http：//myfj. cnjournals. com myfj_sadri@163. com （021）20866796

DOI：10. 19416/j. cnki. 1674-9804. 2024. 02. 024

基于指标体系的相似故障问题推荐方法研究

俞立群 *

（上海飞机设计研究院，上海 201210）

摘　要：针对试飞故障问题数量大、内容多、无法高效利用其信息价值的问题，提出基于运行阶段、ATA 章节号、故障文本和 CAS、OMS 信息的指标体系相似故障问题推荐方法。依据各指标特征，设计不同的相似度计算方法。此外，利用正则表达式，提出一种自动从故障文本描述中提取 CAS 信息和 OMS 信息的方法。最后通过实例验证相对于只选取运行阶段、ATA 章节号和故障文本作为指标的推荐算法以及只使用 TF-IDF 方法的算法，本文所提算法在准确率上分别提高了 25% 和 28%，在召回率上分别提高了约 27% 和 31%，可以为相似试飞故障问题的推荐提供参考。

关键词：故障推荐；指标体系；CAS-OMS 信息；TF-IDF

中图分类号：TP391. 1　　**文献标识码**：A　　**OSID**：

0　引言

民用飞机在试飞阶段积累了大量的故障和纠正措施等文本数据，这些经验数据蕴含了很多有价值的信息，在很大程度上能为故障诊断提供重要支持，同时在试飞阶段尚未有成熟的手册供工程师直接查询。因此，工程师在遇到故障时可以从已有的经验库中查询是否发生过相似故障及对应解决措施。然而，由于飞机试飞故障案例文本数量大且内容多，人工查找花费的时间长，无法高效利用这些文本的信息价值。为了解决这个问题，可以采用故障案例智能推荐方法。

在现有的研究中，机器学习技术在故障匹配中得到了广泛应用。刘科研[1]等提出利用支持向量机比较历史数据和实时数据进行故障判别和匹配；王永坚[2]等提出建立故障朴素贝叶斯网络用于对船舶发动机故障模式进行匹配；徐涯昕[3]等基于故障记录文本数据构建 CNN-BiLSTM 网络对中小数控机床发生故障的原因进行类别匹配；ZHOU Faguo[4]等基于规则算法对故障案例标题进行了匹配；Shah[5]等则提出聚类算法对故障案例进行匹配和分类。以上研究仅针对故障类别进行匹配以进一步缩小故障原因的判别范围，但未细化到单个故障问题的匹配和推荐。

针对单个故障问题的匹配，杨祎[6]等提出改进 BERT 的故障案例智能匹配方法用于电网故障案例的匹配；唐瑞春[7]等在手机故障案例相似度的任务中提出了一种新的算法；祖月芳[8]等综合考虑词性、语义和词所处位置等因素，提出一种衡量故障文本相似度的算法。上述研究都只从故障文本描述本身出发，未充分考虑设备特性等其他因素。王峻洲[9]等人提出基于相似体系的民机结构超手册维修案例分析方法，通过建立案例相似度体系进行故障案例匹配，但是关键信息需要人工从故障文本描述中进行提取补充。

TF-IDF 方法是文本相似度衡量中一种经典方法，因其简单有效且可解释性好，在工程领域得到了广泛应用。TF-IDF 方法最早由 Dierk[10]在 1972 年提出，后续的改进主要是在词频的基础上考虑位置[11]和语义信息[12]等因素。

推荐算法主要有协同过滤推荐和基于内容的推荐[13]，其核心思想就是提取出相似度较高的项目

* 通信作者. E-mail：yuliqun@comac. cc

引用格式：俞立群. 基于指标体系的相似故障问题推荐方法研究[J]. 民用飞机设计与研究，2024（2）：153-158. YU L Q. Research on recommendation method for similar fault problems based on indicator system[J]. Civil Aircraft Design and Research，2024（2）：153-158（in Chinese）.

集合,而衡量相似度的计算方式包括 Pearson 相关系数和余弦相似度[14]。

民用飞机试飞时,故障记录包括故障文本描述、美国航空运输协会(air transport association America,简称 ATA)章节号、运行阶段和解决措施等信息,其中故障文本描述中可能含有机组告警系统(crew alerting system,简称 CAS)、机载维护系统(onboard maintenance system,简称 OMS)中显示的告警信息,这些告警信息在很大程度上代表这条故障文本的含义,也是故障描述文本的关键词句。因此,在利用 TF-IDF 方法计算故障文本相似性的基础上,构建包含 CAS、OMS 信息、ATA 章节号和运行阶段等因素的指标体系,并建立正则表达式方法将 CAS、OMS 信息从故障文本中自动取出,提出一种基于指标体系的相似故障问题推荐方法。此外考虑到民用飞机设计领域存在许多专有词汇,因此在分词时引入民用飞机设计领域专有词汇库,以提高故障案例匹配度。

1　指标体系构建与相似度计算

1.1　指标选取

试飞阶段一份故障记录信息通常包含架机号、任务编号、任务名称、运行阶段、故障件名称、故障件件号、问题现象描述、ATA 章节号、纠正措施、问题关闭状态等字段,故障记录字段含义如表 1 所示。

表 1　故障记录字段含义

指　标	说　明
架机号	区分飞机代号
任务编号	执行的试验任务编号
任务名称	执行的试验任务名称
运行阶段	故障发生时飞机所处的阶段
故障件名称	发生故障的部件
故障件件号	发生故障的部件件号
ATA 章节号	故障所属专业
故障现象描述	描述故障信息的详细文本
纠正措施	消除故障采取的方法手段
故障状态	包括关闭和开口两种状态

选择的指标应当具备代表性、独立性和可及时获取性。代表性是考虑对相似故障判断是否有显著影响;独立性是指两个指标之间是否强关联;可及时获取性是考虑信息在故障发生时能否直观获取,例如故障件及件号在故障发生时大概率无法直接判断。最终选取的指标包括运行阶段、ATA 章节号和故障现象描述,选取这些指标的原因如表 2 所示。

表 2　相似故障判别指标

指　标	选取原因
运行阶段	相近运行阶段发生的故障模式相似且易于获取
ATA 章节号	相同 ATA 故障相似性高且易于获取
故障现象描述	文本描述相似的故障相似性高且易于获取

1.2　运行阶段相似度计算

参考中国民用航空局发布的《运行阶段和地面阶段》[15],飞机可能发生故障的运行阶段有行前准备、推出/牵引和起动、滑出、起飞、初始爬升、航路爬升、巡航、下降、等待、进近、着陆、滑入和航后,各阶段主要特征如表 3 所示。

表 3　运行阶段及主要特征

指　标	主要特征
行前准备	自机组登上飞机起,至飞机准备放行
推出/牵引和起动	自飞机不依靠自身动力受牵引力作用从廊桥、机坪或停机位等机场活动区移动起,至开始依靠自身动力移动前
滑出	自飞机依靠自身动力离开廊桥、机坪或停机位等机场活动区,至在起飞位置机组开始为实际起飞而使用动力前
起飞	自机组松刹车、为实际起飞而使用动力时起,经滑跑直至到达距跑道标高之上 35 ft
初始爬升	自离地 35 ft 高度起,至飞机距跑道标高之上 1 500 ft 或建立干净构型
航路爬升	自飞机距跑道标高之上 1 500 ft 或建立干净构型,至达到初始巡航高度
巡航	自飞机达到初始巡航高度后开始以固定速度、高度平飞,至开始向目的地下降前
下降	自最后巡航高度下降,至等待高度或离地 1 500 ft
等待	在巡航高度以下以固定高度(10 000 ~ 20 000 ft)、在指定空域做预先确定的机动飞行以等待进一步指令

表3(续)

指　标	主要特征
进近	自离地 1 500 ft 高或起始进近定位点,至开始着陆拉平
着陆	自开始着陆拉平,直至脱离着陆跑道或在跑道上停止
滑入	自离开跑道,直至到达机位、机坪等机场活动区并且停止依靠自身动力移动为止
航后	自机组开始关闭机载设备,至机组和乘组均离开飞机

一般来说相似故障发生的运行阶段也相似,差别较大的运行阶段,如地面和空中发生的故障,差别也较大,因此对各个运行阶段按照表 4 进行赋值。

表 4　各个运行阶段赋值

指　标	赋　值
行前准备	1
推出/牵引和起动	2
滑出	4
起飞	6
初始爬升	8
航路爬升	11
巡航	13
下降	12
等待	9
进近	10
着陆	7
滑入	5
航后	3

根据公式(1)计算运行阶段的相似度:

$$S_{\text{phase}}=\frac{1}{e^{0.2(|p_i-p_j|)}} \tag{1}$$

式中:S_{phase} 表示运行阶段相似度;p_i 表示阶段 i 赋值;p_j 表示阶段 j 赋值。

1.3　ATA 章节号相似度计算

ATA 章节号由系统、分系统和单元体三个层次六位数字组成。在试飞故障问题记录中,ATA 章节号一般只填写了系统层次的两位数字。相似故障问题的 ATA 章节号是相同的。依据公式(2)计算 ATA 章节号相似度,当 ATA 章节号一致时,ATA 章节号相似度取值为 1,当 ATA 章节号不一致时,ATA 章节号相似度取值为 0:

$$s_{\text{ATA}}=\begin{cases}1,\text{当 ATA 章节号一致时}\\0,\text{当 ATA 章节号不一致时}\end{cases} \tag{2}$$

式中:s_{ATA} 表示 ATA 章节号相似度,取值范围为 0~1。

1.4　故障文本相似度计算

故障文本是试飞故障问题记录中核心的字段,是对所发生故障的文字描述,相似故障的文本描述一般也是相似的,所以衡量故障文本的相似度是识别相似故障的重要组成部分。本文使用 TF-IDF 方法建立故障文本向量模型,采用余弦相似度衡量故障文本向量之间的相似度。

故障文本相似度计算步骤如图 1 所示,包括文本预处理、分词、去停用词、制作语料库、TF-IDF 建模和相似度计算。文本预处理包含去除前后空格和删除换行符等。在分词时考虑到民用飞机设计领域存在许多专有词汇,因此引入民用飞机设计领域专有词汇库。分词完成后根据常用停用词库去除停用词,例如“啊”、“的”等。将所有词合并成一个关键词集合形成语料库,在此语料库的基础上利用 TF-IDF 方法建模,最后用余弦相似度进行计算,得到故障文本相似度值。

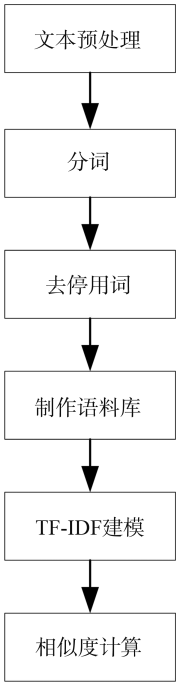

图 1　故障文本相似度计算步骤

TF 的计算方法如公式(3)所示：

$$F=\frac{m}{M} \tag{3}$$

式中：F 表示词频；m 表示关键词集合中的一个词在一份文档中出现的次数；M 表示一份文档的总词数。

根据文献[16]，IDF 的计算方法如公式(4)所示：

$$I=\log\left(\frac{D}{d+1}\right) \tag{4}$$

式中：I 表示逆文档频率；D 表示语料库的文档总数；d 表示包含该词的文档数。

TF-IDF 的计算方法如公式(5)所示：

$$T=FI \tag{5}$$

式中：T 表示 TF-IDF 的值。

根据 TF-IDF 的值，将每条故障文本描述转化为向量，再用余弦相似度进行计算，计算如公式(6)所示：

$$s_{\text{text}}=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{6}$$

式中：s_{text} 表示故障文本相似度；n 表示关键词集合的长度；i 表示词频向量的第 i 个值；A 表示故障 A 描述的词频向量；B 表示故障描述 B 的词频向量。

部分故障文本描述中会包含有 CAS、OMS 等信息，CAS 作为民用飞机机组告警系统，提供综合化机组告警功能。CAS 信息在故障文本中是一串英文字符，一般由四部分组成：具体的子系统或位置标识、系统名称、系统的功能或设备和状况的性质，例如“速度超过 330 kn 后出现 REFUEL DOOR NOT CLSD 的蓝色 CAS 信息”。OMS 作为机载维护系统，主要用于故障检测和状态监控等。OMS 信息在故障文本中也是一串英文字符，例如“ECU 供电后，OMS 报 49-91010 APU LOP Indicator Fault”。CAS 信息和 OMS 信息相同，代表所发生的故障是相似的，因此 CAS 信息和 OMS 信息作为故障文本的关键词句，当它们存在的时候，应基于 TF-IDF 方法提高 CAS 信息和 OMS 信息在计算故障文本时的权重。因此，利用正则表达式，设计一种在故障文本中自动提取 CAS 信息和 OMS 信息的脚本程序。伪代码如图 2 所示。

```
#正则表达式规则，若故障文本描述中存在-、/、0-9数字、a-z小写字母和A-Z大写字母，自动提取出
pattern = '[-/0-9a-zA-z ]+'
results = re.findall(pattern, input_text)

#根据CAS信息和OMS信息格式，判断提取出的信息是否存在空格且第一次出现，是的话作为CAS信息或OMS信息提取出来
for info in results:
    info = info.strip()
    if info.find(" ") != -1 and info not in cas_results:
        cas_results.append(info)
```

图 2　CAS、OMS 信息自动提取伪代码

根据 Jaccard 相似度计算思想[17]，依据公式(7)计算故障文本中 CAS 和 OMS 信息相似度：

$$s_{\text{info}}=\begin{cases}\sum_{j=0}^{j<c}\frac{1}{l_m}\sum_{j=0}^{j<c}\frac{1}{l_k}, \exists k_j\subset m\\ 0, \forall k_j\not\subset m\end{cases} \tag{7}$$

式中：s_{info} 表示 CAS 和 OMS 信息相似度；j 表示从 0 开始依次递增 1 的整数；c 表示数组 k 中存在于数组 m 的信息数量；l_m 表示数组 m 的长度；l_k 表示数组 k 的长度；k 表示存有 CAS 和 OMS 信息的数组；k_j 表示数组 k 中任意一个值；m 表示另一存有 CAS 和 OMS 信息的数组。

1.5　综合相似度计算

综合考虑运行阶段、ATA 章节号和故障文本等指标，最终故障相似度的计算方法如公式(8)所示：

$$s=\alpha s_{\text{phase}}+\beta s_{\text{ATA}}+\mu s_{\text{text}}+\lambda s_{\text{info}} \tag{8}$$

式中：s 表示故障综合相似度；α,β,μ,λ 表示影响因子。结合层次分析法和专家打分，最终确定当故障文本中含有 CAS 或 OMS 信息时，α 取值 0.1，β 取值 0.1，μ 取值 0.2，λ 取值 0.6。当故障文本中不包含 CAS 或 OMS 信息时，α 取值 0.2，β 取值 0.3，μ 取值 0.5，λ 取值 0。

2　相似故障问题推荐

2.1　整体流程

相似故障问题推荐的整体流程主要包括新发生故障文本输入、基于指标体系的综合相似度计算、按相似度大小排序、选取 TOP-N 历史故障案例形成推荐集。整体流程如图 3 所示。

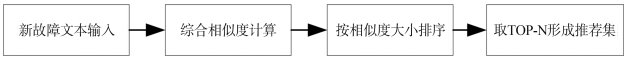

图 3　文本分类流程

2.2　TOP-N 推荐集选取

TOP-N 推荐集选取工作即选取按照综合相似度 s 从大到小排序后前 N 个故障。N 的实际大小可人为规定，同时考虑综合相似度大小，根据实际工程经验，若综合相似度小于 0.3 时，可认为两个故障之间不具备相似性。那么当最大综合相似度值小于 0.3

的时候,可认为在历史故障中不具备相似故障。

2.3　评价指标

故障问题推荐结果评价指标为准确率 A。准确率表示的是推荐集中的故障属于验证集的比例,一般来说随着推荐集数量 N 增大而变大,但是 N 太大会导致推荐集中有许多相似度较小的故障。召回率对推荐集中推荐了多少正确的相似故障的衡量,即验证集中有多少故障在推荐集中,一般来说随着推荐集数量 N 增大而变小,但是 N 太小会导致推荐的故障不够全面。

准确率 A 的计算公式[18]为:

$$A=\frac{U\cap V}{V} \tag{9}$$

式中:U 表示推荐集;V 表示验证集。

召回率 R 的计算公式[18]为:

$$R=\frac{U\cap V}{V} \tag{10}$$

3　实例验证

3.1　数据集构建

飞机试飞故障问题数据集来源于某型号飞机试飞过程中记录的故障问题。首先由于数据来源渠道众多存在重复问题,因此需要对获取的数据去重,再对数据进行清洗,包括去除前后空格和无意义不明字符等,最后得到 2 325 条历史故障数据。选取 3 条故障数据作为测试数据,并人工从历史数据中分别找出这 3 条数据的相似故障数据作为验证集。

3.2　N 值选取

N 值分别选取 1 到 50,对 3 条数据进行相似度大小计算,从大到小排序后取前 N 条数据,再计算准确率和召回率,最后取准确率和召回率的三次平均值。图 4 表示准确率和召回率随着 N 值变化的情况。

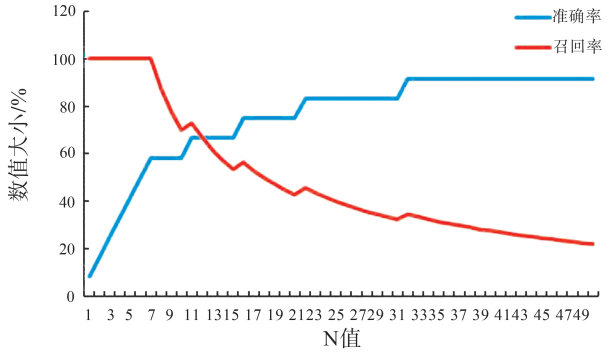

图 4　准确率和召回率随着 N 值变化情况

结合计算结果和图像可知,当 N 取 11 的时候,准确率和召回率的值综合来看是较优的。

3.3　算法对比

将本文提出的基于指标体系的相似故障问题推荐方法与去除 CAS、OMS 指标体系的推荐方法和只采用 TF-IDF 推荐方法进行比较,比较结果如表 5 所示。相较于其他两种方法,本文提出的算法在准确率上分别提高了约 25% 和 28%,在召回率上分别提高了约 27% 和 31%。

表 5　算法结果对比

算　法	准确率	召回率
指标体系推荐	66.67%	72.73%
去除 CAS、OMS 指标体系推荐	41.67%	45.46%
TF-IDF 推荐	38.43%	41.56%

4　结论

1）针对试飞故障问题文本数量大、内容多、人工查找花费时间长、无法高效利用大量的飞机试飞故障案例的信息价值的问题,提出了基于飞行阶段、ATA 章节号、故障文本和 CAS、OMS 信息的指标体系相似故障问题推荐方法,通过实例验证相对于只选取飞行阶段、ATA 章节号和故障文本作为指标的推荐算法以及只使用 TF-IDF 方法的算法,本文所提算法在准确率和召回率指标上都有较大提升,在准确率上分别提高了 25% 和 28%,在召回率上分别提高了约 27% 和 31%。

2）通过设置不同的 N 值,比较准确率和召回率,确定当 N 选取 11 的时候,两个指标的值均表现较好。

3）利用正则表达式,提出了一种自动从故障文本描述中提取 CAS 信息和 OMS 信息的方法。

参考文献:

[1] 刘科研, 董伟杰, 肖仕武,等. 基于电压数据 SVM 分类的有源配电网故障判别及定位[J]. 电网技术, 2021,45(6): 2369-2379.

[2] 王永坚, 陈丹, 戴乐阳. 信息融合与贝叶斯集成的船用中高速发动机磨损故障诊断[J]. 集美大学学报: 自然科学版, 2018, 23(3): 205-211.

[3] 徐涯昕, 何泽恩, 徐绪堪. 基于 CNN-BiLSTM 网络的数控机床故障文本自动分类[J]. 计算机与现代化, 2023, 332(4): 7-14.

[4] ZHOU F G, ZHANG F, YANG B R. Research on Chinese text summarization algorithm based on statistics and

rules[C]//Proceeding of the International Conference on Asian Language Processing. [S. l. :s. n.], 2009.
[5] SHAH K, PATEL H, SANGHVI D. A comparative analysis of logistic regression, random forest and KNN models for the text classification[J]. Augmented Human Research,2020,5:12.
[6] 杨祎, 崔其会, 秦佳峰,等. 改进 BERT 的故障案例智能匹配方法[J]. 山东电力技术, 2022, 49(2): 47-53.
[7] 唐瑞春, 张肖南, 郭双乐, 等. 一种基于粗糙集和欧式距离的手机故障案例匹配算法[J]. 中国海洋大学学报:自然科学版, 2015, 45(12): 125-130.
[8] 祖月芳, 凌海风, 吕永顺. 基于 NLP 技术的装备故障文本匹配算法研究[J]. 兵器装备工程学报,2021, 42(11):204-208.
[9] 王峻洲, 王华伟, 侯召国. 基于相似体系的民机结构超手册维修案例分析[J]. 系统工程与电子技术, 2022,44(9):2978-2985.
[10] DIERK S F. The SMART retrieval system:experiments in automatic document processing[J]. IEEE Transactions on Professional Communication, 1972, PC-15(1):17.
[11] 张瑾. 基于改进 TF-IDF 算法的情报关键词提取方法[J]. 情报杂志,2014,33(4):153-155.
[12] 黄承慧,印鉴,侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011,34(5):856-864.
[13] RESNICK P, LACOVOU N, SUCHAK M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]//Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work. [S. l. :s. n.], 1994:175-186.
[14] 薛鹏. 基于协同过滤理论的民机智能故障诊断方法[J]. 中国民航大学学报, 2014, 32(4): 23-26.
[15] 中国民用航空局. 运行阶段和地面阶段[EB/OL]. (2014-01-16)[2023-09-04]. https://max. book118. com/html/2019/0212/716012201300200a lgorithm 6. shtm.
[16] JOACHIMS T. A probabilistic analysis of the rocchio with TFIDF for text categorization[C]//Proceedings of the Fourteenth International Conference on Machine Learning. [S. l.]:Douglas H. Fisher,1997.
[17] JACCARD P. The distribution of the flora in the alpine zone[J]. New Phytologist, 1912, 11(2):37-50. DOI: 10. 1111/j. 1469-8137. 1912. tb05611. x.
[18] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.

作者简介

俞立群 男,硕士,工程师。主要研究方向:飞机数据管理与分析。E-mail:yuliqun@ comac. cc

Research on recommendation method for similar fault problems based on indicator system

YU Liqun *

(Shanghai Aircraft Design and Research Institute, Shanghai 201210, China)

Abstract: In response to the problem of a large number of test flight fault problems and the inability to efficiently utilize their information value, a recommendation method for similar fault problems based on the index system of flight stages, ATA, fault text, CAS and OMS information was proposed. Similarity calculation methods were proposed for each indicator based on their characteristics. In addition, a method is proposed to automatically extract CAS and OMS information from fault text descriptions using regular expressions. Finally, an example was used to validate the recommendation algorithm that only selects flight stage, ATA, and faults text as indicators, as well as the algorithm that only uses TF-IDF method. The algorithm proposed in this article has improved 25% and 28% in accuracy, 27% and 31% in recall. It can provide reference for the recommendation of similar flight test fault problems.

Keywords: fault recommendation; indicator system; CAS-OMS information; TF-IDF

* Corresponding author. E-mail: yuliqun@ comac. cc