(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116910175 B
(45)授权公告日 2023.12.01

(21)申请号 202311174976.6

(22)申请日 2023.09.13

(65)同一申请的已公布的文献号
申请公布号 CN 116910175 A

(43)申请公布日 2023.10.20

(73)专利权人 人工智能与数字经济广东省实验室(广州)
地址 510330 广东省广州市海珠区新港东路2429号首层自编051房
专利权人 华南理工大学

(72)发明人 钟昊阳 陆璐 汪烜烨 万童 邹全义 冼允廷

(74)专利代理机构 广州科粤专利商标代理有限公司 44001
专利代理师 邓潮彬

(51)Int.Cl.
G06F 16/31(2019.01)
G06F 40/194(2020.01)
G06F 40/289(2020.01)
G06F 16/34(2019.01)
G06F 16/35(2019.01)

(56)对比文件
KR 102123522 B1,2020.06.16
CN 114880584 A,2022.08.09
CN 115345158 A,2022.11.15
CN 115687925 A,2023.02.03
CN 116541510 A,2023.08.04
Tianyu Gao等.SimCSE:Simple Contrastive Learning of Sentence Embeddings.《arXiv:2104.08821v4》.2022,第1-17页.

审查员 杜锦锦

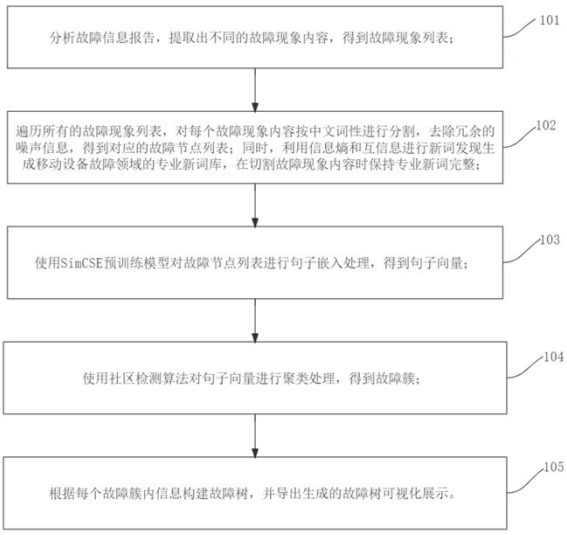
权利要求书2页 说明书6页 附图2页

(54)发明名称

自动化移动设备故障层级树构建方法、装置及储存介质

(57)摘要

针对故障报告无任何标注,同时因为涉及到的故障种类繁多,难以找到合适的规则进行处理的问题,本发明提供公开了自动化移动设备故障层级树构建方法、装置及储存介质,该方法首先汇总多渠道反馈的故障报告,提取报告中不同的故障现象;其次利用新词发现和中文词性对故障现象进行切割操作得到故障节点列表,使用SimCSE预训练模型对故障节点列表完成句子嵌入处理;最后使用社区检测算法并对故障句子完成聚类操作并构建对应的层级故障树。本发明基于中文词性和新词发现算法切割故障现象可以得到较清晰的故障层级关系,通过社区检测算法聚集故障簇并将每个故障簇内信息可视化至Excel表格方便后续测试人员分析故障信息。



CN 116910175 B

1.一种自动化移动设备故障层级树构建方法,其特征在于,所述方法包括如下步骤:

步骤101、分析故障信息报告,提取出不同的故障现象内容,得到故障现象列表;

步骤102、遍历所有的故障现象列表,对每个故障现象内容按中文词性进行分割,去除冗余的噪声信息,得到对应的故障节点列表;同时,利用信息熵和互信息进行新词发现生成移动设备故障领域的专业新词库,在切割故障现象内容时保持专业新词完整;

步骤103、使用SimCSE预训练模型对故障节点列表进行句子嵌入处理,得到句子向量;

步骤104、使用社区检测算法对句子向量进行聚类处理,得到故障簇;

步骤105、根据每个故障簇内信息构建故障树,并导出生成的故障树可视化展示;

所述分析故障信息报告,提取出不同的故障现象内容包括:

根据从用户、开发人员、供应商反馈得到的故障信息汇总故障报告;

利用正则表达式提取故障报告中和故障现象有关的信息,初步过滤部分无关的故障噪声数据;

所述使用SimCSE预训练模型对故障节点列表进行句子嵌入处理,得到句子向量,包括:

对于给定一个故障节点列表$\{X_i\}^m$令$X^+_i=X_i$,SimCSE模型使用独立的dropout作为掩码来获得增强的正样本对;

样本的嵌入向量生成表示为$h^z_i=f_\theta(x_i,z)$,其中z是随机的dropout掩码;

SimCSE模型通过将相同的样本输入编码器,并应用不同的dropout掩码z、z',获得相同样本的不同增强样本;

样本与不同增强样本最终的对比损失函数如下所示,

$$L_i=-log\frac{e^{sim(h_i^{z_i},h_i^{z_i})/\tau}}{\sum_{j=1}^{N}e^{sim(h_i^{z_i},h_j^{z_j})/\tau}}$$

其中,h_i表示初始样本,h_j表示增强样本,log表示对数,e表示自然常数,N表示样本数量,sim表示余弦相似度函数;

所述使用社区检测算法对句子向量进行聚类处理,得到故障簇,包括:

通过社区检测算法结合余弦相似度匹配方法对故障节点列表完成聚类操作,将具有相同故障特征的故障节点列表聚集在一个故障簇中;

引入余弦相似度将社区检测算法产生的散乱句子和大簇的平均向量比较相似度,根据相似度的大小将散乱的句子归类到合适的簇中;

所述余弦相似度表达式为:

$$\cos\theta=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}\ (2);$$

A代表散乱句子的特征向量,B代表故障大簇的代表特征向量;

所述根据每个故障簇内信息构建故障树,并导出生成的故障树可视化展示,包括:

根据故障簇内的节点列表内容构建成一颗包含所有故障节点的故障等级树,从故障树根节点到某个子节点的路径即为单条故障现象内容;

依次将不同故障簇对应的故障树导出至Excel可视化展示。

2.如权利要求1所述的自动化移动设备故障层级树构建方法,其特征在于,在步骤102和103之间还包括步骤:

步骤102’、对故障节点列表中的单个动词节点以及与单个动词节点相邻的名词节点合并,并去除列表中的特殊符号,所述特殊符号包括标点符号、空字符。

3.一种自动化移动设备故障层级树构建装置,包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序,其特征在于,所述处理器执行所述计算机程序时实现如权利要求1至2中任一所述方法的步骤。

4.一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至2中任一所述方法的步骤。

自动化移动设备故障层级树构建方法、装置及储存介质

技术领域

[0001] 本发明涉及一种实体抽取技术领域,具体涉及一种基于无监督聚类的自动化移动设备故障层级树构建方法、装置及存储介质。

背景技术

[0002] 实体抽取的方法分为3类:基于规则的方法、基于统计机器学习的方法、基于深度学习的方法。

[0003] 早期的实体抽取是在限定文本领域、限定语义单元类型的条件下进行的,主要采用的是基于规则与词典的方法,例如使用已定义的规则,抽取出文本中的人名、地名、组织机构名、特定时间等实体。选用的特征包括统计信息、标点符号、关键字、指示词和方向词、中心词等方法,以模式和字符串相匹配为主要手段。

[0004] 基于统计机器学习的方法主要包括隐马尔可夫模型(HiddenMarkovMode,HMM)、最大熵(MaxmiumEntropy,ME)、支持向量机(Support VectorMachine,SVM)、条件随机场(ConditionalRandom Fields,CRF)等。在基于统计的这四种学习方法中,最大熵模型结构紧凑,具有较好的通用性,主要缺点是训练时间长复杂性高,有时甚至导致训练代价难以承受,另外由于需要明确的归一化计算,导致开销比较大。而条件随机场为命名实体识别提供了一个特征灵活、全局最优的标注框架,但同时存在收敛速度慢、训练时间长的问题。一般说来,最大熵和支持向量机在正确率上要比隐马尔可夫模型高一些,但隐马尔可夫模型在训练和识别时的速度要快一些,主要是由于在利用 Viterbi 算法求解命名实体类别序列时的效率较高。隐马尔可夫模型更适用于一些对实时性有要求以及像信息检索这样需要处理大量文本的应用,如短文本命名实别。

[0005] 基于深度学习的方法利用深度学习非线性的特点,从输入到输出建立非线性的映射。相比于线性模型(如线性链式CRF、隐马尔可夫模型),深度学习模型可以利用巨量数据通过非线性激活函数学习得到更加复杂精致的特征。传统的基于特征的方法需要大量的工程技巧与领域知识;而深度学习方法可以从输入中自动发掘信息以及学习信息的表示,而且通常这种自动学习并不意味着更差的结果。深度NER模型是端到端的;端到端模型的一个好处在于可以避免流水线(pipeline)类模型中模块之间的误差传播;另一点是端到端的模型可以承载更加复杂的内部设计,最终产出更好的结果。目前常用的方法为BiLSTM+CRF组合的结构进行标签预测。BiLSTM+CRF 是目前比较流行的序列标注算法,其将 BiLSTM 和 CRF 结合在一起,使模型即可以像 CRF 一样考虑序列前后之间的关联性,又可以拥有 LSTM 的特征抽取及拟合能力。

[0006] 现有基于规则的方法需要大量的人工工作来创建所有可能的规则,必须为每个关系类型创建规则。基于机器学习和深度学习的方法大多对数据有较高要求,需要标注大量的数据。

发明内容

[0007] 针对故障报告无任何标注,同时因为涉及到的故障种类繁多,难以找到合适的规则进行处理的问题,本发明提供一种基于无监督聚类的自动化移动设备故障层级树构建方法、装置及存储介质。

[0008] 为实现上述目的,本发明的技术方案是:

[0009] 第一方面,本发明提供一种自动化移动设备故障层级树构建方法,所述方法包括如下步骤:

[0010] 步骤101、分析故障信息报告,提取出不同的故障现象内容,得到故障现象列表;

[0011] 步骤102、遍历所有的故障现象列表,对每个故障现象内容按中文词性进行分割,去除冗余的噪声信息,得到对应的故障节点列表;同时,利用信息熵和互信息进行新词发现生成移动设备故障领域的专业新词库,在切割故障现象内容时保持专业新词完整;

[0012] 步骤103、使用SimCSE预训练模型对故障节点列表进行句子嵌入处理,得到句子向量;

[0013] 步骤104、使用社区检测算法对句子向量进行聚类处理,得到故障簇;

[0014] 步骤105、根据每个故障簇内信息构建故障树,并导出生成的故障树可视化展示。

[0015] 进一步地,在步骤102和103之间还包括步骤:

[0016] 步骤102’、对故障节点列表中的单个动词节点以及与单个动词节点相邻的名词节点合并,并去除列表中的特殊符号,所述特殊符号包括标点符号、空字符。

[0017] 进一步地,所述分析故障信息报告,提取出不同的故障现象内容包括:

[0018] 根据从用户、开发人员、供应商反馈得到的故障信息汇总故障报告;

[0019] 利用正则表达式提取故障报告中和故障现象有关的信息,初步过滤部分无关的故障噪声数据。

[0020] 进一步地,所述使用SimCSE预训练模型对故障节点列表进行句子嵌入处理,得到句子向量,包括:

[0021] 对于给定一个故障节点列表$\{X_i\}^m$另$X^+_i=X_i$,SimCSE模型使用独立的dropout作为掩码来获得增强的正样本对;

[0022] 样本的嵌入向量生成表示为$h^z_i=f_\theta(x_i,z)$,其中z是随机的dropout掩码;

[0023] SimCSE模型通过将相同的样本输入编码器,并应用不同的dropout掩码z、z’,获得相同样本的不同增强样本。

[0024] 进一步地,最终的对比损失函数为:

[0025] $$L_i = -\log \frac{e^{sim\ (h_i^{z_i}, h_i^{z_i'})/\tau}}{\sum_{j=1}^{N} e^{sim\ (h_i^{z_i}, h_i^{z_i'})/\tau}} \quad (1)。$$

[0026] 进一步地,所述使用社区检测算法对句子向量进行聚类处理,得到故障簇,包括:

[0027] 通过社区检测算法结合余弦相似度匹配方法对故障节点列表完成聚类操作,将具有相同故障特征的故障节点列表聚集在一个故障簇中;

[0028] 引入余弦相似度将社区检测算法产生的散乱句子和大簇的平均向量比较相似度，根据相似度的大小将散乱的句子归类到合适的簇中。

[0029] 进一步地，所述余弦相似度表达式为：

[0030] $$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (2);$$

[0031] A代表散乱句子的特征向量，B代表故障大簇的代表特征向量。

[0032] 进一步地，所述根据每个故障簇内信息构建故障树，并导出生成的故障树可视化展示，包括：

[0033] 根据故障簇内的节点列表内容构建成一颗包含所有故障节点的故障等级树，从故障树根节点到某个子节点的路径即为单条故障现象内容；

[0034] 依次将不同故障簇对应的故障树导出至Excel可视化展示。

[0035] 第二方面，本发明提供一种自动化移动设备故障层级树构建装置，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如上任一所述方法的步骤。

[0036] 第三方面发明提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时实现如上任一所述方法的步骤。

[0037] 本发明与现有技术相比，其有益效果在于：

[0038] 本发明基于中文词性和新词发现切割故障现象可以得到较清晰的故障层级关系，通过社区检测算法聚集故障簇并将每个故障簇内信息可视化展示，方便后续测试人员分析故障信息。

附图说明

[0039] 图1为本发明实施例1提供的自动化移动设备故障层级树构建方法流程图；

[0040] 图2为本发明实施例1提供的自动化移动设备故障层级树构建方法流程图；

[0041] 图3为本发明实施例2提供的自动化移动设备故障层级树构建装置组成示意图。

具体实施方式

[0042] 下面结合附图和实施例对本发明的技术方案做进一步的说明。

[0043] 实施例1：

[0044] 参阅图1所示，本实施例提供的自动化移动设备故障层级树构建方法主要包括如下步骤：

[0045] 步骤101、分析故障信息报告，提取出不同的故障现象内容，得到故障现象列表。

[0046] 步骤102、遍历所有的故障现象列表，对每个故障现象内容按中文词性进行分割，去除冗余的噪声信息，得到对应的故障节点列表；同时，利用信息熵和互信息进行新词发现生成移动设备故障领域的专业新词库，在切割故障现象内容时保持专业新词完整。

[0047] 也就是说，在此步骤中，按照中文词性和新词发现对节点列表进行层级切割，最终按照故障层次等级逐层递进故障信息。

[0048] 步骤103、使用SimCSE预训练模型对故障节点列表进行句子嵌入处理,得到句子向量。

[0049] 步骤104、使用社区检测算法对句子向量进行聚类处理,得到故障簇。

[0050] 在此步骤中,通过使用社区检测算法对经过句子嵌入的故障信息进行聚类操作,整个过程基于无监督技术,不需要任何人工标签。

[0051] 步骤105、根据每个故障簇内信息构建故障树,并导出生成的故障树可视化展示。

[0052] 也就是说,在此步骤中,完成故障聚类之后,将得到的单个故障簇内的信息构建为一颗故障多叉树,给予测试人员清晰的故障可视化展示。

[0053] 由此可见,本方法基于中文词性和新词发现切割故障现象可以得到较清晰的故障层级关系,通过社区检测算法聚集故障簇并将每个故障簇内信息可视化展示,方便后续测试人员分析故障信息。

[0054] 由于分割后的故障列表中包含大量的单个动词节点,单个的动词节点无法准确表示故障信息,为此在一优选实施例中,如图2所示,在步骤102和103之间还包括步骤:

[0055] 步骤102’、对故障节点列表中的单个动词节点以及与单个动词节点相邻的名词节点合并,以获得更丰富的故障节点表示;

[0056] 另外,由于切割操作不可避免的会产生大量单个的特殊符号:标点符号、空字符等,为此在此步骤中,还利用正则表达式清理故障节点列表中的这些特殊符号。

[0057] 在一具体实施例中,上述步骤101包括:

[0058] (1)根据从用户、开发人员、供应商等反馈得到的故障信息汇总故障报告。

[0059] (2)利用正则表达式提取故障报告中和故障现象有关的信息,初步过滤到一些其他无关的故障噪声数据。

[0060] 在一具体实施例中,上述步骤103包括:

[0061] 使用SimCSE预训练模型对收集到的节点列表进行句子嵌入处理得到句子向量的过程如下:对于给定一个故障节点列表$\{X_i\}^m$,另$X^+_i=X_i$,SimCSE模型使用独立的dropout作为掩码来进一步获得增强的正样本对。在标准的Transformer训练过程中会有多个dropout掩码,因此样本的嵌入向量生成表示为$h^z_i=f_\theta(x_i,z)$,其中z是随机的dropout掩码。SimCSE通过将相同的样本输入编码器,并应用不同的dropout掩码z、z’,从而获得相同样本的不同增强样本,样本与不同增强样本最终的对比损失函数如下所示,

[0062] $$L_i=-\log\frac{e^{sim\ (h_i^{z_i},h_i^{z_i'})/\tau}}{\sum_{j=1}^{N}e^{sim\ (h_i^{z_i},h_i^{z_i'})/\tau}} \quad (1)$$其中,

h_i表示初始样本,h_j表示增强样本,log表示对数,e表示自然常数,N表示样本数量,sim表示余弦相似度函数。

[0063] 在一具体实施例中,上述步骤104包括:

[0064] 使用社区检测算法对故障句子进行聚类操作,在实际使用中根据故障报告信息调

整相似度阈值以及单个故障簇内的最少故障信息数量以达到更好的效果。

[0065] 尽管设置了单个故障簇内最少故障信息数量,但社区检测算法仍会产生很多零碎句子,引入余弦相似度将产生的散乱句子和大簇的代表向量比较相似度,根据相似度的大小将散乱的句子归类到合适的簇中,其中故障大簇的代表向量通过取簇内所有向量平均值得到。其中,余弦相似度表达式为:

[0066] $$\cos\theta=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (2);$$

[0067] A代表散乱句子的特征向量,B代表故障大簇的代表特征向量。

[0068] 在一具体实施例中,上述步骤105包括:

[0069] 根据故障簇内的节点列表内容构建成一颗包含所有故障节点的故障等级树,从故障树根节点到某个子节点的路径即为单条故障现象内容。

[0070] 依次将不同故障簇对应的故障树导出至Excel可视化展示。

[0071] 综上,本发明与现有技术相比,具有如下优点和有益效果:

[0072] 1、本发明使用社区检测算法对经过句子嵌入的故障信息进行聚类操作,整个过程基于无监督技术,不需要任何人工标签。

[0073] 2、本发明按照中文词性和新词发现算法对节点列表进行层级切割,最终按照故障层次等级逐层递进故障信息。

[0074] 3、完成故障聚类之后,将得到的单个故障簇内的信息构建为一颗故障多叉树,最终导出成Excel表格,给予测试人员清晰的故障可视化展示。

[0075] 实施例2:

[0076] 参阅图3所示,本实施例提供的自动化移动设备故障层级树构建装置包括处理器31、存储器32以及存储在该存储器32中并可在所述处理器31上运行的计算机程序33,例如自动化移动设备故障层级树构建程序。该处理器31执行所述计算机程序33时实现上述实施例1步骤,例如图1所示的步骤。

[0077] 示例性的,所述计算机程序33可以被分割成一个或多个模块/单元,所述一个或者多个模块/单元被存储在所述存储器32中,并由所述处理器31执行,以完成本发明。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机程序指令段,该指令段用于描述所述计算机程序33在所述自动化移动设备故障层级树构建装置中的执行过程。

[0078] 所述自动化移动设备故障层级树构建装置可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述自动化移动设备故障层级树构建装置可包括,但不仅限于,处理器31、存储器32。本领域技术人员可以理解,图3仅仅是自动化移动设备故障层级树构建装置的示例,并不构成自动化移动设备故障层级树构建装置的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件,例如所述自动化移动设备故障层级树构建装置还可以包括输入输出设备、网络接入设备、总线等。

[0079] 所称处理器31可以是中央处理单元(Central Processing Unit,CPU),还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC) 、现成可编程门阵列

(FieldProgrammable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0080] 所述存储器32可以是所述自动化移动设备故障层级树构建装置的内部存储元,例如自动化移动设备故障层级树构建装置的硬盘或内存。所述存储器32也可以是所述自动化移动设备故障层级树构建装置的外部存储设备,例如所述自动化移动设备故障层级树构建装置上配备的插接式硬盘,智能存储卡(SmartMedia Card,SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)等。进一步地,所述存储器32还可以既包括所述自动化移动设备故障层级树构建装置的内部存储单元也包括外部存储设备。所述存储器32用于存储所述计算机程序以及所述自动化移动设备故障层级树构建装置所需的其他程序和数据。所述存储器32还可以用于暂时地存储已经输出或者将要输出的数据。

[0081] 实施例3:

[0082] 本实施例提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,所述计算机程序被处理器执行时实现实施例1所述方法的步骤。

[0083] 所示计算机可读介质可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例(非穷尽性列表)包括以下:具有一个或多个布线的电连接部(电子装置),便携式计算机盘盒(磁装置),随机存取存储器(RAM),只读存储器(ROM),可擦除可编辑只读存储器(EPROM或闪速存储器),光纤装置,以及便携式光盘只读存储器(CDROM)。另外,计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质,例如通过对纸或其他介质进行光学扫描,接着进行编辑、解译或必要时以其他合适方式进行处理再以电子方式获得所述程序,然后将其存储在计算机存储器中。

[0084] 上述实施例只是为了说明本发明的技术构思及特点,其目的是在于让本领域内的普通技术人员能够了解本发明的内容并据以实施,并不能以此限制本发明的保护范围。凡是根据本发明内容的实质所做出的等效的变化或修饰,都应涵盖在本发明的保护范围内。

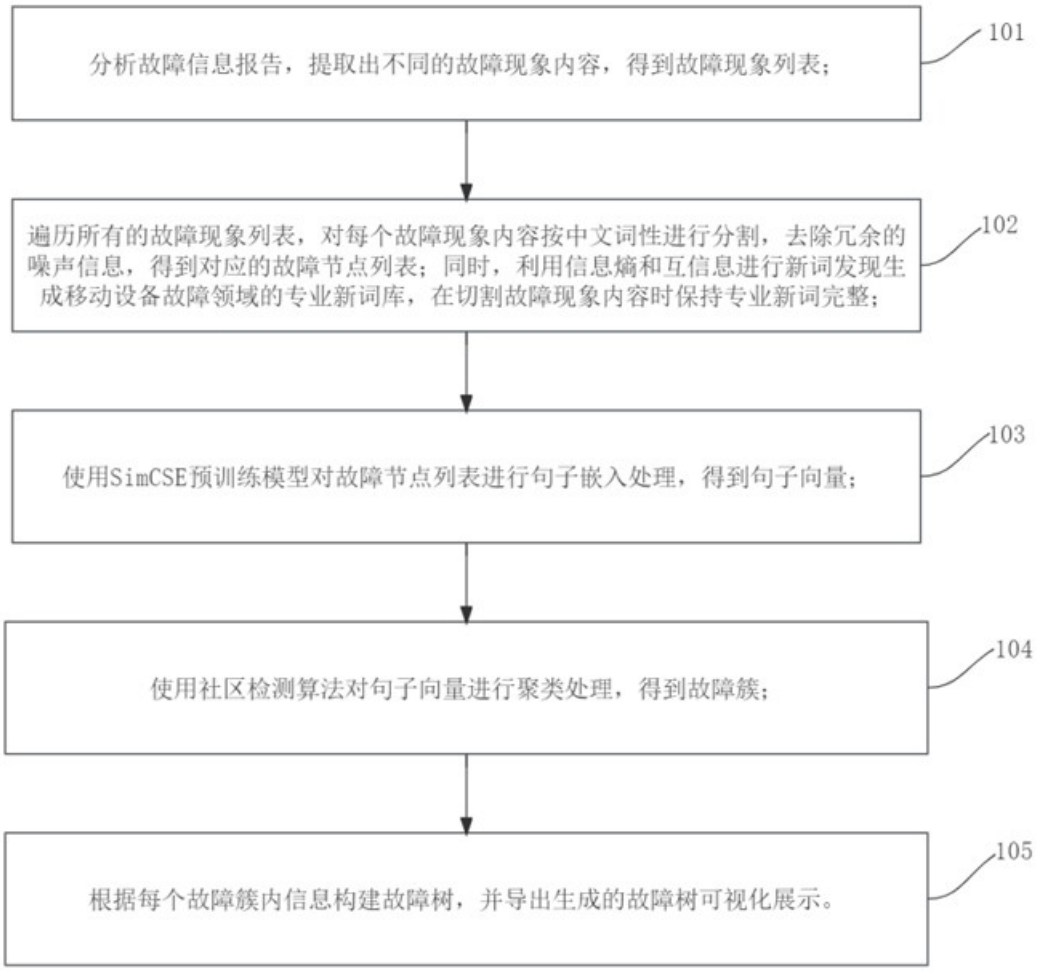

分析故障信息报告，提取出不同的故障现象内容，得到故障现象列表；
101
遍历所有的故障现象列表，对每个故障现象内容按中文词性进行分割，去除冗余的噪声信息，得到对应的故障节点列表；同时，利用信息熵和互信息进行新词发现生成移动设备故障领域的专业新词库，在切割故障现象内容时保持专业新词完整；
102
使用SimCSE预训练模型对故障节点列表进行句子嵌入处理，得到句子向量；
103
使用社区检测算法对句子向量进行聚类处理，得到故障簇；
104
根据每个故障簇内信息构建故障树，并导出生成的故障树可视化展示。
105

图 1

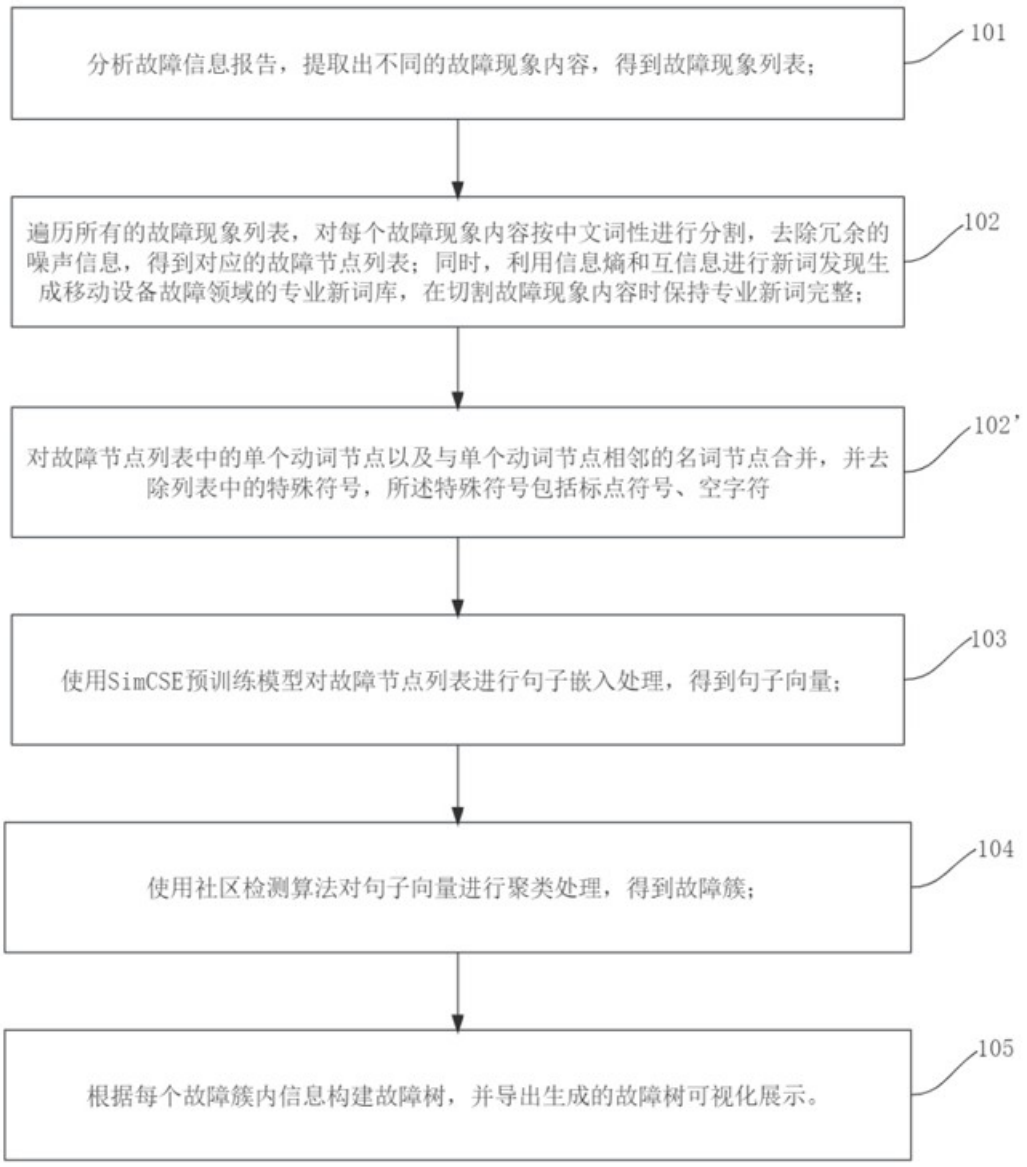

分析故障信息报告，提取出不同的故障现象内容，得到故障现象列表；
101
遍历所有的故障现象列表，对每个故障现象内容按中文词性进行分割，去除冗余的噪声信息，得到对应的故障节点列表；同时，利用信息熵和互信息进行新词发现生成移动设备故障领域的专业新词库，在切割故障现象内容时保持专业新词完整；
102
对故障节点列表中的单个动词节点以及与单个动词节点相邻的名词节点合并，并去除列表中的特殊符号，所述特殊符号包括标点符号、空字符
102'
使用SimCSE预训练模型对故障节点列表进行句子嵌入处理，得到句子向量；
103
使用社区检测算法对句子向量进行聚类处理，得到故障簇；
104
根据每个故障簇内信息构建故障树，并导出生成的故障树可视化展示。
105

图 2

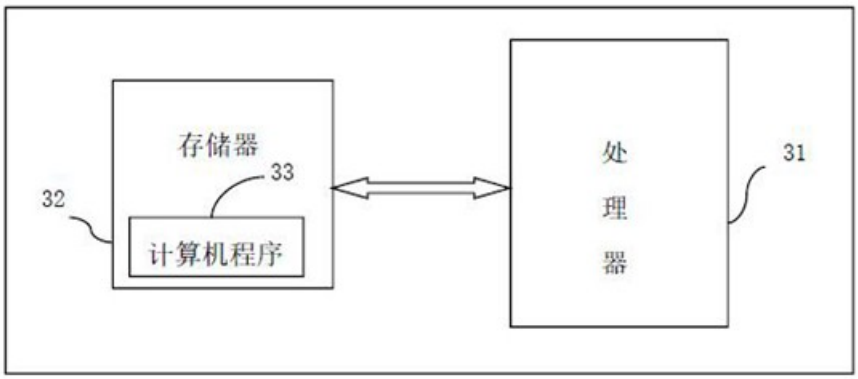

存储器
33
32
计算机程序
处
理
器
31

图 3