DOI：10.19416/j.cnki.1674－9804.2017.03.014

面向持续适航工程数据的大数据采集和处理技术初探

A Preliminary Study on Big-data Acquisition and Processing Technology for Continuous Airworthiness Engineering Data

郑海飞　徐有成　郝　莲　陆　军　杨　波／ZHENG Haifei　XU Youcheng　HAO Lian　LU Jun　YANG Bo
（上海飞机设计研究院，上海 201210）
（Shanghai Aircraft Design and Research Insitute，Shanghai 201210，China）

摘　要：
国产某型飞机在持续适航阶段开展事件收集、风险评估、工程调查和措施制定等工作时，需要机型资料、机队信息、运营记录、事故事件、局方信息等工程数据提供输入、参考以及辅助分析。研究了持续适航工程数据与大数据之间的关系，初步规划了面向持续适航工程数据的大数据系统架构，并通过以下自动化方式实现了相关数据的采集与处理，形成了持续适航工程数据库：首先利用网络爬虫数据采集技术实时准确地获取一些国内外公开数据；其次应用 VBA 语言对已获得数据进行整理与自定义处理；最后基于大数据的映射分析方法对这些工程数据进行分析。该持续适航工程数据库已有效应用于某型国产飞机持续适航体系的日常运行工作。

关键词：持续适航；工程数据；大数据；风险评估；数据抓取；数据处理；映射分析
中图分类号：V221^{+}.91　　**文献标识码：**A

[**Abstract**] When an aircraft in continuous airworthiness phase, the airplane maker needs to do the work of event collection, risk assessment, engineering survey and measure establishment, and then needs plenty of engineering data provided as input information, reference, also can provide the auxiliary analysis, these engineering data such as model data, fleet information, operation records, accident and events, and Airworthiness Directive. This paper studies the relationship between the continued airworthiness engineering data and Big Data, preliminary planning for the continued airworthiness engineering data system structure based on Big Data concept, and through the following automated way to realize the data acquisition and processing, to form the continuous airworthiness engineering database: first of all, using web crawler data acquisition technology to accurately obtain some public data. Secondly, the data is collected and processed by VBA language. Finally, analyzes the engineering data based on the mapping analysis method. The continuous airworthiness engineering database has been used for the daily operation of the continuous airworthiness system.
[**Keywords**] continuous airworthiness; engineering data; big data; risk assessment; data acquisition; data processing; mapping analysis method

1　持续适航工程数据的大数据论述

国产某型飞机在持续适航阶段开展事件收集、安全风险评估、工程调查和制定改正改进措施等工作时，需要工程数据库为风险评估工作提供数据输入、数据参考以及分析结论。目前，应用大数据采

2017年第3期 总126期
ISSN 1674-9804 CN 31-1614/V

集和处理技术，持续适航体系已初步建立“持续适航工程数据库”，并在持续完备中。该数据库包含多种数据类型：(1) 针对特有机型的数据：设计需求数据、设计要求数据、设计规范(英文)、详细设计报告、结构或系统 CATIA 数模、审定计划，适航符合性报告(MOC1-MOC9)、交联系统——专项审定计划、技术出版物以及飞机在运营阶段的事件、事故症候和事故，以及相关的客户服务文件、适航指令；(2) 针对国内外不同机型的数据：国内外不同机型的事故、事故症候数据，国内外不同机型的适航指令数据。

该数据库的主要使用目的为：(1) 为持续适航事件的风险评估提供飞机型号数据输入，包括设计数据、三维结构数据、安全性分析数据、运营维修数据等，为风险评估中的潜在不安全状态分析提供数据依据；(2) 利用反证法或引证法，为持续适航事件的风险评估提供参考依据；(3) 实现内网搜索查询、实现跨库搜索查询和支持辅助分析的功能。

持续适航工程数据库的数据量化指标如表 1 所示。

表 1　持续适航工程数据库的数据量化指标

序号	工作项目	具体情况
1	网络数据库抓取个数	12 个
2	抓取数据条数	303 201 条
3	数据存储量	33GB

综上所述，持续适航工程数据的特征与大数据的特征极为相似[1]。大数据的五大特征与特性为：数据体量大(Volume)、数据处理速度快(Velocity)、数据类别多(Variety)、数据真实性强(Veracity)、数据潜在价值高(Value)，简称为“5V”特征[2-8]。为了确保航空器的安全运营以及运营的经济性，持续适航体系要求针对航空器的运营过程中出现的任何事件，要做出及时、迅速的风险评估，以及分析出航空器的潜在不安全状态，因此要求快速地对持续适航工程数据进行处理与分析，并得出有效的措施建议；持续适航工程数据与大数据类似，其数据来源于航空器的设计、制造、试飞与日常运营过程中，均为原始数据，因此具有很强的真实性。可以看出，持续适航工程数据是大数据在民航制造业、民航运输业的具体体现，站在大数据层面，从大数据的视角，在持续适航工程数据的采集、存储及分析处理方面引入大数据理念，可以更好、更快、更有效地支持并服务于持续适航体系的运行。

2　大数据系统的系统架构

基于大数据的理念分析处理持续适航工程数据，就需要一个完备的持续适航工程数据大数据系统。持续适航体系提出了关于持续适航工程数据的大数据系统的系统架构，主要包括数据架构和管理架构两部分。持续适航工程数据库基础架构的数据架构和管理架构如图 1、图 2 所示。

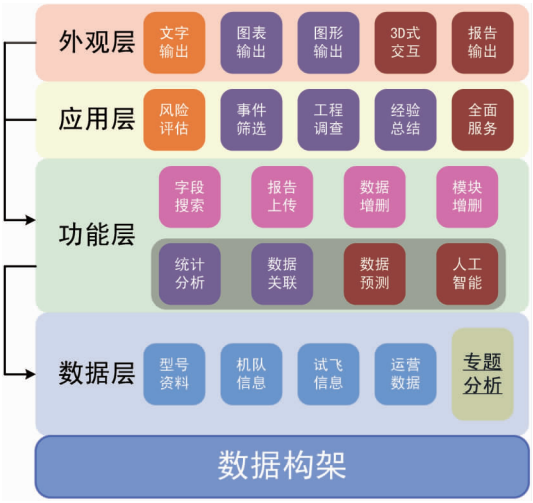

图 1　持续适航工程数据库基础架构的数据架构

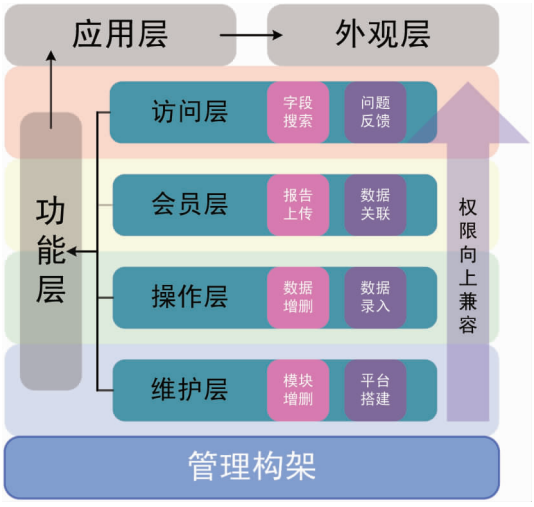

图 2　持续适航工程数据库基础架构的管理架构

2.1　数据架构

主要基于数据层、功能层、应用层和外观层搭建持续适航工程数据库的数据架构，实现数据库的

工程数据输入、输出和处理及人机交互等功能。

1）数据层包括型号资料数据模块、机队信息数据模块、试飞信息数据模块和运营数据模块，以及可输出分析数据的专题分析数据模块。

2）功能层包括字段搜索（含数据下载功能）、报告上传、数据增删和模块增删四个基础功能模块，数据统计分析和数据关联两个延伸功能模块，数据预测和人工智能两个高阶功能模块，功能层总共包含八个功能模块。

字段搜索功能模块（含数据下载功能）对数据层进行关键字段检索，从而输出相应持续适航工程数据；报告上传功能模块仅针对专题分析数据模块进行工程数据分析报告的上传；数据增删功能模块针对数据层中数据的实效性、正确性进行更新、更正的操作；模块增删功能模块针对数据层中数据模块的实效性、正确性进行更新、更正及拓展的操作；统计分析功能模块具有数据的科学统计、处理分析功能，并将分析处理后的数据套用专题分析数据模块中的分析报告模板进行分析报告输出；数据关联功能模块将关联数据之间的相似性、矛盾性等具有工程研究价值的关系与映射；数据预测和人工智能将通过对现有数据及数据之间的关联的运算与分析，得出具有工程研究价值和实际操作价值的结论和建议。

3）应用层包括风险评估、事件筛选、工程调查和经验总结等服务项目，服务于持续适航体系的运营。

4）外观层包括文字输出、图表输出、图形输出、3D 人机交互及报告输出五个应用模块。3D 人机交互是为了实现持续适航工程数据的 3D 交互式可视化，高效、便捷地为持续适航体系运行提供数据支持。

2.2 管理架构

主要基于访问层、访问层（会员）、操作层以及维护层搭建持续适航工程数据库的管理架构，实现工程数据输入、输出和处理以及数据库更新、维护的流程化和权限化管理（权限向上兼容）。

1）访问层具有对数据库的字段搜索（含数据下载权限）权限和统计分析权限，可以得到文字、图表、图形及报告的输出；

2）会员具有对数据库的报告上传及统计分析权限，对专题分析数据库进行数据补充；工程数据库会员拥有个人账户，可以实现对自己已有数据的关联，亦可对已关联的其他会员数据的关联，即多层次的数据关联；

3）操作层对持续适航工程数据进行实时跟踪，具有对数据库数据的增加和删减权限；

4）维护层具有对数据库数据模块的增加、删减和拓展权限。

3 大数据系统的技术方案

建立持续适航工程数据的大数据系统，首先需要获得数据，利用网络爬虫的数据采集技术可以高效准确地获取一些国内外公开数据；其次是对已获得数据的整理与处理，以便后续分析使用；最后是利用大数据的分析方法，例如线性回归、决策树、支持向量机、贝叶斯网络、k 均值以及 Apriori 关联等算法[9-10]，对系统中的数据进行分析，因此需要建立针对不同应用场合的辅助分析方法。已实现的大数据技术方案如下所述。

3.1 基于网络爬虫的数据采集技术

持续适航的工程数据具有体量大的特点，包括飞机型号数据和国外相似机型数据。因此数据的采集不能靠人工来实现。

利用网络数据抓取技术，包括开源网络数据爬虫抓取技术或者 Python 语言自编译数据抓取技术，建立高效、准确的抓取规则，执行持续适航工程数据的抓取工作。基于网络爬虫数据采集的关键技术是抓取规则的建立，其技术方案如图 3 所示。

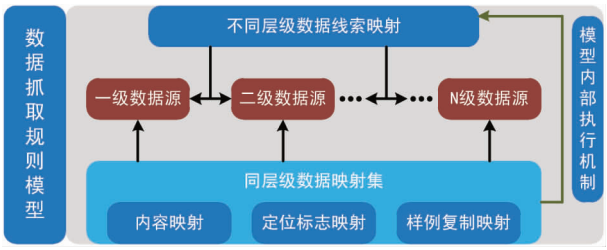

图 3 数据抓取规则创建机理及其模型

持续适航工程数据是不断更新的，不能使用不具有时效性的数据。利用网络数据抓取技术以及上述的数据抓取规则模型，建立数据更新的抓取规则，实现持续适航工程数据的定时更新。采集的数据均来自 FAA、EASA 以及 NTSB 的公开数据，因此数据的真实性是可以保证的。

针对采集数据的重复性，在应用 3.2 节的数据

2017年第3期 总126期
ISSN 1674-9804 CN 31-1614/V

处理技术之后，利用 Excel 工具中自带的“删除重复数据”功能，即可实现数据的去重复性。

3.2 基于 VBA 语言的数据处理技术

持续适航工程数据的数据量较大，数据存储量已达到数十 GB，数据条数已有三十多万条，人工手动处理是不现实的，必须开发数据自动化批处理技术。因此本文提出了基于 VBA 语言的数据处理技术，其关键技术是数据处理的核心代码，如图 4 和图 5 所示。

```
Sub 数据自动化批处理()
'
'    数据自动化批处理 宏
'
    Dim x As Long'整型变量定义
    Dim y As Long'整型变量定义
    Dim i As Integer '整型变量定义
    Dim j As Long'整型变量定义
    Dim Data(1 To 9) As String '数组变量定义
    Dim myPath$, myFile$, am As Workbook '数据文件路径变量、数据文件名称变量定义
    myPath = Dir(ThisWorkbook.Path & "\*.*") '数据文件路径赋值
    Do While myPath <> "" '循环读取需要打开的多个数据文件
        If myPath <> ThisWorkbook.Name Then Workbooks.Open Filename:=myPath'打开
数据文件
            Windows(myPath).Activate'激活数据文件窗口
            For x = 1 To 30 '定位待提取数据的纵坐标
                If Cells(2, x) = "'需自定义'" Then
                    Exit For
                End If
            Next
            For i = 1 To 220'定位待提取数据的横坐标
                If Cells(i, x) = "'需自定义'" Then
                    Data(j) = Cells(i, x + 1) '将数据存储于数组当中
```

图 4 基于 VBA 语言的数据处理核心代码（上）

```
                    Exit For
                End If
            Next
        .....

            Windows(ThisWorkbook.Name).Activate'激活汇总文件窗口
            For i = 1 To 9'将提取的数据统一存储于汇总文件中
                Cells(y, i) = Data(i)
            Next

            y = y + 1
            Cells(1, 1) = z'计数，显示当前已处理的文件个数
            z = z + 1
        Next

        Windows(myPath).Activate
        ActiveWindow.Close'关闭当前已打开的数据文件
        myPath = Dir'打开下一个数据文件

        If myPath = ThisWorkbook.Name Then Exit Do'如果待打开文件为汇总文件，跳出
循环
    Loop
End Sub
```

图 5 基于 VBA 语言的数据处理核心代码（下）

图 4 和图 5 展示了核心代码主体框架，中间具体算法已省略；可根据数据结构的变化，制定不同的算法，从而实现基于 VBA 语言的数据处理技术。

3.3 基于映射分析的数据精确定位方法

持续适航工程数据系统的数据量已达到数十万条，同时具备了基本的查询、关键词搜索功能。但是，查询和搜索的精度还较低。查询和搜索结果经常为数十条或者几百条，而真正与持续适航事件切合的数据就隐藏在这些查询结果中，往往需要人工查阅这些查询结果才能进行有效数据的定位，不利于工作效率的提升，不符合持续适航事件快速处理的原则。因此开展“基于映射分析的持续适航工程数据精确定位方法”的研究。

持续适航工程数据系统中的外部数据绝大部分为英文数据，而在持续适航事件风险评估工作过程中接触的数据均为中文数据，因此需要应用“映射分析法”，建立 BigTable 数据库——映射关系数据库。该数据库中包含多种数据映射表，例如“飞机结构/系统部件数据映射关系表”、“飞机安全性数据映射关系表”等多种数据映射表。

“基于映射分析的持续适航工程数据精确定位方法”的实现路径有两种：一是基于某型飞机结构或者系统部件的映射分析数据精确定位法；二是基于某型飞机安全性数据的映射分析数据精确定位法。如图 6 所示。

1）方法一

根据接收到的某型飞机持续适航事件，从事件中提取出涉及的飞机结构或者系统相关的数据，依据提取的数据在“某型飞机机型数据”中进行飞机结构或者系统数据的精确、全面定位，为风险评估做充分输入。

依据提取的飞机结构或系统的精确、全面数据，使用“BigTable 数据库”，定位国内外其他机型的映射关系，为在“持续适航工程数据外部数据”中进行数据定位提供输入。“BigTable 数据库”中的飞机结构/系统部件映射关系表如图 7 所示。以防冰系统为例。

从图中可以看出，映射关系表不仅包含某型飞机中英文映射关系，还包含常见机型（737、A320）的映射关系。根据映射关系查询结果，使用“持续适航工程数据外部数据”，可进行国内外相关机型事故、事故症候的初步精确定位——标题、概述定位；在初步定位结果不满足需求的情况下，可进行深度精确定位——调查报告、适航指令报告的全文搜索定位。从而实现数据的精确定位功能，为持续适航事件的风险评估工作提供验证性依据。

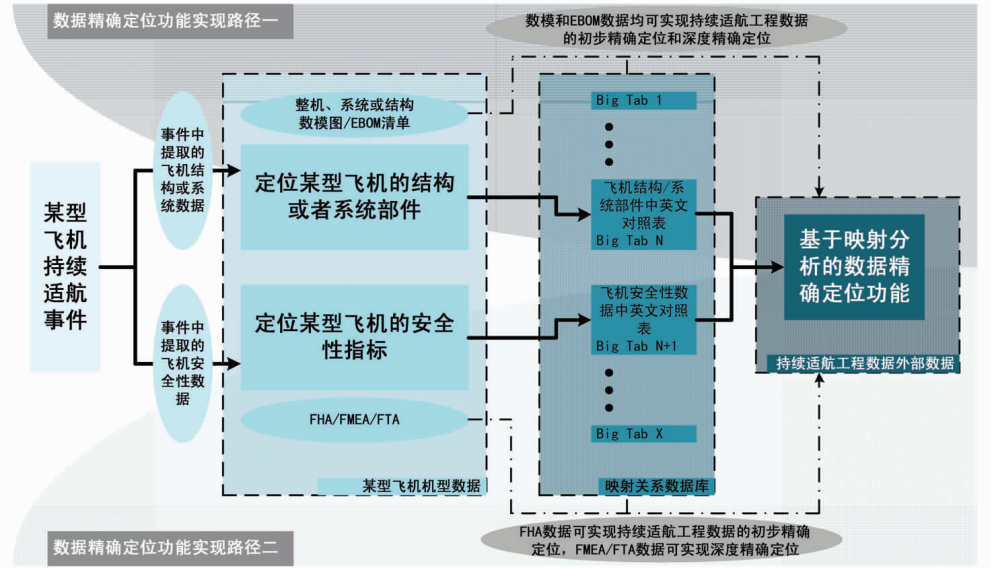

图 6　数据精确定位方法模型

中文（某型飞机）	英文（某型飞机）	英文（737）	英文（A320）
部件1	Part1	Part1	Part1
部件2	Part2	Part2	Part2
部件3	Part3	Part3	Part3
部件4	Part4	Part4	Part4
部件5	Part5	Part5	Part5
部件6	Part6	Part6	Part6
部件7	Part7	Part7	Part7
部件8	Part8	Part8	Part8
部件9	Part9	Part9	Part9
部件10	Part10	Part10	Part10
部件11	Part11	Part11	Part11
部件12	Part12	Part12	Part12
部件13	Part13	Part13	Part13
部件14	Part14	Part14	Part14
部件15	Part15	Part15	Part15
部件16	Part16	Part16	Part16
部件17	Part17	Part17	Part17
部件18	Part18	Part18	Part18
部件19	Part19	Part19	Part19
部件20	Part20	Part20	Part20
部件21	Part21	Part21	Part21
部件22	Part22	Part22	Part22
部件23	Part23	Part23	Part23
部件24	Part24	Part24	Part24
部件25	Part25	Part25	Part25
部件26	Part26	Part26	Part26
部件27	Part27	Part27	Part27
部件28	Part28	Part28	Part28
部件29	Part29	Part29	Part29
部件30	Part30	Part30	Part30

某型飞机机翼防冰系统 / 某型飞机短舱防冰系统 / 某型飞机结冰探测系统 / 某型飞机风挡加热系统 / 某型飞机水、废水防冰系统

图 7　飞机防冰系统部件数据映射关系表

2）方法二

根据接收到的某型飞机持续适航事件，从事件中提取出涉及的飞机安全性指标相关的数据，依据提取的数据在“某型飞机机型数据”中进行飞机安全性分析数据的精确、全面定位，为风险评估做充分输入。

依据提取的飞机安全性分析数据的精确、全面数据，使用“BigTable 数据库”，定位国内外其他机型的映射关系，为在“持续适航工程数据外部数据”中进行数据定位提供输入。“BigTable 数据库”中的飞机安全性分析数据映射关系表如图 8 所示。以防冰系统为例。

从图中可以看出包含某型飞机座舱压调系统 FHA 的中英文映射关系。根据映射关系查询结果，

2017年第3期 总126期
ISSN 1674-9804 CN 31-1614/V

FHA中文	FHA英文
功能危险1	Functional Hazard Assessment1
功能危险2	Functional Hazard Assessment2
功能危险3	Functional Hazard Assessment3
功能危险4	Functional Hazard Assessment4
功能危险5	Functional Hazard Assessment5
功能危险6	Functional Hazard Assessment6
功能危险7	Functional Hazard Assessment7
功能危险8	Functional Hazard Assessment8
功能危险9	Functional Hazard Assessment9
功能危险10	Functional Hazard Assessment10
功能危险11	Functional Hazard Assessment11
功能危险12	Functional Hazard Assessment12
功能危险13	Functional Hazard Assessment13
功能危险14	Functional Hazard Assessment14
功能危险15	Functional Hazard Assessment15
功能危险16	Functional Hazard Assessment16
功能危险17	Functional Hazard Assessment17
功能危险18	Functional Hazard Assessment18
功能危险19	Functional Hazard Assessment19
功能危险20	Functional Hazard Assessment20
功能危险21	Functional Hazard Assessment21
功能危险22	Functional Hazard Assessment22
功能危险23	Functional Hazard Assessment23
功能危险24	Functional Hazard Assessment24
功能危险25	Functional Hazard Assessment25
功能危险26	Functional Hazard Assessment26
功能危险27	Functional Hazard Assessment27
功能危险28	Functional Hazard Assessment28
功能危险29	Functional Hazard Assessment29
功能危险30	Functional Hazard Assessment30
某型飞机座舱压调系统	

图 8　飞机防冰安全性数据映射关系表

使用“持续适航工程数据外部数据”，可进行国内外相关机型事故、事故症候的初步精确定位——标题、概述定位，例如根据 FHA 数据可进行初步精确定位；在初步定位结果不满足需求的情况下，可进行深度精确定位——调查报告、适航指令报告的全文搜索定位，例如根据 FMEA、FTA 数据进行深度精确定位。从而实现数据的精确定位功能，为持续适航事件的风险评估工作提供验证性依据。

4　结论

依据确定的持续适航工程数据大数据系统的系统架构，利用基于网络爬虫的数据采集技术、基于 VBA 语言的数据处理技术和基于映射分析的数据精确定位方法，已形成了大数据系统的初步形态，持续适航工程数据库已有效应用于国产某型飞机持续适航体系的日常运行工作中。

在之后的工作，需要在以下两个方面开展研究工作：（1）数据的更新与迭代，系统内的数据必须能够体现出当前民航运输和航空工业的最新状态；（2）针对持续适航技术工作的，更深层次的大数据辅助分析方法的研究，包括事件相似度研究、风险识别方法研究、风险后果严重性等级辅助分析方法研究和风险概率辅助分析方法研究。

参考文献：

[1] 郑海飞，陆军. 民机持续适航工程数据的大数据视角[J]. 航空科学技术，2017，28(5)：53－58.
[2] 靳小龙，王元卓，程学旗. 大数据的研究体系与现状[J]. 信息通信技术，2013，6：35－43.
[3] 马建光，姜巍. 大数据的概念、特征及其应用[J]. 国防科技，2013，34(2)：10－17.
[4] 张引，陈敏，廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展，2013，50(9)：216－233.
[5] 李战怀，王国仁，周傲英. 从数据库视角解读大数据的研究进展与趋势[J]. 计算机工程与科学，2009，35(10)：1－11.
[6] 程学旗，靳小龙，王元卓，郭嘉丰，张铁赢，李国杰. 大数据系统和分析技术综述[J]. 软件学报，2011，25(9)：1889－1908.
[7] 钟瑛，张恒山. 大数据的缘起、冲击及其应对[J]. 现代传播(中国传媒大学学报)，2013，7：104－109.
[8] 何非，何克清. 大数据及其科学问题与方法的探讨[J]. 武汉大学学报(理学版)，2014，60(1)：1－12.
[9] 袁梅宇. 数据挖掘与机器学习[M]. 北京：清华大学出版社，2016.
[10] 范明，范宏建. 数据挖掘导论[M]. 北京：人民邮电出版社，2011.

作者简介：

郑海飞　男，博士，工程师。主要研究方向：持续适航工程数据的采集、处理技术以及分析方法，E-mail：zhenghaifei@comac.cc

徐有成　男，硕士，研究员。主要研究方向：适航技术与管理研究，E-mail：xuyoucheng@comac.cc

郝　莲　女，硕士，研究员。主要研究方向：适航技术研究，E-mail：haolian@comac.cc

陆　军　男，博士，高级工程师。主要研究方向：持续适航事件风险评估方法与技术，E-mail：lujun1@comac.cc

杨　波　男，硕士，助理工程师。主要研究方向：SMS 安全管理系统，E-mail：yangbo2@comac.cc