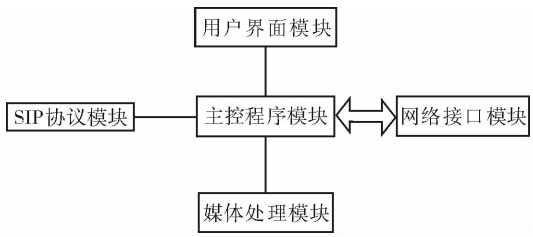


图 2 系统组成模块

3)媒体处理模块[7] 主要采用多线程,音频数据采集线程将从麦克风采集到的音频数据放入原始音频数据队列中;音频编码线程将音频数据从队列中取出并进行编码,并将编码之后的数据放入编码音频数据队列中;RTP 线程将压缩后的音频数据发送给目标用户.

4)SIP 协议模块主要完成发送线程和接收线程. 发送线程从消息队列中读取用户控制命令,并根据 SIP 协议栈将之转化为 SIP 消息后发出;接收线程则循环监听 SIP 消息并进行接收,然后将消息写入消息队列中,供主控程序模块读取并做出响应.

软件工作流程如图 3 所示.

本系统中采用 OSIP2 和 eXosip2 组合的方式实现 SIP 协议栈,采用 eXosip_register_build_initial_register 实现用户注册,采用 eXosip_mes-sage_build_request 实现即时消息发送,采用 eXosip_call_build_initial_invite 实现会话发起,采用 eXosip_call_build_answer 实现会话接收,采用 eXosip_call_terminate 实现会话终结或拒绝.

网络模块实现基本的信息交互,SIP 终端的信令信息和媒体信息均采用 UDP 传输,采用数据报套接字进行网络编程.

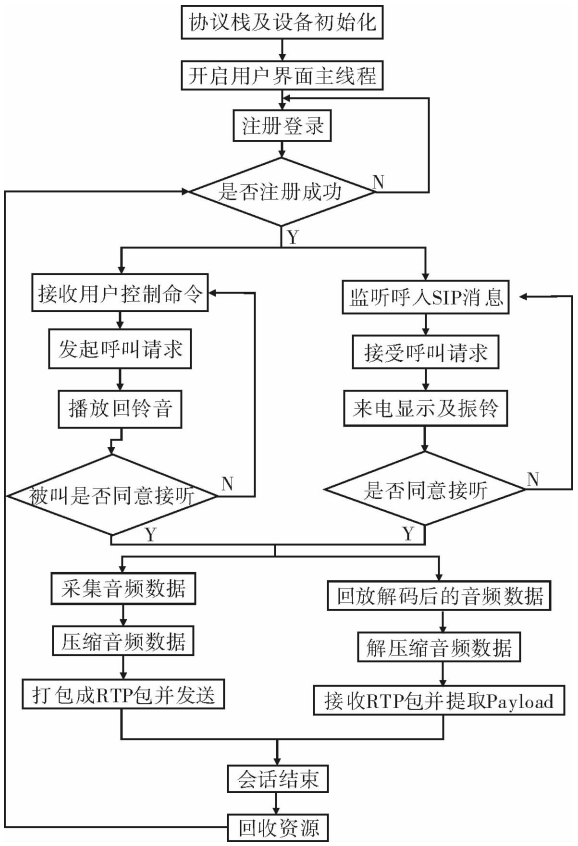


图 3　软件工作流程

4　实验及分析

为了验证系统通话成功与否,本文采用 Ethereal 软件对网络通话过程进行抓包分析,如图 4 所示. 对所抓呼叫流程的数据包的报头协议分析如图 5 所示.

通过对系统进行测试并分析可知,采用本文的设计方式可以实现简单的语音通信功能,较好地完成用户注册、会话控制等功能,性能稳定,但在语音质量及用户界面设计方面还存在着一些不足,需进一步改进.

5　结论

本文将嵌入式技术与 SIP 技术相结合,采用模块化的设计思想,实现了基于 S3C2440 的嵌入式手持终端的设计. 该手持终端系统与传统的 IP 电话相比,有许多优点: 1)采用 Linux 操作系统,可剪裁,占用内存小,功能强大,运行稳定,系统健壮; 2)采用 SIP 协议,具有较强的灵活性及可扩展性,利于业务的增添; 3)将语音的采集、播放、编解码及实时传输并行处理,保证了通话的连续性,具有推广价值.

Source	Destination	Protocol	Info
192.168.85.203	192.168.85.90	SIP/SD	Request: INVITE sip:0371@192.168.85.90:6060,
192.168.85.90	192.168.85.203	SIP	Status: 407 Proxy Authentication Required
192.168.85.203	192.168.85.90	SIP	Request: ACK sip:0371@192.168.85.90:6060
192.168.85.203	192.168.85.90	SIP/SD	Request: INVITE sip:0371@192.168.85.90:6060,
192.168.85.90	192.168.85.203	SIP	Status: 100 Trying
192.168.85.90	192.168.85.202	SIP/SD	Request: INVITE sip:0371@192.168.85.202:6060
192.168.85.90	192.168.85.203	SIP	Status: 180 Ringing
192.168.85.202	192.168.85.90	SIP	Status: 100 Trying
192.168.85.202	192.168.85.90	SIP	Status: 180 Ringing
192.168.85.202	192.168.85.90	SIP/SD	Status: 200 OK, with session description
192.168.85.90	192.168.85.202	SIP	Request: ACK sip:0371@192.168.85.202:6060
192.168.85.90	192.168.85.203	SIP/SD	Status: 200 OK, with session description
192.168.85.203	192.168.85.90	SIP	Request: ACK sip:0371@192.168.85.90:6060
192.168.85.203	192.168.85.90	SIP	Request: BYE sip:0371@192.168.85.90:6060
192.168.85.90	192.168.85.203	SIP	Status: 200 OK
192.168.85.90	192.168.85.202	SIP	Request: BYE sip:0371@192.168.85.202:6060
192.168.85.202	192.168.85.90	SIP	Status: 200 OK

图 4　呼叫抓包过程

```
⊟ Session Initiation Protocol
 ⊟ Request-Line: INVITE sip:0371@192.168.85.90:6060 SIP/2.0
    Method: INVITE
    [Resent Packet: False]
 ⊟ Message Header
    Via: SIP/2.0/UDP 192.168.85.203:6060;branch=z9hG4bKR245DZJeIS
  ⊞ From: "0372" <sip:0372@192.168.85.90:6060>;tag=300513766
  ⊞ To: <sip:0371@192.168.85.90:6060>
    Call-ID: 2E33624B-4D4F-7490-9E99-16EB3AB603B2@192.168.85.203
    CSeq: 5 INVITE
  ⊞ Contact: <sip:0372@192.168.85.203:6060>
    Max-Forwards: 70
    User-Agent: Star-Net V2.2.3E
    Content-Type: application/sdp
    Content-Length: 251
```

图 5　语音数据包报头分析

参考文献:

[1]　陈赜,秦贵和. AMR9 嵌入式技术 Linux 高级实践教程[M]. 北京:北京航空航天大学出版社,2005:125 - 134.

[2]　张智江,张云勇,刘韵洁. SIP 协议及其应用[M]. 北京:电子工业出版社,2005:34 - 50.

[3]　胡凌凌,彭荣修. SIP 协议在一个 IP 电话模型中的实现[D]. 武汉:华中科技大学,2005.

[4]　司端锋,韩心慧,尤勤,等. SIP 标准中的核心技术与研究进展[J]. 软件学报,2005,16(2):239.

[5]　孙建勇. 基于 SIP 协议软终端的研究与实现[D]. 北京:北京邮电大学,2004.

[6]　黄勋,唐慧强. 嵌入式平台 ARM-Linux 的构建与应用开发[J]. 武汉理工大学学报:交通科学与工程版,2006,27(6):174.

[7]　何彬,张国清. SIP 可视电话系统的信令流分析[J]. 计算机工程与应用,2005,41(5):157.

文章编号:2095-476X(2012)06-0062-03

脉冲激光应答机测距精度研究

魏龙超[1], 何子杰[2], 陈静[1], 张文平[1], 范杰平[1]

(1. 中国电子科技集团公司第二十七研究所, 河南 郑州 450047;
2. 河南中医院, 河南 郑州 450002)

摘要:针对脉冲激光应答测距易产生传输延迟时间抖动进而影响测距精度的问题,选用LD激光器,采取自动温度控制、自动功率控制和高精度的脉冲时刻检测等技术,研制了极小时延抖动脉冲激光应答机. 光源试验结果表明,该设计的时延抖动<1 ns,有效提高了测距精度.

关键词:脉冲激光应答机;测距精度;恒比定时
中图分类号:TN24 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.017

Study on the ranging measurement accuracy of pulse laser responder

WEI Long-chao[1], HE Zi-jie[2], CHEN Jing[1], ZHANG Wen-ping[1], FAN Jie-ping[1]

(1. 27*th Institute of Electronics, China Electronics Technology Group Corporation, Zhengzhou* 450047, *China*;
2. *He'nan Provincial Hospital of Traditional Chinese Medicine, Zhengzhou* 450002, *China*)

Abstract: Aiming at the problem of the translate time delay's jitter with pulse laser ranging, further the jitter affects the accuracy of the ranging. The minimal time's delay pulse laser responder were designed adopting some techniques such as the diode laser, automatic temperature control, automatic power control and the detection method of the high-accuracy pulse time. The results of light experiment indicated the pulse of jitter less then 1 ns, the accuracy of ranging was effectively improved.

Key words: pulse laser responder; measurement accuracy of ranging; constant fraction discriminator

0 引言

随着空间技术科研活动范围的不断扩大,目标之间的距离也在加大,测卫、测月、测火星以及将来更远距离的测量活动,都对目标的距离测量技术提出了挑战,激光测距已成为重要手段之一. 目前激光测距多采用测量目标反射回波的方法来实现目标测量,其缺点是在现有技术条件下,测量能力无法满足目标活动范围不断扩大的要求. 脉冲式激光应答机可以实现较远距离的测量,在空间测距研究中具有广泛的应用前景[1].

目前已有对激光应答测距方案进行的研究[2],但是针对影响脉冲激光应答测距精度的因素的研究不多. 鉴于此,本文拟对脉冲激光应答机测距精度进行研究,并提出提高测距精度的具体方案.

1 脉冲激光应答机原理

图1为脉冲激光应答机的工作原理框图. 其工作过程为:当光电探测传感器接收到对方的激光信号时,将接收信号转换为电信号,通过放大器和滤波电路处理,产生一个本地激光控制信号送给激光器驱动电路,并驱动激光器输出脉冲式激光信号.

收稿日期:2012-09-04

作者简介:魏龙超(1982—),男,河南省宝丰县人,中国电子科技集团公司工程师,主要研究方向为激光技术及其应用.

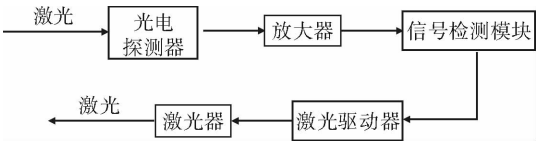


图 1　脉冲激光应答机原理框图

2　影响脉冲激光应答机测距精度的内外界条件分析

在应答式测距中,影响脉冲激光应答机测距精度的内外界条件有接收对方激光脉冲到达时刻的测量延迟时间和测量精度、本地激光触发延迟时间和稳定精度.

2.1　接收激光脉冲信号到达时刻的测量精度分析

在脉冲激光应答系统的设计中,测距精度主要取决于时点判别电路的设计. 由于激光传输过程中会发生畸变和衰减,接收到的脉冲与发射脉冲在幅度和形状上都会发生很大的变化,因此很难确定光脉冲回波信号的到达时刻,由此引起的测量误差称为漂移误差. 另外,由输入噪声引起的时间波动也会给测量带来误差. 为了尽可能减小漂移误差和时间抖动,本文采用恒比定时法测量. 恒比定时的原理是以某固定的脉冲高度比进行计时,该方法简单、精度高,只要波形变化不大,即使幅度变化很大,定时精度也是极高的.

2.2　激光脉冲触发延迟时间的抖动分析

激光脉冲触发延迟时间的抖动主要来源于系统的时间统一基准与激光器工作的时延,且延迟时间并不稳定,即延迟时间的抖动[3].

3　脉冲激光应答机设计方案及试验

3.1　系统设计方案

3.1.1　激光器的选择　在脉冲激光发射中,体现时延的指标主要是延迟时间的稳定性. 影响激光器触发延迟抖动的因素主要有激光器的工作形式、环境稳定性和功率稳定性等因素. 由于不同的激光器工作过程存在差异,激光触发到输出的延迟与抖动原因也有所区别. 因此笔者对其他项目所用不同激光器的时延抖动进行了测量,结果表明,大功率的 DPL 激光器触发延迟时间长且抖动大,而 LD 延迟时间较短且抖动小, LD 激光器的时延抖动 <1 ns. 在应答式激光测距中,即使较小的激光功率也可实现较远距离的测量,因此在脉冲激光应答机系统设计时,可首先考虑使用 LD 激光器.

3.1.2　稳定激光器延时抖动的应对方法　1)保持温度稳定性. LD 激光器属于功率器件,在不同的工作温度条件下,工作性能有所改变. 一般来讲,激光器在低温条件下的输出功率较高,为此,可以考虑使用制冷技术对激光器进行恒温制冷,最大限度地改善激光器的工作状态. 2)保持功率稳定性. 输出功率的改变会造成输入电流功率和其他工作参数的改变,从而影响激光传输延迟时间的稳定. 为了使激光器的输出功率稳定,可以采用自动功率控制技术,保持功率的稳定性,改善激光传输延迟时间的精度.

3.1.3　激光应答机方案设计　根据对关键技术的分析和试验,设计了极小时延抖动脉冲激光应答机,主要进行应答延迟抖动时间的测量,其原理如图 2 所示.

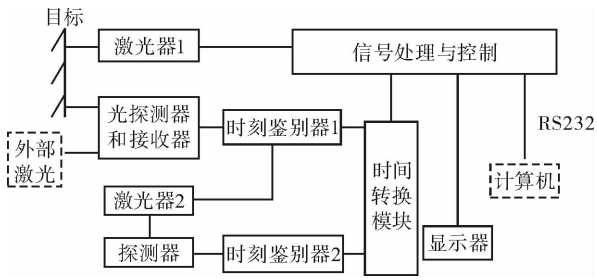


图 2　极小时延抖动脉冲激光应答机原理框图

脉冲激光应答样机系统工作时,首先由本地信息处理与控制系统控制激光器 1 发射激光,经外部目标反射后的信号进入接收光学及传感器系统转换成电信号,经时刻鉴别器 1 检测出信号到达时刻后,分成 2 路,一路送 TDC 转换系统启动时间转换,另一路驱动激光器 2 发射激光,经激光探测和时刻鉴别 2 检测出激光出光时间后,送 TDC 转换系统停止时间转换,此时 TDC 转换系统送出的数据即为应答延迟时间数据[4-7].

3.2　试验结果与分析

3.2.1　内部光源试验　使用内部光源试验时,放置一合作目标,自身光源发射,通过目标反射光信号进行系统测试. 试验采用不同重频的激光发射,数据处理结果见图 3. 从试验数据可以看出,排除由于系统噪声干扰出现的异常大数据,时延抖动 <1 ns.

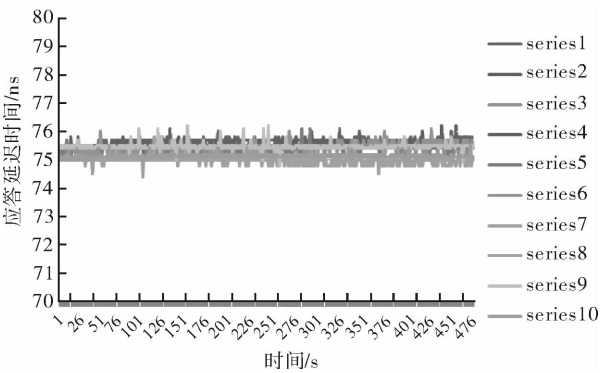


图 3 内光源试验数据处理图

3.2.2 外部光源试验 将脉冲激光应答样机和外部激光光源的放置距离间隔 10 m 以上,外部光源试验框图如图 4 所示. 试验结果表明,同一光源在10 ~ 30 m 距离变化时,测量精度 <1 ns.

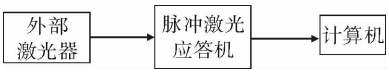


图 4 外部光源试验连接框图

4 结论

本文选用 LD 激光器,采取自动温度控制、自动功率控制和高精度的脉冲时刻检测等技术,有效地控制并减小了脉冲激光应答机的传输延迟时间抖动,其数值 <1 ns,从而保证了激光测距精度. 但在本设计方案中,由于恒比定时电路的保精度动态范围比较小,一般在 20 ~ 30 dB 量级,而在联试过程中信号的变化比较大,使得恒比输出精度变差,因此下一步工作将重点针对恒比电路进行优化设计,以实现更高精度的脉冲激光应答机距离测量.

参考文献:

[1] 唐嘉,高昕,邢强林,等. 异步应答激光测距技术测量精度实验[J]. 红外与激光工程,2011,40(5):939.

[2] 彭孝祥,张兴敢. 一种改进的脉冲式激光测距仪的设计[J]. 电子测量技术,2008,31(6):133.

[3] 陈霞. 数字脉冲应答机的时延补偿方法[J]. 电讯技术,2010,50(4):65.

[4] 王勇新,章恩耀,方仲彦,等. 时幅转换技术及其在激光测距系统中的应用[J]. 光学技术,2001,27(2):132.

[5] 霍玉晶,陈千颂,潘志文. 脉冲激光雷达的时间间隔测量综述[J]. 激光与红外,2001,31(3):136.

[6] 韦卫东,孙晓泉,孙晓军. 一种激光脉冲相关检测的改进方法[J]. 电光技术应用,2011,26(3):4.

[7] 王洪喆,辛德胜,张剑家,等. 脉冲激光测距时间间隔测量技术[J]. 强激光与离子束,2010,22(8):1751.

文章编号:2095-476X(2012)06-0065-04

USB 3.0设备中并行循环冗余校验的研究与实现

滕立伟[1], 李小花[2]

(1. 中国海洋大学 信息科学与工程学院, 山东 青岛 266100;
2. 昆明理工大学 信息工程与自动化学院, 云南 昆明 650500)

摘要:针对串行循环冗余校验(CRC)算法不适于高速传输且不易于硬件实现的问题,结合USB 3.0设备中CRC的特点,推导出一种并行CRC算法,并用Verilog硬件编程语言加以实现. 仿真试验表明,并行CRC校验算法具有更高的数据吞吐率,能降低时钟频率,易于硬件实现.

关键词:USB 3.0设备;循环冗余校验;并行算法
中图分类号:TN919 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.018

Research and implementation of the parallel CRC in USB 3.0 device

TENG Li-wei[1], LI Xiao-hua[2]

(1. *College of Information Science and Engineering, Ocean University of China, Qingdao* 266100, *China*;
2. *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500, *China*)

Abstract: Aiming at the problem that the serial CRC check algorithm is not suitable for high-speed transmission and not easy to hardware realization, a parallel CRC algorithm was deduced base on the characteristics of the CRC in the USB device, and was implemented with Verilog hardware programming language. The simulation results showed that parallel CRC algorithm has a higher data throughput rate and can reduce the clock frequency. It's easier to implement in hardware.

Key words: USB 3.0 device; cyclic redundancy check(CRC); parallel algorithm

0 引言

为了提高数据传输的有效性,通常在发送数据时对数据进行编码,并加入校验位来检测接收数据是否正确. 常用的校验码有奇偶校验码、海明码和循环冗余校验CRC(cyclic redundancy check)码. 这些编码方式都是按照一定的编码规则,在信息位后增加冗余位后一起发送. 接收端在接收到信息位后,按照相同的编码规则得到校验位,与接收的校验位比较即可得知信息位是否正确. 其中,CRC是USB 3.0协议采用的数据校验方式. CRC码不仅具有很强的检测能力,而且实现简单,在数据传输中被广泛应用. 由于串行CRC算法本身的局限性,以提高时钟频率为代价来提高数据吞吐率已不能满足高速传输的要求,且不易于硬件实现,因此需要采用更快的并行算法. 本文在CRC原理的基础上,针对USB 3.0设备的具体应用进行研究,由串行CRC算法推导出一种并行CRC算法,以提高数据吞

收稿日期:2012-07-17

作者简介:滕立伟(1987—),男,山东省烟台市人,中国海洋大学硕士研究生,主要研究方向为多媒体通信网络与系统设计.

吐率,降低时钟频率.

1 USB 中的 CRC 算法

1.1 CRC 原理

CRC 码是一种截短循环码,属于线性分组码. 其编码的过程为:当发送一个 k 位的信息序列 $T(x)$ 时,根据 k 可以使用一个特定的 $r+1$ 位生成多项式 $G(x)$,再将信息序列 $T(x)$ 左移 r 位,低位补零,用 $T(x)x^r$ 除以生成多项式 $G(x)$ 得到余数 $R(x)$ 和商 $Q(x)$,其中 $R(x)$ 即是校验位. 将 $R(x)$ 附加在信息位后一并发送. 接收端接收到数据(包括校验位)后,用同一个生成多项式 $G(x)$ 去除,若余数为零则可判断所接收到的信息位正确;否则,表明传输过程中发生了错误.

由上述过程可知

$$T(x)x^r = Q(x)G(x) + R(x) \qquad ①$$

余数 $R(x)$ 作为校验位附加在信息位 $T(x)$ 后. 此时实际发送的数据

$$T(x) = T(x)x^r + R(x) \qquad ②$$

由式①②可得

$$T(x) = Q(x)G(x) + R(x) + R(x) \qquad ③$$

由于采用的是模 2 运算,所以接收端校验时相当于用 $Q(x)G(x)$ 除以生成多项式 $G(x)$. 如果信息序列在发送、传输和接收过程中没有发生错误,校验时余数为 0.

移位算法是常用的 CRC 编码算法,其编码电路是一种串行结构的除法器[1],图 1 为 CRC-16 的硬件电路框图.

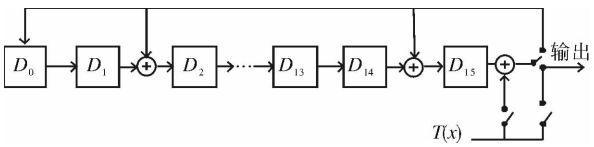


图 1 CRC-16 硬件电路实现框图

CRC 编码电路工作过程如下:

1)在 $T(x)$ 输入时就可以进行除法运算,同时,将信息位 $T(x)$ 送到输出,形成编码的前半段;

2)所有的信息序列 $T(x)$ 输入完后,移位寄存器中的数据即是校验位,将校验位紧接着信息序列输入即可完成编码.

1.2 USB 3.0 的包结构

USB 3.0 协议中有超高速、高速、全速和低速 4 种传输模式. 高速、全速和低速模式传输的数据包有 Token 令牌包、Data 数据包和握手包 3 种[2]. 令牌包分为 IN 令牌、OUT 令牌、SETUP 令牌和 SOF 令牌,其中 SOF 令牌包与前三者的包结构略有不同,但数据位数相同. 各数据包的结构如图 2 所示.

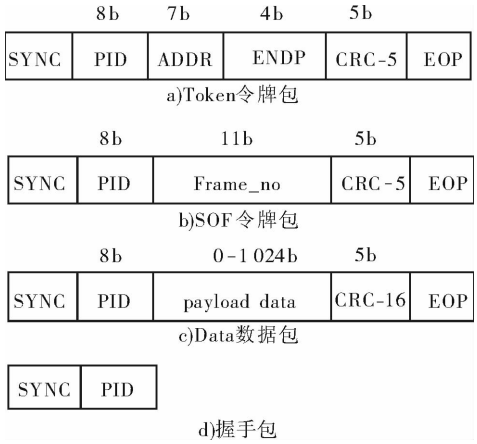


图 2 高速模式中的 4 种包结构

CRC 码用来在令牌包和数据包中保护所有的非 PID 字段. PID 字段存在自身的校验方式,故其不在 CRC 的校验范围内. 其中,令牌包采用 CRC-5 对 ADDR,ENDP 或者 Frame_no 字段提供 11 b 的数据校验,其生成多项式为 $G(x) = x^5 + x^2 + 1$;Data 数据包采用 CRC-16 对有效数据位提供校验,其生成多项式为 $G(x) = x^{16} + x^{15} + x^2 + 1$;握手包用来报告数据事务的状态,能表示数据成功接收、命令的接收或拒绝等,仅由一个 8 b 的 PID 构成,不需要校验.

超高速(SS)模式规定了 4 类型的包:Link Management Packet(LMP),Transaction Packet(TP),Data Packet(DP)和 Isochronous Timestamp Packet(ITP)[3]. 所有的包都有一个 16 B 包头,其中 TP,LMP 和 ITP 仅由 1 个包头构成;DP 由 Data Packet Header(DPH)和 Data Packet Payload(DPP)组成,其中 DPH 为 16 B 的包头,DPP 是 0—1 024 B有效数据. 图 3 给出了 DP 的结构图,其他 3 种类型的包结构与 DPH 的结构类似.

在 DPH 中包含 12 B 头信息、2 B CRC-16 校验位和 2 B 链路控制字. 其中 2 B 的 CRC-16 校验位用于保护 12 B 头信息. 链路控制字中包含 7 b 的信息位和 5 b 的 CRC-5,CRC-5 校验位用于保护 7 b 的信息位.

在 DPP 中包含 0—1 024 B 的数据位和 4 B 的 CRC-32 校验位,其生成多项式为

$$G(x) = x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^8 + x^7 + x^5 + x^4 + x^2 + x + 1$$

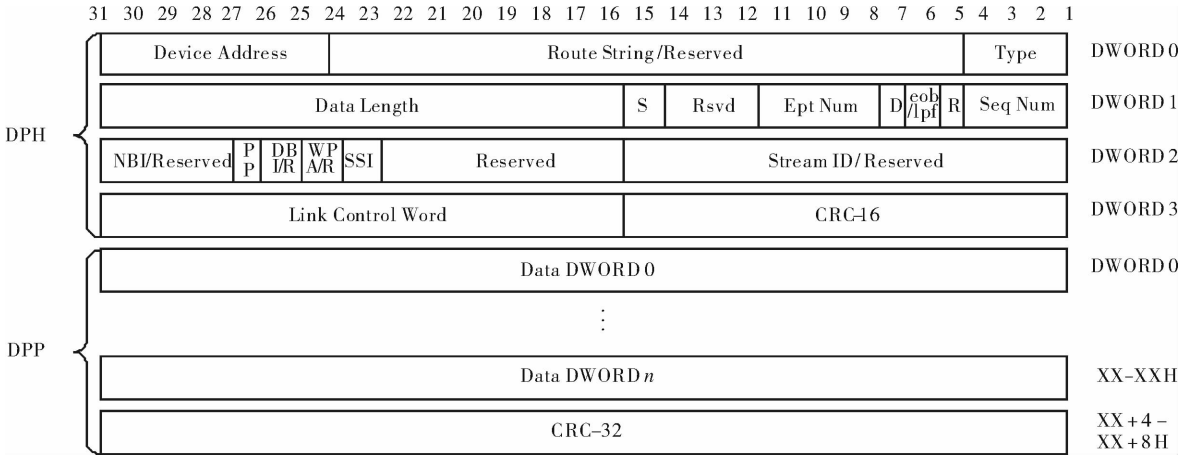


图 3　DP 结构图

2　并行 CRC 在 USB 3.0 中的实现

2.1　并行 CRC 算法的推导

图 1 所示的 CRC-16 硬件电路可以实现 USB 3.0数据传输中的 CRC 算法,但速度较慢. 在高速模式下传输速度为 480 Mb/s,超高速模式下传输速度高达 5 Gb/s,难以用硬件电路实现串行 CRC 算法. 故本文拟将串行算法改进为并行 CRC 算法. 下面以串行 CRC-16 算法推导并行 CRC 算法. 由图 1 所示电路可以得到在输入 1 位数据 din 后寄存器 crc15 和 crc14 的计算式,即

$$crc15' = crc15 \oplus crc14 \oplus din \quad ④$$

$$crc14' = crc13 \quad ⑤$$

当输入第 2 位数据后,用输入数据、crc14′和 crc15′计算得到新的 crc15. 连续输入 8 位数据,寄存器的结果根据式④⑤重复 8 次运算,最终得到串行输入 8 位数据后的 crc15.

其计算式为

crc15 = din[7]^din[6]^din[5]^din[4]^din[3]^
din[2]^din[1]^din[0]^crc7^crc8^crc_in9^crc10^
crc11^crc12^crc13^crc14^crc15

按照上述方法可以推导出其他位寄存器在串行输入 8 位数据后的计算式. 为了提高处理速度,设计中直接采用该计算式对并行输入的 8 位数据进行计算,得到 16 位的校验结果,再将新的校验结果存入寄存器,与下一次输入的并行数据进行计算,得到新的校验结果. 重复此操作,直到接收数据全部校验结束. 这样可实现并行的 CRC 校验,即在一个时钟周期内完成并行 8 位数据的校验.

2.2　并行 CRC 算法的实现

为了保护数据包开头的 0 数据位,初始时需将 CRC 生成器移位寄存器全部置 1,相当于在数据前添加一组固定的数据位[4]. 如果初始值为 0,移位寄存器内的数据将与输入数据做异或运算,数据包开头的 0 与寄存器内的 0 异或,移位寄存器中的数据仍为 0,导致数据 0 得不到保护. 同理,在接收端也以同样的方式设置移位寄存器的初始值,这样可以消除移位寄存器置 1 对 CRC 校验的影响.

在数据位校验完后,移位寄存器中的校验位要按位取反,然后高位在前进行发送,这样可以保护数据包末尾的 0 数据位. 另外,USB 设备中 CRC 采用并行方式(并行 CRC 方式将在下文介绍).

由于发送端对移位寄存器初始值置 1 和对校验位取反,在接收端有 2 种校验方式,2 种校验方式移位寄存器初始值与发送端相同,均置 1. 不同之处在于接收到的校验位是否进入校验器进行校验. 一种校验方式是与发送端的处理过程相同,仅对数据位做校验,不对校验位进行校验,将得到的校验结果取反后与接收到的校验位比较. 如果相同说明接收成功,否则说明数据位发生错误. 然而这种方式仅限于数据位数确定的情况,因为数据位确定时才可以提取出接收到的校验位. Token 令牌包中需校验的数据长度固定为 11 b,容易提取出校验位,然而对于 Data 数据包,其有效数据长度为 0—1 024 B,难以从 Data 数据包中取出校验位. 因此,此方法仅适用于 CRC-5 校验 Token 令牌包.

另一种校验方式是将接收到的数据位连同校验位一起送入校验器进行校验. 由于对校验位取反,相当于添加了一组固定的数据,这样在接收端校验的结果总是一个恒定值[1]. 对于接收的 CRC-32 数据包,这个恒定值是 32′hc704dd7b;对于接收的 CRC-16 数

据包,这个恒定值是16′h800d;对于 CRC-5 数据包,校验结果是5′h06. 接收端只需要判断校验后的结果是否与这个恒定值相同,即可判断接收数据是否正确. 设计中 CRC-16,CRC-32 采用此方法.

表1 是 CRC-16 校验器的端口说明. 设计中 crc_out[15:0]为 16 个校验寄存器,crc_in[15:0]是校验寄存器 crc1—crc15 的输入, crc_in[15:0]的值等于 crc_out[15:0],其初始值为16′hffff,din[7:0]是输入的 8 位待校验的数据. 根据并行 CRC 算法,此处给出 crc_out[0]的计算表达式为

$$crc\_out[0] = din[7] \wedge din[6] \wedge din[5] \wedge din[4] \wedge din[3] \wedge din[2] \wedge din[1] \wedge din[0] \wedge crc8 \wedge crc9 \wedge crc10 \wedge crc11 \wedge crc12 \wedge crc13 \wedge crc14 \wedge crc15$$

当前计算得到的 crc_out 送给 crc_in 作为下一次计算的寄存器值. 在发送端 CRC-16 校验时,需要将 crc_out 取反后发送;在接收端 CRC-16 校验时,不需要将 crc_out 取反,只需将其与 16′h800d 进行比较即可. crc_out[0]的硬件电路如图 4 所示,其他寄存器输出与 crc_out[0]类似.

表 1 CRC-16 校验器端口说明

端口	位宽/b	输入/输出	说明
crc_in	16	Input	校验寄存器
din	8	Input	输入数据
crc_out	16	Output	校验寄存器输出

CRC-32 的校验方式与 CRC-16 的校验方式相同. CRC-5 的设计与 CRC-16 类似,CRC-5 采用 11 b 并行输入,对接收的 Token 令牌包只需要一个时钟周期就可以得到校验结果;然后,将此校验结果取反与发送的校验结果相比较来判断数据是否正确.

3 仿真结果

本设计采用 Verilog 硬件编程语言编写,并使用 modlesim 仿真软件进行仿真. 图 5 是对 CRC-16 接收校验的仿真结果. 仿真过程中 USB 设备接收 DATA 包数据. 接收模块接收到 8 位串行数据后,将串行数据转换成8 位并行数据送给 CRC-16 校验器进行校验. 在图中竖线右侧 din = 39 为接收到的最后的 8 位数据. 在时钟的上升沿,将前一拍的输出 crc_out 传递给 crc_in 作为当拍的输入. crc_in 与并行输入数据 din(39)计算得到校验结果 crc_out 位800d. 表明数据接收正确.

将设计使用 Quartus Ⅱ进行综合,将得到门级网表下载到 ALTERA 公司的 Cyclone Ⅱ系列的 FPGA 进行验证,得到的结果如下:串行算法最大工作频率 180 MHz,吞吐率峰值 210 Mb/s,使用的 slice 为 60;并行算法最大工作频率 130 MHz,吞吐率峰值 600 Mb/s,使用的 slice 为 128.

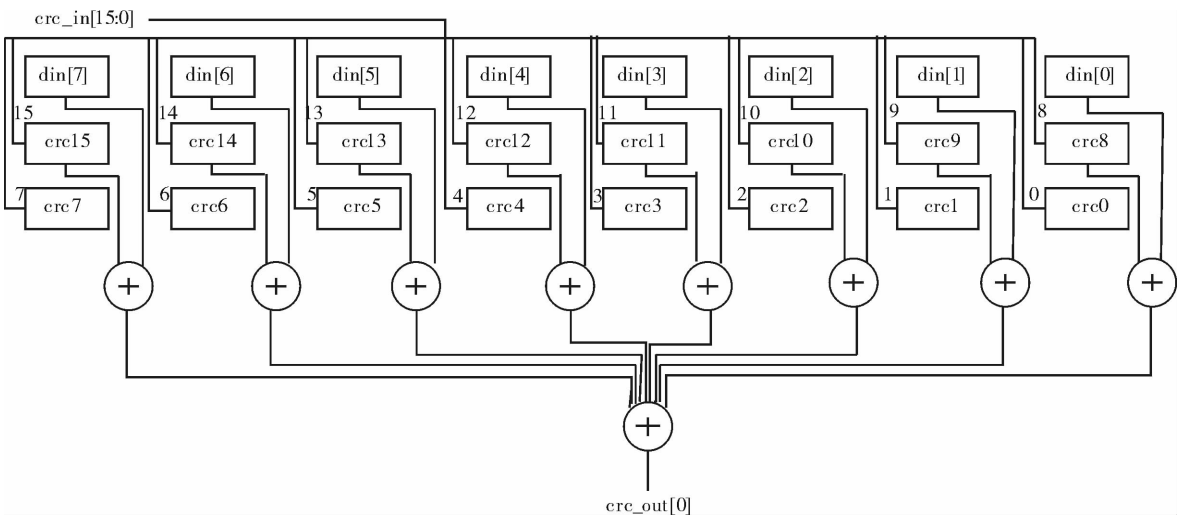


图 4 并行 crc_out[0]硬件电路

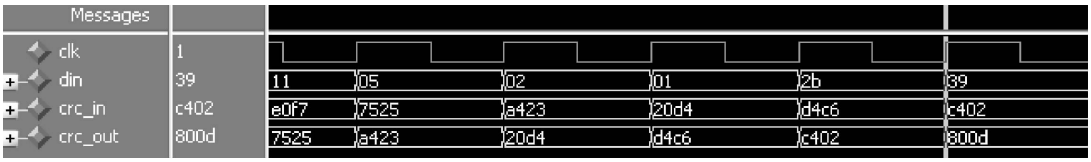


图 5 CRC-16 接收校验仿真结果

(下转第 72 页)

文章编号:2095-476X(2012)06-0069-04

应用BP神经网络实现基于等高线图像的CFD地形网格

甘勇[1,2], 刘新新[1], 郑远攀[1,2]

(1.郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001;
2.应急平台信息技术河南省工程实验室, 河南 郑州 450001)

摘要:针对目前生成CFD地形网格时计算量大且精度不高等问题,提出一种基于BP神经网络的CFD地形网格生成方法:将扫描数字化和线状要素提取相结合获取的等高线二维矢量矩阵作为训练样本,生成BP神经网络模型;通过Matlab实现BP神经网络的构建、训练与仿真,拟合出未在等高线上的点的高程值,建立高程DEM栅格矩阵,获得地形三维数据;在网格划分软件Gambit中进行网格生成与优化,生成精细的三维地形网格.实例验证表明,基于BP神经网络模型生成的CFD地形网格精确度高,建模效率也较高,适合CFD模拟工程的应用.

关键词:BP神经网络;等高线图像;Matlab;CFD地形网格
中图分类号:TP274 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.019

Constructing CFD terrain mesh based on contours image by applying BP neural network

GAN Yong[1,2], LIU Xin-xin[1], ZHENG Yuan-pan[1,2]

(1. *College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*;
2. *Engineering Lab of He'nan Province for Emergency Management Platform Oriented Information Technology, Zhengzhou* 450001, *China*)

Abstract: Aiming at the problem of huge calculation and poor accuracy when generating the CFD terrain mesh, a method for generating CFD terrain mesh based on the BP neural network was put forward, which put 2D vector matrix as the training sample, generating the BP neural network model. The matrix combines with the methods of scanning digitizing and linear elements extraction. And the model was structured, trained, simulated by Matlab. Based on the method, DEM elevation grid matrix was established by fitting the elevation that is not in contours. And then the 3D terrain mesh is generated in the software Gambit based on the data from the matrix. The example validation showed that applying the method can generate high accuracy and efficient CFD terrain mesh, it is suitable for the CFD simulation applications.

Key words: BP neural network; contour image; Matlab; CFD terrain mesh

---

收稿日期:2012-05-30

基金项目:河南省科技攻关计划项目(102102310030);郑州轻工业学院博士基金项目(2010BSJJ006);郑州市科技创新团队计划项目(112PCXTD344)

作者简介:甘勇(1965—),男,湖南省株洲市人,郑州轻工业学院教授,博士,主要研究方向为分布式计算机系统、计算机网络.

0 引言

作为计算流体力学 CFD(computational fluid dynamics)模拟的前处理过程,地形网格是数值模拟的一项重要课题,国内外已有众多学者对构建三维地形 CFD 网格开展了一系列研究.

杨长强等[1]通过 B 样条曲面建模方式实现了三维地形的构建,郭晓刚等[2]提出了基于辅助线的等高线地形算法,程雪玲等[3]通过编程提取出等高线图像中地形高程数据,并用 Gambit 的 Journal 功能生成 CFD 网格.

针对目前众多学者对 CFD 地形网格生成的研究中计算量大以及精确度不高等问题,本文提出一种基于 BP 神经网络的生成 CFD 地形网格的方法.

1 基于 BP 神经网络的三维地形预测建模

1.1 样本准备

本文采用扫描数字化和线状要素提取相结合[4]的方法获取单条等高线,并经过数据冗余处理,得到单条等高线的二维矢量矩阵 $\boldsymbol{A}_h$(h 为高程),其中每个像素点元素

$$a_{i,j} = \begin{cases} h & a_{i,j} \text{ 在等高线上} \\ 0 & a_{i,j} \text{ 不在等高线上} \end{cases}$$

然后将所有 $\boldsymbol{A}_h$ 矩阵求和,得到最终的所有等高线矢量矩阵 $\boldsymbol{C}$,$c_{i,j}$ 为每个像素点元素,即

$$\boldsymbol{C} = \sum_{h=20}^{100} \boldsymbol{A}_h$$

1.2 建立 BP 神经网络模型

1.2.1 基本原理 由于 BP 神经网络算法具有简单易行、计算量小、并行性强等优点,目前仍是网络训练的首选方法之一. BP 网络的结构由 1 个输入层、1 个或多个隐含层和 1 个输出层组成,各层由若干个神经元(节点)构成,上下层之间实现全连接,而每层神经元之间无连接,其拓扑结构见图 1. 每个节点的输出值与输入值是由作用函数和阈值决定,神经元可以实现输入与输出之间的任意非线性映射[5-6].

假设输入层有 n 个输入信号,输入向量为 $\boldsymbol{X} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$,隐含层节点数为 h,输出向量为 $\boldsymbol{Y} = (y_1, y_2, \cdots, y_h)^{\mathrm{T}}$,输出层有 m 个神经元,输出向量为 $\boldsymbol{O} = (o_1, o_2, \cdots, o_m)^{\mathrm{T}}$,期望输出向量为 $\boldsymbol{D} = (d_1, d_2, \cdots, d_m)^{\mathrm{T}}$. 输入层到隐含层之间的连接权矩阵 $\boldsymbol{V} = (v_1, v_2, \cdots, v_j, \cdots, v_h)$,其中列向量 $\boldsymbol{V}_j$ 为隐含层第 j 个神经元对应的权值,隐含层到输出层之间的连接权矩阵 $\boldsymbol{W} = (w_1, w_2, \cdots, w_k, \cdots, w_m)^{\mathrm{T}}$,其中列向量 $\boldsymbol{w}_k$ 为输出层第 k 个神经元对应的权向量.

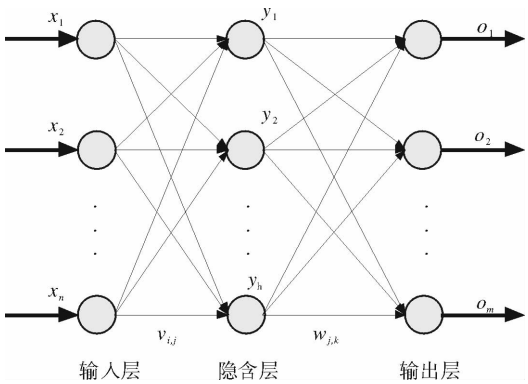


图 1 BP 神经网络拓扑结构图

输出层第 k 个神经元的输入输出分别为

$$o_k = f(net_k) \qquad k = 1, 2, \cdots, m$$

$$net_k = \sum_{j=1}^{h} w_{jk} y_j \qquad k = 1, 2, \cdots, m$$

隐含层第 j 个神经元的输入输出分别为

$$y_j = f(net_k) \qquad k = 1, 2, \cdots, h$$

$$net_j = \sum_{i=1}^{h} w_{ij} y_j \qquad j = 1, 2, \cdots, h$$

其中 $f(x)$ 为激励函数,连续且可导.

1.2.2 模型参数设置 1)模型各层函数设置. Kolmogorov 定理[4,7]证明了3层 BP 神经网络能够以任意精度逼近任何非线性信号或系统. 增加隐含层能够提高模型的精度,对于拥有 3 个及以上层数的前馈网络,通过增加层数的方式可以提高模型精度,但会带来额外的计算时间. 鉴于三维地形的复杂性,笔者决定隐含层采用 S 型激活函数 $f(x) = \dfrac{1}{1 + e^{-x}}$,输出层采用线性激活函数.

2)学习效率和初始化权值设置. 学习效率的一般取值范围为0.01 ~ 0.8,但在实际设计中,仍然要使用不同的学习效率对网络进行训练比较[5],以保证 $\sum E^2$ 快速下降. 若 $\sum E^2$ 出现震荡现象,则需要对学习效率做相应调整.

初始化权值一般选取一些不同的小随机数,以保证网络不会因为权值过大而进入饱和状态,从而导致训练失败. 根据实际应用,本研究选取 0.1.

3)隐含层节点数设置. 在实际应用中,一个比较好的神经网络不但其训练误差要达到一定的精度,而且要有比较好的泛化能力. 选用合适的隐含层节点数可有效防止训练中的过拟合现象,但隐含层节点的确定缺乏普遍、科学的方法,一般都是靠经验函数进行选择[6].

如公式 $n_1 = \sqrt{n+m} + a$,其中 n 为输入层节点数,m 为输出层节点数,a 为 1 ~ 10 的常数. 本文采用误差最小评价标准,经过多次训练得到最佳隐含层数目.

4)训练次数和目标误差设置. 根据神经网络训练终止条件可知[8]:最大训练次数与训练目标,只要满足其中任意一个条件,都将终止训练. 为达到目标误差要求,本研究选取训练次数为 1 000,训练终止目标误差为 1.0×10^{-6}.

1.2.3　模型训练　本研究为获取相邻等高线间未在等高线上的地形点的高程值,将等高线矩阵 $\boldsymbol{C}$ 赋值给训练样本 x_n, y_n, z_n,其中

$$x_n = i \qquad y_n = j \qquad z_n = c_{i,j} \qquad n = 1,2,3\cdots$$

将 (x_n, y_n) 作为第 n 个输入样本,将 z_n 作为输出样本,由此可选取输入层节点数为 2,输出层节点数为 1,从而计算出合适的隐含层节点数为 2—11. 根据样本数据进行训练,训练所需步数和均方误差见表 1.

表 1　样品数据训练实验数据

节点数	均方误差	训练步数
2	1.894	1 108
3	1.901	908
4	1.857	745
5	1.925	402
6	1.824	324
7	1.890	305
8	1.885	289
9	1.905	276
10	1.899	254
11	1.865	218

由表 1 可知,当隐含层神经元节点数为 6 时,均方误差最小,其泛化能力最好,训练步数不大,且没有出现过拟合.

2　三维地形建模及网格生成

2.1　地形建模

借助 Matlab 实现 BP 神经网络的构建、训练与仿真,通过实验证明上步所得到的 BP 神经网络训练模型有效可用. 应用训练好的 BP 神经网络模型拟合等高线图中未在等高线上的高程数据,获得更精细的高度 DEM 矩阵.

定义 $\boldsymbol{D}$ 为 DEM 栅格矩阵,$x_n = i, y_n = j(n = 1, 2, \cdots)$ 作为输入层输入值进行拟合,结果为 $c_{i,j} = z_n(n = 1,2,\cdots)$,得到所有高程 DEM 栅格矩阵. 借助 Matlab 软件提取出 DEM 栅格矩阵中所有坐标数据,并将其保存为 DAT 格式文件.

2.2　网格生成及优化

将上步得到的三维地形数据导入 Gambit 软件,应用 Gambit 的 Journal[3] 功能,通过命令行方式生成三维地形网格,即可构建地形网格图(见图 2). 图 2 中 X, Y, Z 分别代表地形的三维坐标.

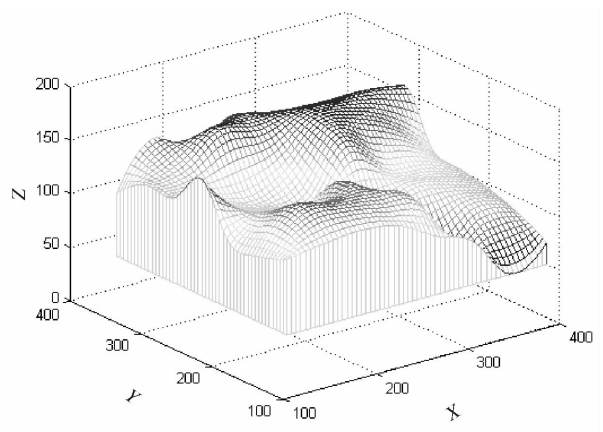


图 2　三维地形网络图

网格质量的好坏将影响到计算收敛速度和网格计算精度. 本研究通过设置网格参数生成不同疏密程度的贴体网格,并对已生成的网格进行优化处理,包括质量检查和多面体网格转换 2 个过程.

1)使用 Gambit 的 examine mesh 功能对网格文件进行检查,然后将其导入 CFD 软件,使用 Fluent 的 grid quality 功能检查网格质量. 结果证明优化后的网格质量较好.

2)在确保 CFD 模型精确与高效的情况下,通过菜单 Mesh > Polyhedra > Covert Skewed Cells 修正计算域内歪曲度较大的六面体网格,进而形成高质量的 CFD 地形网格.

3　结论

本文在常用等高线图的基础上,经过数据处理得到等高线矢量矩阵,应用基于 BP 神经网络训练

方法,提取出地形数据,使用软件 Gambit 生成了高精度三维地形网格. 通过实验证明此方法有效提高了地形建模的效率,适合 CFD 模拟工程的应用.

参考文献:

[1] 杨长强,郑永果,郑作亚. 利用 B 样条实现基于等高线的三维地形图[J]. 信息技术与信息化,2006,37(1):59.

[2] 郭晓刚,黄先祥,仲启媛. 基于辅助线的等高线生成三维地形算法研究[J]. 系统仿真学报,2011,23(6):1191.

[3] 程雪玲,胡非. 复杂地形网格生成研究[J]. 计算力学学报,2006, 23(3):313.

[4] EI-Fouly T H M, EI-Saadany E F, Salama M M A. Improved grey predictor rolling models for wind power prediction[J]. IET Generation, Transmission and Distribution, 2007, 1(6): 928.

[5] 张文霄. 基于 PSO 优化的 BP 神经网络股票预测模型[D]. 哈尔滨:哈尔滨工业大学,2010.

[6] 卢昕昀. 基于 BP 神经网络的超市选址评估研究[D]. 上海:上海交通大学,2008.

[7] Louka P, Galanis G, Siebert N, et al. Improvements in wind speed forecasts for wind power prediction purposes using Kalman filtering[J]. Journal of Wind Engineering and Industrial Aerodynamics,2008,96(12):2348.

[8] 蒋兴恒,朱素蓉. 基于 Levenberg-Marquardt 算法改进 BP 神经网络的卷烟销量预测模型研究[J]. 经济与管理,2011,17(5):81.

(上接第 68 页)

通过验证,并行算法时钟频率明显低于串行算法,数据吞吐率约为串行算法的 3 倍,而占用的 slice 仅仅比串行算法提高了约 1 倍. 因此,采用并行算法可以在少量增加资源的情况下获得较高的校验速率.

4 结语

本文在串行 CRC 算法的基础上推导出了一种并行 CRC 算法,并用 Verilog 语言对此算法进行实现. 仿真结果表明,此校验算法数据吞吐率高,可以有效降低电路的工作频率,易于用硬件电路实现.

参考文献:

[1] 于工. 信息与编码简明教程[M]. 北京:国防工业出版社,2007:67 - 80.

[2] Inter. Universal Serial Bus 2.0 Specification[EB/OL]. (2008 - 01 - 09)[2012 - 07 - 01]. http://download.csdn.net/detail/wchzh318/330131.

[3] Inter. Universal Serial Bus 3.0 Specification[EB/OL]. (2008 - 11 - 18)[2012 - 07 - 01]. http://down.tech.sina.com.cn/content/42037.html.

[4] 蒋安平. 循环冗余校验码(CRC)的硬件并行实现[J]. 微电子学与计算机,2007,24(2):107.

[5] 常天海,胡鉴. 基于 FPGA 的 CRC 并行算法研究与实现[J]. 微处理机,2010(2):45.

[6] 程超,程善美. Unfolding 算法实现的高速并行 CRC 电路的 VLSI 设计[J]. 微电子学与计算机,2002(12):68.

[7] Shukla S, Bergman N W. Signle bit error correction implementation in CRC-16 on FPGA[C]//Field Programmable Technology, Piscataway: The Institute of Electrical and Electronics Engineers,2004.

文章编号:2095-476X(2012)06-0073-05

一种基于点云数据的复杂地形CFD网格生成方法

刘新新[1], 甘勇[1,2], 郑远攀[1,2]

(1.郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001;
2.应急平台信息技术河南省工程实验室, 河南 郑州 450001)

摘要:针对CFD网格生成中地形数据精确度不高的问题,提出一种基于点云数据的复杂地形CFD网格生成方法.该方法引入七参数求解方法生成地方坐标点云文件,采用点—线—面—体方式生成地形网格,并基于多面体网格转换方法进行网格优化.实例验证结果表明,生成的地形网格与实际地形吻合度高、数据精确度高,适合用于后期的CFD精细模拟应用.

关键词:网格生成;七参数;复杂地形;多面体网格转换
中图分类号:TP399 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.020

A method for grids formation based on point cloud data complex terrain CFD

LIU Xin-xin[1], GAN Yong[1,2], ZHENG Yuan-pan[1,2]

(1. *School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*;
2. *Engineering Laboratory of He'nan Province for Emergency Management Platform Oriented Information Technology, Zhengzhou* 450001, *China*)

Abstract: Aiming at the problem that terrain data accuracy is not high in CFD grid generation, a grid generation method based on point cloud data complex terrain CFD was put forward. This method generates the local coordinates of the point cloud file through the introduction of a seven-parameter solution method, and by the point-line-face-volume method for mesh generation, further polyhedral mesh conversion, the terrain mesh is optimized. The example validation showed that the application of method can generate high accuracy data and consistent with the actual terrain mesh, is suitable for the later stage of the fine CFD simulation applications.

Key words: mesh generation; seven-parameter; complex terrain; polyhedron mesh conversion

0 引言

近年来,由于石油泄漏、有害气体扩散等污染而导致的事故所造成的经济损失在国内外呈现逐年上升的态势.计算流体力学CFD(computational fluid dynamics)应用于气体泄漏扩散过程的数值模拟,将对进一步科学预防事故性泄漏的发生、指导紧急救灾具有重要理论价值和实践意义.复杂地形网格生成作为CFD模拟的前处理过程,是CFD软件进行数值离散模拟的前置条件,由此也就成为环

收稿日期:2012-09-25
基金项目:河南省科技攻关计划项目(102102310030);郑州轻工业学院博士基金项目(2010BSJJ006);郑州市科技创新团队计划项目(112PCXTD344)

作者简介:刘新新(1987—),女,河南省济源市人,郑州轻工业学院硕士研究生,主要研究方向为计算机图形学.

境能源发展中解决能源利用以及有效处理包括自然因素(复杂地形地貌、气象活动等)和人为因素(工业过程中为谋求社会福利而进行的活动等)在内所引起危害社会财产及人身安全的突发事件的一项重要课题.

国内外众多学者对构建复杂地形 CFD 网格开展了一些研究. 如 F. Scargiali 等[1]研究了大范围复杂地形中有害气体泄漏扩散的过程;程雪玲等[2]通过编程提取出等高线图中的地形高程数据,并用 Gambit 的 Journal 功能生成 CFD 网格;史明昌等[3]研究了数值高程模型 DEM 网格尺寸对 DEM 精度的影响. 针对目前大多数研究所生成的地形网格精度不够的问题,本文研究的主要是:提取航拍原始图像三维坐标数据,根据数据地理坐标的不同,选择实地参照点,使用七参数验证的方式将其转化为 CFD 前处理器 Gambit 能识别的地方平面直角坐标点云,再使用编程方式,应用 Gambit 的 Journal 功能完成复杂地形的 CFD 网格生成,并对网格进行优化处理,为后期 CFD 精细数值模拟使其可应用于微尺度区域下的能源评估打下基础.

1 数据的获取和转换

1.1 数据的获取

本研究数据来源为美国太空总署(NASA)和国防部国家测绘局(NIMA)联合测量发布的 SRTM(shuttle radar topography mission)数据. 数据使用的水平基准面是 WGS84 椭球模型,覆盖范围为北纬 60°至南纬 56°,绝对水平和高程精度分别为 20 m 和 16 m. SRTM 地形数据按精度可以分为 SRTM1 和 SRTM3,对应的分辨率精度分别为 30 m 和 90 m 数据(目前公开数据为 90 m 分辨率的数据). 通过文献[4]对 SRTM 与不同比例地形图生成的数字高程模型 DEM(digital elevation model)的地形表达能力的比较得出:SRTM 虽然对地形的反映能力不能达到基于1∶100 000地形图建立的 25 m 分辨率的 DEM(DEM25)和基于1∶50 000地形图建立的 10 m 分辨率的 DEM(DEM10),但明显优于基于1∶250 000地形图建立的 100 m 分辨率 DEM,适合于大比例尺的地形数据使用. SRTM 可以免费下载,易得到高精度数据,所以在地形网格生成中得到了广泛的应用. 本实例通过查询所研究区域的经纬度,下载 SRTM 数据,为真实地形网格生成提供数据来源. 图 1 为某地形卫星航拍图.

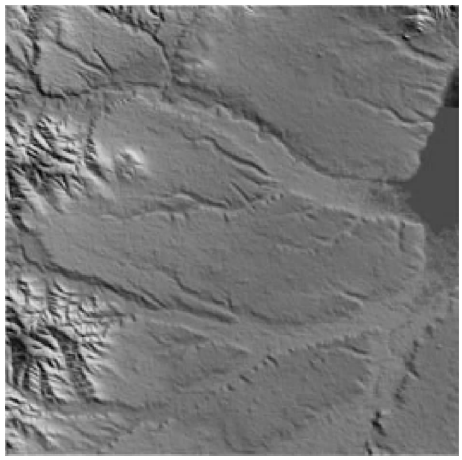


图 1 某地形卫星航拍图

1.2 数据的转换

为保证地形网格原始数据的准确性,本次研究采用地方坐标系统数据作为地形网格生成的数据基础.

SRTM 数据采用 WGS84 大地基准地理坐标系统,我国采用的多是西安 80 坐标系统,因此在从事地形网格生成时,需要将 SRTM 数据成果转换成我国的地方坐标系统数据. 首先将 WGS84 大地基准地理坐标系统转换为西安 80 坐标系统,在此基础之上再转换为带参照点的地方坐标系统.

1.2.1 数据准备 根据坐标转换精度要求,本算例采用空间直角坐标转换方法,将 SRTM 数据格式(.tif)用 Globalmapper 软件读取,并输出 WGS84 空间直角坐标. 通常所下载的 SRTM 数据包括的面积要大于研究区域,因此数据导出时需选择重点关注的地形区域. 本算例选取 2 km×2 km 重点区域导出 WGS84 空间直角坐标值,部分数据见表 1.

1.2.2 七参数坐标转换模型 坐标重合点采用在 2 个坐标系下均有的坐标成果点. 使用坐标转换软件 Coord MG 首次求得七参数,通过反复计算重合点坐标残差值来确定最终重合点,直到所有数据点坐标残差小于 3 倍中误差,即可确定最终七参数值.

由于坐标重合点存在误差,其点位的几何分布及点数量将影响求得的转换参数精度,因而为求得较好的转换参数,应选择一定数量精度较高且分布较均匀,并有较大覆盖面的重合点.

本研究选取了计算区域内均匀分布的 15 个已有的 WGS84 坐标对应西安 80 坐标的坐标成果点作为重合点,首先以此成果点作为公共点,计算七参数,源坐标类型为 WGS84 空间坐标(*XYZ*),目标坐

表 1　某地区提取的空间直角坐标点　　m

X	Y	Z
406 008.501 74	130 238.500 35	1 049.889
406 018.501 74	130 458.500 35	940.833
406 038.501 74	130 548.500 35	1 024.500
406 258.501 74	130 498.500 35	952.000
406 278.501 74	130 138.500 35	951.417
406 348.501 74	130 478.500 35	948.917
406 388.501 74	130 738.500 35	949.583
406 468.501 74	130 708.500 35	947.250
406 628.501 74	130 848.500 35	935.167
406 718.501 74	130 858.500 35	933.000
406 768.501 74	130 268.500 35	912.500
406 838.501 74	130 488.500 35	924.972
406 898.501 74	130 228.500 35	916.833
406 948.501 74	130 868.500 35	907.333
406 968.501 74	130 708.500 35	897.000

标类型为西安 80 空间坐标(*XYZ*),求布尔莎七参数,误差数据见表 2. 剔除不符合精度的公共点后,确定最终的 6 个公共点,源坐标类型为 WGS84 空间坐标(*XYZ*),目标坐标类型为西安 80 空间坐标(*XYZ*),求得最终的布尔莎七参数[5-8],误差见表 3.

使用误差最小的七参数进行数据坐标转换,获取西安 80 坐标值,以 MAPGIS 软件为平台进行投影转换,进行四参数转换,转换后的数据在一定程度上更接近实际地形. 由于数据高程为椭球高,测量中使用的高程为正常高,因此高程存在一定的差异,需要进行拟合. 为保证整个地形网格生成的精准性,从 SRTM 数据提取大量数据点,进行高程曲面拟合,最终得到国家地方坐标(*XYZ*)数据. 将地方坐标(*XYZ*)中 *XY* 值的最低点分别设为 *XY* 轴零点,*Z* 值的近似最低点为 *Z* 轴零点,以此为基准,转换获取其他各点坐标,形成 Gambit 所能识别的三维坐标文件.

2　网格生成

本文采用基于三维点云数据,按照点—线—面—体方式生成地形网格的方法,在软件 Gambit 中进行质量检查和适当调整优化,以提高网格的精确度和实用性.

2.1　网格点的生成

应用编程方式实现 Gambit 绘制点命令 vertex create " vertex name" coordinates,结合 Gambit 的 Journal[2] 文件功能,建立计算区域所有点并进行编号,结果如图 2 所示.

2.2　网格线的生成

根据前一步生成点的编号,建立绘制线的命令文件 edge create "line name" straight 和南北方向与东西方向的连线规则. 借助 Gambit 的 Journal 功能,调用生成网格线的程序文件,生成地形网格线,将所有的点进行四边形连接(如图 3 所示).

2.3　网格面的生成

在网格线生成的基础上,使用 Gambit 中的面生成命令行语句 face create "face name" wireframe 建

表 2　15 个公共点求布尔莎七参数误差数据表　　m

源坐标 *X* 误差	源坐标 *Y* 误差	源坐标 *Z* 误差	源坐标中误差	目标坐标 *X* 误差	目标坐标 *Y* 误差	目标坐标 *Z* 误差	目标坐标中误差
0.044	0.067	0.034	0.088	0.045	0.066	0.036	0.087
−0.024	−0.079	0.061	0.103	0.024	0.081	−0.059	0.103
0.161	0.057	0.034	0.174	−0.162	−0.055	−0.033	0.174
−0.037	−0.011	−0.003	0.039	0.036	0.013	0.005	0.039
0.049	0.083	−0.038	0.104	−0.049	−0.081	0.040	0.103
0.044	0.100	−0.054	0.122	−0.045	−0.098	0.055	0.121
−0.003	−0.072	0.064	0.096	0.002	0.074	−0.063	0.097
0.004	−0.069	0.056	0.089	−0.004	0.071	−0.054	0.090
−0.229	−0.036	−0.106	0.255	0.229	0.038	0.108	0.255
−0.027	−0.103	0.090	0.139	0.026	0.104	−0.089	0.139
−0.118	0.024	−0.100	0.157	0.117	−0.023	0.102	0.157
−0.010	−0.039	0.051	0.065	0.009	0.041	−0.050	0.065
0.164	0.081	0.022	0.184	−0.164	−0.079	−0.021	0.184
−0.023	−0.086	0.071	0.114	0.023	0.088	−0.070	0.115
−0.035	0.149	−0.121	0.195	−0.036	−0.147	0.123	0.195

表 3 6 个公共点求布尔莎七参数误差情况表 m

源坐标 X 误差	源坐标 Y 误差	源坐标 Z 误差	源坐标中误差	目标坐标 X 误差	目标坐标 Y 误差	目标坐标 Z 误差	目标坐标中误差
0.054	0.091	-0.046	0.116	-0.080	-0.090	0.047	0.115
-0.028	-0.066	0.041	0.083	0.028	0.068	-0.040	0.084
-0.051	-0.011	-0.017	0.055	0.050	0.013	0.019	0.055
0.031	0.076	-0.046	0.094	-0.031	-0.074	0.047	0.094
0.002	-0.060	0.050	0.078	-0.003	0.062	-0.048	0.078
-0.011	-0.019	0.025	0.034	0.011	0.021	-0.024	0.033

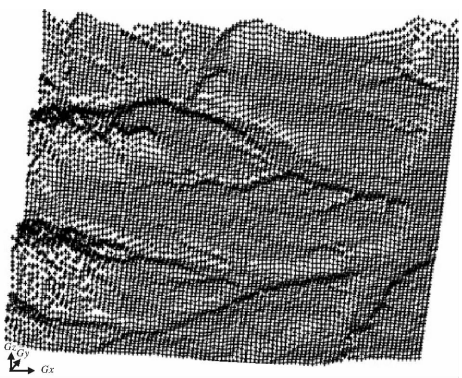


图2 Gambit 某计算区域点的绘制效果

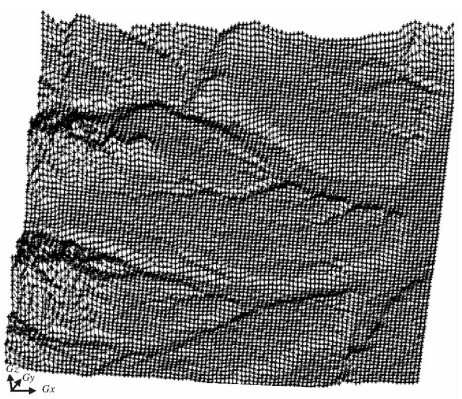


图 3 网格线的绘制效果

立程序文件,经过 Gambit 的简单调用即可生成网格面,形成地形基本轮廓面,如图 4 所示.

2.4 网格体的生成

为生成所需的计算空间并进行网格划分,还需以地面为顶面构建一个包含下部空间的体,包括构建体的几何体和网格体的生成 2 个阶段.

1)构建体的几何体. 根据生成的网格面在 XY 面上投影的大小确定建体所需长宽,高度的设定主要考虑最低点和最高点的海拔,而在进行数据点坐标转换时,所有数据点的海拔高度都经过减去接近于最低点海拔高度的处理. 本算例设定高度近似为最高点与最低点的海拔差,绘制的体的底面顶点及线条,如图 5 所示.

2)虚面、虚体的生成. 为生成结构体网格,使用了 Gambit 中的虚面功能,应用 Gambit 的 Merge faces 功能将所有的四边形小面融合生成虚面[3],即将原本不在同一面上的点拟合到同一曲面上,得到光滑的地形,从而生成便于计算的结构网格. 虚体的生成(见图 6),是在确定绘制的体的底面顶点坐标之后,根据坐标生成底面及生成体所需的 4 个侧面,并

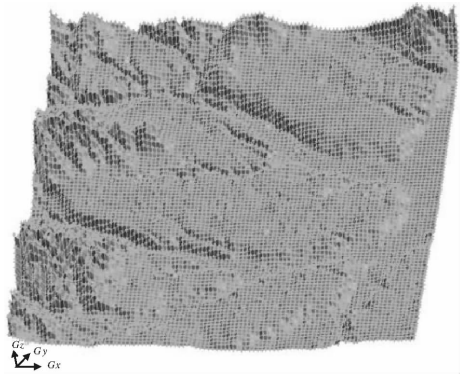


图 4 网格面的绘制效果

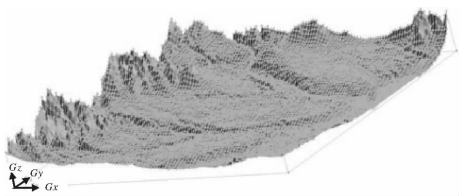


图 5 网格体的范围效果

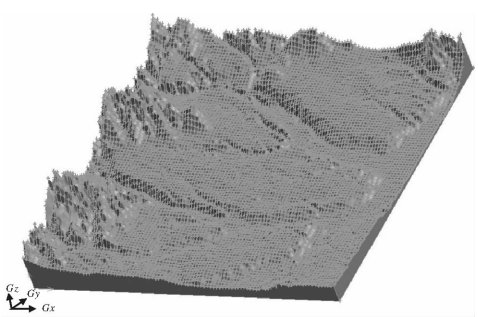


图 6 网格虚体的生成效果

最终将整个地形曲面与底面侧面聚合生成网格体. 至此,完成了 CFD 网格地形的构建,为网格划分奠定了基础.

3　网格优化

在一定程度上,网格质量的好坏将影响到计算收敛和网格计算精度. 考虑到近壁黏性效应,为形成高质量的近地面贴体网格,采用等比数列分布方法对网格体竖直边进行网格划分,设置网格参数生成不同疏密程度的贴体网格,并对已生成的网格进行优化处理,包括质量检查和多面体网格转换 2 个过程.

首先将网格文件使用 Gambit 的 examine mesh 功能进行检查,然后导入到 CFD 软件中,使用 Fluent 的 grid quality 功能检查网格质量,结果证明网格质量较好[9-10].

许多研究表明,多面体网格转化可显著减少网格的数量,提高计算速度及收敛性[11-12],尤其针对大量六面体网格的 CFD 地形具有明显的优化效果. 在确保 CFD 模型精确与高效的情况下,通过菜单 Mesh > Polyhedra > Covert Skewed Cells 修正计算域内歪曲度较大的六面体网格,形成更有效的 CFD 地形网格,效果如图 7 所示.

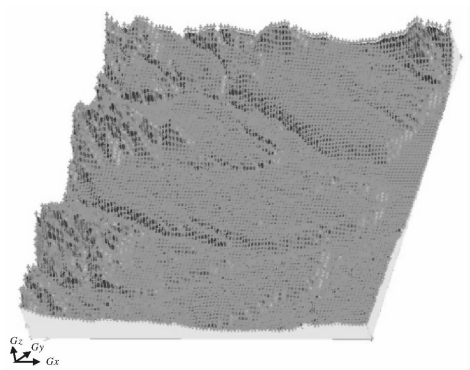


图 7　优化后效果图

4　结论

本文通过七参数求解方法将获取的 SRTM 三维地形坐标点数据转换成地方坐标点,根据 Gambit 软件中的命令行规则,使用 C 语言编程获取执行文件,并应用其 Journal 功能生成地形网格,不仅提高了地形数据精度,而且有效地提高了网格生成效率. 网格生成的实例表明本文提出的方法是一种可行的地形网格生成方法,降低了人工操作的复杂性,为 CFD 模拟应用的前处理问题提供了解决途径.

参考文献:

[1] Scargiali F,di Rienzo E,Ciofalo M,et al. Heavy gas dispersion modelling over a topographically complex mesoscale:A CFD based approach[J]. Process Safety and Environmental Protection,2005,83(3):242.

[2] 程雪玲,胡非. 复杂地形网格生成研究[J]. 计算力学学报,2006,23(3):313.

[3] 史明昌,沈晶玉. 不同地貌起伏状况下网格尺寸与 DEM 精度关系研究[J]. 水土保持研究,2006,13(3):35.

[4] 蔡清华, 杨勤科. SRTM 与地形图生成 DEM 的地形表达能力对比[J]. 水土保持通报,2009,29(3):183.

[5] 谢鸣宇,姚宜斌. 三维空间与二维空间七参数转换参数求解新方法[J]. 大地测量与地球动力学,2008,28(2):104.

[6] 柳光魁,王振禄,赵永强,等. 西安 80 坐标系与 WGS84 坐标系转换方法及精度分析[J]. 测绘与空间地理信息,2006,29(6):167.

[7] 赵宝锋. GPS 坐标向地方坐标转换模型的合理选择[J]. 城市勘测,2009,24(2):224.

[8] 刘亚平,郑若奇,曹立强. GPS 定位中两种七参数坐标转换方法的误差分析[J]. 中国港湾建设,2002,22(5):24.

[9] Coroneo M,Montante G,Paglianti A,et al. CFD prediction of fluid flow and mixing in stirred tanks: numerical issues about the RANS simulations[J]. Computersand Chemical Engineering,2011,35(10):1959.

[10] Gousseau P,Blocken B,Stathopoulos T,et al. CFD simulation of near-field pollutant dispersion on a high-resolution grid:A case study by LES and RANS for a building group in downtown Montreal[J]. Atmospheric Environment,2011,45(2):428.

[11] Zhang Bo,Chen Guo-ming. Quantitative risk analysis of toxic gas release caused poisoning:A CFD and dose-response model combined approach[J]. Process Safety and Environmental Protection,2010,88(4):253.

[12] Zhang Bo,Chen Guo-ming,Kong Ling-zhen. Toxic gas dispersion modelling over complex terrains[C]//Proceedings of 2009 International Conference on Energy and Environment Technology,Guilin:IEEE Computer Society,2009:135-138.

文章编号:2095-476X(2012)06-0078-04

一种适应于远程红外监控的图像增强方法

夏永泉, 董向滢, 王慧敏

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:针对远距离捕获的红外图像对比度低、亮度低以及感兴趣的目标比较小等问题,提出了一种新的自适应直方图均衡的方法.该方法利用图像直方图的2种不同信息,自动生成混合累积直方图,达到对红外图像进行自动均衡化的目的.试验结果表明,该方法可以有效地提高红外图像的对比度,对红外图像中小目标的增强效果更为明显.

关键词:远程红外图像监控;直方图均衡;混合累积直方图

中图分类号:TP391.41 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.021

An image enhancement approach for long-range surveillance based on infrared

XIA Yong-quan, DONG Xiang-ying, WANG Hui-min

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*)

Abstract: The infrared images that are captured at long range usually have low contrast, low brightness, and small hot objects of interest. A novel adaptive histogram-based equalization was put forward to enhance the contrast of infrared images in order to overcome the defects of infrared image itself. It can automatically generate a hybrid cumulative histogram based on two different pieces of information about the image histogram. The proposed adaptive equalization approach can automatically realize image equalization. Experimental results demonstrated that the approach proposed in this paper can improve the contrast of infrared images, and the effect is more obvious to the small objects embedded in infrared images.

Key words: long-range infrared image surveillance; histogram equalization; hybrid cumulative histogram

0 引言

近年来红外图像广泛应用于军事、医疗、工业等各个领域,其中最重要的一项应用是远程监控.然而,远距离捕获的红外图像具有对比度低、亮度低,以及感兴趣的目标比较小的特点.对于提升远程红外监控设备的性能来说,提高红外图像质量十分必要.一般来说,红外图像包括一些感兴趣的目标点和许多背景区域.在远程红外监控图像中,不仅缺少先验信息,而且包含许多噪声、杂波以及固定的非目标物体[1-2],由于感兴趣的目标比较小,所以在红外图像中区分感兴趣的目标和背景区域是十分困难的.本文拟提出一种包含混合累积直方图的自适应直方图均衡化方法来克服红外图像自身

收稿日期:2012-09-24

基金项目:国家科技支撑计划项目(2006BAK01A38);河南高校青年骨干教师资助计划项目(2010GGJS—114);郑州轻工业学院博士基金项目(2007BSJJ005)

作者简介:夏永泉(1972—),男,辽宁省绥中县人,郑州轻工业学院副教授,博士,主要研究方向为图像处理、计算机视觉、模式识别与人工智能.

存在的缺点,并提高图像增强的可靠性、鲁棒性和自适应性.

1　直方图均衡化

直方图均衡化是一种常用的图像增强方法,它根据图像的累积直方图[3]进行灰度调整,以达到增强图像的效果. 其灰度级调整方法是:在直方图中,像素数多而且分布密集的灰度级之间的间隔变大,使对比度得到提高;像素数少、分布较稀疏的灰度级间的间隔变小,甚至为 0(灰度级被合并),降低对比度. 若用该方法对红外图像进行增强处理,将会导致背景和噪声的灰度级偏多,而目标的灰度级偏少,这相当于提高了背景和噪声的对比度,反而降低了目标的对比度. 为了克服直方图均衡化算法的不足,本文提出基于混合累积直方图的自适应直方图均衡化算法. 该算法包括 2 个阶段,分别为自适应阈值选择和混合累积直方图的生成. 第一阶段选择合适的阈值,用于将直方图分为热物体部分和背景部分;第二阶段以直方图的 2 种不同信息为基础,生成 2 种不同的累积直方图,一个增强热物体,另外一个增强背景.

2　自适应阈值的选择

文献[4 - 5]介绍了大量阈值选择方法并根据图像信息进行归类,比如直方图形状、空间聚类尺寸、熵、物体属性、空间相关性和局部灰度面. 由于固定阈值不适合修正不同红外图像的物体,所以采用迭代阈值的方法[6]来自适应地选择阈值. 该方法步骤如下:

1)选择一个初始阈值 $Th(1)$ $(0 < Th(1) < 255)$.

2)利用选择的阈值 Th 对图像进行分割,根据图像像素的灰度值,可以将图像分割为背景和目标 2 个部分.

3)第 k 次迭代,计算 $\mu_B(k)$ 和 $\mu_O(k)$,分别代表背景和目标的灰度级.

$$\mu_B = \sum_{(i,j)\in \text{背景}} f(i,j)/N_B$$

$$\mu_O = \sum_{(i,j)\in \text{目标}} f(i,j)/N_O$$

这里 N_B 和 B_O 分别代表背景和目标的像素数目.

4)计算第 $k+1$ 次迭代得到的新阈值

$$Th(k+1) = \frac{\mu_B(k) + \mu_O(k)}{2}.$$

5)如果 $Th(k+1) = Th(k)$,则停止迭代,选择出的阈值为 Th;否则,返回执行步骤 2).

3　混合累积直方图的生成

与大面积背景区域相比,直方图均衡化对小目标的增强效果不那么明显. 为了克服直方图均衡的这种不足,提出基于混合累积直方图的自适应直方图均衡化的方法,它的主要优势是对小目标和背景产生 2 种增强效果:对小目标的强效果和对背景的弱效果. 生成混合累积直方图的过程如下.

首先定义均衡的累积直方图(如公式 ①) 以及相关参数.

$$T(r_k) = \sum_{j=0}^{k} p_r(r_j) = \sum_{j=0}^{k} \frac{n_j}{n}$$
$$k = 0,1,\cdots,L-1 \qquad ①$$

$$p_r(r_k) = \frac{n_k}{n} \qquad k = 0,1,\cdots,L-1 \qquad ②$$

$$S_k = \text{int}((L-1)\cdot T(r_k)) \qquad ③$$

其中,r_k 表示源图像直方图的灰度,k 表示灰度级,n 是图像的总像素数,n_k 是灰度级为 r_k 的像素数量,$p_r(r_k)$ 是灰度级为 r_k 的概率,S_k 是已处理图像直方图的像素灰度,灰度级数目 $L = 256$,$T(r_k)$ 是源图像产生的均衡的累积直方图, int( )是取整函数.

基于自适应选择的阈值 Th,该直方图被分为 2 部分:像素灰度小于 Th 属于背景部分;像素灰度大于 Th 的部分属于热目标. 用背景直方图概率密度函数(PDF)(函数表达式为④)来建立混合累积直方图. 为提高对小目标的增强效果,定义强度密度函数(IDF)(函数表达式为⑥). 式⑥中的对数函数可防止对小目标的增强效果过强. 使用小目标直方图的 IDF 产生增强小目标的累积直方图(式⑤). 由表达式④和⑤可以构建混合累积直方图.

$$T_{hb}(r_k) = \begin{cases} \sum_{j=0}^{k} Arg\ \min\{p_r(r_j), p_r(r_{Th})\} \\ 0 \leqslant k < Th \qquad ④ \\ \sum_{j=Th}^{k} p_r(r_j) \times \log_2 j \\ Th \leqslant k \leqslant L-1 \qquad ⑤ \end{cases}$$

$$IDF = p_r(r_k) \times \log_2 k \qquad 0 \leqslant k \leqslant L-1 \qquad ⑥$$

其中,$T_{hb}(r_k)$ 是混合累积直方图,$p_r(r_{Th})$ 是像

素灰度级为 Th 的概率.

表达式 ④ 是将概率 $p_r(r_j)$ 和 $p_r(r_{Th})$ 的最小值求和,生成背景的累积直方图;表达式 ⑤ 是将概率 $p_r(r_j)$ 的 IDF 累加求和,生成热目标的累积直方图.

Δy 和 Δx 分别代表函数在 Y 轴和 X 轴上的增量,用 Δy 和 Δx 之比定义函数的导数. 其中,Y 代表函数的输出变量,X 代表函数的输入变量. 该导数代表了累积直方图的增强效果. 累积直方图的系数大,表示对图像增强的效果强;反之,增强效果弱.

将式 ① 与式 ④ 进行比较,当背景的灰度级是 r_k 时,很明显 ④ 的导数小于等于累积直方图的导数. ⑤ 代表小目标的累积直方图. 将式 ① 与式 ⑤ 比较,当小目标的灰度级是 r_k 时,⑤ 的导数大于等于 ① 的导数. 因此,小目标由累积直方图产生的增强效果更强,背景的增强效果更弱. 该结果由一阶导数的后向差分近似法基础上证明得到. 根据一阶导数的后向查分近似法,均衡的累积直方图一阶导数可以表述为

$$T'(r_k) = (T(r_k) - T(r_{k-1}))/(r_k - r_{k-1}) = \left(\sum_{j=0}^{k} p_r(r_j) - \sum_{j=1}^{k-1} p_r(r_j)\right)/1 = p_r(r_k)$$

这里,$T'(r_k)$ 是累积直方图 $T(r_k)$ 的一阶导数.

用同样的方法,混合累积直方图的一阶导数可以表述为

$$T'_{hb}(r_k) = (T_{hb}(r_k) - T_{hb}(r_{k-1}))/(r_k - r_{k-1}) = \begin{cases} Arg\ \min\{p_r(r_k), p_r(r_{Th})\} & 0 < t < Th \\ p_r(r_k) \cdot \log_2 r_k & Th \leqslant k < L-1 \end{cases}$$

这里,$T'_{hb}(r_k)$ 是 HCH $T_{hb}(r_k)$ 的一阶导数,如果背景的灰度级是 r_k,$0 < k < Th$ 时,则产生

$$Arg\ \min\{p_k(r_k), p_r(r_{Th})\} \leqslant p_r(r_k)$$

所以,$T'_{hb}(r_k) \leqslant T'(r_k)$.

由混合累积直方图 ④ 对背景的增强效果小于等于由均衡的累积直方图产生的效果. 当目标的灰度级是 r_k,$Th \leqslant k < L-1$ 时,则产生 $p_r(r_k) \cdot \log_2 r_k \leqslant p_r(r_k)$. 所以 $T'_{hb}(r_k) \geqslant T'(r_k)$. 这说明由混合累积直方图 ⑤ 对目标的增强效果大于等于由均衡的累积直方图产生的效果. 由此得到结论,自适应直方图的增强效果大于等于由均衡的累积直方图产生的效果. 另外,自适应直方图均衡化对于背景的增强效果小于等于直方图均衡化的效果. 在小目标相对于背景区域非常小的情况下,这种方法是十分有效的. 因此,本文提出的包含混合累积直方图的自适应直方图均衡化算法可以弥补直方图均衡化的不足.

此外,为了避免饱和变换像素的均衡灰度级,变换像素的最大灰度级必须小于或等于灰度级的最大值. 自适应直方图均衡算法的标准化表示为

$$T_n(r_k) = T_{hb}(r_k)/T_{hb}(r_{L-1}) \qquad 0 \leqslant k \leqslant L-1$$

$T_n(r_k)$ 是 $T_{hb}(r_k)$ 的标准化形式. 该方法使用迭代阈值选择方法从而得到自适应阈值,它适用于各种红外图像的直方图.

4 试验结果

为了验证文中提出的方法,用红外灰度图像作为测试样本,图像中的每个像素用 8 个比特位表示,所以每个像素有 256 个灰度级. 图 1 所示为原始图像. 为了使红外图像中的小目标更容易、快速、准确地识别出来,必须对图像进行增强处理.

用自适应阈值选择方法对图 1 中的原始图像进行处理,处理结果见图 2,得到自适应选择的阈值 $Th=37$;基于选择出的自适应阈值 Th,将原始图像分为背景和目标 2 个部分,用传统的直方图均衡化方法处理图 1,结果见图 3,利用本文提出的包含混合

图 1 原始图像

图 2 自适应阈值选择法处理结果($Th=37$)

累积直方图的自适应直方图均衡方法的处理结果见图 4. 图 5 和图 6 分别是经过直方图均衡化以及改进方法处理后的图像的直方图. 通过比较可以看出,对于小目标,改进的新方法比传统的直方图均衡化有更好的增强效果.

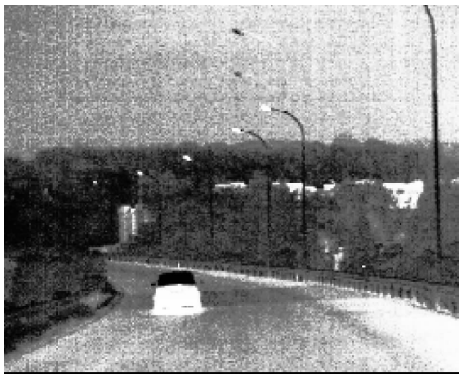


图 3　直方图均衡化后的图像

图 4　改进算法处理后的图像

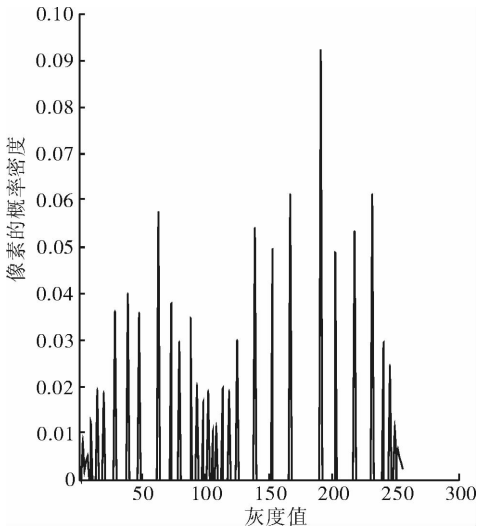


图 5　均衡化后的直方图

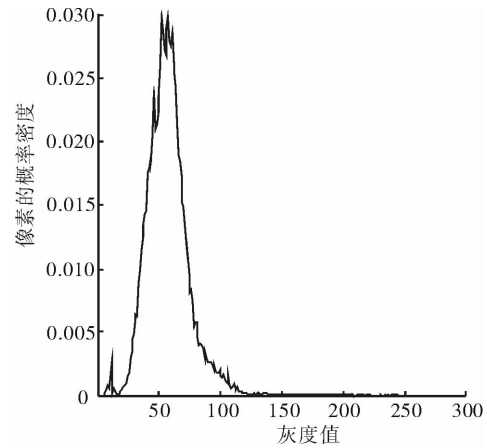


图 6　改进算法的直方图

5　结论

本文针对远距离捕获的红外图像对比度低、亮度低以及感兴趣的目标比较小等问题,提出了一种新颖的自适应直方图均衡的方法. 该方法利用图像直方图的 2 种不同信息,自动生成混合累积直方图,达到对红外图像进行自动均衡化的目的,有效地提高了红外图像的质量. 该方法有 2 个优势: 1) 含混合累积直方图的自适应直方图均衡方法是在直方图的概率密度函数和强度密度函数的基础上增强红外图像,与大背景相比,其对小物体的增强效果更强; 2) 使用该方法无需关于红外图像的先验信息,也无需手动预设参数. 实验结果证明该方法可获得高质量的红外图像.

参考文献:

[1] Dawound A, Alam M S, Bal A, et al. Target tracking in infrared imagery using weighted composite reference function-based decision fusion[J]. IEEE Transactions on Image Processing, 2006, 15(2): 404.

[2] Bal A, Alam M S. Automatic target tracking in FLIR image sequences using intensity variation function and template modeling[J]. IEEE Transaction on Instrumentation and Measurement, 2005, 54(5): 1846.

[3] 范新南,郭建甲. 一种新的自适应工程图像分割算法[J]. 计算机测量与控制, 2006, 14(3): 395.

[4] Qiao Yu, Hu Qingmao, Qian Guoyu, et al. Thresholding based on variance and intensity contrast[J]. Pattern Recognition, 2007, 40(2): 596.

[5] Bazi Y, Bruzzone L, Melgani F. Image thresholding based on the EM algorithm and the generalized Gaussian distribution[J]. Pattern Recognition, 2007, 40(2): 619.

[6] Sonka M, Hlavac V, Boyle R. Image Processing, Analysis, and Machine Vision[M]. 2ed. New York: PWS Publishing, 1999: 128 – 130.

文章编号:2095-476X(2012)06-0082-04

比赛视频中球员号码的定位与识别

徐晓煜, 黄欢, 杨小娜, 何冠雄

(昆明理工大学 信息工程与自动化学院, 云南 昆明 650500)

摘要:提出了比赛视频中球员号码的定位与识别方法:通过图像分割方法提取球员球衣上的号码区域,利用内部轮廓检测对号码定位,然后对提取出的号码采用基于概率统计的贝叶斯分类器进行识别. 经实验验证此方法实用且有效.

关键词:图像分割;轮廓检测;分类识别;球员号码识别

中图分类号:TP391.41 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.022

Player number localization and recognition in sports video

XU Xiao-yu, HUANG Huan, YANG Xiao-na, HE Guan-xiong

(*Faculty of Information Engineering and Automation*,
Kunming University of Science and Technology, *Kunming* 650050, *China*)

Abstract: A method for player number localization and recognition was proposed. It can achieve the recognition of player identity. This method extracts the number region on the jersey by image segmentation and locates the number using internal contours detection. Finally, it can recognize the extracted number based on the probability statistics Bayesian classifier. The experiment verified that this method is practical and effective.

Key words: image segmentation; contours detection; classification recognition; player number recognition

0 引言

基于内容的视频分析方法,在体育运动视频研究领域具有很大的商业价值. 探索基于比赛视频的自动分析方法,也因此成为备受关注的研究热点[1]. 而现有方法并不能简单快速地检测出球员位置,对球员号码的定位与识别的准确率也不高. 因此,寻求自动识别球员的方法,成为目前亟待解决的问题. 在球类比赛中,球员的自动识别对比赛视频检索和提供某位球员球场表现信息具有重要作用[2-3],而球员球衣上的号码为区分与识别球员提供了可能性.

本文拟在现有方法基础上,提出一种比赛视频中球员号码的定位与识别方法,以期为特定球员跟踪以及二维动画的形成奠定基础.

1 球员号码定位

本文首先利用图像分割的方法,对比赛视频进行帧图像处理,提取出球衣上的号码区域,再根据球衣与号码区域在HSV颜色空间的不同表现[4],使用图像分割技术结合内部轮廓检测的方法,对球员的号码进行定位.

1.1 图像分割

图像分割的质量直接决定着号码定位的效果,因此,合适的图像分割方法至关重要. 依照分割时所依据的图像特性不同,图像分割方法大致可分为

收稿日期:2012-07-05

作者简介:徐晓煜(1986—),男,陕西省岐山县人,昆明理工大学硕士研究生,主要研究方向为图像处理和模式识别.

阈值法、边界分割法以及区域提取法[5-7].

由于球员的球衣与号码之间存在明显的色彩差异,结合球场与球员球衣的颜色特点,本文采用聚类[5]的阈值分割算法对其进行区分.

所谓聚类方法,是采用模式识别中的聚类思想,以类内保持最大相似性以及类间保持最大距离为目标,通过迭代优化获得最佳的图像分割阈值. 在进行图像分割之前,首先对灰度化的图像进行对比度的展宽,提高图像的对比度,使下一步的分割具有更好的效果. 聚类法阈值分割的具体步骤如下.

1)给定一个初始阈值 $Th = Th_0$(默认为 128),将原图像分为 C_1 和 C_2 两类.

2)分别计算两类的类内方差

$$\sigma_i^2 = \sum_{(x,y)\in C_i} (f(x,y) - \mu_i)^2 \qquad i = 1,2$$

$$\mu_i = \frac{1}{N_{C_i}} \sum_{(x,y)\in C_i} f(x,y) \qquad i = 1,2$$

其中,μ_1 和 μ_2 分别为 C_1 和 C_2 的中心灰度值,N_{C_i} 为第 i 类中的像素个数.

3)进行分类处理,如果

$$| f(x,y) - \mu_1 | \leqslant | f(x,y) - \mu_2 |$$

则 $f(x,y) \in C_1$,否则 $f(x,y) \in C_2$.

4)对上一步重新分类后得到的 C_1 和 C_2 中的所有像素,分别计算其各自的均值与方差.

5)计算两类的发生概率分别为

$$P_1 = \sum_{i=0}^{Th} P_i \qquad P_2 = 1 - P_1$$

其中,P_i 为图像所有像素的分布概率. 如果有

$$[P_1\sigma_1^2 + P_2\sigma_2^2] |_{Th(t-1)} \leqslant [P_1\sigma_1^2 + P_2\sigma_2^2] |_{Th(t-2)}$$

则输出计算得到的阈值 $Th(t-1)$,否则返回 4).

采用聚类的阈值分割方法对图 1 所示灰度图像进行处理,得到分割阈值 $Th = 235$,分割图像如图 2 所示. 由图 2 可知,该方法可以很好地将球衣与号码区域区分开来,但如果不能得到最佳的分割阈值,将会降低球员号码定位的准确率.

1.2　内部轮廓检测

通过对图像进行分割,将球员的球衣与号码区域分开后,再对号码区域进行内部轮廓检测,从而实现对球员号码的位置定位. 图像的轮廓不同于边缘,图像的边缘信息包含所有的轮廓信息,而轮廓包含比位置更多的信息,从图像的轮廓可识别大量的物体[8].

轮廓提取在许多智能视觉系统中特别是模式识别中是非常重要的过程. 轮廓提取方法主要是利用边缘检测算子进行边缘的提取,然后根据目标物体的轮廓特点去除杂散的冗余边缘并进行边缘的修补. 通过对分割后的图像进行轮廓检测,轮郭检测效果如图 3 所示.

对轮廓检测后的图像再进行数字特征匹配,使其中相关系数最大,这样就可以把号码与非号码区域区别开来,从而实现对球员号码的定位,并为下一步号码识别奠定基础,号码定位效果如图 4 所示.

图 1　原始图像

图 2　分割图像

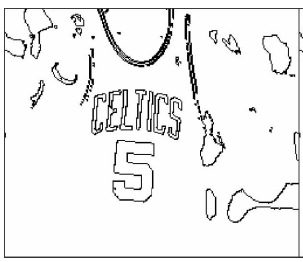


图 3　轮廓检测

图 4　号码定位

2　球员号码识别

根据号码定位,结合分类识别的方法,即可实现球员的号码识别. 识别之前的预处理对识别的正确率有重要影响. 若直接把提取出来的号码作为分类器的输入量进行分类计算,数据量太大;同时由于摄像机角度问题可能出现球员号码的倾斜,会对号码的识别产生一定的影响. 因此,首先需要对号码图像进行校正,可通过图像旋转来实现;然后,通过特征提取,把号码的结构特征提取出来,把反应数字特征的关键信息提供给分类器,这样就大大地减少了数据量.

2.1　图像旋转

号码识别效果如图 5 所示. 在图中确定号码区域后,计算号码上端和下端重心坐标分别为 $A(x_1, y_1)$ 和 $B(x_2, y_2)$,则其旋转角度为

$$\theta = -\tan^{-1}\left(\frac{x_1 - x_2}{y_2 - y_1}\right)$$

其中,负号表示逆时针旋转. 经计算,图 5 中 a)图的旋转角度 $\theta = -4.5°$,旋转效果如图 5 中 b)图所示.

2.2 特征提取

号码识别的特征提取极大程度地影响着分类器的设计和性能[9],以及识别的效果和效率. 对数字号码特征提取的方法很多,通常有逐像素特征提取法、骨架特征提取法及垂直方向数据统计特征提取法等. 本文采用简单的模板法对待测样本提取特征,其特征提取模板的建立步骤如下:

1)搜索数据区,找出号码数字的上下左右 4 个边界;

2)将搜索到的号码区域平均分成 5 × 5 共 25 个小区域;

3)计算 5 × 5 的每个小区域中黑像素所占比例,第 1 行的 5 个比例值是特征的前 5 个,第 2 行对应着特征的 6—10 个,依此类推.

为方便后续的处理,对定位后的号码区域图像进行反色处理,使号码区域为黑色,背景区域为白色. 然后再对反色后的图像进行旋转以及特征提取处理,便于下一步的识别. 特征提取效果如图 5 中的 c)图所示.

a)号码的提取 b)旋转 c)特征提取

图 5 号码识别效果图

2.3 分类识别

对号码进行特征提取之后,即可进行号码的识别. 由于数字本身的特点,本文采用基于最小错误率的贝叶斯分类器[9]进行号码数字的分类. 其具体过程如下.

1)求每一类号码数字样本的均值

$$\bar{X}_i = \frac{1}{N}\sum_{j=1}^{N_i} X_{ij} = (\bar{x}_{i1}, \bar{x}_{i2}, \cdots, \bar{x}_{in})^{\mathrm{T}}$$

$$i = 0,1,\cdots,9$$

其中,N_i 代表 ω_i 类的样品个数,n 代表特征数目.

2)求每一类的协方差矩阵

$$s_{jk}^i = \frac{1}{N_{i-1}}\sum_{i=1}^{N_i}(x_{lj} - \bar{x}_j)(x_{lk} - \bar{x}_k)$$

$$j,k = 1,2,\cdots,n$$

其中,l 代表样品在 ω_i 类中的序号,$l = 0,1,\cdots,N_i$;x_{lj} 代表 ω_i 类第 l 个样品的第 j 个特征值;$\bar{x}_j$ 代表 ω_i 类 N_i 个样品第 j 个特征的平均值;x_{lk} 代表 ω_i 类的 l 个样品第 k 个特征值;$\bar{x}_k$ 代表 ω_i 类 N_i 个样品第 k 个特征值的平均值. ω_i 类的协方差矩阵为

$$\boldsymbol{S}_i = \begin{Bmatrix} s_{11}^i & s_{12}^i & \cdots & s_{1n}^i \\ s_{21}^i & s_{22}^i & \cdots & s_{2n}^i \\ \cdots & \cdots & \cdots & \cdots \\ s_{n1}^i & s_{n2}^i & \cdots & s_{nn}^i \end{Bmatrix}$$

3)计算出每一类的协方差矩阵的逆矩阵 $\boldsymbol{S}_i^{-1}$ 以及协方差矩阵的行列式 $|\boldsymbol{S}_i|$.

4)求每一类的先验概率

$$P(\omega_i) \approx N_i/N \qquad i = 0,1,\cdots,9$$

其中,$P(\omega_i)$ 为类别为号码数字 i 的先验概率,N_i 为号码数字 i 的样品数,N 为样品总数.

5)将各个数值代入判别函数

$$h_i(x) = \frac{1}{2}(X - \bar{X}_i)^{\mathrm{T}}\boldsymbol{S}_i^{-1}(X - \bar{X}_i) + \ln P(\omega_i) - \frac{1}{2}\ln|\boldsymbol{S}_i|$$

6)判别函数最大值所对应的类别就是号码数字的类别.

应用基于最小错误率的贝叶斯分类器,对号码识别的结果见图 6.

图 6 球员号码识别结果

3 试验结果及分析

为了验证本文提出的球员号码定位与识别方法的有效性,选取了 NBA 凯尔特人队球员主客场比赛视频的帧图像,在 Visual C + +6.0 下编程实现.

本次试验对于主客场不同颜色的球衣,分别选取了 100 帧图像对本文提出的方法进行验证. 客场为绿色队服白色号码,定位率 78%,识别率 53%;主

场为白色队服绿色号码,定位率81%,识别率57%.结果表明,本文提出的方法成功地识别出了球员的号码,且与文献[4]的方法相比较,提高了定位与识别的准确率.

相比之下,主场视频的号码定位与识别效果比客场视频的识别效果稍好,这是由于球衣本身的特征以及灯光对主客场不同颜色球衣的影响所致.

4 结语

本文利用图像分割及轮廓检测方法,对球类比赛视频中球员号码进行定位,然后通过分类识别技术实现对球员号码的识别,从而实现了对球员身份的鉴别,为特定球员的跟踪、比赛的自动分析以及二维动画的形成奠定了基础.但是,由于在激烈的比赛中,球员的快速移动、身体的扭曲及球员之间的遮挡,再加上球场的照明问题等,会给球员号码的定位和识别造成一定的困难.在这种情况下,只有采用多个同步的摄像头,通过不同的视角,对同一时刻不同视角的帧图像加以分析,才能实现对每一名球员号码的提取,从而更准确地对球员身份进行鉴别.

参考文献:

[1] Ye Qixiang, Huang Qingming, Jiang Shuqiang, et al. Jersey number detection in sports video for athlete identification [C]//Visual Communications and Image Processing 2005 (VCIP 2005), Beijing: SPIE, 2005: 1599.

[2] Bertini M, Bimbo A D, Nunziati W. Player identification in soccer videos[C]//Processings of the 7th ACM SIGMM International Workshop on Mutimedia Information Retrieral, NewYork: ACM, 2005: 25 - 32.

[3] Kumar R K, Grundmann M, Kihwan K, et al. Player localization using multiple static cameras for sports visualization [C]//2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco: IEEE Press, 2010: 731.

[4] Matko, Hrvoje, Vladan, et al. Player number localization and recognition in soccer video using HSV color space and internal contours[C]//The International Conterence on Signal and Image Precessing, Heidelberg: WASET, 2008: 531 - 535.

[5] 朱虹. 数字图像处理基础[M]. 北京: 科学出版社, 2005: 125.

[6] 林开颜, 吴军辉, 徐立鸿. 彩色图像分割方法综述[J]. 中国图像图形学报, 2005, 10(1): 1.

[7] 刘堂海, 程小平. 篮球视频中球员的分割与跟踪算法[J]. 计算机工程与应用, 2009, 45(35): 243.

[8] 邹柏贤, 林京壤. 图像轮廓提取方法研究[J]. 计算机工程与应用, 2008, 44(25): 161.

[9] 冯伟兴, 唐墨, 贺波, 等. Visual C++数字图像模式识别技术详解[M]. 北京: 机械工业出版社, 2010: 239 - 255.

文章编号:2095-476X(2012)06-0086-04

Windows下BIOS Bootkit检测系统设计

王文冰[1], 范乃梅[1], 刘胜利[2]

(1.郑州轻工业学院 软件学院, 河南 郑州 450001;
2.解放军信息工程大学 信息工程学院, 河南 郑州 450002)

摘要:为了对新型高隐藏性木马BIOS Bootkit实现快速检测、准确定位,提出一种BIOS Bootkit检测方案:IBBDS存放于引导盘,以尽早获取系统的执行权限,通过对IVT模块、ISA模块和HOOK INT 13H模块的检测,在系统的启动过程即实现对BIOS Bootkit的捕获.试验验证了该检测方法的有效性.

关键词:BIOS Bootkit;Windows;安全防护;高隐藏性木马
中图分类号:TP393.08 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.023

Design of detection system for BIOS Bootkit in Windows

WANG Wen-bing[1], FAN Nai-mei[1], LIU Sheng-li[2]

(1. *College of Software, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*;
2. *College of Information Engineering, People's Liberation Army Information Engineering University, Zhengzhou* 450002, *China*)

Abstract: In order to quickly detect and accurately locate a new deeply concealed Trojan Horse Bootkit, the design of exclusive detecting system to BIOS Bootkit—IBBDS was put forward: IBBDS deposited in the bootable disk, to get the system implementation authority as soon as possible, the BIOS Bootkit capture was realized in the system start-up through the detection for IVT, ISA and HOOK INT 13H module. The validity of this detection method was verified with experiment.

Key words: BIOS Bootkit; Windows; security protection; deeply concealed Trojan Horse

0 引言

在不断发展的恶意代码技术中,Rootkit是新型、高级、隐蔽、强壮的代名词,其推动了恶意代码的不断更新,使恶意代码种类在短时间内迅速壮大[1],且隐藏更加深入、危害更为严重.在2008年左右,以卡巴斯基和瑞星为代表的安全软件已经能够成功识别并防御大量Rootkit,使得Rootkit的生存空间日渐狭小.Rootkit的没落并没有使攻击者放弃.恶意代码技术与检测的较量其实是系统权限的争夺战,谁加载启动得越早就越占优势.攻击者不断挑战极限,向系统更底层深入,将加载运行Rootkit的时机置于内核启动之前,利用开机启动过程注入并隐藏恶意代码,导致比Rootkit更高级、利用开机启动过程进行系统内核注入并隐藏代码的BIOS Bootkit的诞生[2].

根据启动时机不同,Bootkit可以分为许多种类.其中,BIOS Bootkit是通过修改系统BIOS内容来实现自身隐藏的一类Bootkit,其危害程度最高,即使重装和升级系统都无法将之去除.BIOS Bootkit作为

收稿日期:2012-10-19
基金项目:河南省科技攻关项目(102102210544)
作者简介:王文冰(1978—),女,河南省郑州市人,郑州轻工业学院讲师,硕士,主要研究方向为信息安全与密码学.

新兴技术,相关参考资料有限,且与计算机底层联系紧密,所以从实现机制到检测技术的研究都存在一定的难度. 但针对 BIOS Bootkit 的反恶意代码技术在近几年也取得一定的进展. 微软公司 MSN 安全组的 Scanmbray Joel 和 McAfee 的 McClure Stuart 提出通过 TPM(trusted platforill module)芯片技术对 BIOS 进行加载检验,有效防御嵌入在 BIOS 中的 Bootkit. 但对于没有采用 TPM 芯片校验的主机不具备保护能力[3]. 此外,中科院高能物理研究中心计算研究所通过研究 BIOS Bootkit 的实现机制[4],针对 BIOS Bootkit 恶意代码设计了相应的检测系统.

相对目前发展较快的 Bootkit 攻击技术,对其进行检测的技术总体表现乏力. 上述针对 BIOS Bootkit 的检测技术各有利弊,检测能力和范围也十分有限. 鉴于此,本文拟深入剖析 BIOS Bootkit 的攻击原理,全面结合特征码检测、动态行为检测、完整性检测和逆向分析方法,设计一种全新的 BIOS Bootkit 检测方案.

1　BIOS 引导过程概述

BIOS Bootkit 加载于计算机系统启动流程中的 BIOS 模块. 为了能对 BIOS Bootkit 进行有效检测,BIOS 引导过程的分析就必不可少. BIOS 被称为基本输入输出系统,其固化于计算机主板的 ROM 芯片,存放了计算机中最重要的输入输出程序、系统设置信息、开机自检程序和系统自启动程序. 在没有操作系统的情况下,BIOS 能提供最基本的操作,启动流程如图 1 所示.

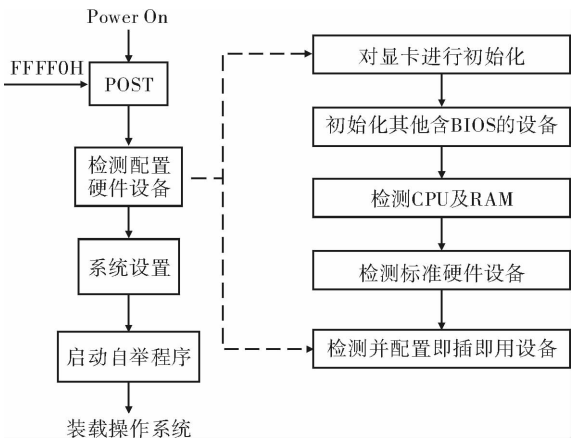


图 1　BIOS 的启动流程

BIOS 引导过程分为 4 步[5]:

1)机器上电稳定后,CPU 马上执行内存 FFFF0H 处的指令,即 BIOS 程序的上电自检测 POST(power on self test)部分,对系统的一些关键设备进行检测,确定其是否存在及能否正常工作.

2)上电自检测结束之后,BIOS 程序会对硬件进行全面检测并进行初始化.

3)进行系统设置,更新储存于 CMOS 中的扩展系统配置数据 ESCD(extended system configuration data).

4)进行操作系统自举. BIOS 程序调用 INT 19H 中断,执行自举程序,根据用户指定的加载顺序启动硬盘、软盘或光盘中的 MBR,将 MBR 加载到内存的 0X7COO 地址处,此时 BIOS 将控制权限移交给 MBR,开始执行加载操作系统等操作.

2　BIOS Bootkit 的攻击原理

BIOS Bootkit 是通过修改系统 BIOS 内容来实现自身隐藏的一类 Bootkit,经过对其样本剖析,总结出 BIOS Bootkit 的攻击原理如图 2 所示.

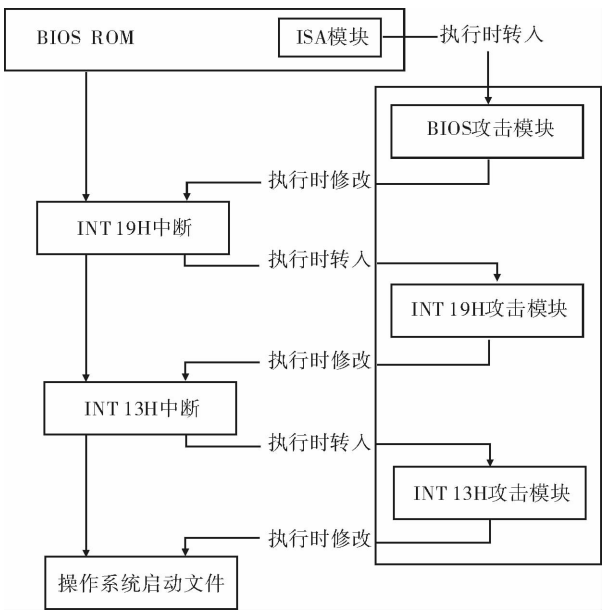


图 2　BIOS Bootkit 的攻击原理

BIOS Bootkit 会以易于编写的 ISA 模块的形式加入到 BIOS 中,当系统 BIOS 对 ISA 模块进行初始化时,BIOS Bootkit 借机获取系统控制权限,加载并执行自身程序. 此后的执行过程中主要由 3 个攻击模块组成,分别是 BIOS 攻击模块、INT 19H 攻击模块和 INT 13H 攻击模块.

系统在初始化 ISA 模块之后,BIOS 会调用 INT 19H 进行自举. 为了顺利地获取和回收控制权,BIOS攻击模块会修改 INT 19H 中断,从而获得 INT 19H 中断执行时的系统控制权限,并将启动流程转

入 INT 19H 攻击模块. 由于此后的启动过程不再调用 INT 19H,所以要使代码重新获得执行,必须进下一级挂钩. 由于在启动过程中需要经常调用 INT 13H读写磁盘,因此可以在 INT 19H 攻击模块中再次挂钩 INT 13H. 这样,当调用 INT 13H 读写磁盘时,代码又可以重新获得执行权,转入 INT 13H 攻击模块. 当启动过程运行到启动文件 NTLDR 时,系统将会切换到 32 位保护模式下,这时将不再使用 INT 13H 服务,代码若需要再次获得执行权,必须对 NTLDR的 32 位部分进行挂钩,即挂钩 OSloader. 这个工作将在 INT 13H 攻击模块中进行,最后通过修改操作系统的启动文件获得系统的执行权限.

由此看来,BIOS Bootkit 通过向 BIOS 中插入 ISA 模块来获得系统启动控制权,然后以 HOOK 的方式将控制权不断地传递下去,篡改操作系统启动模块,执行对整个系统启动过程的修改,加载并执行自身的恶意功能模块,实现恶意功能.

3 BIOS Bookit 检测系统设计

通过 BIOS Bootkit 的攻击原理的分析可以发现:BIOS Bootkit 一般通过截获系统正常工作流程,进而插入自身代码来实现特定功能. 因此获取系统执行权限并成功隐藏自身是决定 BIOS Bootkit 能否成功实现攻击的关键. 反之,定位并获取 BIOS Bootkit 篡改的模块,使用特征码检测法或完整性检测法判断恶意代码的存在性是检测者能否成功检测 BIOS Bootkit 的制胜点. 基于上述分析,本文提出一个针对 BIOS Bootkit 的检测系统 IBBDS(inserted-attack-oriented BIOS Bootkit detection system)的设计方案. 为了保证 IBBDS 的启动早于 BIOS Bootkit,IBBDS 需存放于引导盘,以尽早获取计算机系统执行权限,从而实现对计算机系统启动过程的实时监控. IBBDS 采用动态跟踪计算机系统启动的方法,对整个启动过程中的行为进行检测,并在启动结束后将检测结果进行分析,判断系统是否遭受了 BIOS Bootkit 的攻击. 相对于普通杀毒软件,IBBDS 对 BOIS Bookit 检测的优势在于其在系统启动过程中就可以捕获 BIOS Bootkit 的行为,具有更强的检测能力且结果更加可信.

3.1 IBBDS 系统框架

假设 IBBDS 存储于光盘中,且此光盘已插入计算机的光驱. 计算机加电开机后,BIOS 程序先被执行,由其加载执行光盘中的 IBBDS. IBBDS 首先对先于自身启动的 BIOS 启动模块进行检测. 由于 IVT 和 ISA 模块往往是 BIOS Bootkit 篡改存储的位置,因此检测系统采用检测 IVT,ISA 模块的方式对 BIOS启动模块进行检测. 为了能够在系统启动过程中始终保持控制权,IBBDS 对 INT 13H 进行 HOOK,执行 HOOK 时对每一个被加载的模块进行检测并记录日志.

基于 IBBDS 的工作流程,按照功能和工作环境划分,检测系统由 3 部分构成.

1)检测系统引导模块. 主要负责将 IBBDS 从光盘引导加载进内存,并为检测系统开辟长期存放的空间.

2)启动检测模块. 主要工作是对计算机系统启动过程中的启动模块 BIOS 进行检查.

3)日志记录模块. 主要负责汇总 IVT 检测模块、ISA 检测模块的检测结果,并将结果通报给用户,从而根据操作系统的版本为用户提供合适的恢复方案.

以上 3 个模块中,承担检测系统是否存在 BIOS Bootkit 攻击重任的是启动检测模块,它分别对系统启动过程中所访问的 IVT 模块、ISA 模块和 HOOK INT 13H 模块进行检查.

3.2 启动检测模块

3.2.1 IVT 模块检测 IVT 即中断向量表. 中断向量是中断处理程序的入口地址,而中断向量表则是中断处理程序入口地址的列表[6]. 中断调用表示为 INT n 的形式,其中 n 是中断向量号. 每个中断向量占用 4 个字节,因此通过计算可以在中断向量表里找到各个中断号对应的内存地址. 攻击者正是利用这一点,将中断表中的地址改掉以达到 HOOK 的目的. 其中 INT 13H 中断在系统启动过程中频繁地被用到,也是经常被攻击者所利用的对象;INT 19H 中断常被 BIOS Bootkit 作为 HOOK 对象用于权限传递. 因此检测系统对 IVT 中的关键地址进行检查是必要的.

对每个中断向量的具体地址进行检查的方法不可取,因为不同操作系统版本下中断向量的地址是不同的. 通过分析发现,中断向量存放在内存的高地址中,也就是 C000:0000 ~ FFFF:0000 的地址之间,可以利用这个条件对 IVT 表进行检测. 将地址 C000:0000 作为阈值,与向量表中的地址进行比较:如果小于 C000:0000,则说明存在 Bootkit;大于或等于,则说明不存在 Bootkit.

3.2.2 ISA 模块检测 ISA(industrial standard architecture)是 8/16 b 的系统总线. 由于传输速率过

低、CPU 占用率高、占用硬件终端资源等原因,ISA 总线标准目前已被 PCI(periphearl component interconnect)总线标准所替代. 为了保持新旧主板的兼容性,BIOS 中仍然保留有对 ISA 调用的接口,这为 BIOS Bootkit 的存在提供了有利条件,同时由于 ISA 模块具有编写简单、向下兼容性好的特点,目前 BIOS Bootkit 以 ISA 模块的形式存在于 BIOS 中. 因此,IBBDS 以 ISA 模块的存在性作为检测 BIOS Bootkit 是否存在的依据.

计算机开机后,主板 BIOS 镜像会先存储于 4 G 内存的高地址,由于此时计算机正处于实模式下,无法访问 4 G 内存的高地址,因此需要转换到保护模式,打开 4 G 内存,才能够实现对 ISA 模块的检查.

进入保护模式后,IBBDS 拥有了 4 G 内存空间的访问权限,而目前主板 BIOS 的大小通常在 1 M 以内,所以只需对 0XFFF00000—0XFFFFFFFF 地址范围内的内容进行检测即可.

对 BIOS 内存范围的检测是为了搜索其是否存在 ISA 模块,下面以 Award BIOS 为例进行说明. Award BIOS 的各功能模块都经过 LHA 类型 5 压缩算法压缩,各模块头结构中具有 -lH5- 标记,并且每个模块具有唯一的 Type ID:0XA440. 所以对 ISA 模块存在性的检测就是基于对 0X6C,0X68,0X35,0XA4,0X40 这 5 个特征码的搜索,其中 0X6C,0X68,0X35 是压缩模块的特征码,对应 lH5. 通过对特征码的搜索可以判断是否存在 ISA 模块,进而可以判断是否存在 BIOS Bootkit.

ISA 模块检测完毕后,检测系统必须关闭 A20 总线,转换回实模式.

3.2.3　HOOK INT 13H 模块　HOOK INT 13H 模块的主要目的是对中断向量表中的第 13 号中断地址进行修改,使得每次 INT 13H 被执行的时候 IBBDS 就会获得控制权,对下一模块进行安装 HOOK 的操作,以便完成下一步的检测工作.

该模块首先将中断向量表中 INT 13H 的地址取出并保存到一个指定的地址里,然后将检测代码的地址写入中断向量表中 INT 13H 的位置,这样就可以使操作系统每次调用 INT 13H 中断时都会先调用检测程序. 由 IBBDS 获得系统控制权限后,利用该权限安装下一启动模块的 HOOK,进而通过安装的 HOOK 代码完成启动模块的检测任务. 由此可见,HOOK INT 13H 模块实现了检测系统权限传递的功能.

基于此设计方案所做的原型系统分别经过适应性、兼容性、有效性 3 个方面的测试,结果如下:适应性方面,IBBDS 适用于 Windows 2000, Windows Vista 和 Windows 7 这 3 种版本的操作系统,并且它的使用对操作系统启动速度的影响在 1 min 左右,可以为用户接受;兼容性方面,IBBDS 作为一款恶意代码的检测工具,可以与市场上主流的检测工具卡巴斯基、诺盾、360 安全卫士等共存;有效性方面,IBBDS 能够对本文选取的 3 款公开的 BIOS Bootkit 样本(eeye,stoned 和 vbootkit)进行有效检测.

4　结论

为了能够成功检测 BIOS Bootkit,对 BIOS Bootkit 攻击原理进行深入分析,设计了具有针对性的 IBBDS 检测系统. IBBDS 存放于引导盘,以尽早获取系统的执行权限,通过对 IVT 模块、ISA 模块和 HOOK INT 13H 模块的检测,在系统的启动过程即实现对 BIOS Bootkit 的捕获. 该系统能够很好地检测 BIOS Bootkit,但对于该系统引导模块与日志记录模块的研究还不够完备,是今后工作中需要完善的重点.

参考文献:

[1]　王雷,凌翔. Windows Rootkit 进程隐藏与检测技术[J]. 计算机工程,2010,36(5):140.

[2]　朱瑜,刘胜利,陈嘉勇,等. 针对插入攻击型 Bootkit 的分析及检测[J]. 小型微型计算机系统,2012,33(7):1462.

[3]　Stuar Mc Clure,Joel Scambray. Hacking Exposed Network Security Secrets and Solutions[M]. New York : McGraw-Hill/Osborne,2012:512 - 576.

[4]　王晓箴,刘宝旭,潘林,等. BIOS 恶意代码实现及其检测系统设计[J]. 计算机工程,2010,36(21):17.

[5]　陈文钦. BIOS 研发技术剖析[M]. 北京:清华大学出版社,2001:20 - 23.

[6]　郭彬. Windows 实时处理中断程序的设计[J]. 微型机与应用,1998,17(7):10.

文章编号:2095-476X(2012)06-0090-03

基于NS2的改善802.11b无线网络传输效果异常仿真研究

刘书如, 冯媛, 蔡增玉

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:针对多速率802.11b无线网络环境的数据传输效果异常现象,提出了一种在多速率无线网络环境下改善802.11b无线网络传输效果异常的新方法.该方法根据不同的数据传输速率同时调整数据帧及竞争窗口大小.仿真结果表明,该方法可以有效改善802.11b无线网络传输效果异常现象.

关键词:无线网络;传输效果异常;NS2;802.11b;数据帧

中图分类号:TP393 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.024

The simulation research of improved 802.11b wireless network transmission performance anomaly based on NS2

LIU Shu-ru, FENG Yuan, CAI Zeng-yu

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*)

Abstract: Aiming at the data transmission performance anomaly in multi-rate wireless local networks 802.11b, a new method was proposed which adjusts the size of data frame and contontion window according to the different data transmission rates to improve the performance anomaly in multi-rate wireless local networks, the simulation results showed that this method can effectively improve the performance anomaly in the 802.11b wireless network.

Key words: wireless network; transmission performance anomaly; NS2; 802.11b; data frame

0 引言

IEEE无线局域网802.11b的带宽最高可达11 Mb/s,也可根据实际情况采用不同的编码方式来支持5.5 Mb/s,2 Mb/s和1 Mb/s带宽,较低的传输速度对于信号的抗干扰性较强,较高的传输速度抗干扰性较弱.在实际传输时选择合适的传输速度的机制称为链路调适,常见的做法是利用不同的信号噪声比来判断网络状况.当信号不良时,选择较低的传输速度,当信号接收良好时则选择较高的传输速度.在无线网络里,距离越远信号越容易衰落,因此通常在距离较远的时候,为了改善信号的传输质量会采用较低的传输速度.这样在同一个无线局域网里,存在采用不同传输速度的节点,这种网络环境称为多重速率无线网络.在多重速率802.11b无线网络中,当一个无线节点的速率发生变化时,节点数据包大小不变,则占用信道的时间长度与自己的传输速度成反比.因此,当该节点的传输速度降低时,在相同的时间里,其他节点只能发送一个数据包,这种一个节点因为另一个节点速度降低而被迫

收稿日期:2012-10-22

基金项目:河南省自然科学基金项目(0611054800);河南省教育厅自然科学研究计划项目(2009A520032)

作者简介:刘书如(1979—),男,河南省桐柏县人,郑州轻工业学院讲师,硕士,主要研究方向为网络移动性管理.

降低传输速度的现象称为效果异常[1-2].

为了解决该问题,参考文献[3-5]提出了相应的解决方法. 主要的思路有 2 种:一种是改变数据帧的大小;另一种是改变初始竞争窗口的大小. 本文提出一种将两者相结合的方法,即同时改变数据帧和初始竞争窗口大小的方法,以期为进一步解决无线网络效果异常问题提供有益的参考.

1　传输效果异常仿真模型

为了验证上面描述的 802.11b 无线网络传输效果异常,在网络仿真软件 NS2 中建立仿真模型,拓扑结构如图 1 所示. 图 1 中共有 3 个节点 A,B,C,节点 C 即 AP. 网络参数设置为:节点间采用 802.11b 无线网络协议,数据包的大小为默认大小1 000 B;节点 B 设置为移动节点,仿真时间为 60 s,移动节点 B 在 0 ~ 20 s 时,传输速度为 11 Mb/s,20 ~ 40 s 的时候随着距离节点 C 越来越远,其传输速度变为 1 Mb/s,40 s 之后节点已经超出了无线网络的通信范围,此时节点 C 只能收到节点 A 的数据.

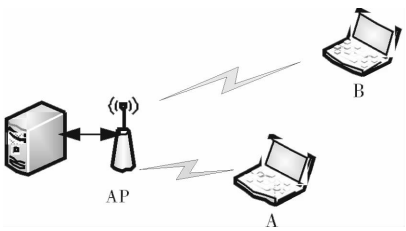


图 1　多重速率无线网络

图 2 为传输效果异常仿真结果. 可以发现,在 0 ~ 20 s,节点 A 和节点 B 发送的速度均为 2 Mb/s,此时,节点 C 收到的数据量分别来自节点 A 和节点 B,整体系统的效果约为 4 Mb/s. 在 20 ~ 40 s 时间内,节点 B 的速度降低为 1 Mb/s,很明显可以看出节点 B 的速度从原本的 2 Mb/s 降低至 1 Mb/s,同时虽然节点 A 的发送速度保持不变,但是由于受到节点 B 的影响,在接收端节点 C 处看来,其传输效果与节点 B 接近,约为 1 Mb/s,而系统整体的传输效果也降低至 2 Mb/s,这种现象就是无线网络传输效果异常. 在 40 s 以后,由于节点 B 超出通信范围,此时只有节点 A 发送,节点 A 又恢复到了原来的发送速度.

2　一种新的改善传输效果异常的方法

对无线网络效果异常现象的仿真可以明显看

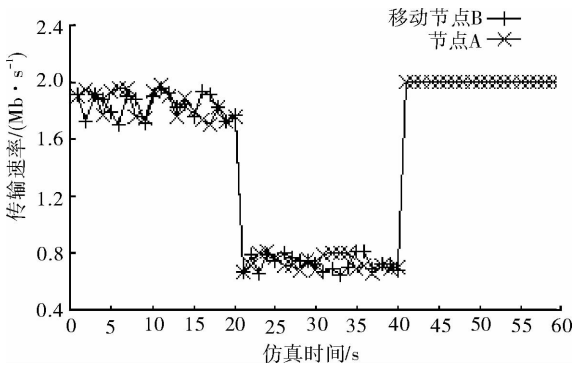


图 2　传输效果异常仿真结果

出,无线网络传输效果异常对无线网络系统的整体性能有着非常大的影响,严重地降低了系统通信的效率[6]. 要想解决该问题,主要的目标有 2 个,一个是要改善系统整体的通信效率,另一个是要考虑到差别化,不同的传输速度应有不同的传输效果,传输速度快的效果应该比传输速度慢的效果好.

2.1　新方法描述

本文提出的方法同时考虑到 2 种因素,首先让传输速度快的节点使用较大的数据帧,而传输速度较慢的节点使用较小的数据帧. 当速度慢的节点竞争到信道的使用权时,这样可以提升整体无线网络的传输性能,而且不会占用信道太多的时间,信道可以很快地开放给其他节点,也就是说随着速率改变数据帧的大小,从而改善无线网络数据传输效果异常现象. 其次是考虑随着无线节点速率的变化来改变竞争窗口大小,以改善 802.11b 无线网络传输效果异常现象. 具体可以描述为

$$C_{throughput} = P(MTU, CW)$$

这里 $C_{throughput}$ 为节点 C 的吞吐量,MTU 为数据帧的大小,CW 为竞争窗口大小. 也就是说,总的 $C_{throughput}$ 可以看作是 MTU 与 CW 的函数.

$$MTU = y \times DATA_{rate}$$

其中,y 是调节系数. 随着发送速度的增加,MTU 也相应变大.

无线节点竞争窗口的大小随不同速率的变化可以描述为

$$CW = CW_{\min} \times \frac{11}{DATA_{rate}}$$

$CW_{\min}$是初始的竞争窗口大小,默认是 32,从公式中可以得到竞争窗口的大小与节点本身的速度成反比. 对于 802.11b 来说,可以得到表 1 所示的竞争窗口大小变化.

表 1 竞争窗口大小变化表

新竞争窗口大小/B	速度/(Mb·s^{-1})	初始竞争窗口大小/B
32	11	32
64	5.5	32
176	2	32
352	1	32

2.2 仿真分析

为了验证本文提出方法的有效性,在采用上述仿真模型的基础上,分别建立 3 种仿真场景.

场景 1:随着移动节点 B 速度从 11 Mb/s 降低为 1 Mb/s,只改变数据帧的大小,仿真结果见图 3.

场景 2:随着移动节点 B 速度从 11 Mb/s 降低为 1 Mb/s,只改变竞争窗口的大小,仿真结果见图 4.

场景 3:随着移动节点 B 速度从 11 Mb/s 降低为 1 Mb/s,按照本文的新方法,仿真结果见图 5.

从仿真结果来看,场景 1 和场景 2 情况下都可以对无线网络的数据传输效果异常有一定改善作用,相比场景 1 和场景 2,场景 3 的仿真结果中明显可以看出,该方法可以有效地改善无线网络的数据传输效果异常,从而也表明该方法的有效性.

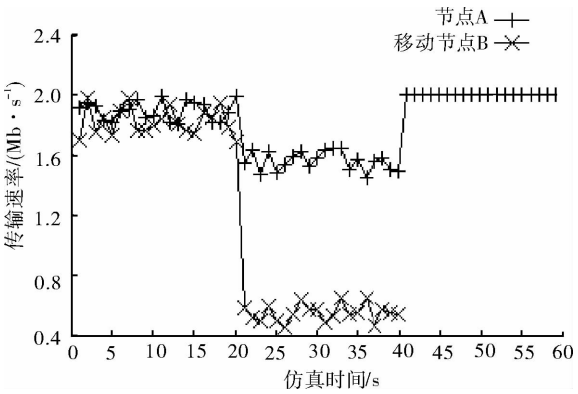


图 3 场景 1 仿真结果

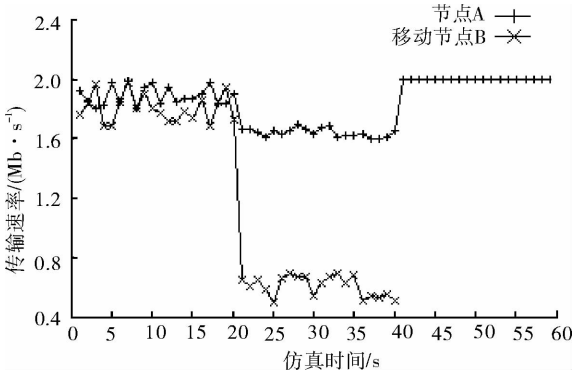


图 4 场景 2 仿真结果

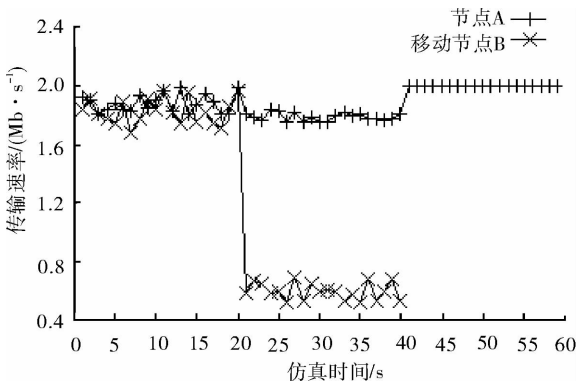


图 5 场景 3 仿真结果

3 结论

本文提出了一种新的改善无线网络效果异常的方案,该方案的关键点是随着速度的变化同时考虑改变数据帧大小及竞争窗口大小 2 个因素. 仿真结果表明该方法可以有效改善 802.11b 无线网络传输效果异常的状况.

参考文献:

[1] Heusse M, Rousseau F, Berger-Sabbatel G, et al. Performance anomaly of 802.11b[C]//INFOCOM 2003 Twenty-Second Annual Joint Conference of the IEEE Computer and Communications, San Francisco: IEEE Societies, 2003:836 – 843.

[2] Barford P, Duffield N, Ron A, et al. Network performance anomaly detection and localization [C]//INFOCOM IEEE, Brozil: Press Riode Janeiro, 2009:1377 – 1385.

[3] Yoo See-hwan, Choi Jin-Hee, Hwang Jae-Hyun, et al. Eliminating the performance anomaly of 802.11b[J]. Lecture Notes in Computer Science, 2005, 3421:1055.

[4] Razafindralambo T, Lassous I G, Iannone L, et al. Dynamic packet aggregation to solve performance anomaly in 802.11 wireless networks[C]//Proceedings of the 9th ACM International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems, New York: ACM, 2006:247 – 254.

[5] Sung J T, Ke C H, Chilamkurti N, et al. Collision avoidance multi-rate MAC protocol: Solving performance anomaly in multi-rate network[C]//Wireless Pervasive Computing ISWPC 2009 4th International Symposium, Melbourne: IEEE Conference Pubications, 2009:1 – 5.

[6] 黄家玮,王建新. 无线局域网中 TCP 公平性问题研究综述[J]. 计算机科学, 2009, 36(2):43.

文章编号:2095-476X(2012)06-0093-04

基于退火遗传算法的无线传感器网络路由优化研究

梁衡[1], 刘新新[2], 郑远攀[2], 徐二锋[3]

(1.许昌学院 计算机与科学技术学院, 河南 许昌 461000;
2.郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001;
3.弘润华夏大酒店, 河南 郑州 450002)

摘要:针对无线传感器中节点能量有限且网络拓扑结构不稳定的问题,提出了一种基于退火遗传算法寻求无线传感器网络最优路径的方法.该方法采用变长路径编码方式,综合考虑节点间通信消耗、通信距离和路径最短等因素,同时选择相应的退火遗传操作算子,通过优化选取种群、计算适应度函数、合理交叉、有效变异和降温退火操作,达到无线传感器网络最优路径的目标.仿真结果表明,基于退火遗传算法的无线传感器网络路由协议能够有效减少节点能耗,延长网络生存周期.

关键词:退火遗传算法;无线传感器网络;路由协议
中图分类号:TP393 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.025

Study on routing optimization for wireless sensor networks based on annealing genetic algorithm

LIANG Heng[1], LIU Xin-xin[2], ZHENG Yuan-pan[2], XU Er-feng[3]

(1. *School of Computer and Science Technology, Xuchang University, Xuchang* 461000, *China*;
2. *College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*;
3. *Hong Embellish the Chinese Hotel, Zhengzhou* 450002, *China*)

Abstract: Aiming at the problem of wireless sensor networks limited nodes energy and unstable network topology structure, a method of searching wireless sensor network optimal path was put forward based on annealing genetic algorithm, which adopts variable-length encoding mode, considers the inter-node communication consumption, communication distance and the shortest path and other factors, and selects the appropriate annealing genetic operators, through the optimal selection of stocks, the calculation of the fitness function, a reasonable cross-effective variation and cooling annealing operation, to achieve the objectives of the optimal path of the wireless sensor network. Simulation results showed that the algorithm can effectively balance node energy consumption, prolong the network survival cycle.

Key words: annealing genetic algorithm; wireless sensor networks (WSNs); routing protocol

0 引言

随着计算机技术以及通信技术的迅速发展,无线传感网络成为众多学者研究的重点.无线传感器网络是由监测区域内大量微型传感器节点组成的多跳自组织网络,能够将逻辑上的信息世界与真实

收稿日期:2012-06-09

作者简介:梁衡(1983—),男,河南省鄢陵县人,许昌学院助教,主要研究方向为计算机应用.

的物理世界结合起来. 由于其较强的容错性能以及自组织能力,通常被应用于人不可到达或危险的区域. 与传统无线网络相比,无线传感网络节点能量有限,网络结构也不稳定,因此其对无线传感器网络路由协议提出了更高的要求.

针对无线传感器网络的特点,国内外已有众多学者对其路由协议做了大量的研究. 张玉等[1]提出了通过改进遗传算法来提高其收敛性的无线传感器网络 QoS 路由算法,虽然能够满足带宽 - 时延要求的路由选择,但忽略了实际路由中路径不一定完全定长的特点. 高德民等[2]采用了变长染色体编码方式,并利用遗传算法实现了全局网络最优路径寻求的方法,延长了网络生存时间,但易陷入局部最优解. 本文根据无线传感网络路由节点的特点,综合遗传算法收敛速度快,模拟退火算法局部搜索能力强的特点,提出一种能够解决多约束优化问题的自适应网络最优算法,应用到无线传感器网络路由中.

1 路由模型

无线传感网络路由优化中,其连通性是首要解决的问题,即任意 2 个节点间必须存在有效路径. 为方便计算,可以将其描述为一种带权无向图[2] $G = \langle V, A \rangle$,源节点 v_1 到目标节点 v_n 间的链路集合表示为 $V = (v_1, v_2, \cdots, v_n)$,节点间各条通信链路的集合表示为 $A = \{a_1, a_2, \cdots, a_m\}$,节点 v_i 到 v_j 的距离表示为 $d(v_i, v_j)$,链路长度表示为

$$L = \sum_{v_i, v_j \in V} d(v_i, v_j)$$

无线传感器网络路由优化的另一个影响因素是能量消耗,如何减少路径中的能量消耗,延长网络生存周期是本文研究的另一个重要问题. 在传感器各个节点的传输过程中,通信能量消耗远远大于计算能量消耗,因此可忽略计算能量消耗,只考虑通信能量消耗[3]. 此外,为防止通信过程中随着通信距离的增加而导致能量消耗急剧增加,本研究采用多跳短距离无线通信方式,其关系表示为

$$E = kd(v_i, v_j)^n$$

其中,k 为常数,$2 < n < 4$. 考虑到发送端和接收端之间的距离不远,但有障碍物阻挡,干扰比较大,又受接收天线性能的影响,选取 $n \approx 4$.

2 算法模型

2.1 编码

针对网络路由中存在路径变长的情况,本文采用变长染色体编码方案,使用基于路径表示的编码方法. 染色体中的基因编号用节点 ID 号表示,用源节点到目标节点所经过的节点号序列表示路径,形成一个染色体. 染色体中的第 1 个基因表示源节点,最后 1 个基因表示目标节点,同一个数据包在同一个节点上只能转发 1 次,因此在进行路由转发时,每个节点最多只能转发 1 次. 事实上,数据包从源节点到基站的传输过程中经过的节点数是不固定的,基站到目标节点便形成不定长染色体,从源节点到目标节点形成的多个不定长染色体组成的个体称为路径种群.

2.2 适应度函数

搜索进化过程中的算法主要是确定适应度函数,其直接影响到算法的收敛速度和最优解的寻找. 一般情况下,根据目标函数来确定适应度函数,种群中总是选取适应度最大的作为遗传的父代[4],因此适应值函数的取值越大越好. 衡量路由算法性能的标准主要有网络延时、可靠性、网络生命周期等,通过多路径可以保证网络可靠性,通过基站性能可以调整网络延时. 延长网络生命周期,只有通过优化路由来减少网络耗能,才能提高适应值.

定义源节点到目标节点间的指示变量

$$x_{i,j} = \begin{cases} 1 & \text{边} \langle v_i, v_j \rangle \text{ 在路径中} \\ 0 & \text{边} \langle v_i, v_j \rangle \text{ 不在路径中} \end{cases}$$

针对路径最短问题,则要求链路长度最短,即

$$\min z(x) = \sum_i \sum_j d(v_i, v_j) x_{i,j}$$

路径耗能最少作为无线传感器网络路径协议的目标之一,即

$$\min E = \min \sum_{i,j} kd(v_i, v_j)^4$$

适应度函数可以表示为

$$F_i = \min z(x) \cdot \min E = \min \sum_i \sum_j d(v_i, v_j) x_{i,j} \cdot \min \sum_{i,j} kd(v_i, v_j)^4$$

2.3 选择算子

选择算子是决定群体中的个体能否被遗传的关键因素,适应度越高,个体被遗传到下一代群体的概率就越大;反之,概率就越小. 为避免陷入局部最优解,本文在初始种群基础上,采用最佳个体保留与轮盘赌选择[5-6]相结合的方法. 首先保留群体中适应度最高的 N 个个体,直接遗传到下一代群体中,然后根据适应度比率 $P_i = F_i \Big/ \sum_{i=1}^{n} F_i$($F_i$ 为第 i 个个体的适应度)计算出个体选择概率,选定作为遗传种群的个体. 这样不仅保留了种群中的优秀个

体,也维持了各代种群的多样性,降低了种群之间的相似性,提高了选择操作的效率[6].

2.4　交叉操作

交叉运算是指两配对染色体依据交叉概率按某种方式相互交换其部分基因,从而形成 2 个新个体的运算方法. 常用的交叉算子有一点交叉、二点交叉和多点交叉等方法. 考虑传感器网络路由的特殊性,本文采用一种特殊的交叉方法:随机产生两个种群个体,设置一个或多个基因相同处为交叉点,将个体染色体分成几个块,根据交叉概率 P_c 交换其交叉点部分染色体,产生新的个体,直到所有个体中不再出现重复基因为止.

2.5　变异

为保持群体的多样性、种群全局最优和局部路径收敛到最优,而且进化不会过早收敛,本文将局部路径看成基因块,采用顺序交换[7] 和随机交换方式来进行基因的变异操作. 把某基因位的值进行变异,以变异概率 P_m 随机改变其值,来改变父代的特性,产生新的个体. 随着变异率的逐步降低,算法收敛的速度逐步加快,算法的局部搜索能力逐步提高.

2.6　退火选择操作

标准遗传算法中,种群进化前期收敛速度较快,能够快速找到最优解,但在进化后期收敛速度将变慢,并易找到次优解从而陷入局部最优解. 在进化后期加入模拟退火算法,加快进化后期收敛速度,有助于获取全局最优解. 主要操作对象为经过变异操作后适应度较低的种群个体. 依据 Metropolis 准则, 以

$$P = \exp\left(\frac{\Delta F}{K \times T_t}\right) > rand[0,1)$$

的概率选择可以作为新种群的个体. F 为定义的适应度函数 F_i 与实际计算的适应度函数 F'_i 的差值,K 为系数,T_t 为第 t 代时的温度. 为保证算法在进化初期能够收敛,在选择适应度较差子代作为新种群个体时,使用较高的选择概率 P. 随着迭代次数 t 的增加,概率将逐渐减小. 直到满足下列条件之一时,结束算法: 1)已完成预设的迭代次数; 2)计算误差未超出预设的误差范围; 3)退火温度达到预设的结束温度. 其中,温度的控制是整个退火操作的关键,它决定了选取适应度较差的子代作为新种群个体的概率[8].

3　算法实现

退火遗传算法的求解过程[9] 如下.

1)初始化种群,设置种群个数 N,最大进化代数 m, 变叉概率 P_c, 变异概率 P_m, 退火控制温度 $T_t = T_0$.

2)计算种群中全部个体的适应度值.

3)经过选择、交叉和变异操作,产生新的种群个体.

4)计算新种群中全部个体的适应度函数值.

5)判断种群是否收敛,跳转到步骤 3,否则设置计数器 $t = 0$ 和退火代数,并对新种群个体进行退火操作:按照 Metropolis 准则来判断当前种群个体是否作为新的遗传种群,若计数器 t 小于预先设置的退火代数,则跳转到步骤 1;否则用本次退火操作后的个体替换种群中适应度最差的个体.

6)如果进化代数达到最大代数,则按照一定方式进行降温处理,将进化代数 $T = T + 1$,并跳转到步骤 2,否则整个优化过程结果,输出最优解. 算法整体流程如图 1 所示.

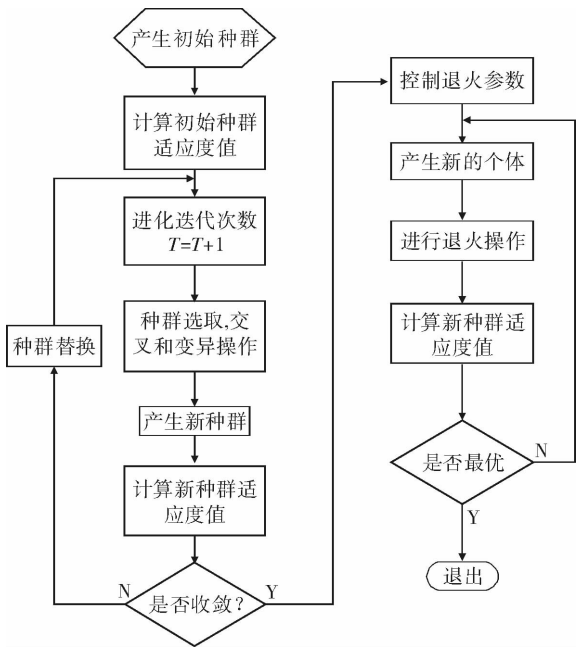


图 1　退火遗传算法整体流程图

4　仿真结果与分析

仿真环境中,随机选择分布在 100 × 100 m^2 平面区域内的 100 个传感器节点,各个节点之间的通信功率根据实际需要选择. 定义每个节点拥有相同初始能量 $E_{start} = 10$ J,交叉概率 $P_c = 0.8$, $\Delta F = 0.001$, $K = 0.01$, 变异概率 P_m 在算法前期选取为 0.15,后期选取为 0.05.

仿真试验使用退火遗传算法、遗传算法和周集

良等[5]提出的基于遗传算法的 WSNs 多路径路由优化模型进行仿真比较,共进行 30 组试验,取平均值作为最终结果. 收敛迭代次数和传输延迟的比较结果见表 1. 以传感器总能量消耗作为衡量指标的 3 种算法模型的能量消耗比较见图 2. 从表 1 和图 2 可以看出,基于遗传算法的 WSNs 多路径路由优化模型的传输路径是在基站进行计算的,信息传输过程中不存在向其他节点发送信息产生的传输消耗,所以不会造成不必要的能量消耗及传输时延. 基于可变长度染色体编码的遗传算法模型与退火遗传算法模型在传输初期的能量消耗接近基于遗传算法的 WSNs 多路径路由优化模型,但在收敛过程中传输链路中包含优越节点,使得整个链路的能量消耗增长速度缓慢,传输时长低于基于遗传算法的 WSNs 多路径路由优化模型. 而退火遗传算法模型在传输路径进化中增加模拟退火算法,使得进化后期路径选择更具准确性,能量消耗及传输时长更优于基于可变长度染色体编码的遗传算法模型.

表 1 3 种算法实验结果比较

算法	收敛迭代次数	传输延迟/s
退火遗传算法	250	200
遗传算法	270	230
最短路径算法	280	240

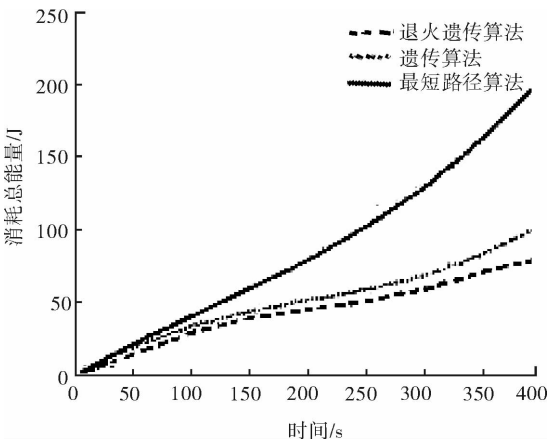


图 2 3 种算法的能量消耗比较

5 结论

本文基于标准遗传算法和模拟退火算法,提出了应用退火遗传算法寻求最优且耗能最少路径的方法,结合无线传感器网络路由特点,采用变长路径编码方式,并综合考虑节点间通信消耗、通信距离和路径最短等因素,同时选择相应的退火遗传操作算子,通过优化选取种群、计算适应度函数、合理交叉、有效变异和降温退火操作,最终寻求到一条最优路径. 仿真结果表明,该算法能够有效减少节点的能耗,延长网络生存周期.

参考文献:

[1] 张玉,蔡红梅. 基于遗传算法的无线传感器网络 QoS 路由优化[J]. 华北水利水电学院学报:自然科学版,2009,30(4):75.

[2] 高德民,钱焕延,汪峥. 基于遗传算法的无线传感器网络路由协议研究[J]. 计算机应用研究,2010,27(17):4226.

[3] Shafiullah G M,Gyasi-Agyei A,Wolfs P J. A survey of energy-efficient and QoS-aware routing protocols for wireless sensor networks[C]//Novel Algorithms and Techniques in Telecommunications, Automation and Industrial Electronics,Netherlands:Springe,2008:352 – 357.

[4] 玄光南,程润伟. 遗传算法与工程优化[M]. 北京:清华大学出版社,2004:157 – 233.

[5] 周集良,李彩霞,曹奇英. 基于遗传算法的 WSNS 多路径路由优化[J]. 计算机应用,2009,29(2):521.

[6] Xiao X P. Traffic engineering with MPLS in the Internet [J]. IEEE Networking,2000,14(2):28.

[7] Thepvilojanapong N,Tobe Y,Sezaki K. An efficient multicast routing protocol for wireless sensor networks[J]. IEIC Technical Report,2005,104(690):419.

[8] 刘彬, 张仁津. 基于退火遗传算法的 NURBS 曲线逼近[J]. 山东大学学报:工学版,2010,40(5):96.

[9] 谭胜兰. 模拟退火遗传算法在网络负载均衡中应用研究[J]. 计算机仿真,2011,28(12):111.

文章编号:2095-476X(2012)06-0097-04

环形无线网络呼叫接入控制模型研究

刘足江, 刘云

(昆明理工大学 信息工程与自动化学院, 云南 昆明 650051)

摘要:针对正交频分复用的无线网络中,当用户靠近基站覆盖区域的边缘且需切换通话时,传统排队模型切换失败率较高的问题,提出了在基站覆盖区域里建立一个多环呼叫接入控制模型的方案:将蜂窝设置在同心圆中,通过从所需的资源量中建立移动用户的信元平衡方程来降低切换失败率. 仿真结果表明,该模型可降低切换失败率,提高资源利用率.

关键词:呼叫接入控制;排队模型;信元平衡方程;环形无线网络

中图分类号:TN929.5 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.026

Study on call admission control model of ring wireless networks

LIU Zu-jiang, LIU Yun

(*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650051, *China*)

Abstract: In the orthogonal frequency division multiplexing wireless network (OFDM), when the user close to the edge of the base station coverage area, if you want to switch the call, the traditional queuing models switching have a higher failure rate. A new call admission control model based on the ring was established in a coverage area with cover of the base station, cellular will set in concentric circles, though the amount of resources required to establish the cell balance equation of the mobile users, to reduce switch failure rate. The simulation results showed that the model reduces switch failure rate in the protection channel, and improves resource utilization.

Key words: call admission control; queuing model; cell balance equation; ring wireless metworks

0 引言

不同于传统的FDMA, TDMA和CDMA网络, OFDM在接入无线网络(如3GPP LTE和WiMAX)时,分配连接资源主要依靠调度方案和信道质量. 比如,资源要求在功率、带宽和时间等方面支持一个固定数据速率,以便更好地定位离基站远的用户. 特别是当手机水平切换时,通常位于蜂窝附近,因此要比其他用户消耗更多的资源. 目前,国内外的研究主要集中在网络性能、无线资源管理以及接入控制算法等方面,在资源分配方面,研究OFDM系统呼叫接入控制(CAC)性能的文献并不多.

在CAC模型中,呼叫通话和结束是利用泊松过程建立简单的M/M/1排队模型[1]. 一些模型考虑信道等待时间和切换,一个电话的接听或者拒绝需要大量准确的资源来满足呼叫拒绝率、切换呼叫失败率以及当前的连接中断率. 而在CAC中,最需要关心的是实时交换的服务质量(QoS)的速度和延迟

收稿日期:2012-04-17

基金项目:国家自然科学基金项目(NSFC10502050)

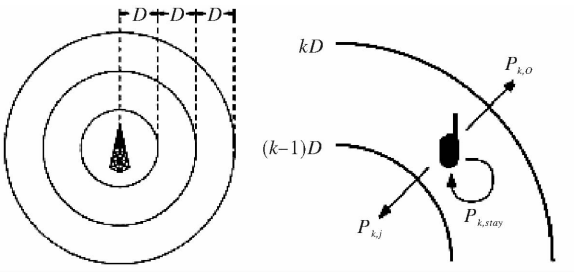
作者简介:刘足江(1988—),男,湖南省长沙市人,昆明理工大学硕士研究生,主要研究方向为无线通信和多媒体通信.

方面的需求. 在简单的分析和匹配应用中,需要考虑恒定比特率的交换[2].

CAC 是网络资源管理的重要手段,是解决用户服务质量和提高网络收益的技术关键. CAC 的目标是在保障已有接纳连接服务质量的前提下,充分利用网络资源,接受尽可能多的新连接. 目前,CAC 方案分为 4 类,即设置等待队列方案、完全分离方案、预留信道方案和基于实测网络状态方案. 设置等待队列方案和完全分离方案的缺点是切换呼叫掉线概率较高,静态保护信道方案虽然降低了呼叫掉线率,但新呼叫只能在系统空闲资源且一个特定的门限值时才能接入,有一定的局限性. 基于实测网络状态方案是目前 CAC 主要采取的方案,随着第 4 代移动通信系统模型的随机性越来越强,对 CAC 策略提出了更高的要求[3]. 因此,本文提出在基站覆盖区域建立一个多环的模型,来降低切换失败率,以适应第 4 代移动通信系统更强的随机性要求,并解决准确估计已有网络资源中的使用量以及有效地结合呼叫要求来判断能否接受新呼叫的问题.

1 系统模型

本移动模型为环型嵌套模型(见图 1). 它将蜂窝设置在同心圆中,从所需的资源量中获取一个移动基站的距离,考虑一般优质信号和多用户的信道质量. 为了提高手机在信元边界时的切换性能,在各种环形区域中建立移动用户的信元平衡方程,并依据切换呼叫失败率确定手机移动到低信道质量地区时因消耗过多基站资源引起的呼叫失败.



a)基站覆盖区域(3 环) b)移动节点的转移概率

图 1 环型嵌套模型

在模型中,蜂窝是分在 K 个同心圆中,每个环的宽度是 D,最里面那个圆的半径是 D,图 1a)展示的是当 $K=3$ 时的布局. n_k 是用户在 k 个环里面的数量,$S=(n_1,n_2,\cdots,n_k)$ 是进入网络的状态,λ 为新呼叫到达速率,P_k 是用户定位在 k 个环里的概率. 由此可以得出

$$P_k=\frac{k\text{ 个环区域}}{\text{信元区域}}=\frac{2k-1}{K^2}$$

$\lambda_k=\lambda P_k$ 是在 k 个环的新呼叫到达速率. 在此假定呼叫等待时间以指数分布,设为 $1/\mu s$,泊松过程切换发生的速率设为 λ_h. 如图 1b)所示,当手机移动进入一个给定的环时,假设在所给的环里的人使用情况符合指数分布,则剩余时间可以设为 $1/\gamma$. 当通话时间结束后,手机向外移动、向内移动或者在原来环里移动的概率分别设为 $P_{k,O},P_{k,I},P_{k,stay}$,并且 $P_{k,O}+P_{k,I}+P_{k,stay}=1$.

如果手机在原来的环里不动,则符合另一个指数分布,可以设为 $1/\gamma$.

设 r_k 为手机定位在 k 个环里的资源需求,如果网络允许交换多种不同类型的服务质量,那么 r_k 就是各个环的值. 本文假设环里只有一种类型交换,那么在各个环有一个简单值 r_k,对基站的总负载记为 $R(S)=\sum_{k=1}^{K}n_kr_k$,基站的容量是 C,并且切换预留资源的数量和移动性是 C_g.

1.1 相关移动系数

现在计算移动系数,手机在临近的环之间移动,产生模型的平衡方程. 在每个环的边界上,向内移动平均速率应该等于向外移动平均速率. 由此,可得

$$P_1\gamma_1=P_2P_{2,I\gamma}$$
$$P_kP_{k,O\gamma}=P_{k+1}P_{k+1,I\gamma}\qquad k=2,\cdots,K-1$$
$$P_KP_{K,O\gamma}=\gamma_R$$

其中,$1/\gamma$ 和 $1/\gamma_R$ 分别为第 1 圈的剩余时间和基站覆盖区域. 假设剩余时间与信元半径成正比,即

$$1/\gamma_R=\alpha R\qquad 1/\gamma_1=\alpha D\qquad 1/\gamma=\alpha\frac{D}{2}$$

其中,$R=KD$ 是信元半径,α 取决于手机移动的常数. 假设用户移动的方向不变,得到 $P_{k,stay}=1/2$,即用户在一个给定的环中,停留在环中的概率是 1/2. 由平衡方程可以得到概率 $\{P_{k,I}\}_{k=1}^{K}$ 和 $\{P_{k,O}\}_{k=1}^{K}$[4]. 此外,切换率定义为 $\lambda_h=\lambda\frac{\gamma R}{\mu},\frac{\gamma R}{\mu}$ 为呼叫等待时间与信元停留时间的比.

1.2 稳态分析

状态转换可能发生以下情况.

呼叫接听:当手机在第 k 个环内有新的呼叫时,如果 $R(S)+r_k \leqslant C-C_g$,那么 $n_k \rightarrow n_k+1$,否则呼叫阻塞.

呼叫结束:当手机呼叫结束在第 k 个环内时,$n_k \rightarrow n_k-1$.

切换接通:呼叫连接和切换接通只在第 k 个环时,如果 $R(S)+r_k \leqslant C$,那么 $n_k \rightarrow n_k+1$,否则切换失败.

向外移动:当用户在第 k 个环里向外移动时,如果 $R(S)+\Delta r_k \geqslant C$,则用户需要保持连接. $\Delta r_k = r_{k+1}-r_k$ 是用户的手机向最外层环移动时所需要增加的资源,用户可以切换到邻近的信元,即 $n_k \rightarrow n_k-1$.

向内移动:当用户由第 k 个环向内移动,$k=2,\cdots,K$,$n_{k-1} \rightarrow n_{k-1}+1$ 和 $n_k \rightarrow n_k-1$. 在稳定状态的概率下,$\{\pi(S)\}s$ 可以从上述所给的状态转换和归一化条件中获得 $\sum all_s \pi(S)=1$. 过渡矩阵为 $\boldsymbol{M}*\boldsymbol{M}$,$\boldsymbol{M}$ 是总的状态数. 从上述描述的转换状态,当非零元素为 $\boldsymbol{M}*(4K+1)$ 时最多,用稀疏计算更容易处理[5].

呼叫拒绝率和切换失败率分别定义如下:

$$P_B=\sum_{k=1}^{K}\sum_{S\in\Omega_k}\pi(S)P_k$$

$$P_f=\sum_{S\in\Gamma}\pi(S)$$

$$\Omega_k=\{S \mid R(S)+r_k>C-C_g\}$$

$$\Gamma=\{S \mid R(S)+r_k>C\}$$ [6]

此外,呼叫拒绝率是由于手机的移动性,所以引入呼叫掉线率 P_D. 一个呼叫被拒绝可能是因为向外移动需要更多的资源. 当用户向内移动,因为用户需要的资源较少,就不会阻塞. 由此,可以定义

$$P_D=\sum_{S\in\Psi_1}\pi(S)P_1+\sum_{k=2}^{K}\sum_{s\in\Psi_k}\pi(S)P_kP_{k,O}$$

$$\Psi_k=\{S \mid R(S)-r_k+r_{k+1}\geqslant C\}$$

预留资源 C_g 可以配置向外移动呼叫,也可以切换呼叫. 否则,即使在移动性低的环境中,P_D 比 P_f 更重要.

2　仿真结果与分析

2.1　模拟参数

首先在每个环上使用模拟来获得所需的资源 r_k,设置基于模拟参数的 3GPP LTE 系统,蜂窝边界设为 $R=300$ m,$K=3$. 基站的总传送功率设为 43 dBm,误码率的目标为 0.1%,噪声功率密度 $N_0=-174$ dBm/Hz,路径损耗 $39.95+43.375\log_{10}(d/10)$,$d$ 是到基站的距离/m. 3GPP 是典型城市区域,将频率设为 6 个开发衰落信道,可用带宽为 5 GHz 频段,其中包含 25 个资源块,所有的资源块由 12 个组成,间距为 15 kHz. 在资源块 j 中,用户 i 的数据速率是 $R_{i,j}=12\log_2(1+\beta SNR_{i,j})$,$\beta=\dfrac{1.5}{-\ln(5BER)}$,$SNR_{i,j}$ 表示 SNR 的用户 i 在资源块 j 中覆盖超过 12,BER 为误码率. 在基站和用户设备(UE)的单天线配置中,CBR 数据速率目标是 10 kb/s[7].

当目标数据速率可以支持部分资源块时(即资源块不需要用于连续时间领域),假设基站用作分时. 在仿真中,模拟给一个用户的资源数量少于给其他用户的资源量,在 CBR 交换调度的基础上研究最大最小法[8]. 所需的资源是衡量支持 CBR 目标率的资源量,假设一个能量分配超过频分带宽,那么证明达到饱和后对多用户调度影响不大.

图 2 为用户在每个环中资源平均需求,观察当环为 500 圈的模拟情况. 在每个成员 r 足够小时注意差别,3 个环足够模拟信元模型. 如果当 $r_1=1$ 时,r 可以近似等于(1,1.5,2),即一个用户在第 3 圈需要的资源可能是最内层用户所需资源的 2 倍.

2.2　性能分析比较

为了更好地证明所提出的基于环的模型,需要考虑当信元 $r=(1,1,1)$ 时的情况,将它同传统的排队模型进行比较,在不同的信道以及不同的环中可以发现,阻塞和切换失败率更符合这 2 种模型.

下面考虑当 $r=(1,1.5,2)$ 以及 $C=30$ 时的情况,给出不同的保护信道,如 $0<C_g<7$. 呼叫到达率 $\lambda=0.08$ calls/s,呼叫等待时间 $1/\mu=100$ s,并且令 $\alpha=1/3$,所以 $\lambda_h=\lambda$. 为了方便比较,本文改变了传统排队模型,得出手机在平均区域需要信元的资源. 图 3 为性能指标 P_B,P_f 和 P_D 在传统的和基于环的排队模型中的状况. 从中可以看出,只有 P_D 可以用于基于环的排队模型,并且误差值很小,因为保护通道可以配置手机,当手机向外移动时可以切换. 当 $C_g=0$,P_f 稍微高于 P_B 所需的资源,所以不能在传统的排队模型中观察. 从图 3 可知,当基础模型 $K=1$ 时,只要 C_g 增加,P_f 低估所关注 3 个环的模

型. 这是因为发生切换呼叫进入信元边界,它需要比在信元内调用更多的资源.

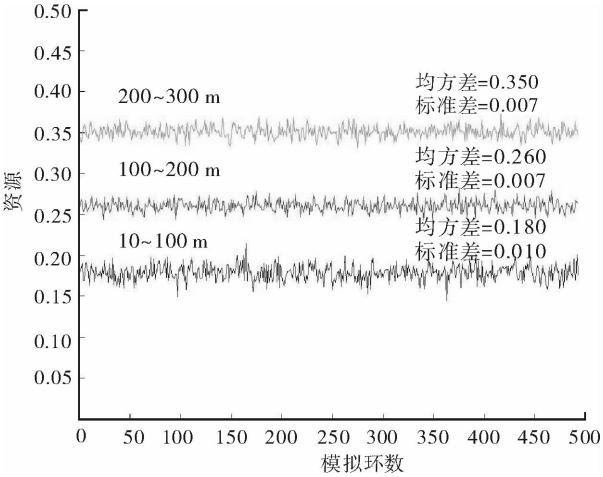


图 2 在信元的 3 层环中所需的资源量

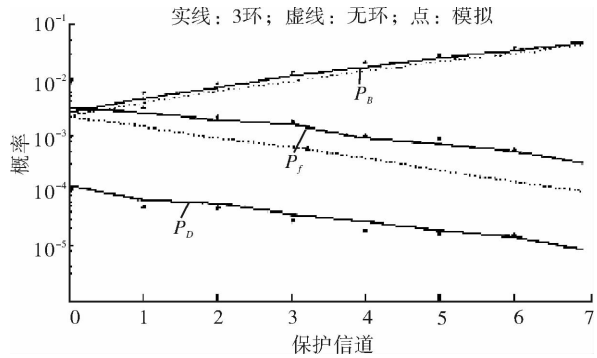


图 3 P_B,P_f 和 P_D 在传统的和基于环的排队模型中的比较

3 结论

本文提出了在基站覆盖区域里建立一个多环移动呼叫接入控制模型,将蜂窝设置在同心圆中,通过基于 3GPP LTE 系统的大量模拟,获得每个环形区域 CBR 交换需要的资源量,以此建立移动用户的信元平衡方程. 与传统的排队模型相比,该模型更适合用于环形区域. 该模型主要是改进性能关系,可以通过足够的保护信道来降低掉线率.

参考文献:

[1] Niyato D, Hossain E. Call admission control for QoS provisioning in 4G wireless networks: Issues and approaches [J]. IEEE Network, 2005, 19(5): 5.

[2] Tao M, Liang Y C, Zhang F. Resource allocation for delay differentiated traffic in multiuser OFDM systems [J]. IEEE Trans Wireless Commun, 2008, 7(6): 2109.

[3] Shen Z, Andrews J G, Evans B L. Adaptive resource allocation in multiuser OFDM systems with proportional fairness [J]. IEEE Trans Wireless Commun, 2005, 4(6): 2726.

[4] Ali S H, Lee Ki-Dong, Leung V C M. Dynamic resource allocation in OFDMA wireless metropolitan area networks [J]. IEEE Wireless Commun, 2007, 14(1): 6.

[5] Rong Bo, Qian Yi, Lu Kejie. Integrated downlink resource management for multiservice WiMAX networks [J]. IEEE Transactions on Mobile Comp, 2007, 6(6): 621.

[6] Ramjee R, Nagarajan R, Towsley D. On optimal call admission control in cellular networks [J]. Wireless Networks, 1997, 3(1): 29.

[7] Majid G, Raouf B. Call admission control in mobile cellular networks a comprehensive survey [J]. Wireless Commun and Mobile Comp, 2006, 6(1): 69.

[8] Rhee W, Cioffi J M. Increase in capacity of multiuser OFDM system using dynamic subchannel allocation [C]//Proc IEEE VTC, Tokyo: IEEE Press, 2000: 1085 - 1089.

文章编号:2095-476X(2012)06-0101-04

一种适用于低速无线传感器网络的LR-MAC机制研究

王明超

(无锡工艺职业技术学院 电子信息系, 江苏 宜兴 214206)

摘要:针对节点数量增加时无线传感器网络会产生数据包冲突增加和饱和吞吐量下降等问题,提出一种新的适用于低速无线传感器网络的LR-MAC机制. 该法采取当站点成功发送数据后,先进行一段时间的退避再参与信道竞争的措施来对L-MAC机制进行改进. 仿真结果表明,当网络规模发生变化时,LR-MAC在性能上明显优于L-MAC.

关键词:无线传感器网络;饱和吞吐量;分组丢弃概率;马尔可夫链模型

中图分类号:TP393 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.027

Study on low-rate wireless sensor network LR-MAC mechanism

WANG Ming-chao

(*Department of Electronic Information*, *Wuxi Institute of Arts & Technology*, *Yixing* 214206, *China*)

Abstract: Aiming at the problem that the increasing of the node in the wireless sensor networks may cause the lower saturation throughput and the higher packet collision. A novel channel access control mechanism for WSN, i.e. Lowe-rate MAC (LR-MAC) was proposed that the site firstly backoff for a moment after the site successfully send the data. The simulation results demonstrated that when the network size varies, LR-MAC performs much better than the L-MAC does.

Key words: wireless sensor network (WSN); saturation throughput; packet dropping probability; Markov chain model

0 引言

随着传感、无线通信、嵌入式系统以及微电子等高新技术的快速发展,具有数据收集、数据传输、信息融合处理等功能特点的传感器及由其构成的无线传感器网络(WSN)引起了人们的极大关注. WSN是由部署在检测区域内数量比较多且廉价的微型传感器节点组成的,它们通过无线通信的方式形成了一个多跳的、自组织网络. 无线传感器节点的这些特性以及连接方式使其能够广泛应用于环境监测、城市交通管理、大型车间以及仓库管理等领域,也吸引了学术界的广大研究者对其进行深入研究[1-5].

作为无线传感器网络协议栈的重要部分,介质访问控制(MAC)协议决定着无线信道的使用方式和信道资源的分配方式,成为WSN网络协议研究的

收稿日期:2012-06-05

基金项目:国家自然科学基金项目(60673185)

作者简介:王明超(1984—),男,江苏省宜兴市人,无锡工艺职业技术学院助教,主要研究方向为无线传感器网络性能分析与评价.

重点. Bianchi 首先提出了一个经典的马尔科夫链模型,该模型创造性地用数学模型的方式分析了无线局域网 MAC 协议在基本模式和 RTS/CTS 模式下的综合性能[1]. J. Zheng 等[2]在文献[1]的基础上提出 S-MAC 协议,该协议主要用来减少空闲侦听时所消耗的能量,但是当网络负载较小时,空闲侦听时间过长会造成系统整体性能下降. Y. Xiao 等[3]引入自适应性占空比,减少了侦听时浪费的能量,但是当网络负载达到一定数量时,系统的整体性能仍不理想. 文献[4]提出一种适用于低速率无线传感器网络 L-MAC,该协议通过在马尔科夫链中加入对空闲状态的建模,从数学分析的角度上对 WSN 进行了分析,但是仿真结果表明,当一定区域中节点数量较多时,WSN 的网络性能表现得很不理想.

无线传感器网络有一个特点,即规模和节点密度都比较大,但是网络所要传输的数据量比较少. 这使得 WSN 站点间的碰撞加剧. 而一般的基于 WSN 的 MAC 层协议在数据传输成功后,站点直接参与竞争信道,当一定区域中的节点数量较多时,就会造成站点间的碰撞加剧,从而导致网络整体性能的下降.

针对上述情况,本文对 L-MAC 机制进行改进,提出了一种新的适用于低速率无线传感器网络 MAC 层机制 LR-MAC(low-rate MAC),以期网络规模发生变化时,其性能明显优于基本 L-MAC 机制.

1 新机制的数学分析模型

为了简化模型,本文做如下假设: 1)网络共有 n 台设备,包括 1 台 PAN 网络协调器和 $n-1$ 台传感器节点; 2)假设网络传输是在理想信道中进行的,并且过程中没有捕获效应.

1.1 马尔科夫链分析模型

为了方便起见,设定随机过程 $w(t)$ 和 $s(t)$,分别表示设备在时隙为 t 时退避窗口的大小和设备在 t 时刻所处的退避阶段的随即过程; $\{s(t),w(t)=-1\}$ 和 $\{s(t),w(t)=-2\}$ 分别代表首次 CCA 检测和第 2 次 CCA 检测的随机过程;概率 α 和 β 分别表示在第 1 次 CCA 检测和第 2 次 CCA 检测到信道中有其他站点传输数据包的概率; p_s 代表数据包成功发送的概率; p_f 代表数据包发送失败的概率; L 代表信道中传输数据包的长度[6].

图 1 是 LR-MAC 协议的二维马尔科夫链模型.

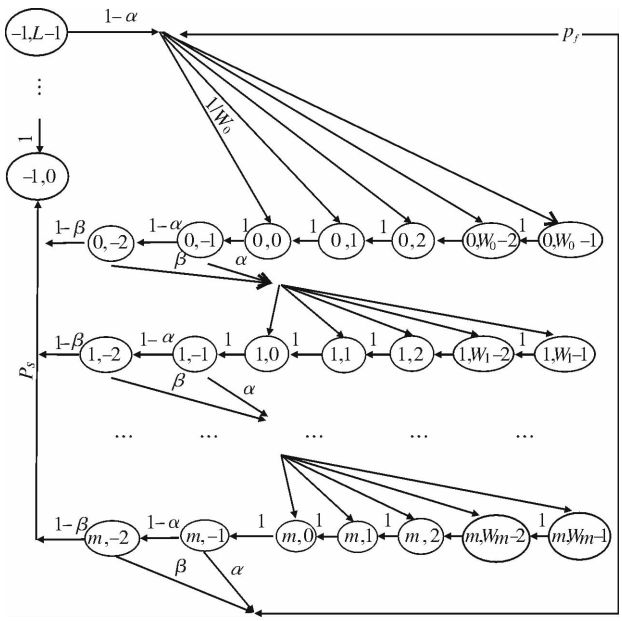


图 1 LR-MAC 协议的二维马尔科夫链模型

从图 1 可以看出,当数据包发送成功后,新机制 LR-MAC 区别于旧机制 L-MAC 的最大特点是:当数据包发送成功后,先要进行一段时间的退避再竞争信道,该时间的长度在模型里面表示为 L,为此可以得出数据流单步转移概率为

$$\begin{cases} P\{i,k\mid i,k+1\}=1 & k\geqslant 0 \\ P\{0,k\mid i,0\}=\dfrac{(1-\alpha)(1-\beta)}{W_0} & i<m \\ P\{i,k\mid i-1,0\}=\dfrac{\alpha+\beta(1-\alpha)}{W_i} & \\ \qquad i\in(1,m) \qquad k\in(0,W_i-1) & \\ P\{0,k\mid m,0\}=\dfrac{(1-\alpha)(1-\beta)}{W_0}+\dfrac{p_f}{W_0} & \end{cases} \quad ①$$

在公式①中,第 1 个等式表明在每个时隙的开始时刻,退避时间减 1;第 2 个等式表明,当退避计数器退避到 0 时,如果要传输数据包,则要经过 2 次 CCA 检测;第 3 个等式表明,在退避阶段 $i-1$,数据包传输发生冲突,竞争窗口 CW 加倍;第 4 个等式表明,达到最大退避阶段 m,如果数据包传输成功,则还要进行一段时间的退避,再竞争信道. 如果传输失败,重新回到初始状态. 设

$$b_{i,k}=\lim_{t\to\infty}P\{s(t)=i,w(t)=k\}$$

$$i\in(-1,m) \qquad k\in(-2,\max(L-1,m-1))$$

为各状态的稳态分布,根据马尔科夫链规则有

$$b_{i-1,0}(\alpha+\beta(1-\alpha))=b_{i,0} \qquad 0<i\leqslant m \quad ②$$

由②式可以推出

$$b_{i,0} = [(\alpha + \beta(1-\alpha))]^i \times b_{0,0} \quad 0 < i \leqslant m \quad ③$$

由公式③以及马尔科夫链规则,可以得出

$$b_{i,k} = \frac{W_i - k}{W_i}[(1-\alpha)(1-\beta)\sum_{j=0}^{m} b_{j,0} + p_f]$$
$$0 < i \leqslant m \quad ④$$

当 $i = 0$ 时,公式④变为

$$b_{i,k} = \frac{W_i - k}{W_i} b_{i,0}$$

由马尔科夫链性质可知,图 1 中各种状态的转移概率总和为 1,可以得出关系式

$$1 = \sum_{i=0}^{m}\sum_{k=0}^{W_i-1} b_{i,k} + \sum_{i=0}^{m} b_{i,-1} + \sum_{i=0}^{m} b_{i,-2} + \sum_{i=0}^{L-1} b_{-1,i} =$$
$$\sum_{i=0}^{m} b_{i,0}\left[\frac{W_i+1}{2} + 1 + (1-\alpha) + L(1-\alpha)(1-\beta)\right] \quad ⑤$$

$$1 = \frac{b_{0,0}}{2}[3 + 2(1-\alpha) + 2L(1-\alpha)(1-\beta)] \cdot \left(\frac{1-(\alpha+\beta-\alpha\beta)^{m+1}}{1-(\alpha+\beta-\alpha\beta)}\right) + W\left(\frac{1-2^{m+1}(\alpha+\beta-\alpha\beta)^{m+1}}{1-2(\alpha+\beta-\alpha\beta)}\right) \quad ⑥$$

式⑤中,$W_i = 2^i W$. 数据包传输失败的概率 p_f 和成功的概率 p_s 分别为

$$p_f = b_{m,0}(\alpha + \beta - \alpha\beta) \quad ⑦$$
$$p_s = \tau(1-\alpha)(1-\beta) \quad ⑧$$

所有的数据传输发生在退避时隙为 0 的时刻,而不管退避阶段,所以站点分组传输概率为

$$\tau = \sum_{i=0}^{m} b_{i,0}$$

把式①⑥⑦⑧代入,可得

$$\alpha = L[1-(1-\tau)^{N-1}](1-\alpha)(1-\beta) \quad ⑩$$
$$P_{\text{send}} = (1-(1-\tau)^{N-1})(1-\alpha)(1-\beta) \quad ⑪$$
$$\beta = \left[1 - \frac{P_{\text{send}}}{P_{\text{send}}\left(1 + \frac{1}{1-(1-\tau)^N}\right)}\right](1-(1-\tau)^N) \quad ⑫$$
$$\tau = 1 - \left(1 - \frac{\beta}{1-\beta}\right)^{\frac{1}{N}} \quad ⑬$$

其中,P_{send} 表示数据包经过 2 次 CCA 检测后信道空闲后成功发送的概率.

1.2　饱和吞吐量分析

设 P_{tr} 为在一个随机选择的时隙内网络中至少有一次分组发送的概率,有

$$p_{tr} = 1 - (1-\tau)^n \quad ⑭$$

饱和吞吐量指当负载持续增加时,系统能够达到的最大吞吐量[7]. 本文将系统的饱和吞吐 S 定义为成功发射有效载荷的信道时间在总的信道时间中所占的比重大小,因此可以得出

$$S = \frac{\text{在一个时隙内用于发送数据分组的有效载荷}}{\text{时隙的长度}}$$

$$S = \frac{p_s T_{E(p)}}{(1-p_{tr})\sigma + p_s T_s + (p_{tr} - p_s)T_c} \quad ⑮$$

把公式⑩—⑭代入公式⑮,则可以得出系统的饱和吞吐量为

$$S = L \cdot N(1-\tau)^{N-1}(1-\alpha)(1-\beta) \quad ⑯$$

从公式⑯可以看出,新机制 LR-MAC 的饱和吞吐量和旧机制 L-MAC 最大的区别是加入了时间长度 L,从而更加精确地模拟了站点成功发送数据包后再进行时间长度为 L 的退避.

1.3　分组丢弃概率分析

由图 1 可知,分组丢弃只会发生在退避阶段$\{i, -2\}$. 令 $P_{i,drop}$ 表示数据包 i 的分组丢弃概率,所以有 $p_{i,drop} = p_i^{L_i+1}$.

2　LR-MAC 协议性能分析与评价

2.1　试验的模拟环境

为验证本文提出的 LR-MAC 机制的有效性,采用数学分析的方法对该协议和 L-MAC 协议进行性能仿真比较.

本文在一个单跳的星型网络场景并且节点处于一个能够相互监听的信道区域内进行仿真试验. 假设发送节点从 10 递增至 100,进行 Matlab 数学分析所采用的系统参数见表 1,假设 MAC 层发送的数据分组大小恒定,无线信道的传输速率为 1 Mb/s. 实验中所计算出的饱和吞吐量、分组丢弃概率都是利用模型中的相关公式和试验所使用的相关参数进行计算得出的.

表 1　参数设置

参数	数值	参数	数值
数据包有效载荷	71 b	时隙 σ	20 symbol
MAC 帧头	14 b	MinBE0	3—5
PHY 帧头	7 b	amaxBE	5
信道速率	250 kb/s	MinBE1	3
SIFS	21.5 symbol	MaxCSMABackoff	5

2.2　结果及分析

图 2 给出了 LR-MAC 机制和 L-MAC 机制在不同站点数目下的饱和吞吐量的变化曲线. 从图 2 可

以看出,LR-MAC 算法在每个站点的饱和吞吐量都高于 L-MAC 算法;随着站点数的增加,2 种机制的饱和吞吐量表现出下降趋势,这是因为随着站点数目的增加,站点间发送数据包发生的碰撞明显增加,所以导致了饱和吞吐量的下降. 但是 LR-MAC 算法的下降幅度明显比 L-MAC 算法平缓,这是因为 LR-MAC 算法加入了发送数据包成功后的一段退避时间,减少了碰撞,因此能很好地提升网络的整体性能.

图 3 给出了 LR-MAC 机制和 L-MAC 机制在站点数从 10 递增到 100 的情况下,站点分组丢弃概率的变化曲线. 从图 3 可以看出,LR-MAC 机制在每一站点的分组丢弃概率始终低于 L-MAC 机制. 随着站

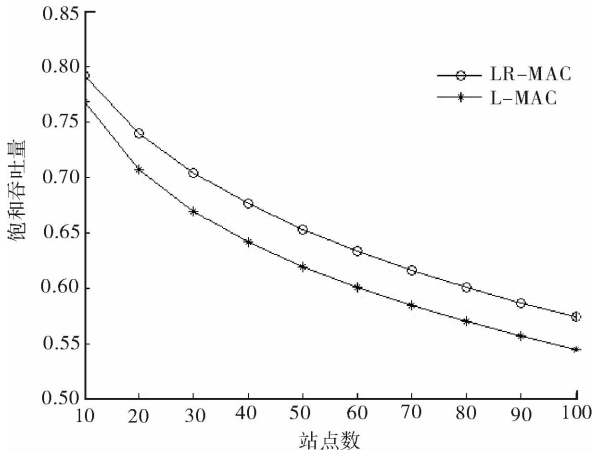


图 2 饱和吞吐量随站点数的变化

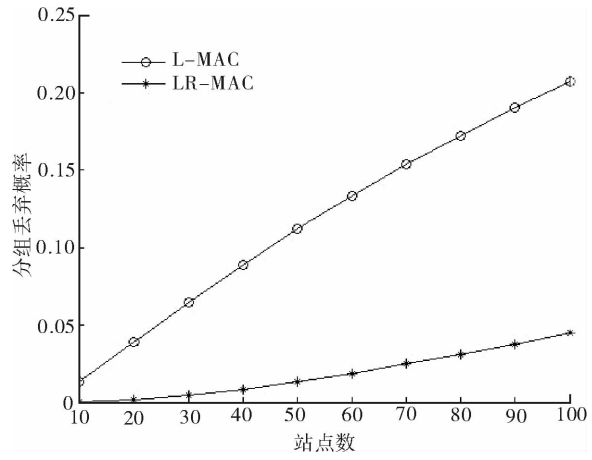


图 3 分组丢弃概率随站点数的变化

点数的增加,LR-MAC 机制和 L-MAC 机制的分组丢弃概率逐渐上升,这是因为随着站点数目的增加,站点间发送数据包发生的碰撞明显增加.

总体而言,本文提出的方法可以提高无线传感器网络的整体性能.

3 结语

针对 WSN 机制中站点成功发送数据包后立刻进行信道竞争所带来的问题,本文提出了一种新的适用于低速无线传感网络 MAC 层机制协议 LR-MAC. 该协议主要做出了如下改进:当站点成功发送数据后,先进行一段时间的退避,再参与信道的竞争. 试验结果表明:当网络规模发生变化时,LR-MAC 的饱和吞吐量明显优于 L-MAC 机制,且分组丢弃概率低于 L-MAC 机制. 下一步的工作将继续深入探讨如何提高高负载下无线传感器网络的综合性能.

参考文献:

[1] Pollin S, Ergen M, Ergen S C, et al. Performance analysis of slotted carrier sence IEEE 802. 15. 4 medium access layer[C]//Proc of IEEE GLOBE-COM, San Francisco: IEEE Press, 2006:1 – 6.

[2] Zheng J, Lee J M. A Comprehensive Performance Study of IEEE 802. 15. 4. Sensor Network Operations[M]. New York: IEEE Press, 2006:218 – 237.

[3] Xiao Y, Pan Y. Differentiation QoS guarantee and optimization for real-time traffic over one-hop ad hoc networks [J]. IEEE Transactions on Parallel and Distributed Systems, 2005, 16(6):538.

[4] 孙利民, 李建中, 陈渝, 等. 无线传感器网络[M]. 北京:清华大学出版社, 2005.

[5] Omprakash G, Rodrigo F. Collection tree protocol[C]// Proc of the 7th ACM Conf on Embedded Networked Sensor Systems, Berkeley: ACM Press, 2009.

[6] 郑国强, 李建东, 周志立. 无线传感器网络 MAC 协议研究进展[J]. 自动化学报, 2008, 34(3):305.

[7] 孙利民, 李波, 周新运. 无线传感器网络的拥塞技术[J]. 计算机研究与发展, 2008, 45(1):63.

文章编号:2095-476X(2012)06-0105-04

工业网络中非标准 VPN 的安全技术研究

朱鹏, 张智斌, 黄昱泽

(昆明理工大学 信息工程与自动化学院, 云南 昆明 650500)

摘要:针对在工业控制领域中,传统的监控与数据采集独立运转且很少配置安全管理的问题,利用N2N为工业网络之间的通信构建一条安全通道,使用数字证书对加入的节点进行身份验证,借助IKEv2协议实现节点之间的协商通信,并通过动态选择加密算法及通信密钥,有效提高了N2N在工业网络通信中的安全性.

关键词:N2N;监控与数据采集;工业网络通信

中图分类号:TP393.08 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2012.06.028

Study on non-standard VPN security technology in the industrial network

ZHU Peng, ZHANG Zhi-bin, HUANG Yu-ze

(*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500, *China*)

Abstract: Aiming at the problem that the traditional SCADA (supervision control and data acquisition) is operated independently with less configuration safety management in current industrial control field, a secure channel was constructed the communication between industrial network using N2N (a layer two peer-to-peer VPN), in which the joining node will be authenticated with digital certificates, and node communication between joint be realized with IKEv2 protocol. As a result, the security of N2N in industrial network communication will be improved efficiently and greatly through dynamic selective encryption algorithm and communication key.

Key words: a layer two peer-to-peer VPN; supervision control and data acquisition; industrial network communication

0 引言

目前,监控与数据采集SCADA(supervision control and data acquisition)系统在工业控制领域中应用广泛,很多工程人员通过其对工业设备进行实时监控管理.随着互联网的普及,工业自动化控制技术的应用从局部范围扩大到整个网络世界,在某些领域的需求越来越强烈.由于传统的工业网络不提供安全管理机制,直接将工业网络与Internet相连,容易受到非法用户的入侵,数据在传输的过程中易被窃取.而对于SCADA系统来说,系统中的很多设备不能安装相关安全客户端软件,为了保证工业网络中的数据安全,可在各个SCADA系统中配置一台N2N(a layer two peer-to-peer VPN)客户端机器,通过对N2N客户端进行功能扩充,使之能收集需要发送到其他SCADA系统的数据,并将其加密后传输;也能够接收其他SCADA系统传来的数据,并将其解密后分发给系统中的工业设备.通过使用N2N

收稿日期:2012-07-05

作者简介:朱鹏(1987—),男,湖南省双峰县人,昆明理工大学硕士研究生,主要研究方向为计算机网络.

在 Internet 上构建安全通道,来保证数据在 Internet 上的安全传输,实现 SCADA 系统从区域到全域的管理. 本文主要对应用工业网络互联的 N2N 安全技术进行研究,通过将 PKI 技术和 IKEv2 协议引入其中,增强 N2N 的安全性,进而保证使用 N2N 进行互联的工业网络的安全性.

1 N2N 在工业网络中的安全问题分析

N2N 是一个通用型的 P2P-VPN 程序[1],使用预共享密钥管理方式,只提供 Twofish 加密算法. 如果将 N2N 直接用来实现工业网络互联,可能会产生较大的安全隐患. 本文主要从认证管理、密钥管理和加解密算法 3 个可待加强的方面进行分析.

1)认证管理. N2N 允许用户自行创建 super node 端和 edge 端,edge 端注册过程如下:用户在启动 edge 端以后向已启动的 super node 端发送注册信息,super node 在收到注册信息以后,就在注册链表中查询该 edge 是否注册,如果没有注册,super node 就将 edge 加入到注册链表中,完成注册. 而对于 SCADA 系统来说,如果 super node 不对 edge 进行严格的身份认证,允许任意 edge 加入网络,这可能会给系统带来毁坏性的破坏. 因此对于用于工业网络互联的 N2N 来说,需要在 edge 注册之初对 edge 进行严格的身份认证,只允许合法用户加入到网络中,从而避免非法用户入侵网络.

2)密钥管理. N2N 中用来加密解密的密钥使用预共享的方式,数据的加密解密都在 edge 端进行. 在启动 edge 时,可以选择手工输入密钥或者从已配置好的密钥文件中读取密钥,只有 edge 双方都使用相同的密钥时,数据信息才能被解密. 然而对于工业网络来说,为了保持网络的稳定性,保障各个网络之间信息的互通性,整个网络长时间使用预共享密钥对数据进行加密解密,会给工业网络带来巨大的风险以及管理上的困难.

3)加密解密算法. 在 N2N 的源程序中,只使用了一个 Twofish 加密算法对数据进行加密解密,虽然 Twofish 的抗差分攻击能力以及抗相关密钥 Slide 攻击的能力都很强,但由于工业网络的特殊性以及被破坏所带来的毁坏性,可在程序中部署多种成熟的加密算法供动态选择,将增强数据的安全性,从而更好地为工业网络保驾护航.

2 N2N 在工业网络中安全技术的设计

对于直接将 N2N 应用于工业网络存在的安全问题,可以使用 edge 向 super node 注册之前进行身份认证、edge 之间通信之前进行密钥协商以及选择加密算法的方式进行解决. 而身份认证与密钥协商都是在注册成功之前或数据发送之前进行的,将认证和协商与通信分离,能更好地为数据提供安全保护.

在各种不同的身份认证机制中,基于数字证书的身份认证是最灵活和安全的. super node 与 edge 之间的身份认证可以结合 PKI 技术来完成,所有需要进入网络的用户设备都需要有确定的身份证书. 在这里,主要是借助认证中心(CA)来为用户提供安全保证. CA 为每个用户或设备发放一张身份数字证书,利用 CA 强大的管理功能,证书的发放、维护和撤销等管理就比较容易. 而由于数字证书具有唯一性,对于移动用户也可以很方便地进行身份认证.

在 N2N 中,数据信息的加密过程都是在 edge 中完成的,当某个 edge 需要和其他 edge 进行数据通信时,该 edge 先与接收信息的 edge 就通信过程中所用的加密算法、通信 SA(security association)进行协商,只有当双方协商成功之后,双方的数据通信才能进行. 改进后的 N2N 安全设计总体框架图如图 1 所示.

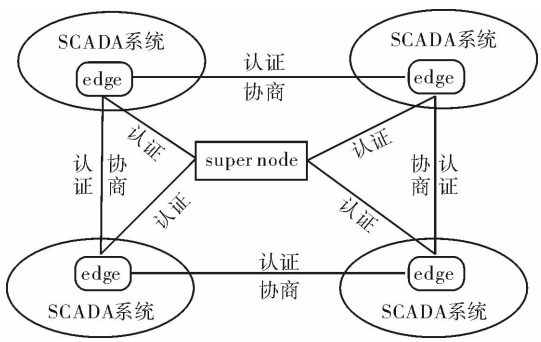


图 1 安全设计总体框架图

2.1 edge 与 super node 数字签名的认证过程

认证结构的 PKI 环境[2]主要由认证中心 CA、注册中心 RA 和 LDAP 服务器组成. CA 是签发和管理证书的实体;RA 负责核查申请证书实体的身份,并完成提交数据正确性验证;LDAP 用来存储签发的属性证书和属性证书撤销列表. 整个系统的认证

模型如图 2 所示.

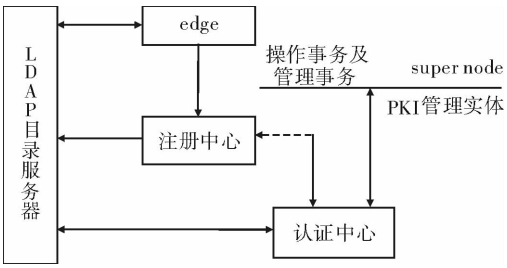


图 2　系统认证模型

edge 认证过程[3]如下:

1)edge E 发起认证请求,将自己的身份、认证信息 AUTH 以及随机数 nonce 发送给 super node S;

2)S 收到 E 发来的验证请求后,对 E 的身份进行验证,验证通过则向 E 发送自己的身份、认证信息 AUTH 以及随机数 nonce;

3)E 验证 S 身份后,向 S 发送注册请求;

4)S 收到 E 的注册请求后,返回一个注册成功数据包.

在整个注册认证过程中,如果 edge 端没有申请数字证书,则向 RA 发送一个证书请求,RA 审核后提交证书请求给认证机构 CA. CA 对证书申请请求进行处理,签署并颁发用户证书,并且登记在证书库中,同时定期更新证书失效列表,供用户查询. edge 在收到 CA 颁发的证书后,将证书封装在注册包中,一起发送给 super node. super node 收到 edge 发来的注册信息,解析注册信息并对其中的数字证书进行验证,如果证书有效,则通过身份认证,完成注册,如认证不通过,终止注册.

2.2　edge 与 edge 通信协商的实现

在 N2N 的源程序中,只要 edge 注册到 super node 以后,就可以向其他 edge 发送信息,若目的 edge 不在发送 edge 的认证链表中,则把信息转发给 super node. super node 在接收到信息后,根据目的 edge 的 MAC 地址在其注册链表中进行查询,如果目的 edge 存在于注册链表中,则向该 edge 转发信息,否则丢弃该信息. 目的 edge 在收到 super node 转发过来的信息后,如果解析到 edge 不在认证链表中,则向该发送 edge 发送一个认证包,发送 edge 在收到目的 edge 发来的认证包后,返回给目的 edge 一个确认认证包,完成认证.

改进后的认证模式是在原有认证过程前加入一个协商过程,使得 edge 与 edge 在进行通信之前,必须对对方的身份、通信的加密算法以及密钥进行协商;通信过程将使用协商的加密解密算法和密钥进行通信. 在协商过程中,主要借助 IKEv2 协议加以实现[4], IKEv2 的协商过程如图 3 所示.

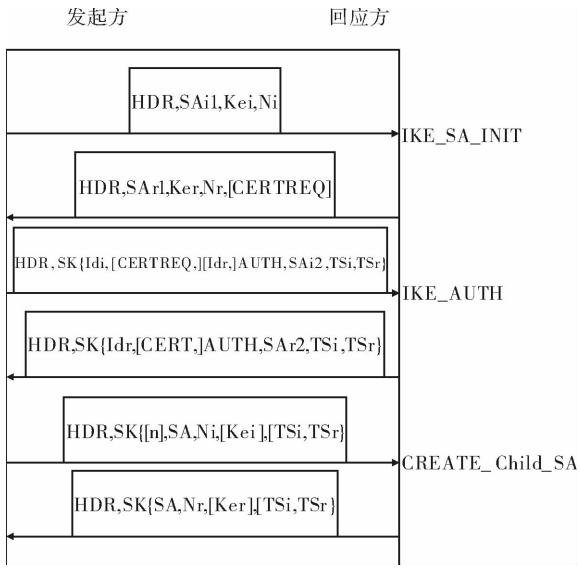


图 3　IKEv2 协商过程

结合 IKEv2 的 edge 与 edge 之间的通信协商[5]过程如下:

1)edge A 发起连接请求,向 edge B 发送安全关联的 SA 提议、DH 交换的临时公共值 KEa 和防重放攻击的随机数 Na.

2)B 在收到 A 的连接请求后,从 A 的提议中选择一个提议,并将自己用于 DH 交换的临时公共值 KEb、防重放攻击的随机数 Nb 以及查询 A 的证书请求返回给 A.

3)A 收到 B 的回应后,向 B 发送用于 N2N 加密的 SA 提议以及自己的身份信息、证明、对方的身份信息请求和流量选择符.

4)B 对 A 的身份进行认证,通过后选择一个 SA 提议,将自己的身份信息 ID,认证载荷 AUTH 及流量选择符返回给 A,完成初始化交互.

5)A 在对 B 进行身份验证后,向 B 发送 SA 提案,交换 nonce、流量选择符 TSi 和 TSr 以及可选的进行 DH 交换值 Kei,发送信息使用前一次通信协商的加密算法和 B 的公钥进行加密.

6)B 在收到 A 的信息后,使用私钥进行解密,并对 A 的 SA 提案和流量选择符进行响应,交换 nonce.

整个协商过程由 6 条消息组成[6]. 其中 1)和 2)两条消息用来协商密码算法、交换 nonce 和 DH 公共值;3)和 4)两条消息用来对前面的消息进行认证和交换各自的身份,并建立第 1 个 Child_SA;5)和6)可由通信的任一方在前 4 条消息交换结束后发起,以生成额外的 Child_SA 或重新进行密钥协商(rekeying).

当 edge A 与 edge B 在前4 条信息交换成功后,就可以选择第 5,6 条消息中生成的 Child_SA 进行通信了. 在通信过程中根据协商好的 Child_SA 选择加密算法,使用计算出的密钥对数据进行加密. 在这个解决方案里,可供 SA 进行协商的密码库采用的是 OpenSSL 提供的密钥算法库. 在整个协商过程中,edge 可以自动选择加密算法,通过创建 Child_SA 来确保在通信过程中实现加密算法及密钥的动态切换,从而避免预共享密钥管理方式以及加密算法单一的安全隐患,更好地保证数据在传输过程中的安全性.

3 结论

本文对应用于工业网络互联的非标准 VPN 的安全技术进行了研究,通过使用数字证书认证,使得 edge 只有在通过 super node 身份认证后才能完成注册,而在 edge 与 edge 通信前使用 IKEv2 协议进行协商,动态选择加密算法及通信密钥,增强 N2N 在工业网络通信过程中的安全性. 然而工业网络是一个非常复杂的网络,它的涉及面非常广,对安全性的要求非常高,因此还有许多问题有待进一步的研究.

参考文献:

[1] Deri L, Andrews R. N2N: A layer two peer-to-peer VPN [J]. LNCS, 2008, 5127: 53 - 64.

[2] 张小波,程良伦. PKI 在虚拟专用网络中的应用[J]. 计算机工程,2011,37(15):113.

[3] 寇晓蕤,王清闲. 网络安全协议——原理、结构与应用[M]. 北京:高等教育出版社,2009.

[4] 邱司川,潘进,刘丽明. IKEv2 协议的分析与改进[J]. 计算机工程,2009,35(15):126.

[5] 韩旭东,汤隽,郭玉东. 新一代 IPSec 密钥交换规范 IKEv2 的研究[J]. 计算机工程与设计,2007,28(11):2549.

[6] 韩明奎,潘进,李波. 一种改进的 IKEv2 协议及其形式化验证[J]. 计算机应用研究,2010,27(2):707.