基于故障模式的装备质量问题文本分类方法

费清春[1],史莹莹[1],曾庆国[2]
(1.南京电子技术研究所,江苏 南京 210039;2.工业和信息化部电子第五研究所,广东 广州 511300)

摘 要:面对大规模的海量装备质量问题文本,如何精准有效地将它们按照故障模式分类具有重要的理论意义。目前,主要以专家人工判定的传统方式开展问题分类费时费力,难以满足实际的应用需求。在此背景下,提出了一种基于故障模式的装备问题自动分类方法。该方法首先利用中文分词技术开展文本切词,生成文本关键词特征向量,进而计算质量问题与故障模式文本特征向量的相似度,最后按照相似度的阈值判定质量问题归属故障模式的种类。采用信息化技术进行装备质量问题分类方法简单易行,实验结果表明效果良好。

关键词:装备质量;质量问题;文本分类;故障模式;相似度
中图分类号:TP311.5 **文献标识码**:A **DOI**:10.19358/j.issn.2096-5133.2021.09.003
引用格式:费清春,史莹莹,曾庆国. 基于故障模式的装备质量问题文本分类方法[J].信息技术与网络安全,2021,40(9):14-18.

Text classification method for equipment quality problems based on failure mode

Fei Qingchun[1], Shi Yingying[1], Zeng Qingguo[2]
(1.Nanjing Research Institute of Electronics Technology, Nanjing 210039, China;
2.The Fifth Electronic Research Institute of Ministry of Industry and Information Technology, Guangzhou 511300, China)

Abstract: In the face of large-scale and massive equipment quality problem texts, how to accurately and effectively classify them according to failure modes has important theoretical significance. At present, the mainly method based on manual judgement is a time-consuming and laborious task, which is difficult to satisfy the real-world application requirements. Under the above background, this paper proposes an automatic classification approach based on failure modes. It firstly utilizes Chinese word segmentation technology to segment text, which is used to generate keyword feature vectors. Then, it calculates the similarity of the quality problem text vectors and failure mode text vectors, and finally determines the type of failure mode that the quality problem belongs to according to similarity threshold. The proposed approach is implemented by information technology that is simple in its implementation for equipment quality problem classification. Experimental results show that the proposed approach has received superior performance on classification for equipment quality problem texts.

Key words: equipment quality; quality problem; text classification; failure mode; similarity

0 引言

随着计算机技术的快速发展,企业建立了产品质量问题处理信息系统,存储了大量的产品质量问题处理历史记录。产品质量改进通常是建立在产品质量问题数据分析的基础上,将质量问题快速、准确地自动归类为不同的故障模式,对于促进企业识别质量问题关键因素,推动产品质量改进具有十分重要的现实意义。如何将成千上万,甚至是几十万条质量问题数据按照故障模式自动分类,单凭专家筛选、甄别和分类,是一个巨量的、难以短时间完成的任务,成为了亟需解决的实际问题。以关键词检索等自动化程度较低的人机协作模式开展质量问题分类,结果存在大量的误报和漏报,不能满足实际使用的需要。

运用大数据技术,分析挖掘产品质量问题数据,能够为产品质量改进的技术创新提供有效的技术

支持[1]。当前，计算机领域已形成了中文分词、文本挖掘等自然语言处理技术，在此背景下，本文重点聚焦装备质量问题文本数据的故障模式自动分类方法展开研究。

1 相关研究

在计算机文本挖掘方面，Kenter 等人[2]合并由相同算法、语料库、参数设置得到的不同维度词向量，训练出分类模型，并利用此分类模型计算短文本问题之间的相似度；Kusner 等人[3]基于词与词之间的最小移动距离，求解问题文本之间的文档相似度；孟繁宇[4]则将基于检索词的摘要提取问题转化为文本聚类问题，利用提取式摘要抽取方法，对文档的主要特征进行向量化抽取和去冗余等操作。

针对装备故障和失效等质量问题分类方法研究，张计晨[5]围绕天气雷达运行工作原理，分析雷达发射系统故障触发机理，形成发射系统故障分类模型。龚俊杰[6]提出航空产品质量问题的三维分类模型，从“过程–问题–性质”三个维度对质量问题的不同分析类别进行定义，再通过每一维度的层次分类，实现对问题的全面分类管理。李擎等人[7]提出基于层叠隐马尔可夫的设备质量风险隐患识别模型，在此基础上统计每类质量问题的出现频度，实现对基于风险等级的质量问题管理方案。谢荣琦[8]则将数据挖掘中的特征聚类算法引入质量特性识别过程中，并与过滤型特征算法相结合，构造面向复杂产品关键质量特征的问题识别模型。张青等人[9]提出基于主题扩展的领域问题分类方法，给出了评价分类的指标。Liu 等人[10]提出了一种基于朴素贝叶斯的分类算法，通过计算描述文本的统计学特征进行分类。洪晟等人[11–15]针对雷达电源系统健康分级分类、车载离子电池的健康状况评价等方面，开展特征数据训练，并引入长–短期记忆网络预测和判别健康状态，在互相依存网络中开展故障关联分类分析、级联失效分类分析等。

上述研究文献启发了笔者通过文本之间的相似度判断问题分类的思路，相对于从装备实时监测状态判定故障模式，本文从自然语言处理的角度，提出一种基于文本特征抽取和相似度计算的装备质量问题自动分类方法，为解决此类问题提供了一个新的路径。

2 装备质量问题文本分类基本定义

定义 1 装备质量问题文本表示为 6 元集合 P，如式(1)所示：

$$P=(p_1,p_2,p_3,p_4,p_5,p_6) \tag{1}$$

其中，p_i 表示质量问题的特定数据项。p_1 表示质量问题唯一编号；p_2 表示质量问题发生的部位；p_3 表示质量问题现象文本；p_4 表示质量问题原因文本；p_5 表示质量问题纠正文本；p_6 表示质量问题纠正措施文本。

定义 2 装备质量问题故障模式表示为 3 元集合 F，如式(2)所示：

$$F=(f_1,f_2,f_3) \tag{2}$$

式中，f_i 表示装备质量问题故障模式的特定数据项。其中，f_1 表示故障模式唯一编号；f_2 表示故障模式名称；f_3 表示故障模式文本描述。

定义 3 装备质量问题分类的结果表示为装备质量问题文本集 P 到装备故障模式集 F 的一个映射关系 $\zeta_{P\to F}$。假设 $\forall x_i\in P$ 均有且仅有一个 $y_i\in F$ 与之对应，即一个质量问题与一个故障模式存在唯一映射关系。

3 装备质量问题文本分类方法

3.1 质量问题文本分类框架与流程

本文提出了质量问题文本分类的框架，如图 1 所示。数据预处理对质量问题和故障模式文本进行中文分词等；数据特征提取对质量问题和故障模式文本提取有用的特征；相似度计算获得质量问题与故障模式的文本相似性；分类判定用以建立质量问题文本与故障模式文本的映射关系；指标评价完成评估质量问题分类方法的性能。

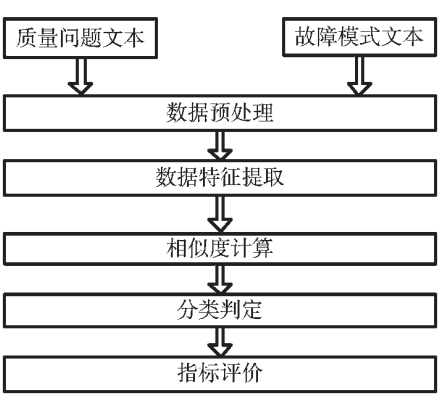

图 1 装备质量问题文本分类框架图

基于故障模式的装备质量问题文本自动分类方法包含 3 个核心部分：(1)文本特征向量构造：利用中文分词技术分别将质量问题和故障模式文本

切词，生成关键词特征向量；(2)质量问题特征向量相似度计算：进行质量问题文本与故障模式文本的特征向量之间的相似度计算；(3)质量问题故障模式判别：依据相似度阈值，自动判定质量问题归属的故障模式种类。装备质量问题文本分类方法的主要流程如图 2 所示。

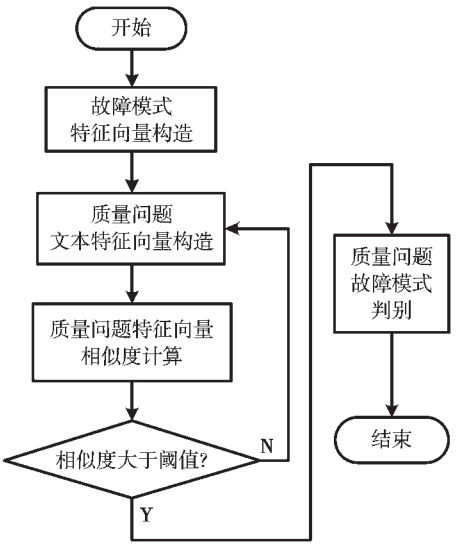

图 2 装备质量问题文本分类流程图

3.2 质量问题文本特征向量构造

在建立映射关系 $\zeta_{P\to F}$ 的过程中，需要同时考虑质量问题文本的多维度信息 p_i 和故障模式文本 F 中的多维度信息 f_i，最大程度地利用多元语义特征，具体步骤包括：

(1)提取装备质量问题文本的语义特征，构造质量问题文本特征向量，创建字符串 $s=p_1+p_2+p_3+p_4+p_5+p_6$，对 s 进行中文分词并构建单词集合 X。

(2)提取故障模式文本的语义特征，创建字符串 $f=f_1+f_2+f_3$，对 f 进行中文分词并构建单词集合 Y，X 和 Y 合并为词典 Z，词典 Z 中单词的总数为 n。

(3)建立质量问题文本的特征向量，记为 v，向量空间长度为 n；建立故障模式的特征向量，记为 w，向量空间长度为 n。

(4)对照文本在 Z 中查字典，按照独热编码方式，完成 v 和 w 特征向量赋值。

3.3 质量问题特征向量相似度计算

装备质量问题文本与质量问题故障模式文本的相似度记为 a，相似度计算的常用方法包括杰卡德相似系数(Jaccard Similarity Coefficient)、余弦相似度(Cosine Similarity)和皮尔逊相关系数(Pearson Correlation Coefficient)等。

(1)杰卡德相似系数通过测量两个有限样本集合之间的重叠，计算它们之间的相似性。给定一个装备质量问题文本分词集合 X，一个故障模式的分词集合 Y，则杰卡德相似系数表示为：

$$a=J(X,Y)=\frac{|X\cap Y|}{|X\cup Y|} \tag{3}$$

(2)余弦相似度通过计算质量问题文本的特征向量 v 和故障模式的特征向量 w 的夹角余弦值来评估它们的相似度。给定一个质量问题特征向量 v，一个故障模式的特征向量 w，则余弦相似度表示为：

$$a=\frac{\sum_{i=1}^{n}(v_i\times w_i)}{\sqrt{\sum_{i=1}^{n}(v_i)^2}\times\sqrt{\sum_{i=1}^{n}(w_i)^2}} \tag{4}$$

其中，v_i 和 w_i 分别表示为装备质量问题文本和故障模式的特征向量中第 i 维特征值，a 是它们之间的余弦相似度。

(3)通过计算装备质量问题文本的特征向量 v 和故障模式的特征向量 w，得到皮尔逊相关系数，表示为：

$$a=\frac{\sum_{i=1}^{n}(v_i-\overline{V})(w_i-\overline{W})}{\sqrt{\sum_{i=1}^{n}(v_i-\overline{V})^2}\sqrt{\sum_{i=1}^{n}(w_i-\overline{W})^2}} \tag{5}$$

其中，v_i 和 w_i 分别表示为装备质量问题文本和故障模式的特征向量中第 i 维特征值，$\overline{V}$ 和 $\overline{W}$ 分别表示为装备质量问题文本和故障模式的特征向量平均值，a 是它们之间的皮尔逊相关系数。

3.4 质量问题故障模式判别

在建立映射关系 $\zeta_{P\to F}$ 的过程中，相似度 a 的值域为 $[0,1]$，在此范围内设置 k 作为质量问题分类故障模式的阈值。一个装备质量问题与所有故障模式文本均进行了相似度计算，假设与第 i 个故障模式的相似度最高，记为 a_i：

(1)当 $a_i\geqslant k$ 时，则映射关系成立，即判定装备质量问题分类至第 i 个故障模式；

(2)当 $a_i<k$ 时，则映射关系不成立，即判定装备质量问题暂无映射的故障模式。

4 实验结果与分析

4.1 实验数据集

以某企业 313 项装备质量问题文本和 6 类故障

模式文本数据开展实验对比与分析。其中，装备质量问题文本包括编号、部位、现象、原因、纠正和纠正措施等维度的短文本，而6类故障模式包括编号、名称和内容等维度短文本。

例如，一个装备质量问题文本编号为Q0001，现象为“雷达扫描线不转动，目标无法显示”，部位为“数据处理分析”，原因为“数据处理死机”，纠正为“重新安装升级后的软件”，纠正措施为“修改代码完善非法数据验证，提高容错性”。与之对应的装备故障模式编号为F001，故障模式名称为“雷达无法探测目标”，故障内容描述为“数据处理软件死机”。故障模式类别及其对应的装备质量问题文本数如表1所示。

表1 装备质量问题样本分类分布

(个)

类1	类2	类3	类4	类5	类6
24	25	8	22	218	16

4.2 评价指标[9]

为了评价基于故障模式的装备质量问题分类方法的性能，采用准确率P、召回率R和$F1$指标(F1-score)作为实验评价指标。其中，准确率P反映了已分类结果的正确性，计算如式(6)所示。召回率R是已正确分类占所有应该正确分类的比例，计算如式(7)所示。$F1$同时兼顾了准确率P和召回率R两个方面的评价指标，它是准确率和召回率的调和平均数，计算如式(8)所示。

$$P=\frac{\text{所有已正确分类的质量问题数}}{\text{测试集中已分类质量问题总数}} \tag{6}$$

$$R=\frac{\text{所有已正确分类的质量问题数}}{\text{所有应正确分类的质量问题总数}} \tag{7}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{8}$$

4.3 实验设计

为了有效验证本文提出的装备质量问题文本分类方法的有效性，设计了3个实验开展分类有效性的比对研究。

实验1：在相同的相似度阈值k下，按照杰卡德相似系数、余弦相似度和皮尔逊相关系数3种相似度计算方式，开展装备质量问题文本自动分类实验，选出性能最优的相似度算法，并开展相关结果分析。

实验2：按照实验1优选的相似度算法，开展装备质量问题文本分类实验，针对在不同的相似度阈值k下的各项指标，选出性能最优的相似度阈值k。

实验3：按照实验1优选的相似度算法，实验2优选的相似度阈值k，开展装备质量问题文本分类实验，依据在6个类别上的评价指标，分析目前存在的差距和改进方向。

4.4 实验结果

在实验1中，针对313项装备质量问题文本，按照杰卡德相似系数、余弦相似度和皮尔逊相关系数3种不同方式计算相似度a，统一设置相似度阈值$k=0.01$，实验1的性能指标结果如表2所示。

表2 实验1的性能测试指标结果

(%)

算法	准确率	召回率	$F1$
杰卡德	78.05	71.57	74.67
余弦	65.05	57.19	61.30
皮尔逊	75.00	0.96	1.89

在实验2中，按照杰卡德相似系数计算相似度a，设置相似度阈值k分别为0.01、0.1、0.2和0.3，实验2的性能指标结果如表3所示。

表3 实验2的性能测试指标结果

(%)

阈值k	准确率	召回率	$F1$
0.01	78.05	71.57	74.67
0.1	77.61	33.23	46.53
0.2	87.50	4.47	8.51
0.3	100.0	0.64	1.27

在实验3中，采用杰卡德系数计算相似度a，设置相似度阈$k=0.01$，在6种故障模式类别下，开展实验比对，实验3的性能指标如表4所示。

表4 实验3的性能测试指标结果

(%)

	准确率	召回率	$F1$
类1	45.45	60.00	51.72
类2	97.25	80.25	87.94
类3	45.65	87.50	60.00
类4	28.57	25.00	26.67
类5	33.33	12.50	18.18
类6	72.73	36.36	48.48

4.5 结果分析

实验1结果表明，采用杰卡德相似系数在准确率、召回率和$F1$值3项评价指标上均优于余弦相似度和皮尔逊相关系数。相似度计算方式优选杰卡

德系数。

实验2结果表明，采用杰卡德相似系数，随着阈值k逐步增加，准确率随之上升，而召回率则随之下降，准确率的提升会带来装备质量问题文本分类中漏报的风险，因此在[0.01，0.4]范围内，相似度阈值k最优为0.01。

实验3结果表明，采用杰卡德相似系数计算相似度，设置相似度阈值k=0.01时，在所有测试样本集上进行装备质量问题文本分类，整体上取得了较好的总体性能，然而，在6个故障模式类别之间性能差距较大，例如在故障模式类别2和故障模式类别4上的分类准确率和召回率具有显著差异性。因此，需要深度挖掘不同类别的质量问题文本特征，改进故障模式判别方式，均衡不同类别的分类差异，进一步优化分类效果。

5 结论

本文针对当前装备质量问题文本的分类方法自动化程度较低，提出了一种基于文本特征提取和相似度计算的分类方法，实现装备质量问题文本与故障模式的自动和有效分类，减少了对专业人员的依赖，极大地降低了分类中的人工工作量，推动了产品质量改进的效率。

在未来工作中，针对装备质量问题文本分类性能尚存在的差距，将采用深度学习模型挖掘质量数据的隐藏语义特征，进一步提升装备质量问题文本特征提取效果，并拓展故障模式库的广度和深度，优化装备质量问题文本分类的各项性能。

参考文献

[1] 刘烨，胡昌平，张国政.基于工业大数据的产品质量改进新模式的探索和研究[J].计算机应用与软件，2019，36(12)：329-333.

[2] KENTER T，RIJKE M D.Short text similarity with word embedding[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management，2015：1411-1420.

[3] KUSNER M J，SUN Y.From word embeddings to document distances[C]//Proceedings of the 32nd International Conference on Machine Learning，2015：957-966.

[4] 孟繁宇.大数据环境下文本聚类与摘要提取[D].北京：北京邮电大学，2015.

[5] 张计晨.新一代天气雷达(CINRAD/CC)发射系统故障触发的分类研究[J].气象水文海洋仪器，2019，36(1)：95-101.

[6] 龚俊杰.基于3P模型的航空产品质量问题分类研究及应用[J].航空标准化与质量，2019(4)：25-28.

[7] 李擎，张秋艳，白磊.一种基于文本挖掘的铁路基础设施设备风险隐患识别模型[J].铁路计算机应用，2018，27(2)：1-4.

[8] 谢荣琦.复杂产品关键质量特性识别问题的数据挖掘模型研究[D].天津：天津大学，2014.

[9] 张青，吕钊.基于主题扩展的领域问题分类方法[J].计算机工程，2016，42(9)：202-207.

[10] LIU J X，TIAN Z L，LIU P B，et al.An approach of semantic web service classification based on Naive Bayes[C]//Proceedings of the IEEE International Conference on Services Computing(SCC)，2015：356-362.

[11] 洪晟，罗无为，周闯，等.雷达电源系统安全运行健康状态评估研究[J].航空工程进展，2020，11(4)：585-590.

[12] HONG S，ZENG Y N.A health assessment framework of lithium-ion batteries for cyber defense[J].Applied Soft Computing，2021，101：107067.

[13] HONG S，YUE T Y，LIU H.Vehicle energy system active defense：a health assessment of lithium-ion batteries[J/OL].International Journal of Intelligent Systems，2020：1-19[2021-05-01].https://doi.org/10.1002/int.22309.

[14] HONG S，WANG B Q，MA X M，et al.Failure cascade in interdependent network with traffic loads[J].Journal of Physics A：Mathematical and Theoretical，2015，48(48)：485101.

[15] HONG S，LV C，ZHAO T D，et al.Cascading failure analysis and restoration strategy in an interdependent network[J].Journal of Physics A：Mathematical and Theoretical，2016，49(19)：195101.

(收稿日期：2021-05-21)

作者简介：

费清春(1982-)，男，硕士，高级工程师，主要研究方向：产品质量管理、可靠性工程技术等。

史莹莹(1984-)，女，硕士，高级工程师，主要研究方向：雷达软件设计等。

曾庆国(1987-)，通信作者，男，硕士，工程师，主要研究方向：产品可靠性工程技术等。E-mail：qgzeng@126.com。

版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部

中国电子信息产业集团有限公司第六研究所