

基于文本信息抽取的列控车载设备故障发现

陈曦, 李润梅, 王剑, 董舒韵

(北京交通大学, 北京 100044)

摘要: 列车运行控制系统车载安全计算机日志文件和地面设备维护工作人员撰写的交接班记录——质量分析工区工作记录单, 包含大量反映车载设备运行状态的信息, 是车载设备故障发现和诊断的重要依据。对这两类文本数据的利用目前还局限于工作人员人工记录、查询和解读, 存在效率低、主观性强、易遗漏出错等问题。针对以上问题, 以车载安全计算机日志文件和质量分析工区工作记录单为研究对象, 设计信息抽取方法及故障发现自动化方案, 以代替现有人工信息抽取和故障记录的工作方式。分析了车载安全计算机日志文件的数据特点, 利用分词、智能信息抽取算法对非结构化的车载设备数据进行处理, 并结合历史质量分析工区工作记录单数据进行故障匹配, 从而构建故障分析字典, 实现基于正则表达式算法的故障语句自动定位。设计实现了故障自动发现软件, 验证了故障发现方案的可行性和软件的功能。

关键词: 车载设备; 车载安全计算机日志文件; 质量分析工区工作记录单; 故障发现; 信息抽取; 故障分析字典
中图分类号: TP181

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202120

Fault discovery based on text information extraction for on-board equipment of CTCS

CHEN Xi, LI Runmei, WANG Jian, DONG Shuyun

Beijing Jiaotong University, Beijing 100044, China

Abstract: The on-board safety computer of Chinese Train Control System generates a large number of Log files during operation every day, which together with the quality analysis record sheet, record the running status of all equipment and provide data support for the fault discovery and processing of the equipment of train control system. The use of these two types of text data is still limited to manual record, query and analysis, which has the problems of low efficiency, strong subjectivity and easy omission and error. The on-board safety computer Log files and the quality analysis record sheet were taken as the research object, and designs an information extraction method and the fault discovery automation scheme to replace the existing manual working mode. The information of Log files and manual recording by the staffs were analyzed, the word segmentation algorithm and information extraction algorithm were selected to find useful information automatically from two kinds of record to avoiding the cumbersome manual search. Then a failure analysis dictionary was built. The automatic fault discovery application was established by using regular expression algorithm. The automatic fault discovery and analysis functions of the application were validated by experiment.

Key words: on-board equipment, on-board safety computer Log file, quality analysis record sheet, fault discovery, information extraction, failure analysis dictionary

收稿日期: 2020-08-02; 修回日期: 2020-11-26

通信作者: 王剑, wangj@bjtu.edu.cn

基金项目: 国家重点研发计划基金资助项目 (No.2018YFB1201500)

Foundation Item: The National Key Research and Development Program of China (No.2018YFB1201500)

1 引言

列车运行控制系统 (Chinese train control system, CTCS) 包括车载子系统和地面子系统。车载子系统包括车载设备和无线车载模块; 地面子系统包括应答器、轨道电路、列车控制中心 (train control center, TCC) / 无线闭塞中心 (radio block center, RBC) 子系统等。CTCS 取消了地面信号机, 车载设备基于安全计算机, 从地面子系统提取信息来控制列车运行^[1-2]。

CTCS-3 级列车运行控制系统基于 GSM-R 无线通信实现列车运行控制, 保证高速列车安全、可靠、高效运行。列车自动保护 (automatic train protection, ATP) 系统是车载设备的重要组成部分, ATP 系统内部的计算系统对列车运行过程中的车载设备状况进行实时记录, 并且以多种格式的文件进行存储, 其中车载安全计算机日志文件 (以下简称日志文件) 就是众多的记录文件之一, 是 CTCS 运行过程设备状态的重要信息载体^[3-5]。

在每天例行设备检修工作时, 工作人员会撰写交接班记录, 形成质量分析工区工作记录单。这些质量分析工区工作记录单是对列车运行过程中发生的问题或故障的概括。日志文件和工作人员的交接班记录之间有着重要的内在联系。工作人员在填写质量分析工区工作记录单之后, 需要手工查阅日志文件, 将故障人工描述和计算机故障语句对应起来, 对故障语句进行解读, 进而对质量分析工区工作记录单记录的故障进行更加全面的描述。因此, 日志文件和质量分析工区工作记录单就成为车载设备优化运营和维护的重要数据, 是车载设备故障发现的信息依据。

目前基于日志文件的人工信息处理和记录模式是车载设备故障发现及维护工作的关键环节, 并已有成熟的工作流程, 但随着 CTCS 的发展, 以下问题越来越凸显^[6]。

(1) 日志文件数据量越来越大, 以本文处理的 CTCS3-300T 动车组的数据为例, 仅 2016 年 4 月的日志文件记录就高达 120 000 多条, 异常记录 300 多条, 人工查阅日志和异常记录的工作量非常大。

(2) 人工描述主观性易导致遗漏、出错和记录的不一致性。质量分析工区工作记录单的填写主要依靠工作人员的经验, 同一故障可能有不同的故障原因, 对同一故障也可能会有多种描述方式, 文本

的规范性和可用性差。

(3) 基于人工进行日志文件和质量分析工区工作记录单的关联和核对, 效率极低, 且容易漏查。

随着智能信息处理算法研究的深入和计算机技术的进步, 很多依赖人工知识和经验的工作逐渐被计算机替代。本文针对 CTCS 日志文件数据处理及故障发现问题, 引入智能信息处理算法, 构建故障分析字典, 开发了故障自动发现和定位软件^[6]。本文的主要工作如下:

- 提出日志文件数据的智能处理算法, 本文选择 TFIDF (term frequency and inverse documentation frequency) 算法^[7]进行信息抽取, 自动定位描述车载设备故障的语句;

- 提出质量分析工区工作记录单文本处理算法, 本文选择最大匹配 (maximum matching, MM) 分词算法^[8]进行分词, 从历史工作记录文本中获取故障类型和故障处理方案, 搭建了一个包括规范的故障名、故障语句、故障处理方案的故障分析字典;

- 根据日志文件非结构化文本^[9]的数据特点, 利用正则表达式 (regular expression) 算法^[10]实现故障语句的自动查询定位, 依托故障分析字典, 开发了故障自动发现和定位软件。

2 问题分析与技术方案

2.1 日志文件分析

本文使用的数据来源于某铁路局动车段 2016 年 4 月 30 天的日志文件, 记录数据 120 000 多条, 工作人员的异常记录 300 多条。局部日志文件如图 1 所示。

每个日志文件包含 250 条设备信息记录, 每条记录分为两部分, 第一部分主要包括时间信息, 第二部分记录这个时间车载设备的运行状态。如果出现故障, 会有故障名称等。如 “13 16-04-01 09:24:19;129 FID: stmm.cpp LID: 3459 TID:SMGM_Log Task 000077199 ATP A 4A29R000 STM reports failure state.”, 该句是一句完整的信息, “13” 是该条信息的序号, 表示该信息是本文档中自上而下的第 13 条信息; “16-04-01 09:24:19” 表示该条信息记录的时间是 2016 年 4 月 1 日的 9 时 24 分 19 秒; “129 FID: stmm.cpp LID: 3459 TID: SMGM_Log Task 000077199 ATP A 4A29R000” 表示该条信息的代码等。每条信息中的该部分都不相同, 但是该部分对

故障的描述没有实际的意义，因此在本文的研究中不予关注。“STM reports failure state.”是该条信息中最有效的部分，描述了列车运行过程中的状态，本条信息表示 STM（specific transmission module）的报告处于故障状态。

```
-----
CSS: RTS: 1.10.2.3
Filename
Time      SUN OCT 26 13:49:06 2014
Address   0x40000
Size      3767200
Checksum  341920339
Comment   3EST120-50 CSS Run time system
RTS_Mon-> ae
LIST OF ALL STORED APPLICATION EVENTS
=====
1 16-04-03 00:52:04;135 FID:   tih_a.cpp LID: 2754 TID:SMGM_LogTask
000041253   ATP A A1C2R000 Vehicle Ready Active (Lifesign 8 OK).
2 16-04-03 00:52:03;638 FID:   tih_a.cpp LID: 2005 TID:SMGM_LogTask
000040955   ATP A A1B1R720 Alive Lifesign 2.
3 16-04-03 00:52:03;637 FID:   tih_a.cpp LID: 1999 TID:SMGM_LogTask
000040655   ATP A A1B0R710 Alive Lifesign 1.
4 16-04-03 00:52:01;620 FID:   stam.cpp LID: 5127 TID:SMGM_LogTask
000038541   ATP A A413R000 STM connection running.
5 16-04-03 00:52:00;620 FID:tih_a_stmdat LID: 461 TID:SMGM_LogTask
000037942   ATP A A069D0B0 Profibus TI-H to TSG connected.
6 16-04-03 00:52:00;119 FID:bih_a_stmdat LID: 468 TID:SMGM_LogTask
000037343   ATP A A4C9D0B0 Profibus BI-H to TSG connected.

Hit any key to continue or ESC to stop ...
7 16-04-03 00:51:59;619 FID:JruHandler.c LID: 1750 TID:SMGM_LogTask
000036745   ATP A BE10R000 Unexpected telegram counter in JRU status
```

图1 局部日志文件

2.2 质量分析工区工作记录单

工作人员在每天例行设备检修工作中，根据设备情况，结合日志文件中异常的信息进行记录和分析，形成质量分析工区工作记录单，详细记录故障并给出处理方式和记录处理结果。表1为某个关于CTCS3-300T的部分记录单。

表1 质量分析工区工作记录单（CTCS3-300T）

项	内容
车号	5542
端号	0
系	b
故障日期	4月1日
故障时间	12:49:24
故障描述	电台超时，C3掉C2，JRJ中：车载ATP在12:49:20收到地面发送的信息包后，在10秒内未再次收到地面消息，导致无线连接超时，C3降级C2
通知情况	已通知南所
处理情况	测试正常、灯显正常，单呼试验，试验结果良好，已更换右侧电台

表1给出了4月1日的工作人员工作记录的前8项。前两项是列车的车号和端号，第3项是出现故障或者问题的车载设备所在的系（有a、b两个

不同的系进行备份运作），第4项和第5项主要是故障产生的时间。前5项的信息一般可以从日志文件中找到对应的描述，因此质量分析工区工作记录单和日志文件就可以建立一一对应的关系。以表1数据为例，通过查找日志文件中车号为5542的文档，找到对应时间的记录为“16-04-01 12:49:24;838 FID: wi_a_RBCcont LID: 856 TID: SMGM_Log Task 004751814 ATP A 5224R000 1 1 1 radio service lost”，从而可以找出与工作人员故障描述对应的故障语句。记录的第6项主要是故障描述，该描述是基于工作人员的经验 and 现场的状况进行的，口语化严重，并且因为不同的交接班记录由不同的工作人员填写，对同一个故障的描述可能存在很大的差异，但是在关键词语上会比较相似。第7项和第8项记录了故障发生之后的处理方式。

值得注意的是，工作人员的质量分析工区工作记录单是不完全记录，也就是说，工作人员并未把车载设备的所有故障或者运行状况记录下来，只是记录了表现出来的故障，而对隐而未现的问题并未记录，因此还需要对日志文件进行更加细致的搜索，才能获取列车运行状态的全面信息。质量分析工区工作记录单虽然不够全面，但是该类文档具有很大的参考价值，是对故障的人工确认和详细描述，可以为车载设备的故障发现提供良好的数据支持。

2.3 两类文本数据的分析与算法应用

日志文件和质量分析工区工作记录单属于两类文本数据，这两类文本数据既紧密联系又有所不同。两者内容不同，日志文件是列车运行中实时记录下来、全面反映列车运行状况的文件，质量分析工区工作记录单是对表现出来的故障的详细信息描述及处理方案。两者形式不同，质量分析工区工作记录单是中文记录，日志文件是非结构化英文记录。因此，对这两类文本数据的分析需要采用不同的算法。

对于中文文本，一般利用分词算法进行处理，常见的分词算法主要有基于字符串匹配、基于理解、基于统计、基于语义的分词算法等^[11-13]。本文采用基于字符串的最大匹配分词算法处理质量分析工区工作记录单。该方法针对给定词表，从句子的一端开始扫描，尽可能给待分词的句子匹配词表中最长的词^[14]。

对于非结构化英文文本，常见的信息抽取算法

有基于文档频率 (document frequency, DF) 法、互信息 (mutual information, MI) 法^[15]和 TFIDF 算法等。本文采用简洁而准确率高的 TFIDF 算法对日志文件进行处理。TFIDF 算法选用词频 (term frequency, TF) 与逆文本频率 (inverse document frequency, IDF) 指数来评估某词条对于一个语料库中的某份文件的重要程度。TF 为 t_i 在文档 d 中出现的频率, 记为 $\text{tf}(t_i, d)$; DF 为特征项的文档频率, 即在全部文档集中出现 t_i 的文档数量, 记为 $\text{df}(t_i)$; IDF 为特征项的逆文档频率, 即包含词条 t_i 的文档越少, IDF 越大, 则词条 t_i 具有更好的类别区分能力, 记为 $\text{idf}(t_i) = \log(n/\text{df}(t_i))$, 其中 n 表示训练样本总数; 则 t_i 的权重 W_i 如式 (1) 所示:

$$W_i = \text{tf}(t_i, d) \times \text{idf}(t_i) \quad (1)$$

TFIDF 算法的原理为: 某词条在某文档中出现的频率较高, 而在其他文档中出现的频率较低, 说明该词条适合用于分类; 包含某个词条的文档越少, 词条区分文档内容的能力越强^[7]。

日志文件和质量分析工区工作记录是非结构化文本数据, 因此本文用正则表达式算法进行文本信息的降维处理, 匹配特征项, 实现日志文件的自动查询、阅读。

3 算法与故障分析字典搭建

基于上述分析, 本文采用智能信息抽取算法对日志文件和质量分析工区工作记录单进行分析处理; 构造故障分析字典, 设计自动故障发现软件, 实现对日志文件和质量分析工区工作记录单信息的挖掘和自动匹配。

3.1 分词算法

对表 1 所示质量分析工区工作记录单中的故障描述语言进行分词, 主要步骤如下。

步骤 1: 收集处理信息时间范围内所有质量分析工区工作记录单故障记录, 构造原始字典, 统计字典的最长词条, 记为 MaxLen, 本研究处理数据的最长词条为 MaxLen=14。

步骤 2: 以标点符号为标识进行分句, 并进行分词。

步骤 3: 以原始字典中的第一个字为初始值, 向后取 n 个字, $n \leq \text{MaxLen}$, 跳转至步骤 4。

步骤 4: 如果取到的 n 个字是一个正确的词语,

则分词成功, 跳转至步骤 6, 否则跳转至步骤 5。

步骤 5: 去掉 n 个字的最后一个字, 令 $n = n - 1$, 跳转至步骤 4, 继续判断和操作, 如此循环, 直到 $n = 1$ 。

步骤 6: 将词条中的词语去掉, 如果该词条为空, 则算法结束, 否则跳转至步骤 3。

步骤 7: 整理结果, 得到较为完整的故障名称和故障处理方案。

基于字符串的最大匹配分词算法可以对工作记录中冗杂数据进行分词, 通过对匹配的内容进行整理, 即可确定质量分析工区工作记录单中常见的故障类型和故障的解决方案。图 2 所示为将某质量分析工区工作记录单进行分词之后的结果, 可以得到如“多普勒雷达故障”和“启机 BSA 故障”等常见故障名称。

```
str=
Columns 1 through 6
'车给地发送136包后' '未收到RBC回复' '多普勒雷达故障'
'jru数据' 'C3掉C2' '报EBFR反馈失败'

Columns 7 through 11
'导致测试第8步失败' 'ATP未死机' '车给地发送136包后'
'未收到RBC回复' '启机BSA故障'
```

图2 分词结果示例

3.2 TFIDF 算法

使用 TFIDF 算法对日志文件进行信息抽取, 步骤如下。

步骤 1: 基于质量分析工区工作记录单的故障记录时间, 确定对应的所有日志文件。

步骤 2: 对日志文件中的每条语句进行首次分词, 得到单个词语。

步骤 3: 计算单个词语在每一个日志文件中出现的次数 TF。

步骤 4: 计算单个词语在所有的日志文件中出现的次数 DF, 计算 IDF。

步骤 5: 根据式 (1) 计算权重, 并将其放大。

步骤 6: 根据权重计算结果确定故障语句中的关键词。

步骤 7: 整理得到的关键词, 构建故障语句词库。

本文以“Speed sensor failure”为例, 计算这个词语在某铁路局动车段 2016 年 4 月 30 天内所有日志文件中出现的权重, 其结果如图 3 所示。

weight=		
5.388 2	5.388 2	5.388 2
15.185	15.185	15.185
0	0	0
0	0	0
4.912 8	0	0
0	0	0
0	0	0
5.694 4	5.694 4	5.694 4
0	167.035 1	0
0	0	0
0	0	0
4.008 8	4.008 8	4.008 8
0	0	0
0	0	167.035 1
0	0	0

图 3 权重计算结果

图 3 是所有日志文件中截取到的“Speed sensor failure”的权重计算结果。其中每一行的数据是在同一天产生的，分别代表“speed”“sensor”“failure”的权重。从结果可以看出，第 1、2、5 等行对应的日期的日志文件记录中都出现过这一故障语句，第 3、4、6 等行对应日期的日志文件记录中没有出现这类故障语句，因此，根据权重计算结果可以查找出对应的故障名称。

3.3 故障分析字典

基于前文提到的分词算法和信息抽取 TFIDF 算法，可以从质量分析工区工作记录单和对应日期的日志文件中提取关键故障信息，以时间和故障名称为索引，建立自动记录信息和人工记录信息的对应关系，从而构造故障分析字典，为自动故障发现奠定了算法和数据基础。

车载设备故障分析字典（部分）见表 2，包括故障关键词、故障描述、故障处理方案 3 个部分。故障关键词是根据日志文件中的记录进行信息抽取得到的结果。故障描述和故障处理方案则是利用分词对质量分析工区工作记录单进行信息抽取得到的关键词。

4 基于文本信息抽取的故障发现实验

4.1 故障发现系统功能需求

本文以 CTCS 车载设备故障发现为研究背景，以日志文件和质量分析工区工作记录单为数据源，依托智能信息抽取算法，开发了故障发现软件，软件需求如下：

- 完备的故障分析字典，基于人工经验的专业故障分析功能；
- 便捷的故障定位功能；
- 良好的人机交互能力，简洁的操作界面。

4.2 故障发现软件的操作流程

本文基于 MATLAB 开发的故障发现软件包括 3 个功能程序：文件读取子程序、故障定位子程序和故障处理子程序，其中故障定位子程序基于正则表达式的程序语言对日志文件中的信息进行降维处理，实现故障语句定位。故障发现软件主界面如图 4 所示，软件操作流程如下。

步骤 1：用户登录系统；新用户需先注册，注册成功后即可登录使用。

步骤 2：用户进入主界面，点击“浏览”，选择并读取需要进行故障信息抽取的日志文件。

步骤 3：从下拉菜单中选择故障语句或故障名称，点击“查询”，则可以在定位界面显示对应故障的位置、记录时间等，该功能可以自动统计所选日志文件中某故障发生的次数。如果出现新的故障

表 2

车载设备故障分析字典（部分）

故障关键词	故障描述	故障处理方案
ATP A LOG MSG SPL_Initialise-104	ATPCU 硬件故障	更换 ATPCU 模块后测试
BSA Permanent Error	启机 BSA 故障	换系重启
Speed sensor failure	多普勒雷达故障	核对轮径值设置是否有误；清洁速度传感器探头；雷达工作若正常，则在雷达参数标定后观察，否则更换
JRU message could not be sent (event size too long)	JRU 硬件故障	重启 ATP 和 JRU 故障未消除，确定为 JRU 硬件故障，予以更换
EBFR state wrong	制动反馈继电器故障	检查 BFB 继电器与底座安装是否紧固没有松动；若无松动重启后无法恢复则更换处理
Network resource not available	网络资源不可用	C3 降级为 C2，更换电台

语句或名称,则跳转至步骤5。

步骤4: 点击“分析”按钮,在故障分析界面显示故障产生的原因和故障的处理方案。

步骤5: 基于质量分析工区工作记录单中新故障的产生时间,查找新故障的故障语句,并将其录入故障分析字典。

步骤6: 分别点击“导出1”和“导出2”按钮,可以导出定位界面和故障分析界面的内容。如点击“导出3”按钮,则导出故障发现与分析记录单,这个记录单可以替代以往基于人工方式得到的质量分析工区工作记录单,同时可以得到故障分析信息,从而实现基于日志文件的故障发现、记录及处理的自动化。

4.3 实验成果

图4所示为故障发现软件使用主界面,包括导入日志文件和故障名称选择的查询界面;日志文件定位故障记录的定位界面以及故障描述和处理的故障分析界面。



图4 故障发现软件主界面

在查询界面点击“浏览”按键,可以打开日志文件并读取信息。查询界面有两个不同的下拉菜单,第一个是日志文件中的故障语句选项,第二个是故障分析字典中的故障中文名称。在进行故障定位时,可以选择适合的下拉菜单,从中选择想要查找的故障语句或者故障名称,查询的同时可以统计故障次数。

本文在选择2016年4月1日的一个日志文件之后,选择“多普勒雷达故障”这一故障名称。点击“分析”按钮,则在定位界面和故障分析界面分别展示了与这一故障相关的信息。

点击“导出2”按钮,可以将故障语句和故障分析描述以Excel格式导出。

主界面的左侧有时间查询的选项,对于新发现的故障名称,可以选择时间作为查询条件,从日志

文件中定位这一故障语句,并填充至故障分析字典。

本文的仿真实验使用了某铁路局动车段2016年4月的日志文件和质量分析工区工作记录单,共有78个车次的数据,日志文件记录120 000多条,工作人员的异常记录300多条。在本实验中,将这300多条记录用于构建词典,形成的故障分析字典记录约100多条。在更多的验证实验中,笔者增加异常记录至千余条,但故障分析字典记录没有明显增加。通过对质量分析工区工作记录单记录故障进行整理可以发现,大部分故障类型可以归为几大类设备或运行故障,虽然总故障数目随着记录的增加而增加,但故障类型没有明显增加,主要原因是车载设备本身具有极高的可靠性,并且设备模块化。

本文提出的算法和软件很好地实现了日志故障语句的自动定位和分析,在基于历史数据的实验中,故障发现准确率为100%,有效提高了故障发现及记录分析的工作效率。

5 结束语

日志文件和地面设备维护工作人员撰写的交接班记录——质量分析工区工作记录单,包含着大量的反映车载设备运行状态的信息,是车载设备故障发现和诊断的重要依据。然而目前对CTCS车载设备进行故障诊断和发现的研究中,利用日志文件和质量分析工区工作记录单的研究还很少见。对这两类文本数据的利用,目前还局限于工作人员人工记录、查询和解读,存在效率低、主观性强等问题。特别是,日志文件的日常分析记录工作关系到CTCS运行的安全,是一项工作量非常大的日常维护任务。人工方式的分析和记录很容易出现记录遗漏和分析差错。本文设计了基于两类信息的故障发现研究方法,构建了基于分词算法、智能信息抽取算法的故障分析字典,利用正则表达式算法实现故障语句的自动定位。最后通过MATLAB编程开发了一个可以适用于Windows操作系统的故障发现软件,该软件通过基于历史质量分析工区记录单信息获得的故障分析字典,对导入的日志文件进行故障语句定位和分析。将两类关于车载设备安全的数据有机结合、延伸应用。基于实际数据的测试实验表明了所提算法和软件的合理性及可应用性。

参考文献:

- [1] 张铁增. 列车运行控制系统[M]. 北京: 中国铁道出版社, 2014.

- ZHANG T Z. Chinese train control system[M]. Beijing: China Railway Publishing House, 2014.
- [2] 王洪良, 王少华, 赵松. 列控设备动态检测系统[M]. 北京: 中国铁道出版社, 2014.
- WANG H L, WANG S H, ZHAO S. Train control equipment dynamic detection system[M]. Beijing: China Railway Publishing House, 2014.
- [3] 赵阳, 徐田华. 基于文本挖掘的高铁信号系统车载设备故障诊断[J]. 铁道学报, 2015, 37(8): 53-59.
- ZHAO Y, XU T H. Text mining based fault diagnosis for vehicle on-board equipment of high speed railway signal system[J]. Journal of the China Railway Society, 2015, 37(8): 53-59.
- [4] 石先明. 探究列控系统对动车组制动系统故障后的安全防护作用[J]. 铁路通信信号工程技术, 2013, 10(1): 5-11.
- SHI X M. Safety protection after the braking system fault of train control system[J]. Railway Signalling and Communication Engineering, 2013, 10(1): 5-11.
- [5] 张友兵, 马麟, 张国振, 等. CTCS-3 级列控车载记录下载器的软件设计[J]. 铁道通信信号, 2012, 48(12): 13-15.
- ZHANG Y B, MA L, ZHANG G Z, et al. Software design of CTCS-3 carrier record downloader[J]. Railway Signal and Communication, 2012, 48(12): 13-15.
- [6] 陈曦. 基于文本信息抽取的高铁车载设备故障发现的理论与方法[D]. 北京: 北京交通大学, 2017.
- CHEN X. The theory and method of fault discovery based on text information extraction for on-board vehicle equipment of high speed railway[D]. Beijing: Beijing Jiaotong University, 2017.
- [7] 施聪莺, 徐朝军, 杨晓江. TFIDF 算法研究综述[J]. 计算机应用, 2009, 29(z1): 167-170, 180.
- SHI C Y, XU C J, YANG X J. Study of TFIDF algorithm[J]. Journal of Computer Applications, 2009, 29(z1): 167-170, 180.
- [8] 吴栋, 滕育平. 中文信息搜索引擎中的分词与检索技术[J]. 计算机应用, 2004, 24(7): 128-131.
- WU D, TENG Y P. Word segment and search techniques for chinese information search engines[J]. Journal of Computer Applications, 2004, 24(7): 128-131.
- [9] 张智雄, 吴振新, 赵琦, 等. 非结构化文本中内容对象抽取的技术方法综述[J]. 数字图书馆论坛, 2008(9): 1-12, 36.
- ZHANG Z X, WU Z X, ZHAO Q, et al. Review of the technologies and methods for extracting content objects from unstructured text[J]. Digital Library Forum, 2008(9): 1-12, 36.
- [10] 陈航宇. 正则表达式匹配算法研究[D]. 秦皇岛: 燕山大学, 2016.
- CHEN H Y. Research on regular expression matching algorithms[D]. Qinhuangdao: Yanshan University, 2016.
- [11] HUANG C N, ZHAO H. Chinese word segmentation: a decade review[J]. Journal of Chinese Information Processing, 2007, 21(3): 8-19.
- [12] MO J W, ZHENG Y, SHOU Z Y. Improved Chinese word segmentation method based on dictionary[J]. Computer Engineering and Design, 2013, 34(5): 1802-1807.
- [13] 揭春雨, 刘源, 梁南元. 论汉语自动分词方法[J]. 中文信息学报,

1989, 3(1): 3-11.

JIE C Y, LIU Y, LIANG N Y. The method of automatic Chinese word segmentation[J]. Journal of Chinese Information Processing, 1989, 3(1): 3-11.

[14] JIN G. Critical tokenization and its properties[J]. Computational Linguistics, 1997, 23(4): 569-596.

[15] NGUYEN X V, CHAN J, ROMANO S, et al. Effective global approaches for mutual information based feature selection[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 512-521.

[作者简介]



陈曦 (1992-), 女, 北京交通大学硕士生, 主要研究方向为轨道交通系统数据处理算法。



李润梅 (1975-), 女, 博士, 北京交通大学副教授, 主要研究方向为智能车路协同系统、无人车控制关键技术、交通大数据处理与预测。



王剑 (1978-), 男, 博士, 北京交通大学教授, 主要研究方向为列车运行控制及卫星导航在铁路领域的理论、技术创新。



董舒韵 (1996-) 女, 北京交通大学硕士生, 主要研究方向为数据处理算法。