(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108509995 A
(43)申请公布日 2018.09.07

(21)申请号 201810287208.4

(22)申请日 2018.04.03

(71)申请人 电子科技大学
地址 611731 四川省成都市高新区(西区)西源大道2006号

(72)发明人 廖伟智 张强 阴艳超 曹奕翎 严伟军

(74)专利代理机构 成都虹盛汇泉专利代理有限公司 51268
代理人 王伟

(51)Int.Cl.
G06K 9/62(2006.01)
G06F 17/27(2006.01)

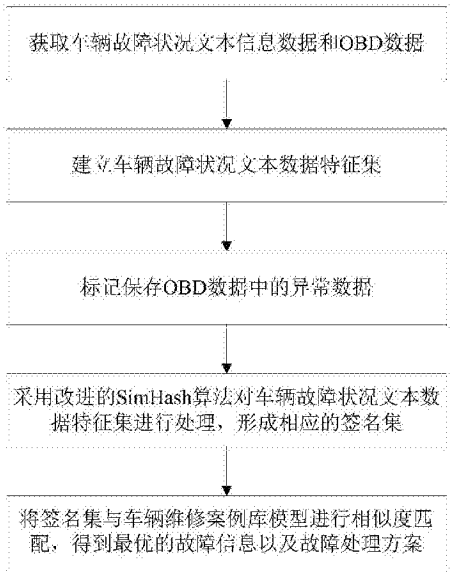
权利要求书1页 说明书5页 附图1页

(54)发明名称

基于中文文本分析和OBD数据处理的汽车故障诊断方法

(57)摘要

本发明公开了一种基于中文文本分析和OBD数据处理的汽车故障诊断方法,其包括获取车辆故障状况文本信息数据和OBD数据,建立车辆故障状况文本数据特征集,标记保存OBD数据中的异常数据,采用改进的SimHash算法对车辆故障状况文本数据特征集进行处理形成相应的签名集,将签名集与车辆维修案例库模型进行相似度匹配,得到最优的故障信息以及故障处理方案。本发明节约了大量的人工处理的时间,实现了经验的数据化,并且通过专业的OBD数据与维修案例经验的结合使得诊断结果更加准确;此外还能方便用户随时掌握汽车运行数据和故障码,实现有效的对车辆进行故障检测。



CN 108509995 A

1.一种基于中文文本分析和OBD数据处理的汽车故障诊断方法，其特征在于，包括以下步骤：

A、获取车辆故障状况文本信息数据和OBD数据；

B、对步骤A中的车辆故障状况文本信息数据进行中文分词、消除停用词、同义词归一化处理，建立相应的车辆故障状况文本数据特征集；

C、将步骤A中的OBD数据与OBD数据库进行对比分析处理，对异常数据进行标记保存，同时将异常数据出现的车辆部位作为特定特征添加到步骤B中的车辆故障状况文本数据特征集中；

D、采用改进的SimHash算法对步骤C中的的车辆故障状况文本数据特征集进行处理，形成相应的签名集；

E、将步骤D中的签名集与车辆维修案例库模型进行相似度匹配，得到最优的故障信息以及故障处理方案。

2.如权利要求1所述的基于中文文本分析和OBD数据处理的汽车故障诊断方法，其特征在于，所述步骤D采用改进的SimHash算法对步骤C中的的车辆故障状况文本数据特征集进行处理，形成相应的签名集具体为：根据车辆故障状况文本数据特征集中的词频信息和词性信息计算每个特征词的权值，采用改进的SimHash算法构建相应的签名集。

3.如权利要求2所述的基于中文文本分析和OBD数据处理的汽车故障诊断方法，其特征在于，所述根据车辆故障状况文本数据特征集中的词频信息和词性信息计算每个特征词的权值的计算公式具体为：

$$w_v = \frac{count(v)}{count(d)} \cdot \log \frac{N}{count(v,D)} + w_p$$

其中，w_v表示特征词v的权值，count(v)表示特征词v在车辆故障状况文本d中的数量，count(d)表示车辆故障状况文本d含有特征词的总数量，N表示车辆故障状况文本集合D的数量，count(v,D)表示车辆故障状况文本集合D包含特征词v的文本数量，w_p表示特征词词性的权值。

4.如权利要求3所述的基于中文文本分析和OBD数据处理的汽车故障诊断方法，其特征在于，所述步骤E将步骤D中的签名集与车辆维修案例库模型进行相似度匹配具体为：计算步骤D中的签名集与车辆维修案例签名集的汉明距离，生成候选集；采用BM25算法对车辆故障状况文本数据特征集与候选集中每个车辆维修案例计算相似度评分，并按评分进行排序。

基于中文文本分析和OBD数据处理的汽车故障诊断方法

技术领域

[0001] 本发明属于汽车故障诊断技术领域，具体涉及一种基于中文文本分析和OBD数据处理的汽车故障诊断方法。

背景技术

[0002] 汽车故障诊断技术，是指在汽车整车不进行拆解和解体的情况下，通过专业设备和汽车表征，确定汽车的技术和工作状况，排查故障原因和故障部位的汽车应用技术。

[0003] 汽车是由许多总成、机构和元器件有序构成的一个复杂技术系统。因而研究汽车故障的产生机理和变化规律，定期甚至实时在线检测汽车的使用性能，从而及时准确的诊断出故障部位并排除故障，就成为汽车实用技术和故障诊断技术的一项重要内容。

[0004] 传统的汽车故障诊断，是需要在特定的场所，通过有经验或者有相关知识背景的专业人士检查、测量、分析、判断等一系列过程完成的。其基本方法主要分为2种：现代仪器设备诊断法和人工经验诊断法。

[0005] (1)现代仪器设备诊断法是在汽车不被解体情况下，利用测试仪器、检验设备和检验工具，检测整车、总成或机构的参数、波形曲线，为分析、判断汽车技术状况提供定量依据的诊断方法：

[0006] (2)人工经验诊断法，是指诊断人员凭一定的理论知识和相关的实战经验，在汽车不解体或局部解体情况下，采用眼观、耳听、手摸等这些直观的感觉、借助简单工具器械，进行检查分析汽车的技术状况，排查故障原因和故障部位的一种诊断方法。

[0007] 现实状况中，传统的汽车故障诊断，往往同时使用上述两种方法，也称综合诊断法。

[0008] 然而，这种传统的故障诊断需要满足许多的条件并且存在一定的局限性：

[0009] (1)经验和理论知识背景；

[0010] (2)昂贵的设备检测仪器；

[0011] (3)维保需要在特定的地点进行；

[0012] (4)车辆故障不能提前进行预警，都是在已经发生或产生影响驾驶行为的结果时才能发现故障；

[0013] (5)只能在定期保养的时间做故障诊断；

[0014] (6)维修保养行为并不完全透明，车主一般并不能了解到汽车发生故障的真实原因。

发明内容

[0015] 本发明的发明目的是：为了解决现有技术中存在的以上问题，本发明提出了一种基于中文文本分析和OBD(On-Board Diagnostic，车载诊断系统)数据处理的汽车故障诊断方法，能够方便的对车辆故障进行实时、快速诊断，辅助车主和维修人员进行相关车辆维修与保养。。

[0016] 本发明的技术方案是：一种基于中文文本分析和OBD数据处理的汽车故障诊断方法，包括以下步骤：

[0017] A、获取车辆故障状况文本信息数据和OBD数据；

[0018] B、对步骤A中的车辆故障状况文本信息数据进行中文分词、消除停用词、同义词归一化处理，建立相应的车辆故障状况文本数据特征集；

[0019] C、将步骤A中的OBD数据与OBD数据库进行对比分析处理，对异常数据进行标记保存，同时将异常数据出现的车辆部位作为特定特征添加到步骤B中的车辆故障状况文本数据特征集中；

[0020] D、采用改进的SimHash算法对步骤C中的的车辆故障状况文本数据特征集进行处理，形成相应的签名集；

[0021] E、将步骤D中的签名集与车辆维修案例库模型进行相似度匹配，得到最优的故障信息以及故障处理方案。

[0022] 进一步地，所述步骤D采用改进的SimHash算法对步骤C中的的车辆故障状况文本数据特征集进行处理，形成相应的签名集具体为：根据车辆故障状况文本数据特征集中的词频信息和词性信息计算每个特征词的权值，采用改进的SimHash算法构建相应的签名集。

[0023] 进一步地，所述根据车辆故障状况文本数据特征集中的词频信息和词性信息计算每个特征词的权值的计算公式具体为：

[0024] $$w_v = \frac{count(v)}{count(d)} \cdot \log \frac{N}{count(v,D)} + w_p$$

[0025] 其中，w_v表示特征词v的权值，count(v)表示特征词v在车辆故障状况文本d中的数量，count(d)表示车辆故障状况文本d含有特征词的总数量，N表示车辆故障状况文本集合D的数量，count(v,D)表示车辆故障状况文本集合D包含特征词v的文本数量，w_p表示特征词词性的权值。

[0026] 进一步地，所述步骤E将步骤D中的签名集与车辆维修案例库模型进行相似度匹配具体为：计算步骤D中的签名集与车辆维修案例签名集的汉明距离，生成候选集；采用BM25算法对车辆故障状况文本数据特征集与候选集中每个车辆维修案例计算相似度评分，并按评分进行排序。

[0027] 本发明的有益效果是：本发明采用中文文本分析方法，通过对文本数据进行分词、特征提取并建立相应模型，同时结合SimHash算法对汽车维修案例数据库进行相似度匹配，筛选出相似度最高的几个结果提供给用户，从而节约了大量的人工处理的时间，实现了经验的数据化，并且通过专业的OBD数据与维修案例经验的结合使得诊断结果更加准确；此外还能方便用户随时掌握汽车运行数据和故障码，实现有效的对车辆进行故障检测。

附图说明

[0028] 图1是本法吗的基于中文文本分析和OBD数据处理的汽车故障诊断方法的流程示意图。

具体实施方式

[0029] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对

本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。

[0030] 如图1所示，为本法吗的基于中文文本分析和OBD数据处理的汽车故障诊断方法的流程示意图。一种基于中文文本分析和OBD数据处理的汽车故障诊断方法，包括以下步骤：

[0031] A、获取车辆故障状况文本信息数据和OBD数据；

[0032] B、对步骤A中的车辆故障状况文本信息数据进行中文分词、消除停用词、同义词归一化处理，建立相应的车辆故障状况文本数据特征集；

[0033] C、将步骤A中的OBD数据与OBD数据库进行对比分析处理，对异常数据进行标记保存，同时将异常数据出现的车辆部位作为特定特征添加到步骤B中的车辆故障状况文本数据特征集中；

[0034] D、采用改进的SimHash算法对步骤C中的的车辆故障状况文本数据特征集进行处理，形成相应的签名集；

[0035] E、将步骤D中的签名集与车辆维修案例库模型进行相似度匹配，得到最优的故障信息以及故障处理方案。

[0036] 在本发明的一个可选实施例中，上述步骤A采用移动端通过用户输入车辆故障状况文本信息数据，车辆故障状况文本信息数据包括多种车辆故障反映出的现象；并采用车载终端硬件通过汽车OBD接口读取汽车的运行信息以及故障码，同时将其通过4G网络上传至云端的OBD数据库和汽车诊断系统；OBD数据包括故障码和汽车的运行信息数据流，汽车的运行信息数据流由多个传感器的监测数据组成。

[0037] 在本发明的一个可选实施例中，上述步骤B利用汽车诊断系统调取车辆维修案例数据库模型中的自然语言处理模块进行中文分词、消除停用词、同义词归一化等处理过程后，建立相应的车辆故障状况文本数据特征集。这里的车辆维修数据库内存储有汽车故障案例、专业的解决方法，且上述数据根据车辆部位、车辆系统和故障等级三方面进行归类、整理和预警分析判断。车辆维修案例数据库模型通过收集以往的汽车维修案例和处理方法的文本数据，针对其短文本的特点，采用相应的自然语言处理方法，包括分词、故障特征提取等。车辆部位包括发动机、车身、底盘；车辆系统包括燃油系统、冷却系统、润滑系统、进排气系统、传动系统、电控系统、转向系统、制动系统等。

[0038] 在本发明的一个可选实施例中，上述步骤C利用汽车诊断系统判断是否侦测到车辆运行信息及故障码信息数据，如侦测到则调取OBD数据库进行数据比对分析处理，对异常数据进行标记保存(标记中备注异常数据出现的车辆具体部位)，如未检测到则调取云端保存的OBD数据按照上述方法进行数据处理。同时将异常数据出现的车辆部位作为特定特征添加到步骤B中的车辆故障状况文本数据特征集中。这里的OBD数据库内包括故障码、中文定义、英文定义、不同车型传感器监测参数、背景知识及解决方案。

[0039] 现有的自然语言处理算法，通常是将词频作为该词语的权值，但仅将词频作为权值必将丢失很多文本信息，尤其对于此类车辆故障信息，通常都是少于500字的短文本，词频数据会非常稀疏，词频的大小不能完全代表词语的重要程度；本发明考虑到在车辆故障信息中，一些特殊的专有名词是故障特征中的重要体现，仅考虑词频信息计算文本相似度的精度并不高，因此结合词频信息和词性信息来综合计算关键词的权值。

[0040] 在本发明的一个可选实施例中，上述步骤D采用改进的SimHash算法对步骤C中的

的车辆故障状况文本数据特征集进行处理，形成相应的签名集具体为：根据车辆故障状况文本数据特征集中的词频信息和词性信息计算每个特征词的权值，采用改进的SimHash算法构建相应的签名集。

[0041] 对于车辆故障来说，车辆中的零件名称是对整个故障描述最重要的，因为它直接描述了出故障的器件，接下来对该器件的描述词语也就是形容词和动词等也应该要比其他词语重要，其权值应该次之，因此本文在对词语计算权值时，考虑加入了词性信息。本发明实施例中词性权值如下表。

[0042] 表1词性权值表

[0043]

词性	特殊词	动词	普通名词	其它
权值	1	0.75	0.6	0.4

[0044] 在表1所设定的词性权值下，本发明设定词频权值和词性权值的加权比为1。本发明在计算特征词的权值时，将词频信息与词性信息相结合，综合考虑了词性和词频信息对特征词的重要程度。根据车辆故障状况文本数据特征集中的词频信息和词性信息计算每个特征词的权值的计算公式具体为：

[0045] $$w_v = \frac{count(v)}{count(d)} \cdot \log \frac{N}{count(v,D)} + w_p$$

[0046] 其中，w_v表示特征词v的权值，count(v)表示特征词v在车辆故障状况文本d中的数量，count(d)表示车辆故障状况文本d含有特征词的总数量，N表示车辆故障状况文本集合D的数量，count(v,D)表示车辆故障状况文本集合D包含特征词v的文本数量，w_p表示特征词词性的权值。

[0047] 由于车辆故障诊断系统中的维修案例数据库模型中都是短文本，语义信息非常稀疏，不适合采用基于语义相似度的计算方法，因此本发明在计算案例之间的相似度时采用基于向量空间的方法。而文本信息处理是有很大内存消耗的，案例库中有大量案例，将其读入计算机内存，并生成向量矩阵会占用巨大的存储空间，且大维度的数据计算时间复杂度也非常高。为了解决该问题，本发明将数据库存放在云端服务器，同时采用SimHash算法。车辆维修数据库和OBD数据库均与云服务器连接并将相关数据传输至云服务器。优选地，本发明设置汉明距离阈值为10。

[0048] 在本发明的一个可选实施例中，上述步骤E首先按照上述步骤对车辆维修案例进行处理形成车辆维修案例签名集，具体包括

[0049] 获得案例特征词集：根据分词词典、停用词表对车辆维修案例进行分词、词性过滤和过滤停用词后，再根据同义词词典归一化同义词，获得车辆维修案例的特征词集；

[0050] 构建Simhash签名：对车辆维修案例特征词集采用上述计算公式计算每个特征词的权值，根据SimHash生成算法构建每个案例的SimHash签名；

[0051] 在获得步骤D中车辆故障状况的签名集和车辆维修案例的签名集后，计算步骤D中的签名集与车辆维修案例签名集的汉明距离，选取汉明距离小于汉明距离阈值的检索案例集作为候选集；采用BM25算法对车辆故障状况文本数据特征集与候选集中每个车辆维修案例计算相似度评分，并按评分进行排序；从而筛选出相似度最高的几个故障信息以及故障处理方案提供给用户。

[0052] 本领域的普通技术人员将会意识到，这里所述的实施例是为了帮助读者理解本发明的原理，应被理解为本发明的保护范围并不局限于这样的特别陈述和实施例。本领域的普通技术人员可以根据本发明公开的这些技术启示做出各种不脱离本发明实质的其它各种具体变形和组合，这些变形和组合仍然在本发明的保护范围内。

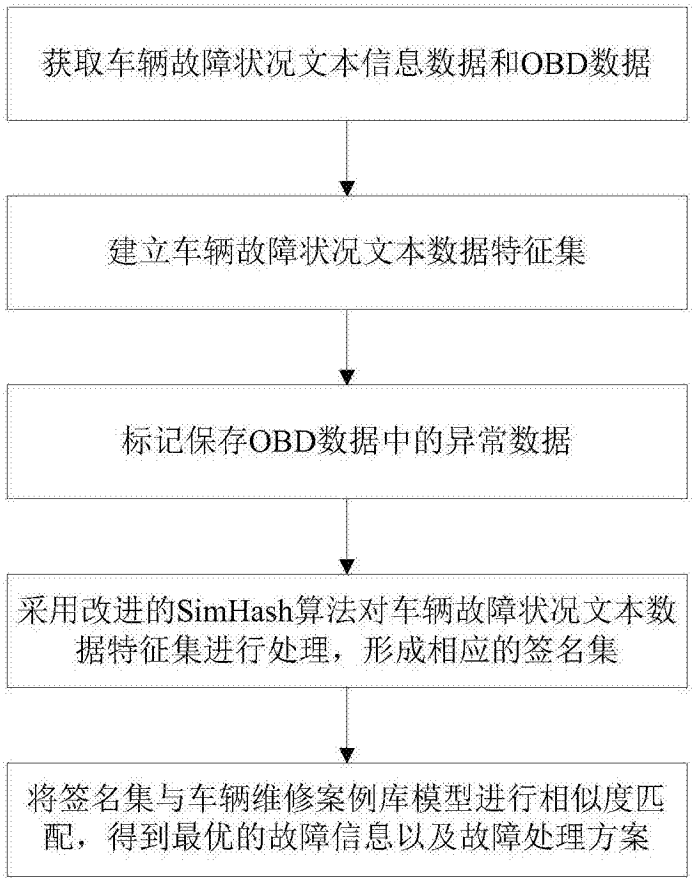

获取车辆故障状况文本信息数据和OBD数据
建立车辆故障状况文本数据特征集
标记保存OBD数据中的异常数据
采用改进的SimHash算法对车辆故障状况文本数据特征集进行处理，形成相应的签名集
将签名集与车辆维修案例库模型进行相似度匹配，得到最优的故障信息以及故障处理方案

图1