消费品多源缺陷线索信息预处理研究

徐思红 张力丹 田晶晶 齐 月 孙 宁

（中国标准化研究院）

摘 要：消费品召回已成为产品质量安全后市场监管最重要的措施之一。消费品缺陷线索是发现潜在缺陷的数据源，准确、及时、有效地收集与分析消费品缺陷线索是快速掌握消费品缺陷并实施召回的基础。如何基于消费品缺陷线索快速定位产品潜在缺陷是召回管理的关键，消费品具有产品类型多、故障模式杂的特点，获取缺陷线索到缺陷线索可深入分析而后初步判定潜在缺陷之间存在一定的差距，本文主要从消费品缺陷线索采集监测、标签字典、数据筛选、数据清洗等环节提出数据预处理的要求，为消费品缺陷线索分析提供参考。

关键词：消费品，缺陷线索，数据预处理

DOI编码：10.3969/j.issn.1674-5698.2023.01.018

Research on the Preprocessing of Clue Information of Consumer Products with Multi Source Defects

XU Si-hong ZHANG Li-dan TIAN Jing-jing QI Yue SUN Ning

（China National Institute of Standardization）

Abstract: Consumer product recall has become one of the most important measures for market supervision after product quality and safety. Consumer product defect clues are the data source for discovering potential defects. Accurate, timely and effective collection and analysis of consumer product defect clues is the basis for quickly grasping consumer product defects and implementing recall. How to quickly locate the potential defects of products based on the defect clues of consumer goods is the key to recall management. Consumer goods are characterized by multiple product types and various failure modes. There is a certain gap between obtaining the defect clues and preliminarily finding the potential defects by analyzing defect clues. This paper mainly puts forward the requirements for data preprocessing from the collection and monitoring of defect clues of consumer goods, label dictionaries, data screening, data cleaning and other links. It provides reference for the clue analysis of consumer goods defects.

Keywords: consumer goods, defect clue information, data preprocessing

---

基金项目：本文受中央基本业务费项目“基于多源信息融合的消费品缺陷线索智能分析关键技术研究与应用”（项目编号：282022Y-9461）资助。

作者简介：徐思红，工程师，主要研究方向为缺陷信息、舆情信息、缺陷产品召回信息等综合数据分析。

张力丹，助理工程师，主要研究方向为缺陷信息采集和处理。

田晶晶，高级工程师，主要研究方向为缺陷产品召回信息管理、数据挖掘应用等。

齐月，助理工程师，主要研究方向为缺陷信息采集和处理。

孙宁，高级工程师，主要研究方向为缺陷产品管理召回、数据挖掘应用、新闻宣传等。

1　引 言

随着社会经济和科学技术的快速发展，消费品的种类以及功能越来越丰富，但是也带来了一系列安全隐患。近年来，消费品的安全性问题引发社会公众的普遍关注。如何通过产品质量安全监管，减少产品安全伤害、保护消费者人身和财产安全，是市场监管的工作方向。缺陷产品召回是产品质量安全监管的国际通行做法，是后市场监管的重要手段[1]。我国消费品召回工作从2004年开始，随着2015年发布的《缺陷消费品召回管理办法》、2020年发布的《消费品召回管理暂行规定》的相继实施以及相关配套文件的出台，我国消费品召回管理工作的法律依据日趋完善。

根据《2021年全国消协组织受理投诉情况分析》，2021年全国消协组织共受理消费者投诉约104.5万件，相较2020年增长6.37%，其中消费者关心的质量问题与使用安全问题占22.9%。消费者对于消费品质量安全的要求越来越高，保护自身权益的意识越来越强。根据《市场监管总局关于2021年全国汽车和消费品召回情况的通告》，2021年受市场监管部门调查影响的消费品召回占全年召回总量的90.5%，而消费者投诉以及其他形式的缺陷线索是引发缺陷调查导致召回最重要的信息源，随着召回制度的逐步完善，我国消费品召回监管已初步形成全国联动工作格局，通过数据交换共享与业务协同，为消费品缺陷调查和召回工作提供了有效支撑。通过多种方式增强消费品多源缺陷线索信息的采集力度，信息量呈现爆发性的增长，与此同时，由于消费品具有种类多、故障模式复杂等特性，导致多源缺陷线索中产品信息不统一、故障描述不准确、缺陷线索信息重复等系列问题。为快速从消费品多源缺陷线索信息中提取有价值、有效的线索信息，采取高效技术措施做好数据预处理工作显得尤为重要。在缺陷线索数据挖掘与发现有潜在缺陷的过程中，消费品多源缺陷线索信息的数据预处理是核心环节之一。在数据预处理过程中，主要解决的数据问题包括：（1）重复性；（2）不完整性；（3）噪音；（4）不一致性；（5）不精简性。

2　消费品缺陷线索采集内容

消费品缺陷线索根据来源不同主要包括：消费者投诉、产品安全网络舆情、境外召回信息、电商平台评价信息、国内召回信息、其他信息等。根据消费品缺陷线索的用途，提出了不同类型缺陷线索采集内容（如图1所示）。

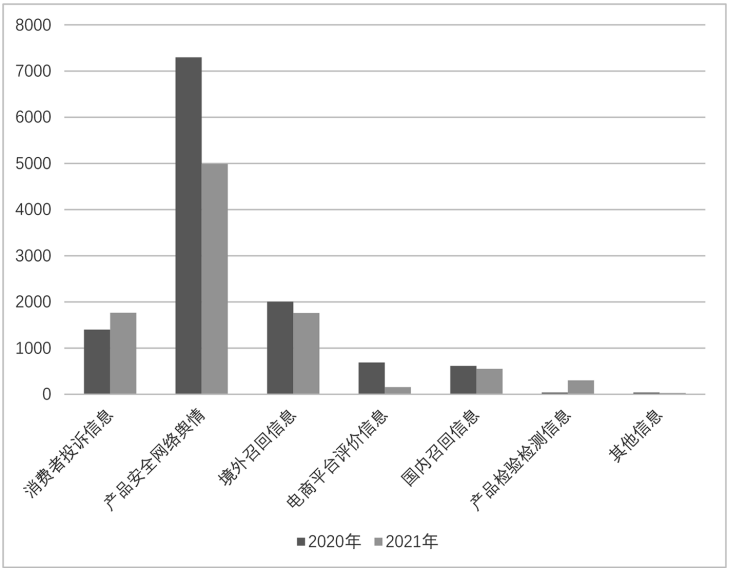

图1　2020年和2021年消费品多源缺陷线索信息数量

（1）消费者投诉信息：产品类别、生产者名称、产品名称、产品品牌、产品型号、产品产地、购买日期、产品应用场景、故障描述、是否造成伤害、联系人及联系方式等。

（2）产品安全网络舆情：产品类别、标题、描述、链接地址、相似新闻条数、发布时间等。

（3）境外召回信息：通报日期、通报国家、产品名称、产品类别、缺陷原因（危险描述）、措施、链接地址等。

（4）电商平台评价信息：评价时间、电商平台、评价内容、链接地址等。

（5）国内召回信息：产品类别、产品名称、品牌、缺陷描述、召回措施、受理单位等。

（6）产品检验检测信息：产品类别、品牌、产品名称、主要不合格项目、检测单位、检测批次等。

（7）其他信息：产品类别、品牌、产品名称、问题描述等。

3 消费品故障标签字典构建

产品故障现象作为消费品综合分析判定产品安全风险缺陷线索案例的基础信息，决定着是否存在安全性问题。多源信息中对于产品故障的描述文字量长短不一、表述随意多样化、故障问题多，针对这些复杂的内容，只有通过数据归纳，减少数据分析的信息量才能在分析挖掘的过程中提高效率。以已有的信息为基础，首先对产品故障现象进行归类，整理出每一类故障描述问题涉及的关键词，然后根据实际需求，分为两级，而后再根据描述提炼提取出同义词，进行同义词扩展，基于多个特征维度对近义词表进行过滤，形成同义描述集合，丰富故障描述特征，形成产品故障现象标签字典[2]（见表1）。

4 消费品缺陷线索预处理

由于消费品缺陷线索信息的数据种类和数据结构模式多元化，关联性较为复杂，在数据分析和信息挖掘环节中存在较大难度。在消费品缺陷线索信息收集和选择的初期环节，通过对数据的重复、缺失、噪音等问题进行预处理，然后将数据中与分析发掘相关性较高的数据通过数据清洗的方法再进行预处理，以获得可靠性较高的有效数据。相关实践证明，数据预处理在数据分析和挖掘过程中所占时间达70%以上，数据预处理的好坏对整个数据分析和挖掘结果有着至关重要的影响[3]。在消费品缺陷线索信息分析前的数据预处理主要包含以下几方面。

4.1 缺陷线索数据筛选

面对消费品多源缺陷线索信息，无论是通过系统被动采集消费者投诉信息、人工记录信函或举报，还是主动通过网络、电商平台主动采集产品安全网络舆情信息、境外召回信息、电商评价信息、检验检测信息，都要确保信息数据的有效性和唯一性，如果在信息数据的初始收集过程中就确保数据的有效性和唯一性，那么相比于先采集后筛选数据更为便捷和高效、准确。在数据选择的初始阶段确保有效性和唯一性，包括以下几项。

（1）有效性。产品信息的品牌、产品分类、类别信息完整，故障描述信息真实描述产品使用中出现的故障，排除消费纠纷、服务质量以及怀疑揣测等问题。如果有联系人信息，联系人手机号码有效。

（2）唯一性。同一来源的信息避免重复。

（3）字体和词性转化。多源信息同一数据字段的信息字体和词性保持一致。

对于不符合上述有效性要求的信息数据不进行采集或是不进行选择和选取。消费者投诉信息通过采集信息页面的必填项、手机号码验证的设置，确保信息有效性，重复性需要系统后台管理人员通过产品、手机号码和故障描述判定；产品安全网络舆情信息根据采集内容数据项采集信息，确保信息有效性，网络舆情信息本身具有随意性和开放性特性，所以真实性待定，故这类信息也是综合判定安全风险缺陷线索案例的辅助信息，应用网络信息爬取工具排除重复信息；境外召回信息主要是监测翻译国外消费品召回主管机构网络发布的召回信息，信息来源本身具有有效性和唯一性；电商评价信息根据采集内容数据项采集信息，确保信息有效性，评价信息是消费者购买和应用产品后真实反馈的内容，具有信息真实性的属性，应用爬取工具排除同一电商平台的重复信息；国内召

表1 消费品故障现象标签字典的构建实例

产品类别	产品名称	产品故障描述	关键词	关键词汇总及扩展	产品故障现象标签二级	产品故障现象标签一级
电子电器	手机	购买手机断网	断网	信号 无服务 上网 联网 网络连接超时 断网 断流 掉线 断线	网络/信号中断	网络/信号故障
电子电器	无人机	无人机失去信号未能自行飞回	失去信号		无网络/信号	网络/信号故障
电子电器	笔记本电脑	电脑开机无显以及多次出现WLAN出现问题的情况	无wland		网络/信号故障	网络/信号故障
电子电器	手机	手机无信号	无信号		无网络/信号	网络/信号故障
电子电器	电视机	电视机经常断网	断网		网络/信号弱	网络/信号故障
电子电器	智能猫眼	网络时常断线	短线		网络/信号故障	网络/信号故障
电子电器	手机	网络不稳定，断流	不稳定，断流		网络/信号中断	网络/信号故障
电子电器	手机	使用时信号不好	信号不好		网络/信号弱	网络/信号故障

回信息和检验检测信息是国家政府机关发布的信息，信息来源本身具有有效性和唯一性；其他来源信息根据采集内容数据项采集信息，确保信息有效性和唯一性，通过与已有数据的产品信息、手机号码和故障描述来判定是否重复，重复数据在已有数据的基础上进行特殊标注。

4.2 数据清洗

美国社会保险号错误纠正是数据清洗技术的最早起源，随着信息业和商业的高速发展，数据清洗也进一步发展，并根据各行各业的不同需求，有着不同的数据清洗方法，消费品缺陷线索信息的数据预处理，根据现有需求以及经验的积累，其中的数据清洗主要包括忽略部分数据项、基础数据核实、故障标签标注、智能与人工结合，各自解决不同的问题以达到缺陷线索信息的预处理数据优化效果。

4.2.1 忽略部分数据项

消费品多源缺陷线索信息各自具有其特殊属性和信息内容，而这些信息内容在综合判定安全风险的缺陷线索案例过程中不是分析的内容，影响分析判定结果的准确性，在信息预处理过程中，不影响消费品多源缺陷线索信息各自数据的基础上，采取忽略元组的方式将这些信息数据进行暂时忽略，忽略多源信息内容的数据项实例见表2。

表2 消费品多源缺陷线索信息数据预处理可忽略数据项实例

来源	可忽略数据项
消费者投诉信息	产品产地、生产者联系电话/邮箱、姓名、手机、所在省份、所在地市等
产品安全网络舆情	链接地址、时间等
境外召回信息	通报日期、链接地址等
电商平台评价信息	时间、平台等
国内召回信息	时间、受理单位等
检验检测信息	检测单位、检测批次等
其他信息	姓名、手机、所在省份、所在地市等

4.2.2 基础数据核实

产品品牌、产品分类、产品类别作为关联消费品多源缺陷线索信息的产品基础数据信息，统一性、标准化对于后续信息数据的分析挖掘尤为重要，消费品品牌繁多、种类复杂、产品多样、，明确这些信息才能确定是哪个产品。消费品品牌信息的研究发现，目前没有相关标准可借鉴，通过对已有信息数据的分析概括、同时借鉴电商平台中经销商对于产品的描述、网络舆情中消费者对于产品的描述，最终再通过平衡学习总结的方法，形成品牌字典。产品分类和产品类别字典可直接应用标准GB/T 36431–2018《消费品分类与代码》，同时借鉴电商平台中经销商对于产品的描述，进行数据的统一和规范（见表3）。

表3 品牌、产品类别、产品分类应用字典统一标注实例

样例	品牌	产品类别	产品分类
原始数据实例1	苹果	电器	电话
原始数据实例2	apple	家用电器	手机
原始数据实例3	Apple iPhone 11	手机	手机
规范数据实例1	苹果	电子电器	手机
规范数据实例2	苹果	电子电器	手机
规范数据实例3	苹果	电子电器	手机

4.2.3 故障标签标注

根据已形成的产品故障现象字典，对采集和选择的消费品多源缺陷线索信息：消费品的消费者投诉信息、产品安全网络舆情、境外召回信息、电商评价信息、国内召回信息、检验检测信息、其他的信息中的故障描述、描述、缺陷原因（危险描述）、评价内容、缺陷描述、主要不合格项目、问题描述的内容分别进行故障标签标注，最终将不规范的故障现象描述数据进行规范（见表4）。

表4 产品故障现象描述应用故障现象字典统一标注实例

	消费者对于产品的故障现象描述	规范故障现象标签标注
1	使用充电器对耳机进行充电，大约半小时耳机盒突然冒烟	冒烟
2	小夜灯，同年9月再拿出来使用，发现电池鼓包	电池鼓包
3	耳机导致外耳道炎。在购买入耳式耳机并使用之后，我的女儿总会感觉到耳痛，并且偶尔会有脓液	材质引起过敏

在产品安全网络舆情信息和电商评价信息的爬取和选择时，将爬取信息的关键词汇设置为需要的品牌、产品分类和产品类别、产品故障现象字典的组合或是产品分类和产品类别、产品故障现象字典的组合，最大限度和精准地采集与消费品安全相关的信息线索。

4.2.4 人工智能修正

在信息数据采集和选择过程中，不可避免地会产生不规范、错误、重复等问题，采用计算机和人工判断结合的方式制定方案，完善或剔除问题信息，最终保留有效信息。消费品的多样性和故障现象的复杂性，在构建品牌字典数据和故障现象标签字典数据

时，必定会存在字典数据不完整的问题，结合消费品各类产品的相关标准以及行业经验，通过计算机和人工专业知识不断完善字典数据，更好地为精准产品、简化故障描述奠定基础。

5 多源缺陷线索信息数据预处理实例

某A品牌耳机过敏的多源缺陷线索信息的数据预处理实例见表5~表7。

6 结 语

随着消费品多源缺陷线索信息数量的不断增加，相信数据预处理一定会越来越重要，为数据分析挖掘提供更加干净、高质量的信息源。消费品多源缺陷线索信息预处理方式完善建议：与专业知识应用融合，且贯穿预处理各环节；严控预处理各环节质量，保证高效[4]；应用计算机智能学习和语义识别技术，解放人工，提高效率和准确性。

表5 多源缺陷线索信息–消费者投诉信息

原始信息							预处理信息							
品牌	产品类别	产品分类	故障描述	是否造成伤害	……	手机号	是否重复	品牌	产品类别	产品分类	故障现象标签（二级）	故障现象标签（一级）	是否造成伤害	忽略其他数据项
A品牌	电子电器	耳机	带着耳机耳朵流脓	是	……	156*******5		A	电子电器	耳机	材质引起过敏	过敏	是	
a品牌	家用电器	无线耳机	这款耳机佩戴后导致耳朵发炎	是	……	187*******0		A	电子电器	耳机	材质引起过敏	过敏	是	
A	电器	A品牌耳机	佩戴引起耳道严重发炎	是	……	187*******0	是							

表6 多源缺陷线索信息–产品安全网络舆情信息

原始信息				预处理信息					
产品类别	标题	描述	……	品牌	产品类别	产品分类	故障现象标签（二级）	故障现象标签（一级）	忽略其他数据项
电子电器	A品牌耳机疑似引发耳部炎症，多人中招！企业回复：正在确认由于耳机结构引发耳内不适的可能性	近日，不少消费者表示，在佩戴A品牌新款无线耳机后，耳朵出现流脓、结痂等症状。具体是什么情况，企业方面又有怎样的回应?		A	电子电器	耳机	材质引起过敏	过敏	

表7 多源缺陷线索信息–电商平台评价信息

原始信息			预处理信息					
产品名称	评价内容	……	品牌	产品类别	产品分类	故障现象标签（二级）	故障现象标签（一级）	忽略其他数据项
A品牌耳机	不知道是不是入耳胶套材质的原因，戴着耳朵是有些痒	……	A	电子电器	耳机	材质引起过敏	过敏	
a品牌无线耳机	戴一会儿耳朵就发痒流水	……	A	电子电器	耳机	材质引起过敏	过敏	

参考文献

[1] 林建军. 浅谈我国消费品召回的特点及其重要意义[J]. 质量与市场, 2020, (20) :46–48.

[2] 姜肇财, 宋黎, 王雯.基于电商评论信息的产品故障标签体系构建研究[J]. 标准科学, 2021, (12) :128–131.

[3] 胡远樟,程小恩,何黎, 等. 一种基于糖尿病的中医数据挖掘预处理方法[J]. CJCM 中医临床研究, 2021, (30) :75–77.

[4] 田桂丰, 谌颃, 尹帮治. 信息熵和灰色关联分析在企业大数据分析中的应用[J]. 信息记录材料, 2021, 22(3):151–152.

[5] 唐成龙,谌颃,唐海春,等. 大数据背景下数据预处理方法研究运用[J]. 信息记录材料, 2021, 22(9):199–200.

[6] 郑杰昌, 谢志利, 王长林. 消费品召回追溯体系研究[J]. 标准科学, 2020, (5):32–52.

[7] 许辉.数据挖掘中的数据预处理[J]. 电脑知识与技术, 2022, (2):27–31.

[8] 李颜平,吴刚. 基于典型数据集的数据预处理方法对比分析[J]. 沈阳工业大学学报, 2022, 44(2):165–192.

[9] 杨忠诚. 数据挖掘工具WEKA及其应用研究[J]. 企业科技与发展, 2018, (9):38–39.

[10] 张治斌,刘威. 浅析数据挖掘中的数据预处理技术[J]. 数字技术与应用, 2017(10):216–217.

（上接第99页）

参考文献

[1] 日光温室发展的适宜地区及优型结构参数[J]. 农业工程技术(温室园艺), 2014(09):18–19.

[2] 崔斌斌,赵玉玲,翟军委,等. 哈茨木霉菌不同施用方式对番茄促生和防病的效果[J]. 园艺与种苗, 2022,42(09):3–4+9.

[3] 李元广,张道敬,罗远婵,等. 药肥兼能的防治土传病害新型高效系列微生物杀菌剂的创制与产业化[C]//.2015年中国化工学会年会论文集.[出版者不详],2015:1951–1952.

[4] 吴琼,王东凯,王雷,等. 荧光假单胞菌对四种植物病原菌的拮抗作用[J]. 黑龙江科学, 2014,5(05):22–23.

[5] 李潇潇,师桂英,张立彭,等. 荧光假单胞菌(Pseudomonas fluorescens)在植物病害生物防治中的研究及展望[J]. 草原与草坪, 2021,41(05):148–156.

[6] 张永利. 农产品质量安全法规对农户生产行为的影响[J]. 世界热带农业信息, 2022(12):45–46.

[7] 卢裕亿. 预冷工艺参数对蔬菜预冷速率及其冷链品质的影响[D]. 上海：上海海洋大学, 2021.

[8] 陈东杰,于怀智,邓秀丽,等. 不同贮藏湿度对番茄贮藏品质的影响[J]. 中国果菜, 2022,42(04):10–15.

[9] 陈瑶. 设施蔬菜高温闷棚与秸秆还田技术[J]. 四川农业科技, 2019(06):18–19.