Proceeding of the 15th International Forum of Automotive Traffic Safety, 2018, pp 1-7
No.ATS.2018.101

Application of Dynamic Report Generating system on Accident Data Quality Control

Jiguang Chen
China Automotive Technology & Research Center, Beijing,China
Email: chenjiguang@catarc.ac.cn

Abstract: Data has permeated into various field and has become a valuable resources and production factors in recent years. While the enterprises are striving to layout its big data ecosystem, data administering and improving data quality turns into an effective way for its core competence enhancement. China In-Depth accident Study (CIDAS) has accumulated a lot of accident data after developed for more than 7 years. The requirement of better data quality has raised up due to the deeper accident analysis and complicated data analytical. Data could be distorted during the process of transferring the accident information into data. This study designed an accident data quality report generating system which consist of two parts. The first part is a R script which can detect the data and sheet completeness, data consistency, data anomalism as well as data plausibility. The second part is report output control and report generating function which can output the detected problems in formatted report. The data recorder can notice the accident data problems according to this report. The system can apply to inspection a single accident case as well as specific sheets or problematic variables. The system with advantages of repeatability and convenience due to unnecessary to recompile the script disregarding various kind of accident type. It improved the efficiency of data quality management massively, and the time cost as well as labor cost can be saved.

Keywords: Dynamic report; data quality management; data cleaning; knitr

动态报告生成系统在事故数据质量管理上的运用

陈吉光
中国汽车技术研究中心，北京，中国，100070
Email: chenjiguang@catarc.ac.cn

摘　要: 近年来，数据已渗透到各个领域并成为了一种极具价值的资源和生产要素，在企业争相布局大数据生态的环境下，进行数据治理、提升数据质量是加强核心竞争力的有效途径。中国交通事故深入研究（CIDAS）在经历 7 年的发展后，积累了大量的事故数据，随着事故研究和数据分析的深入，对数据质量要求越来越高，由于在事故相关信息转化为数据的过程中容易引起失真，降低了基于事故数据分析结果的可靠性。为了提高数据质量，本研究设计了一套能够针对事故数据生成动态报告的系统，该系统由两部分组成，第一部分为使用 R 语言编写的 CIDAS 数据质量控制的脚本文件，能够识别出数据和表格的完整性、数据的不一致性、数据的异常性和数据的逻辑性等数据的有效性问题；第二部分为 knitr 的报告输出控制部分和报告生成部分，能够将脚本文件所检测到的问题以格式化报告的形式输出，数据录入人员可根据动态报告的内容进行数据核查和修改。该系统可对单个的事故案例数据进行报告生成，也可以针对特定的车辆信息或人员信息等生成质量报告，该系统能够适用不同类型的事故，具有可重复性和快捷性。此系统的运用能够大大提高数据质量管理的效率，降低时间和人力成本，适合推广到其他数据质量管理领域。

关键词: 动态报告；数据质量管理；数据清洗；knit

1 引言

数据是企业的宝贵资产，基于数据的相关技术和应用也在快速的改变人们的生产生活方式，互联网加速了数据的积累，对数据相关的开发技术、应用场景和商业模式得到了极大的发展，然而关系到数据生命线的质量问题却成为制约企业发展的瓶颈，IDC 对中国数据集成和数据质量市场的调查结果显示[1]，大部分中国企业数据集成项目难以达到预期的主要原因在于数据质量问题，这也是我国复杂的数据生态环境导致的必然结果，调查还表明 72%的企业存在数据重复和 60%的企业存在数据不完整的情况，因此构建数据质量管理体系是加强产品竞争力的必备条件。数据质量管理 DQM（Data Quality Management）是一项艰巨、长期的任务，进行数据清洗（data cleaning）是提高数据质量的必经之路，同时也是 ETL（extract、transform、load）最重要的组成部分[2]，然而容易忽略的是，进行数据清洗前人们往往不了解数据中存在什么样的问题，这样导致了经过清洗后的数据仍然不能有效使用。Erhard Rahm 等人总结在数据清洗过程中发现的一些常见现象[3]，将数据质量问题分为单数据源问题和多数据源问题，并列出了每一类型中的典型示例（如图 1 所示）。陈孟婕等人设计了分层次的缺陷数据识别方法和相应的数据清洗策略，扩大了数据问题识别的范围，提高了数据清洗的自动化程度[4]。

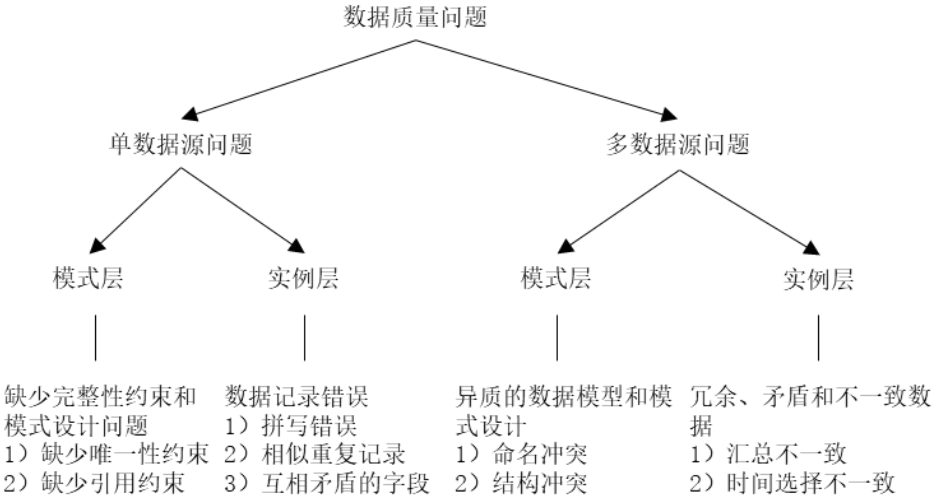

图 1 数据质量问题分类

本文利用数据质量管理体系的概念和方法，结合中国交通事故深入调查数据库，用 R 语言编写了关于事故数据抽取、数据清洗与数据质量识别的脚本程序，利用 knitr 可生成动态报告的特点，开发了一套可重复使用的数据质量动态报告系统，该系统覆盖了数据库中绝大部分变量，对各个变量的数据进行有效性进行检测，针对数据库中实时更新的数据生成数据质量报告，并在报告中体现出所识别参数的变量名和存在的数据问题，帮助数据录入人提高数据质量，为深入的数据分析和数据挖掘提供了基础保障。

2 研究对象与方法工具

2.1 研究对象

真实交通事故数据对车辆安全技术开发的价值和意义已得到了国内外的普遍认可，丰富详细的交通事故数据也是制定汽车交通安全相关法规标准的重要依据，许多国家和地区建立了专业的团队进行深入事故调查[5-9]。本研究所依托的中国交通事故深入研究项目，已连续 7 年在全国多个城市开展了深入的事故现场调查，每起案例都包括碰撞前、碰撞中和碰撞后三个阶段的信息，收集涉及道路环境、事故涉及人员和事故车辆等层次的数据信息 2000 多条，在此基础上搭建了基于 ORACLE web 前端应用开发的在线交通事故数据库，案例中所涉及的不同层级的信息被合理的架构到不同表格中，如图 2 所示。目前，数据库中包含了 4700 余起真实的事故案

例。其中的数据类型包括了文本型数据、整型数据、浮点型数据、逻辑型数据与日期型数据，数据之间的表格跨度（31 种表格）和层次（车辆、人员、伤口、事件、碰撞编号等）跨度较为复杂。

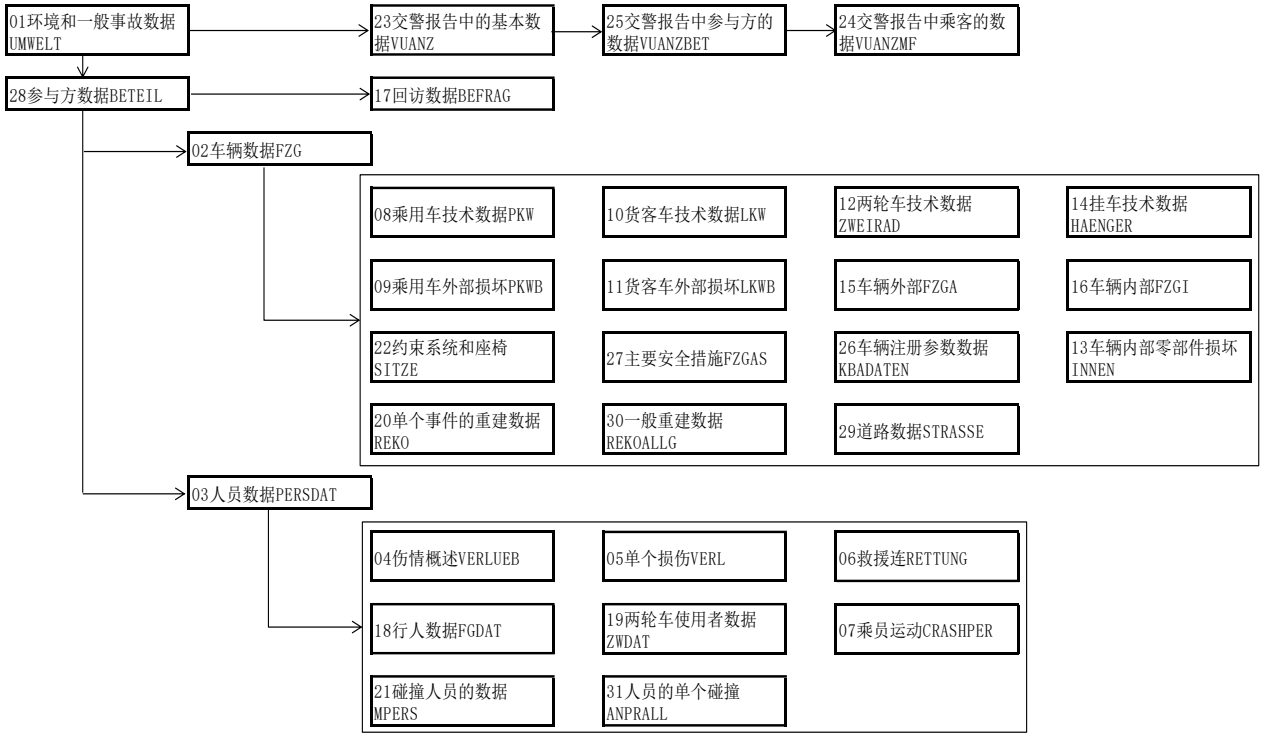

图 2 CIDAS 数据层次关系图

2.2 方法和工具

2.2.1 R 语言与 knitr

R 语言是一种广泛运用于数据清洗与计算、统计分析和作图的解释型语言 [10-11]，具有直观、易用、低成本的特点，同时 R 也属于 GNU 系统的源代码开放型软件，最开始运用于统计学方面，由于其完全免费、开源的特点，越来越多的使用者通过编制具有特定功能的包来丰富 R 的功能，其他用户则可以通过 CRAN（comprehensive R Archive Network）下载所需要的包，在 R 语言里以加载包的方式来获得所需要的函数和功能。Knitr 是一种能够实现代码语言与输出结果互联的 R 包，在进行可重复研究时能够根据所输入的数据而生成与数据相匹配的报告内容[12]，亦即动态报告，其主要设计思想来自于文学化编程范式（Literate Programming）[13]，即使得计算机语言和人类语言按照一定的格式进行编排，与传统的 Sweave 相比，knitr 能够结合 HTML 和 Markdown 语法生成格式化的文本与图片[14]，其常见的文本输出格式控制参数如表 1 所示。

表 1 knitr 段落和文本控制参数

代码名称	控制类型	功能
highlight	逻辑	是否高亮代码
tidy	逻辑	是否整理代码
eval	逻辑	是否执行代码
size	字符	文本字体大小
echo	逻辑和数值	显示/隐藏代码
results	字符	装裱输出、原样输出和隐藏
comment	字符	输出结果前缀符号
split	逻辑	是否剥离代码和文本到外部文件

2.2.2 语言与语法

CIDAS 数据库属于 ORACLE 关系型数据库，在动态报告生成系统中需要将数据处理脚本与数据库连接起来，以便进行数据的读取，SQL（Structured Query Language）语言是一种专门用于数据库查询、更新和管理的 ANSI 标准计算机语言，能够与诸如 MS Access、MS SQL Server、Oracle 等数据库程序协同工作[15]。本研究利用 SQL 查询语句对数据库进行抽取，但在使用该功能前需要加载“RODBC”包，然后通过 odbcConnect 函数进行数据库连接，连接建立后，可以通过下列函数对数据库和连接本身进行查询、修改、保存等操作：odbcClose、sqlColumns、sqlCopy、sqlDrop、sqlFetch、sqlQuery、sqlSave、sqlTables、sqlTypeInfo。Markdown 是一种轻量级的排版标记语言，相比 LaTex 和 OLE 排版系统，其特点是书写快速、管理方便、样式可复用等，本系统的动态报告在 R markdown 文档创建的基础上，使用简易的 markdown 语法使得报告更加美观，增强报告的可阅读性。

3 实现过程

3.1 数据准备

由于数据库所含的表格数较多，因此将各个表格的执行代码存放在不同的块（chunk）中，本系统一共包含 32 个块，其中有 31 个块用于存放 31 个表格的代码，而第一个块则用于数据的准备工作，其主要内容包括以下三方面：

1）加载函数所需的包。R 软件会自带一些常用的基本函数包，但在使用一些特殊的函数前需要使用 library 来加载相应的包，因为本系统在代码执行过程中需要对日期、字符和因子进行转化与处理，因此需要加载下列包："knitr"、"stringr"、"RODBC"、"lubridate"、"plyr"。

2）建立数据库连接并对数据进行抽取。在此过程中 R 可以直接预读取数据库中的数据，也可以将数据导入到本地中。

3）重新建立各表主键。在数据库中，涉及到车辆、座椅、人员、伤口、人体碰撞部位、事件编号等信息的表格主键由多个键组成，为了方便计算和快速定位，重新建立了这些表格的主键，即通过组合数字的方式，值得注意的是，在进行组合前需要对数据格式进行统一，主要通过 formatC 和 paste 函数来实现，如图 3 所示。值得注意的是，不是每个表格都可以直接建立主键，对于衍生表格和不同的事故类型，需要先判断表格是否存在。

```
E03$PSKZ < −formatC(E03$PSKZ, flag = 0, width = 2); E03$newpskz < − paste(E03$BETNR, E03$PSKZ, sep = "")
E05$PSKZ < − formatC(E05$PSKZ, flag = 0, width = 2); E05$NR < − formatC(E05$NR, flag = 0, width = 2)
E05$newpskz < − paste(E05$BETNR, E05$PSKZ, sep = ""); E05$injuryNR < − paste(E05$newpskz, E05$NR, sep = "")
E20$newevent < − paste(E20$BETNR, E20$KNR, sep = "")
if (length(E31$FALL) != 0){
        E31$PSKZ  < − formatC(E31$PSKZ, flag = 0, width = 2)
        E31$newpskz  < − paste(E31$BETNR, E31$PSKZ, sep = "")
        E31$newimpact  < − paste(E31$BETNR, E31$PSKZ, E31$ANR, sep = "")
}
```

图 3 建立表格主键代码示例

3.2 数据清洗与识别

在数据读取完成后，需要对各个表格的参数进行识别和核查，但对于一些复杂的参数，还需要进行数据清洗。以时间序列为例，时间序列在事故数据中有着非常重要的地位，是标记事故相关事件的刻度，CIDAS 数据库中涉及到时间的变量包括事故日期、事故星期、月份、年份、碰撞时间、报警时间、通知事故调查小组的时

间、第一辆救护车到达现场的时间、医生达到现场的时间、事故调查小组达到现场的时间、事故调查小组离开现场的时间、伤者离开现场的时间、伤者到达医院的时间、伤者到达医院到得到救治的时间、人员出生日期、人员年龄、车辆注册日期等，这些时间有年月日类型、时分类型、整型数值。首先要检查这些变量的数据是否缺失，然后过滤掉特殊值（其他和未知项），再对其格式进行规范化，时间格式化函数包括 format、as.Data、as.POSIXct 等，时间切割和运算的函数包括 strptime、strftime、difftime、lubridate 系列等。不但要对时间的先后顺序进行判别，而且要对时间差的合理性进行判别，如图 4 所示，将离散的时、分变量组合成规范的时间格式，在排除碰撞时间和报警时间未知的情况后，求出两个时间的时间差并转化为分钟格式，分别对时间顺序矛盾和时间跨度异常进行提示，提示信息分别存放到向量 a 和向量 b 中，其中对时间的格式化、时间差值和时间切割嵌套在一个语句中，精简代码的同时也减少了运算时间。

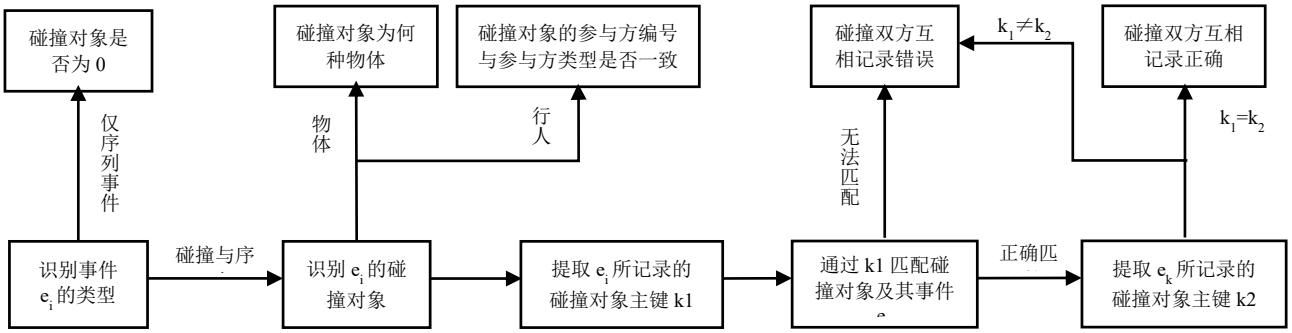

图 4 时间类型数据清洗与核查示例

为了更深入的研究碰撞的前后过程，CIDAS 将每个车辆的碰撞过程分割为数个事件，事件可分为含碰撞事件或仅序列事件，为了清晰的记录每个参与方碰撞的对象，在数据库里设置了记录碰撞对方的参与方编号和事件编号的变量，因此参与同一次碰撞的所有参与方均会记录对方的编号及其正在进行的事件。为了确保每个事件所记录的碰撞对方信息正确，需要对每个参与方的每次事件进行循环识别，其判断流程如图 5 所示。

```
E01$collisiontime < − paste(E01$UZEITH, E01$UZEITM, sep=":")
E01$warningtime < − paste(E01$MELDUNGH, E01$MELDUNGM, sep = ":")
if ((!88 %in% c(E01$UZEITH, E01$UZEITM, E01$MELDUNGH, E01$MELDUNGM) | !99 %in% c(E01$UZEITH, E01$UZEITM,
E01$MELDUNGH, E01$MELDUNGM))){
  if (str_extract_all(difftime(as.POSIXct(E01$warningtime, format = "%H:%M"), as.POSIXct(E01$collisiontime, format =
"%H:%M"), units = "mins"), "[0-9]+") < 0) print(a)
 if (str_extract_all(difftime(as.POSIXct(E01$warningtime, format = "%H:%M"), as.POSIXct(E01$collisiontime, format =
"%H:%M"), units = "mins"), "[0-9]+") > 60) print(b)
}
```

图 5 碰撞事件流程

在数据清洗和问题识别的过程中，应将同种类型的参数归纳到一个控制流中，一方面可以减少报告生成的耗时，另一方面也能减少代码编译的工作量。例如在对人体伤情概述和单个损伤进行处理时，头部、面部、颈部等 10 个身体部位可以处于一个控制流内，在完成一个部位的清洗和识别后只需将其他身体部位的变量名称进行替换即可完成所有的工作，类似的情况包括同种参与方类型、同种类型的车辆。为了简化工作量和提高效率，在编译过程中应使用一些高效率的函数，当现有函数不能满足要求而需要多次使用该功能时，可以通过 function 和 source 来自定义函数和调用自定义函数，表 2 列出了本系统中常用的一些函数。

表 2 事故数据处理常用函数

功能类型	函数名称
缺失值处理	is.na、na.omit、na.rm、complete.case
字符串处理	nchar、substr、strsplit、gsub、grep、charmatch、chartr
类型转化	as.numeric、as.character、as.vector、as.factor、as.logical、as.data.frame
数据框运算	apply 系列、aggregate、reshape、melt、cast、plyr
数据管理	length、rev、rep、seq、subset、rbind、cbind

3.3 生成报告

为了确保系统能够适应所有形态的事故，在脚本语言编译完成后需要利用一些特殊的事故案例进行测试，这些特殊案例包括单车事故、多车事故、同一车辆发生多次碰撞的事故、多种类型 VRU 同时参与的事故和人员伤口数较多的事故，借此发现系统的漏洞、完善系统功能。该系统可以生成三种形式的动态报告，包括 PDF、Word 和 HTML，系统按照块（chunk）的顺序逐个进行处理，如果其中任何一个参数出现数据错误，则会中断报告生成，因此在进行数据清洗和问题识别的时候要合理设计结构，在完成所有块的处理后会自动弹出报告结果，如图 6 所示。报告可以用文字和图表的形式来表达所检测到的数据质量问题，录入人可以根据报告所提示的信息对自己录入的案例进行核查和修改，以此形成固定的工作流程。

图 6 系统自动生成的案例质量报告

4 总结和展望

数据质量控制需要覆盖数据的整个生命周期，但质量监管的重点是数据获取的源头和数据产生的过程，结合交通事故数据的不可逆和信息隐藏的特点，CIDAS 数据质量的管理应从两个方面着手，一方面是数据收集过程，另一方面是数据录入过程。动态报告生成系统的运用能够切实提高数据的质量，将数据录入与数据质量监控紧密结合起来，及时发现问题和解决问题，该方法还可以运用到类似的常态化报告制作中，同时也为探索事故数据和事故分析报告可重复性研究提供了指导方向。在系统构建过程中，笔者总结了以下关于数据质量管理的思考：

1） 正确认识数据的质量问题。数据的质量问题是任何数据相关企业都无法回避的，数据的产生、维护、传输、使用过程都可能引发数据质量问题，找到问题来源，明确责任主体，才能提出相应的解决策略。

2） 建立数据标准体系和质量管控规范。数据要遵循标准先行的规则，只有制定了完善的数据标准体系才能获取到高质量的数据，而数据质量管控规范则有助于明确相关人员在数据产生、存储、应用整个生命周期中工作内容和工作流程，形成统一的数据质量管理体系。

3） 加深数据的挖掘才能突破数据质量的瓶颈。数据是信息的一种表现形式，深入挖掘数据和使用数据才能发现信息之间的内在联系，从而指导数据质量的提升和数据价值的开发。

参考文献

[1] 刘飞. 中国企业数据集成与数据质量市场白皮书 [R]. 北京: IDC 中国 , 2008.
[2] 李磊. 基于 ETL 的数据集成及交换系统的实现与优化[D].北方工业大学,2018.
[3] Rahm, E., Do, H.H. Data cleaning: problems and current approaches. IEEE Data Engineering Bulletin, 2000,23(4):3~13.
[4] 陈孟婕.数据质量管理与数据清洗技术的研究应用. 北京: 中国邮电大学, 2013.
[5] Jeya Padmanaban, R. Ravishankar, Ajit Dandapani. Methodology to Derive National Estimates of Injuries and Fatalities in Road Traffic Crashes in India. SAE International. 2017-26-0016
[6] Otte (2002). Scientific Approach and Methodology of a New In-Depth-Investigation Study in Germany so called GIDAS.
[7] T. Kiuchi, T. Motomura, T. Nishimoto, and H. Ishikawa. The in-depth accident study to evaluate the advanced automatic collision notification system in Japan. 18th International Conference Road Safety on Five Continents. 2018.
[8] Michael R. Elliott, Alexa Resler, Carol A. Flannagan, Jonathan D. Rupp. Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. Accident Analysis & Prevention, Volume 42, Issue 2. 2010. Pages 530-539.
[9] M. Bougueroua, L. Carnis. Economic development, mobility and traffic accidents in Algeria. Accident Analysis & Prevention, Volume 92, 2016. Pages 168-174 ISSN 0001-4575.
[10] 杨霞,吴东伟.R 语言在大数据处理中的应用[J].科技资讯,2013(23):19-20.
[11] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.
[12] Yihui Xie. knitr: A general-purpose package for dynamic report generation in R, 2013. R package version 1.1.8.
[13] Anthony Rossini. Literate statistical analysis. In Proceedings of the 2nd International Workshop on Distributed Statistical Computing, pages 15-17, 2002.
[14] Baumer B, Cetinkaya-Rundel M, Bray A, Loi L, Horton NJ. R Markdown: Integrating a reproducible analysis tool into introductory statistics. arXiv preprint arXiv:1402.1894. 2014 Feb 8.
[15] 王俊峰.浅析 SQL 语言在数据库中的应用[J].计算机光盘软件与应用,2013,16(07):102+107.