

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号

WO 2020/207167 A1

(43) 国际公布日
2020 年 10 月 15 日 (15.10.2020)

(51) 国际专利分类号:
G06F 16/35 (2019.01)

(21) 国际申请号: PCT/CN2020/078389

(22) 国际申请日: 2020 年 3 月 9 日 (09.03.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910297133.2 2019 年 4 月 12 日 (12.04.2019) CN

(71) 申请人: 深圳前海微众银行股份有限公司 (WEBANK CO., LTD) [CN/CN]; 中国广东省深圳市南山区沙河西路 1819 号深圳湾科技生态园 7 栋 A 座, Guangdong 518052 (CN)。

(72) 发明人: 张威 (ZHANG, Wei); 中国广东省深圳市南山区沙河西路 1819 号深圳湾科技生态园 7 栋 A 座, Guangdong 518052 (CN)。 杨永

帮 (YANG, Yongbang); 中国广东省深圳市南山区沙河西路 1819 号深圳湾科技生态园 7 栋 A 座, Guangdong 518052 (CN)。

(74) 代理人: 深圳市世纪恒程知识产权代理事务所 (CENFO INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市南山区西丽街道松坪山社区松坪山路 3 号奥特迅电力大厦 201, Guangdong 518052 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) Title: TEXT CLASSIFICATION METHOD, APPARATUS AND DEVICE, AND COMPUTER-READABLE STORAGE MEDIUM

(54) 发明名称: 文本分类方法、装置、设备及计算机可读存储介质

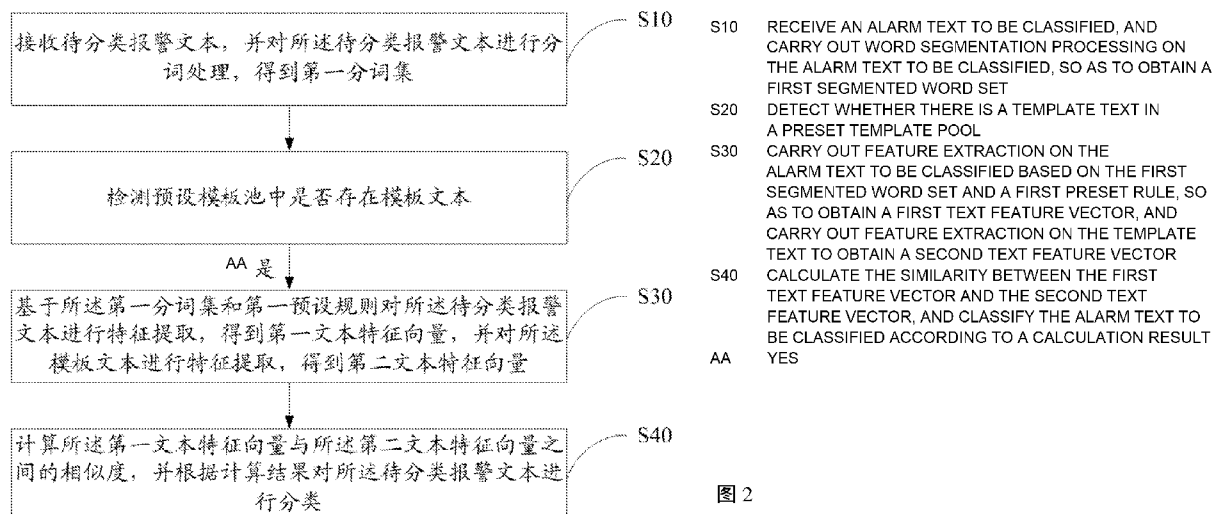


图 2

(57) Abstract: Disclosed are a text classification method, apparatus and device, and a computer-readable storage medium. The text classification method comprises: receiving an alarm text to be classified, and carrying out word segmentation processing on the alarm text to be classified, so as to obtain a first segmented word set (S10); detecting whether there is a template text in a preset template pool (S20); if there is a template text in the preset template pool, carrying out feature extraction on the alarm text to be classified based on the first segmented word set and a first preset rule, so as to obtain a first text feature vector, and carrying out feature extraction on the template text to obtain a second text feature vector (S30); and calculating the similarity between the first text feature vector and the second text feature vector, and classifying the alarm text to be classified according to a calculation result (S40). The method can solve the problem that the accuracy of alarm text classification for an operation and maintenance scenario is relatively low.

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种文本分类方法、装置、设备及计算机可读存储介质, 该文本分类方法包括: 接收待分类报警文本, 并对所述待分类报警文本进行分词处理, 得到第一分词集(S10); 检测预设模板池中是否存在模板文本(S20); 若预设模板池中存在模板文本, 则基于所述第一分词集和第一预设规则对所述待分类报警文本进行特征提取, 得到第一文本特征向量, 并对所述模板文本进行特征提取, 得到第二文本特征向量(S30); 计算第一文本特征向量与第二文本特征向量之间的相似度, 并根据计算结果对所述待分类报警文本进行分类(S40)。上述方法能解决现有针对运维场景的报警文本分类准确性较差的问题。

文本分类方法、装置、设备及计算机可读存储介质

本申请要求于2019年4月12日提交中国专利局、申请号为201910297133.2、
5 发明名称为“文本分类方法、装置、设备及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在申请中。

技术领域

本申请涉及金融科技(Fintech)技术领域，尤其涉及一种文本分类方法、
10 装置、设备及计算机可读存储介质。

背景技术

随着金融科技(Fintech)，尤其是互联网科技金融的不断发展，企业业务
数据中心的规模也日益庞大，在金融科技的日常运维中经常会出现许多报警
15 信息。运维人员需定期对报警文本进行统计分析，以分析报警产生的根本原因，从而确定对应的解决方案。由于金融领域的报警信息的复杂性和庞大性，因此，需通过计算机系统对报警文本进行智能分类挖掘。然而，由于运维场景的报警文本中存在大量专业术语，大多都未被收入到字典中，如果采用现有的文本挖掘方法对运维场景的报警文本进行文本挖掘，会出现报警文本分
20 类准确性较差的问题。因此，亟需一种可提高运维场景中报警文本分类准确性的方法。

发明内容

本申请的主要目的在于提供一种文本分类方法、装置、设备及计算机可
25 读存储介质，旨在解决现有针对运维场景的报警文本分类准确性较差的问题。

为实现上述目的，本申请提供一种文本分类方法，所述文本分类方法包括：

接收待分类报警文本，并对所述待分类报警文本进行分词处理，得到第一分词集；

30 检测预设模板池中是否存在模板文本；

若预设模板池中存在模板文本，则基于所述第一分词集和第一预设规则
对所述待分类报警文本进行特征提取，得到第一文本特征向量，并对所述模板文本进行特征提取，得到第二文本特征向量；

计算所述第一文本特征向量与所述第二文本特征向量之间的相似度，并

根据计算结果对所述待分类报警文本进行分类。

此外，为实现上述目的，本申请还提供一种文本分类装置，所述文本分类装置包括：

5 分词处理模块，用于接收待分类报警文本，并对所述待分类报警文本进行分词处理，得到第一分词集；

模板检测模块，用于检测预设模板池中是否存在模板文本；

10 第一提取模块，用于若预设模板池中存在模板文本，则基于所述第一分词集和第一预设规则对所述待分类报警文本进行特征提取，得到第一文本特征向量，并对所述模板文本进行特征提取，得到第二文本特征向量；

文本分类模块，用于计算所述第一文本特征向量与所述第二文本特征向量之间的相似度，并根据计算结果对所述待分类报警文本进行分类。

15 此外，为实现上述目的，本申请还提供一种文本分类设备，所述文本分类设备包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的文本分类程序，所述文本分类程序被所述处理器执行时实现如上所述的文本分类方法的步骤。

20 此外，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有文本分类程序，所述文本分类程序被处理器执行时实现如上所述的文本分类方法的步骤。

25 本申请提供一种文本分类方法、装置、设备及计算机可读存储介质，通过接收待分类报警文本，并对待分类报警文本进行分词处理，得到第一分词集，并检测预设模板池中是否存在模板文本；若预设模板池中存在模板文本，则基于第一分词集和第一预设规则对待分类报警文本进行特征提取，得到第一文本特征向量，并对所述模板文本进行特征提取，得到第二文本特征向量；计算第一文本特征向量与第二文本特征向量之间的相似度，并根据计算结果对待分类报警文本进行分类。通过上述方式，本申请基于分词处理后的待分
30 类报警文本和预设模板池中的模板文本分别提取得到第一文本特征向量和第二文本特征向量，并通过计算两者之间的相似度，可精确地对待分类报警信息进行分类，即可提高运维场景中报警文本分类的准确性。同时，本申请通过智能且准确地对运维场景中的报警文本进行分类，还可以提高运维工作人员的工作效率。

附图说明

图1为本申请实施例方案涉及的硬件运行环境的设备结构示意图；

图2为本申请文本分类方法第一实施例的流程示意图；

图3为本申请第一实施例中步骤S30的细化流程示意图；

5 图4为本申请文本分类方法第二实施例的流程示意图；

图5为本申请文本分类装置第一实施例的功能模块示意图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

10

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

15 参照图 1 ,图 1 为本申请实施例方案涉及的硬件运行环境的设备结构示意图。

本申请实施例文本分类设备可以是 PC(Personal Computer ,个人计算机),也可以是服务器、平板电脑、便携计算机、智能手机等终端设备。

20 如图 1 所示，该文本分类设备可以包括：处理器 1001，例如 CPU，通信总线 1002，用户接口 1003，网络接口 1004，存储器 1005。其中，通信总线 1002 用于实现这些组件之间的连接通信。用户接口 1003 可以包括显示屏 (Display) 输入单元比如键盘 (Keyboard)，可选用户接口 1003 还可以包括标准的有线接口、无线接口。网络接口 1004 可选的可以包括标准的有线接口、无线接口 (如 Wi-Fi 接口)。存储器 1005 可以是高速 RAM 存储器，也可以是稳定的存储器 (non-volatile memory)，例如磁盘存储器。存储器 1005 可选的还可以是独立于前述处理器 1001 的存储装置。

25 本领域技术人员可以理解，图1中示出的文本分类设备结构并不构成对文本分类设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

30 如图1所示，作为一种计算机存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及文本分类程序。

在图 1 所示的终端中，网络接口 1004 主要用于连接后台服务器，与后台服务器进行数据通信；用户接口 1003 主要用于连接客户端，与客户端进行数据通信；而处理器 1001 可以用于调用存储器 1005 中存储的文本分类程序，并执行以下文本分类方法的各个步骤。

35

基于上述硬件结构，提出本申请文本分类方法的各实施例。

本申请提供一种文本分类方法。

参照图 2，图 2 为本申请文本分类方法第一实施例的流程示意图。

5 在本实施例中，该文本分类方法包括：

步骤 S10，接收待分类报警文本，并对所述待分类报警文本进行分词处理，得到第一分词集；

本实施例的文本分类方法是由文本分类设备实现的，该设备以服务器为例进行说明。在本实施例中，服务器先接收各业务系统（本实施例中，可选为银行机构的业务系统或金融理财机构的业务系统，当然也可以为其他类型的系统）发送的待分类报警文本，然后对待分类报警文本进行分词处理，得到第一分词集。其中，分词处理可以通过分词工具实施，例如汉语词法分析系统 ICTCLAS，中文词法分析程序 THULAC，语言技术平台 LTP 等。分词主要是根据中文语言的特点，将所述样本数据中的每条中文文本切割成一个一个的单词。

步骤 S20，检测预设模板池中是否存在模板文本；

在通过分词处理得到第一分词集后，检测预设模板池中是否存在模板文本，其中，模板文本是指对分类完成的报警文本，在分类得到的每一类中所选取出的一报警文本，用于与后续接收到的待分类报警文本进行比较，以对后续接收到的待分类报警文本进行分类。

若预设模板池中存在模板文本，则执行步骤 S30：基于所述第一分词集和第一预设规则对所述待分类报警文本进行特征提取，得到第一文本特征向量，并对所述模板文本进行特征提取，得到第二文本特征向量；

若检测到预设模板池中存在模板文本，则以待分类报警文本和模板文本为一个集合，进行文本特征向量提取，即：基于第一分词集和第一预设规则对待分类报警文本进行特征提取，得到第一文本特征向量，并对所述模板文本进行特征提取，得到第二文本特征向量。具体的，参照图 3，步骤 S30 包括：

步骤 S31，对所述模板文本进行分词处理，得到第二分词集；

先对模板文本进行分词处理，得到第二分词集，其中，分词处理方法可参照上述对待分类报警文本的分词处理方法，此处不作赘述。可以理解的是，由于模板文本也是经过分词处理、文本特征向量提取和分类等过程后，筛选得到的，因此在将模板文本保存至预设模板池中时，也可以将模板文本对应的分词集与模板文本进行关联保存，从而此时无需重新进行分词处理，可直接获取模板文本所对应的分词集。

35 步骤 S32，基于所述第一分词集和所述第二分词集计算各预设特征词的第

一属性值，并分别计算所述待分类报警文本中各预设特征词的第二属性值和所述模板文本中各预设特征词的第三属性值；

然后，基于第一分词集和第二分词集计算各预设特征词的第一属性值，并分别计算待分类报警文本中各预设特征词的第二属性值和模板文本中各预设特征词的第三属性值。其中，预设特征词是运维工作人员根据实际分类需要设定的，可以包括对报警文本的分类有实质性影响的词汇。第一属性值的计算公式为：

$$F_{s1} = \begin{cases} 0, \log \frac{n_s}{df(t,s)} < \alpha \\ \log \frac{n_s}{df(t,s)}, \log \frac{n_s}{df(t,s)} \geq \alpha \end{cases};$$

第二属性值的计算公式为：

$$F_{s2} = \frac{t_{s1}}{t_{total1}};$$

第三属性值的计算公式为：

$$F_{s3} = \frac{t_{s2}}{t_{total2}};$$

其中， F_{s1} 为预设特征词 s 的第一属性值，该 F_{s1} 反映了预设特征词 s 在所有文本中出现的频率，如果一个词在很多的文本中出现，那么 F_{s1} 值应该较低， n_s 为所述待分类报警文本和所述模板文本的文本总数（即待分类报警文本与模板文本的数量之和）， $df(t,s)$ 为待分类报警文本和模板文本中包含预设特征词 s 的文本数量， α 为一预设值（可根据实际情况进行设定，此处不作限定）， F_{s2} 为待分类报警文本中预设特征词 s 的第二属性值， F_{s2} 表示预设特征词 s 在当前待分类报警文本中的词频， t_{s1} 为预设特征词 s 在当前待分类报警文本中出现的次数， t_{total1} 为当前待分类报警文本中的分词总数， F_{s3} 为模板文本中预设特征词 s 的第三属性值， F_{s3} 表示预设特征词 s 在当前模板文本中的词频， t_{s2} 为预设特征词 s 在当前模板文本中出现的次数， t_{total2} 为当前模板文本中的分词总数。需要说明的是，本申请实施例中，对于第一属性值 F_{s1} 的计算，采用分段计算的方式，在 $\log \frac{n_s}{df(t,s)} < \alpha$ 时，定义 $F_{s1} = 0$ ，在 $\log \frac{n_s}{df(t,s)} \geq \alpha$ 时，定义 $F_{s1} = \log \frac{n_s}{df(t,s)}$ ，通过上述计算方式，可使得各预设特征词的词频（即第二属性值）所对应的权重（即第一属性值）更具有区分度，可进一步提高后续分类的准确性。

步骤 S33，根据所述第一属性值和所述第二属性值计算所述待分类报警文本中各预设特征词的第一特征值，并对所述第一特征值进行拼接，得到第一文本特征向量；

步骤 S34，根据所述第一属性值和所述第三属性值计算所述模板文本中各预设特征词的第二特征值，并对所述第二特征值进行拼接，得到第二文本特征向量。

在得到第一属性值、第二属性值和第三属性值后，根据第一属性值和第二属性值计算待分类报警文本中各预设特征词的第一特征值，并对第一特征

值进行拼接，得到第一文本特征向量，同时，根据第一属性值和第三属性值计算模板文本中各预设特征词的第二特征值，并对第二特征值进行拼接，得到第二文本特征向量。需要说明的是，步骤 S33 和 S34 的执行顺序不分先后。

其中，第一特征值的计算公式为： $V_{s1}=F_{s1} \cdot F_{s2}$ ，第二特征值的计算公式为：

5 $V_{s2}=F_{s1} \cdot F_{s3}$ 。 V_{s1} 为待分类报警文本中预设特征词 s 的第一特征值， V_{s2} 为模板文本中预设特征词 s 的第二特征值， F_{s1} 为预设特征词 s 的第一属性值， F_{s2} 为待分类报警文本中预设特征词 s 的第二属性值， F_{s3} 为模板文本中预设特征词 s 的第三属性值。其中，第一文本特征向量的获取方式为：对待分类报警文本中各预设特征词的第一特征值进行拼接，例如，假设预设特征词有 y 个，计算得到预设特征词 1-y 的第一特征值分别为 V_{11} 、 V_{21} 、... V_{s1} 、... V_{y1} ，可拼接得到第一文本特征向量为 $\{V_{11}, V_{21}, \dots, V_{s1}, \dots, V_{y1}\}$ 。第二文本特征向量与第一文本特征向量的获取方式相似，此处不再赘述。

步骤 S40，计算所述第一文本特征向量与所述第二文本特征向量之间的相似度，并根据计算结果对所述待分类报警文本进行分类。

15 在得到第一文本特征向量和第二文本特征向量之后，计算第一文本特征向量与第二文本特征向量之间的相似度，并根据计算结果对待分类报警文本进行分类。具体的，步骤 S40 可以包括：

步骤 a1，计算所述第一文本特征向量与各所述第二文本特征向量之间的欧氏距离，并根据计算结果判断是否存在大于预设阈值的欧式距离；

20 作为其中一种相似度的计算方式，可通过计算欧式距离来表征相似度。具体的，计算第一文本特征向量与各第二文本特征向量之间的欧氏距离，并根据计算结果判断是否存在大于预设阈值的欧式距离。其中，欧氏距离又称为欧几里得距离或欧几里得度量，是指欧几里得空间中两点之间的直线距离。当第一文本特征向量为 $\{V_{11}, V_{21}, \dots, V_{s1}, \dots, V_{y1}\}$ ，第二文本特征向量为 $\{V_{12}, V_{22}, \dots, V_{s2}, \dots, V_{y2}\}$ 时，可计算得到两者之间的欧氏距离为：

$$d = \sqrt{(V_{11} - V_{12})^2 + (V_{21} - V_{22})^2 + \dots + (V_{s1} - V_{s2})^2 + \dots + (V_{y1} - V_{y2})^2}。$$

当然，在具体实施例中，还可以通过计算第一文本特征向量与各第二文本特征向量之间的余弦相似度、或杰卡德 Jaccard 距离、或编辑距离，来表征两者之间的相似度。

30 步骤 a2，若存在大于预设阈值的欧式距离，则将所述待分类报警文本与所述大于预设阈值的欧式距离对应的模板文本分为同一类；

若计算结果中存在大于预设阈值的欧式距离，则说明模板文本对应的第二文本特征向量中存在与待分类报警文本对应的第一文本特征向量较为相似的特征向量，即模板文本中存在一与待分类报警文本相似的文本，此时，可

将两者划分为同一类，即将待分类报警文本与该大于预设阈值的欧式距离所对应的模板文本分为同一类。此外，该预设阈值可根据实际需要进行设定，此处不作限定。

步骤 a3，若不存在大于预设阈值的欧式距离，则将所述待分类报警文本另分为一个新类，并将所述待分类报警文本保存至所述预设模板池中，作为一新模板文本。

若计算结果中不存在大于预设阈值的欧式距离，则说明该待分类报警文本与模板文本中的任一文本均不相似，此时，将待分类报警文本另分为一个新类，并将待分类报警文本保存至预设模板池中，作为一新模板文本，以用于对后续接收到的报警文本进行分类。

本申请实施例提供一种文本分类方法，通过接收待分类报警文本，并对待分类报警文本进行分词处理，得到第一分词集，并检测预设模板池中是否存在模板文本；若预设模板池中存在模板文本，则基于第一分词集和第一预设规则对待分类报警文本进行特征提取，得到第一文本特征向量，并对所述模板文本进行特征提取，得到第二文本特征向量；计算第一文本特征向量与第二文本特征向量之间的相似度，并根据计算结果对待分类报警文本进行分类。通过上述方式，本申请实施例基于分词处理后的待分类报警文本和预设模板池中的模板文本分别提取得到第一文本特征向量和第二文本特征向量，并通过计算两者之间的相似度，可精确地对待分类报警信息进行分类，即可提高运维场景中报警文本分类的准确性。同时，本申请通过智能且准确地对运维场景中的报警文本进行分类，还可以提高运维工作人员的工作效率。

为进一步提高运维场景中报警文本分类的准确性，在上述步骤“对所述待分类报警文本进行分词处理，得到第一分词集”之前，该文本分类方法还可以包括以下步骤：

对所述待分类报警文本进行词汇替换处理；

此时，步骤“对所述待分类报警文本进行分词处理，得到第一分词集”包括：对经词汇替换处理的待分类报警文本进行分词处理，得到第一分词集。

在本实施例中，由于报警文本一般较为简短，各词汇在整个报警文本中所占的权重较大，为避免对整体语义判断扰动比较大的词汇影响后续特征值的计算结果及分类的准确性，需对同性质非标准的词汇进行预处理，将含有类似信息的词汇替换成相应的词汇标签或类名，从而降低词汇对语义判断的扰动，进而提高文本分类的准确性。本实施例中，通过对待分类报警文本进行词汇替换处理，以提高文本分类的准确性。具体的，步骤“对所述待分类报警文本进行词汇替换处理”包括：

步骤 b1，接收待分类报警文本，检测所述待分类报警文本中是否存在预设目标词汇；

先接收待分类报警文本，并检测待分类报警文本中是否存在预设目标词汇。

5 步骤 b2，若所述待分类报警文本中存在预设目标词汇，则将所述待分类报警文本中所存在的预设目标词汇替换成对应的词汇标签。

若待分类报警文本中存在预设目标词汇，则将待分类报警文本中所存在的预设目标词汇替换成对应的词汇标签。具体的，可预先构建预设目标词汇与词汇标签之间的映射关系，在检测到待分类报警文本中存在预设目标词汇时，根据该映射关系确定对应的词汇标签，并进行替换。例如，对于报警文
10 本 1：XXX 系统的 CPU 占用率达到 98.7%，报警文本 2：XXX 系统的 CPU 占用率达到 90.1%，其中报警文本 1 和 2 中的具体数值，在整个报警文本中所占的权重较大，但对报警文本的分类无实质性的作用，因此，可将数值类的词汇替换成词汇标签<num>。此外，由于 XXX 系统命名差异巨大，同一（类）
15 系统可能有不同的系统命名，因此可以设置将其进行替代成词汇标签<SUBSYS>。当然，上述例子仅作举例，并不用于限定本申请，本申请中对待分类报警文本进行预处理，是为了通过将同性质非标准的词汇替换为同一词汇标签，以助于待分类报警文本中纯粹的对后续分类有实质影响的信息被提取，降低同类词汇对语义判断的扰动，从而提高文本分类的准确性。

20 在对待分类报警文本进行词汇替换之后，然后对经词汇替换处理的待分类报警文本进行分词处理，得到第一分词集，进而执行后续步骤，具体过程可参照上述第一实施例，此处不作赘述。需要说明的是，在后续基于第一分词集和第二分词集计算各预设特征词的第一属性值，并分别计算待分类报警文本中各预设特征词的第二属性值和模板文本中各预设特征词的第三属性值
25 时，预设特征词除可包括上述所说的对报警文本的分类有实质性影响的词汇外，还可以包括上述词汇替换后的词汇标签。

本申请实施例中通过对待分类报警文本进行词汇替换处理，将同性质非标准的词汇替换为同一词汇标签，降低同类词汇对语义判断的扰动，可有助于待分类报警文本中对分类有实质影响的特征信息被提取，从而可进一步提
30 高报警文本分类的准确性。

进一步的，参照图 4，图 4 为本申请文本分类方法第二实施例的流程示意图。

35 基于图 2 所示的第一实施例，在步骤 S20 之后，该文本分类方法还包括：若预设模板池中不存在模板文本，则执行步骤 S50：基于所述第一分词集

和第二预设规则对所述待分类报警文本进行特征提取，得到第三文本特征向量；

在本实施例中，在检测预设模板池中是否存在模板文本之后，若该预设模板池中不存在模板文本，此时，则以待分类报警文本为一集合，对其进行聚类，进而分类。具体的，先基于第一分词集和第二预设规则对待分类报警文本进行特征，得到第三文本特征向量。步骤 S50 包括：

步骤 c1，基于所述第一分词集计算各预设特征词的第四属性值和第五属性值；

先基于第一分词集计算各预设特征词的第四属性值和第五属性值，其中，第四属性值的计算公式为：

$$F_{s4} = \begin{cases} 0, \log \frac{n_s'}{df(t,s)'} < \alpha \\ \log \frac{n_s'}{df(t,s)'}, \log \frac{n_s'}{df(t,s)'} \geq \alpha \end{cases};$$

第五属性值的计算公式为：

$$F_{s5} = \frac{t_{s3}}{t_{total3}};$$

其中， F_{s4} 为预设特征词 s 的第四属性值， n_s' 为所述待分类报警文本的文本总数， $df(t,s)'$ 为待分类报警文本中包含预设特征词 s 的文本数量， α 为一预设值（可根据实际情况进行设定，此处不作限定）， F_{s5} 为待分类报警文本中预设特征词 s 的第五属性值， F_{s5} 表示预设特征词 s 在当前待分类报警文本中的词频， t_{s3} 为预设特征词 s 在当前待分类报警文本中出现的次数， t_{total3} 为当前待分类报警文本中的分词总数。

步骤 c2，根据所述第四属性值和所述第五属性值计算所述待分类报警文本中各预设特征词的第三特征值，并对所述第三特征值进行拼接，得到第三文本特征向量。

然后，根据第四属性值和第五属性值计算待分类报警文本中各预设特征词的第三特征值，并对第三特征值进行拼接，得到第三文本特征向量。其中，第三特征值的计算公式为： $V_{s3} = F_{s4} \cdot F_{s5}$ ， V_{s3} 为待分类报警文本中预设特征词 s 的第三特征值， F_{s4} 为预设特征词 s 的第四属性值， F_{s5} 为待分类报警文本中预设特征词 s 的第五属性值。第三文本特征向量与第一文本特征向量的获取方式相似，可参照上述第一实施例，此处不再赘述。

步骤 S60，对所述第三文本特征向量进行聚类，并根据聚类结果对所述待分类报警文本进行分类；

在得到第三文本特征向量之后，对第三文本特征向量进行聚类，并根据聚类结果对待分类报警文本进行分类。其中，聚类和分类方式可以包括但不限于：1) 计算各第三文本特征向量之间的欧式距离，将欧氏距离大于预设阈

值的第三文本特征向量所对应的待分类报警文本分类同一类；2) 计算各第三文本特征向量之间的余弦相似度(或 Jaccard 距离等其他可表征相似度的值)，根据计算结果对待分类报警文本进行分类；3) 通过预设聚类算法(如 K-Means(K 均值)聚类、层次聚类算法等)对第三文本特征向量进行聚类，根据聚类结果对所述待分类报警文本进行分类，例如，聚为 n 类，则将聚成的每一类作为一分类，即得到 n 个类型的报警文本。

步骤 S70，根据分类结果从每一分类中随机选取一待分类报警文本，作为模板文本，并保存至所述预设模板池中。

然后，根据分类结果从每一分类中随机选取一待分类报警文本，作为模板文本，并保存至预设模板池中，以便于对后续接收到的报警文本进行分类。

本申请实施例中介绍了预设模板池中不存在模板文本时，待分类报警文本的分类方法。通过对待分类报警文本进行特征提取，进而对提取得到的第三文本特征向量进行聚类，并根据聚类结果对待分类报警信息进行分类，可提高运维场景中报警文本分类的准确性。同时，本申请通过智能且准确地对运维场景中的报警文本进行分类，还可以提高运维工作人员的工作效率。

需要说明的是，在本实施例中，在分词处理之前，可先对待分类报警文本进行词汇替换处理，从而通过词汇替换处理，将同性质非标准的词汇替换为同一词汇标签，降低同类词汇对语义判断的扰动，可有助于待分类报警文本中对分类有实质影响的特征信息被提取，以进一步提高报警文本分类的准确性。其中，具体的词汇替换处理过程可参照上述实施例，此处不作赘述。

本申请还提供一种文本分类装置。

参照图 5，图 5 为本申请文本分类装置第一实施例的功能模块示意图。

所述文本分类装置包括：

分词处理模块 10，用于接收待分类报警文本，并对所述待分类报警文本进行分词处理，得到第一分词集；

模板检测模块 20，用于检测预设模板池中是否存在模板文本；

第一提取模块 30，用于若预设模板池中存在模板文本，则基于所述第一分词集和第一预设规则对所述待分类报警文本进行特征提取，得到第一文本特征向量，并对所述模板文本进行特征提取，得到第二文本特征向量；

文本分类模块 40，用于计算所述第一文本特征向量与所述第二文本特征向量之间的相似度，并根据计算结果对所述待分类报警文本进行分类。

进一步的，所述第一提取模块 30 包括：

分词处理单元，用于对所述模板文本进行分词处理，得到第二分词集；

第一计算单元，用于基于所述第一分词集和所述第二分词集计算各预设

特征词的第一属性值，并分别计算所述待分类报警文本中各预设特征词的第二属性值和所述模板文本中各预设特征词第三属性值；

第一拼接单元，用于根据所述第一属性值和所述第二属性值计算所述待分类报警文本中各预设特征词的第一特征值，并对所述第一特征值进行拼接，得到第一文本特征向量；

第二拼接单元，用于根据所述第一属性值和所述第三属性值计算所述模板文本中各预设特征词的第二特征值，并对所述第二特征值进行拼接，得到第二文本特征向量。

进一步的，所述第一属性值的计算公式为：

$$F_{s1} = \begin{cases} 0, \log \frac{n_s}{df(t,s)} < \alpha \\ \log \frac{n_s}{df(t,s)}, \log \frac{n_s}{df(t,s)} \geq \alpha \end{cases};$$

所述第二属性值的计算公式为：

$$F_{s2} = \frac{t_{s1}}{t_{total1}};$$

所述第三属性值的计算公式为：

$$F_{s3} = \frac{t_{s2}}{t_{total2}};$$

所述第一特征值的计算公式为：

$$V_{s1} = F_{s1} \cdot F_{s2};$$

所述第二特征值的计算公式为：

$$V_{s2} = F_{s1} \cdot F_{s3};$$

其中， F_{s1} 为预设特征词 s 的第一属性值， n_s 为所述待分类报警文本和所述模板文本的文本总数， $df(t,s)$ 为所述待分类报警文本和所述模板文本中包含预设特征词 s 的文本数量， α 为一预设值， F_{s2} 为所述待分类报警文本中预设特征词 s 的第二属性值， t_{s1} 为预设特征词 s 在当前待分类报警文本中出现的次数， t_{total1} 为当前待分类报警文本中的分词总数， F_{s3} 为所述模板文本中预设特征词 s 的第三属性值， t_{s2} 为预设特征词 s 在当前模板文本中出现的次数， t_{total2} 为当前模板文本中的分词总数， V_{s1} 为所述待分类报警文本中预设特征词 s 的第一特征值， V_{s2} 为所述模板文本中预设特征词 s 的第二特征值。

进一步的，所述文本分类模块 40 包括：

第二计算单元，用于计算所述第一文本特征向量与各所述第二文本特征向量之间的欧氏距离，并根据计算结果判断是否存在大于预设阈值的欧式距离；

第一分类单元，用于若存在大于预设阈值的欧式距离，则将所述待分类报警文本与所述大于预设阈值的欧式距离对应的模板文本分为同一类；

第二分类单元，用于若不存在大于预设阈值的欧式距离，则将所述待分

类报警文本另分为一个新类，并将所述待分类报警文本保存至所述预设模板池中，作为一新模板文本。

进一步的，所述文本分类装置还包括：

词汇替换模块，用于对所述待分类报警文本进行词汇替换处理；

5 其中，所述词汇替换模块包括：

文本检测单元，用于检测所述待分类报警文本中是否存在预设目标词汇；

词汇替换单元，用于若所述待分类报警文本中存在预设目标词汇，则将所述待分类报警文本中所存在的预设目标词汇替换成对应的词汇标签；

10 所述分词处理模块 10，具体用于对经词汇替换处理的待分类报警文本进行分词处理，得到第一分词集。

进一步的，所述文本分类装置还包括：

第二提取模块，用于若预设模板池中不存在模板文本，则基于所述第一分词集和第二预设规则对所述待分类报警文本进行特征提取，得到第三文本特征向量；

15 向量聚类模块，用于对所述第三文本特征向量进行聚类，并根据聚类结果对所述待分类报警文本进行分类；

模板选取模块，用于根据分类结果从每一分类中随机选取一待分类报警文本，作为模板文本，并保存至所述预设模板池中。

进一步的，所述第二提取模块包括：

20 第三计算单元，用于基于所述第一分词集计算各预设特征词的第四属性值和第五属性值；

第三拼接单元，用于根据所述第四属性值和所述第五属性值计算所述待分类报警文本中各预设特征词的第三特征值，并对所述第三特征值进行拼接，得到第三文本特征向量。

25 其中，上述文本分类装置中各个模块的功能实现与上述文本分类方法实施例中各步骤相对应，其功能和实现过程在此处不再一一赘述。

30 本申请还提供一种计算机可读存储介质，该计算机可读存储介质上存储有文本分类程序，所述文本分类程序被处理器执行时实现如以上任一项实施例所述的文本分类方法的步骤。

本申请计算机可读存储介质的具体实施例与上述文本分类方法各实施例基本相同，在此不作赘述。

35 需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者系

统不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

5 上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个存储介质(如 ROM/RAM、
10 磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，空调器，或者网络设备等)执行本申请各个实施例所述的方法。

 以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是
15 利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1、一种文本分类方法，其中，所述文本分类方法包括：

接收待分类报警文本，并对所述待分类报警文本进行分词处理，得到第
5 一分词集；

检测预设模板池中是否存在模板文本；

若预设模板池中存在模板文本，则基于所述第一分词集和第一预设规则
对所述待分类报警文本进行特征提取，得到第一文本特征向量，并对所述模
板文本进行特征提取，得到第二文本特征向量；

10 计算所述第一文本特征向量与所述第二文本特征向量之间的相似度，并
根据计算结果对所述待分类报警文本进行分类。

2、如权利要求 1 所述的文本分类方法，其中，所述基于所述第一分词集
和第一预设规则对所述待分类报警文本进行特征提取，得到第一文本特征向
15 量，并对所述模板文本进行特征提取，得到第二文本特征向量的步骤包括：

对所述模板文本进行分词处理，得到第二分词集；

基于所述第一分词集和所述第二分词集计算各预设特征词的第一属性值，
并分别计算所述待分类报警文本中各预设特征词的第二属性值和所述模板文
本中各预设特征词的第三属性值；

20 根据所述第一属性值和所述第二属性值计算所述待分类报警文本中各预
设特征词的第一特征值，并对所述第一特征值进行拼接，得到第一文本特征
向量；

根据所述第一属性值和所述第三属性值计算所述模板文本中各预设特征
词的第二特征值，并对所述第二特征值进行拼接，得到第二文本特征向量。

25 3、如权利要求 2 所述的文本分类方法，其中，所述第一属性值的计算公
式为：

$$F_{s1} = \begin{cases} 0, \log \frac{n_s}{df(t,s)} < \alpha \\ \log \frac{n_s}{df(t,s)}, \log \frac{n_s}{df(t,s)} \geq \alpha \end{cases};$$

所述第二属性值的计算公式为：

$$F_{s2} = \frac{t_{s1}}{t_{total1}};$$

所述第三属性值的计算公式为：

$$F_{s3} = \frac{t_{s2}}{t_{total2}};$$

所述第一特征值的计算公式为：

$$V_{s1}=F_{s1} \cdot F_{s2} ;$$

所述第二特征值的计算公式为：

$$V_{s2}=F_{s1} \cdot F_{s3} ;$$

其中， F_{s1} 为预设特征词 s 的第一属性值， n_s 为所述待分类报警文本和所述模板文本的文本总数， $df(t,s)$ 为所述待分类报警文本和所述模板文本中包含预设特征词 s 的文本数量， α 为一预设值， F_{s2} 为所述待分类报警文本中预设特征词 s 的第二属性值， t_{s1} 为预设特征词 s 在当前待分类报警文本中出现的次数， t_{total1} 为当前待分类报警文本中的分词总数， F_{s3} 为所述模板文本中预设特征词 s 的第三属性值， t_{s2} 为预设特征词 s 在当前模板文本中出现的次数， t_{total2} 为当前模板文本中的分词总数， V_{s1} 为所述待分类报警文本中预设特征词 s 的第一特征值， V_{s2} 为所述模板文本中预设特征词 s 的第二特征值。

4、如权利要求 1 所述的文本分类方法，其中，所述计算所述第一文本特征向量与所述第二文本特征向量之间的相似度，并根据计算结果对所述待分类报警文本进行分类的步骤包括：

计算所述第一文本特征向量与各所述第二文本特征向量之间的欧氏距离，并根据计算结果判断是否存在大于预设阈值的欧式距离；

若存在大于预设阈值的欧式距离，则将所述待分类报警文本与所述大于预设阈值的欧式距离对应的模板文本分为同一类；

若不存在大于预设阈值的欧式距离，则将所述待分类报警文本另分为一个新类，并将所述待分类报警文本保存至所述预设模板池中，作为一新模板文本。

5、如权利要求 1 所述的文本分类方法，其中，所述对所述待分类报警文本进行分词处理，得到第一分词集的步骤之前，还包括：

对所述待分类报警文本进行词汇替换处理；其中，所述对所述待分类报警文本进行词汇替换处理的步骤包括：

检测所述待分类报警文本中是否存在预设目标词汇；

若所述待分类报警文本中存在预设目标词汇，则将所述待分类报警文本中所存在的预设目标词汇替换成对应的词汇标签；

所述对所述待分类报警文本进行分词处理，得到第一分词集的步骤包括：对经词汇替换处理的待分类报警文本进行分词处理，得到第一分词集。

6、如权利要求 1 所述的文本分类方法，其中，所述检测预设模板池中是否存在模板文本的步骤之后，还包括：

若预设模板池中不存在模板文本，则基于所述第一分词集和第二预设规则对所述待分类报警文本进行特征提取，得到第三文本特征向量；

对所述第三文本特征向量进行聚类，并根据聚类结果对所述待分类报警文本进行分类；

5 根据分类结果从每一分类中随机选取一待分类报警文本，作为模板文本，并保存至所述预设模板池中。

10 7、如权利要求6所述的文本分类方法，其中，所述基于所述第一分词集和第二预设规则对所述待分类报警文本进行特征提取，得到第三文本特征向量的步骤包括：

基于所述第一分词集计算各预设特征词的第四属性值和第五属性值；

根据所述第四属性值和所述第五属性值计算所述待分类报警文本中各预设特征词的第三特征值，并对所述第三特征值进行拼接，得到第三文本特征向量。

15

8、一种文本分类装置，其中，所述文本分类装置包括：

分词处理模块，用于接收待分类报警文本，并对所述待分类报警文本进行分词处理，得到第一分词集；

模板检测模块，用于检测预设模板池中是否存在模板文本；

20

第一提取模块，用于若预设模板池中存在模板文本，则基于所述第一分词集和第一预设规则对所述待分类报警文本进行特征提取，得到第一文本特征向量，并对所述模板文本进行特征提取，得到第二文本特征向量；

文本分类模块，用于计算所述第一文本特征向量与所述第二文本特征向量之间的相似度，并根据计算结果对所述待分类报警文本进行分类。

25

9、如权利要求8所述的文本分类装置，其中，所述第一提取模块包括：

分词处理单元，用于对所述模板文本进行分词处理，得到第二分词集；

30

第一计算单元，用于基于所述第一分词集和所述第二分词集计算各预设特征词的第一属性值，并分别计算所述待分类报警文本中各预设特征词的第二属性值和所述模板文本中各预设特征词的第三属性值；

第一拼接单元，用于根据所述第一属性值和所述第二属性值计算所述待分类报警文本中各预设特征词的第一特征值，并对所述第一特征值进行拼接，得到第一文本特征向量；

35

第二拼接单元，用于根据所述第一属性值和所述第三属性值计算所述模板文本中各预设特征词的第二特征值，并对所述第二特征值进行拼接，得到

第二文本特征向量。

10、如权利要求 9 所述的文本分类装置，其中，所述第一属性值的计算公式为：

$$F_{s1} = \begin{cases} 0, \log \frac{n_s}{df(t,s)} < \alpha \\ \log \frac{n_s}{df(t,s)}, \log \frac{n_s}{df(t,s)} \geq \alpha \end{cases};$$

所述第二属性值的计算公式为：

$$F_{s2} = \frac{t_{s1}}{t_{total1}};$$

所述第三属性值的计算公式为：

$$F_{s3} = \frac{t_{s2}}{t_{total2}};$$

所述第一特征值的计算公式为：

$$V_{s1} = F_{s1} \cdot F_{s2};$$

所述第二特征值的计算公式为：

$$V_{s2} = F_{s1} \cdot F_{s3};$$

其中， F_{s1} 为预设特征词 s 的第一属性值， n_s 为所述待分类报警文本和所述模板文本的文本总数， $df(t,s)$ 为所述待分类报警文本和所述模板文本中包含预设特征词 s 的文本数量， α 为一预设值， F_{s2} 为所述待分类报警文本中预设特征词 s 的第二属性值， t_{s1} 为预设特征词 s 在当前待分类报警文本中出现的次数， t_{total1} 为当前待分类报警文本中的分词总数， F_{s3} 为所述模板文本中预设特征词 s 的第三属性值， t_{s2} 为预设特征词 s 在当前模板文本中出现的次数， t_{total2} 为当前模板文本中的分词总数， V_{s1} 为所述待分类报警文本中预设特征词 s 的第一特征值， V_{s2} 为所述模板文本中预设特征词 s 的第二特征值。

11、如权利要求 8 所述的文本分类装置，其中，所述文本分类模块包括：

第二计算单元，用于计算所述第一文本特征向量与各所述第二文本特征向量之间的欧氏距离，并根据计算结果判断是否存在大于预设阈值的欧式距离；

第一分类单元，用于若存在大于预设阈值的欧式距离，则将所述待分类报警文本与所述大于预设阈值的欧式距离对应的模板文本分为同一类；

第二分类单元，用于若不存在大于预设阈值的欧式距离，则将所述待分类报警文本另分为一个新类，并将所述待分类报警文本保存至所述预设模板池中，作为一新模板文本。

12、如权利要求 8 所述的文本分类装置，其中，所述文本分类装置还包

括：

词汇替换模块，用于对所述待分类报警文本进行词汇替换处理；

其中，所述词汇替换模块包括：

文本检测单元，用于检测所述待分类报警文本中是否存在预设目标词汇；

5 词汇替换单元，用于若所述待分类报警文本中存在预设目标词汇，则将所述待分类报警文本中所存在的预设目标词汇替换成对应的词汇标签；

所述分词处理模块，具体用于对经词汇替换处理的待分类报警文本进行分词处理，得到第一分词集。

10 13、如权利要求 8 所述的文本分类装置，其中，所述文本分类装置还包括：

第二提取模块，用于若预设模板池中不存在模板文本，则基于所述第一分词集和第二预设规则对所述待分类报警文本进行特征提取，得到第三文本特征向量；

15 向量聚类模块，用于对所述第三文本特征向量进行聚类，并根据聚类结果对所述待分类报警文本进行分类；

模板选取模块，用于根据分类结果从每一分类中随机选取一待分类报警文本，作为模板文本，并保存至所述预设模板池中。

20 14、如权利要求 13 所述的文本分类装置，其中，所述第二提取模块包括：

第三计算单元，用于基于所述第一分词集计算各预设特征词的第四属性值和第五属性值；

25 第三拼接单元，用于根据所述第四属性值和所述第五属性值计算所述待分类报警文本中各预设特征词的第三特征值，并对所述第三特征值进行拼接，得到第三文本特征向量。

15、一种文本分类设备，其中，所述文本分类设备包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的文本分类程序，所述文本分类程序被所述处理器执行时实现如下步骤：

30 获取待见证的数据，计算所述待见证的数据的哈希值并使用预先保存的见证发起方私钥对所述哈希值进行签名，根据经所述见证发起方私钥签名后的所述哈希值和预设的多方见证完成条件信息生成见证请求消息；

将所述待见证的数据和所述见证请求消息发送至预设的见证参与方节点，并接收各见证参与方节点返回的见证响应消息；

35 从所述见证响应消息中获取各见证参与方节点使用自身私钥对所述哈希

值进行签名的签名信息，形成签名列表；

发起基于所述签名列表的链上共识；

当所述链上共识达成时，判定所述待见证的数据完成了多方见证。

5 16、如权利要求 15 所述的文本分类设备，其中，所述文本分类程序被所述处理器执行时还实现如下步骤：

 对所述模板文本进行分词处理，得到第二分词集；

 基于所述第一分词集和所述第二分词集计算各预设特征词的第一属性值，并分别计算所述待分类报警文本中各预设特征词的第二属性值和所述模板文本中各预设特征词的第三属性值；

10 根据所述第一属性值和所述第二属性值计算所述待分类报警文本中各预设特征词的第一特征值，并对所述第一特征值进行拼接，得到第一文本特征向量；

 根据所述第一属性值和所述第三属性值计算所述模板文本中各预设特征词的第二特征值，并对所述第二特征值进行拼接，得到第二文本特征向量。

 17、如权利要求 15 所述的文本分类设备，其中，所述文本分类程序被所述处理器执行时还实现如下步骤：

20 计算所述第一文本特征向量与各所述第二文本特征向量之间的欧氏距离，并根据计算结果判断是否存在大于预设阈值的欧式距离；

 若存在大于预设阈值的欧式距离，则将所述待分类报警文本与所述大于预设阈值的欧式距离对应的模板文本分为同一类；

25 若不存在大于预设阈值的欧式距离，则将所述待分类报警文本另分为一个新类，并将所述待分类报警文本保存至所述预设模板池中，作为一新模板文本。

 18、一种计算机可读存储介质，其中，所述计算机可读存储介质上存储有文本分类程序，所述文本分类程序被处理器执行时实现如下步骤：

30 获取待见证的数据，计算所述待见证的数据的哈希值并使用预先保存的见证发起方私钥对所述哈希值进行签名，根据经所述见证发起方私钥签名后的所述哈希值和预设的多方见证完成条件信息生成见证请求消息；

 将所述待见证的数据和所述见证请求消息发送至预设的见证参与方节点，并接收各见证参与方节点返回的见证响应消息；

35 从所述见证响应消息中获取各见证参与方节点使用自身私钥对所述哈希值进行签名的签名信息，形成签名列表；

发起基于所述签名列表的链上共识；
当所述链上共识达成时，判定所述待见证的数据完成了多方见证。

19、如权利要求 18 所述的计算机可读存储介质，其中，所述文本分类程序被处理器执行时还实现如下步骤：

对所述模板文本进行分词处理，得到第二分词集；

基于所述第一分词集和所述第二分词集计算各预设特征词的第一属性值，并分别计算所述待分类报警文本中各预设特征词的第二属性值和所述模板文本中各预设特征词的第三属性值；

根据所述第一属性值和所述第二属性值计算所述待分类报警文本中各预设特征词的第一特征值，并对所述第一特征值进行拼接，得到第一文本特征向量；

根据所述第一属性值和所述第三属性值计算所述模板文本中各预设特征词的第二特征值，并对所述第二特征值进行拼接，得到第二文本特征向量。

20、如权利要求 18 所述的计算机可读存储介质，其中，所述文本分类程序被处理器执行时还实现如下步骤：

计算所述第一文本特征向量与各所述第二文本特征向量之间的欧氏距离，并根据计算结果判断是否存在大于预设阈值的欧式距离；

若存在大于预设阈值的欧式距离，则将所述待分类报警文本与所述大于预设阈值的欧式距离对应的模板文本分为同一类；

若不存在大于预设阈值的欧式距离，则将所述待分类报警文本另分为一个新类，并将所述待分类报警文本保存至所述预设模板池中，作为一新模板文本。

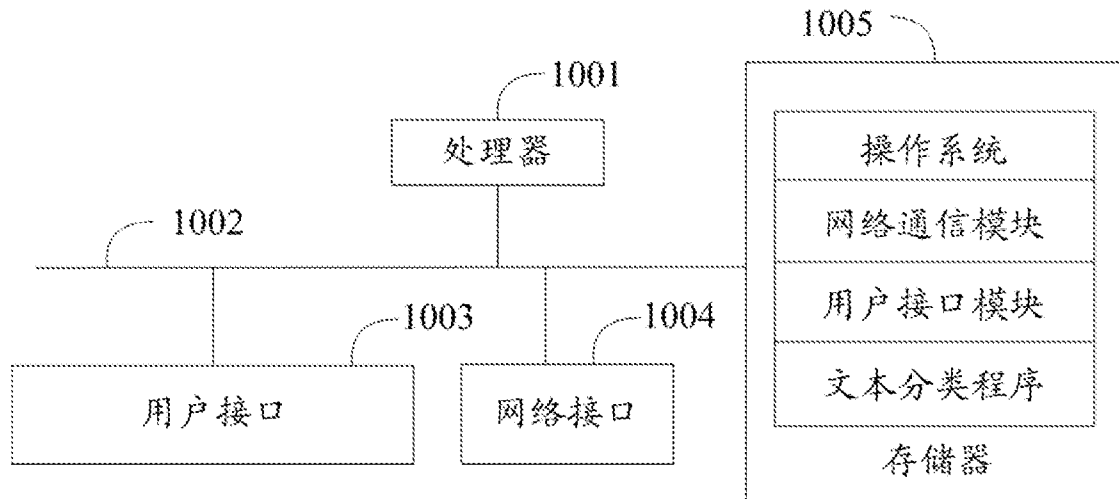


图 1

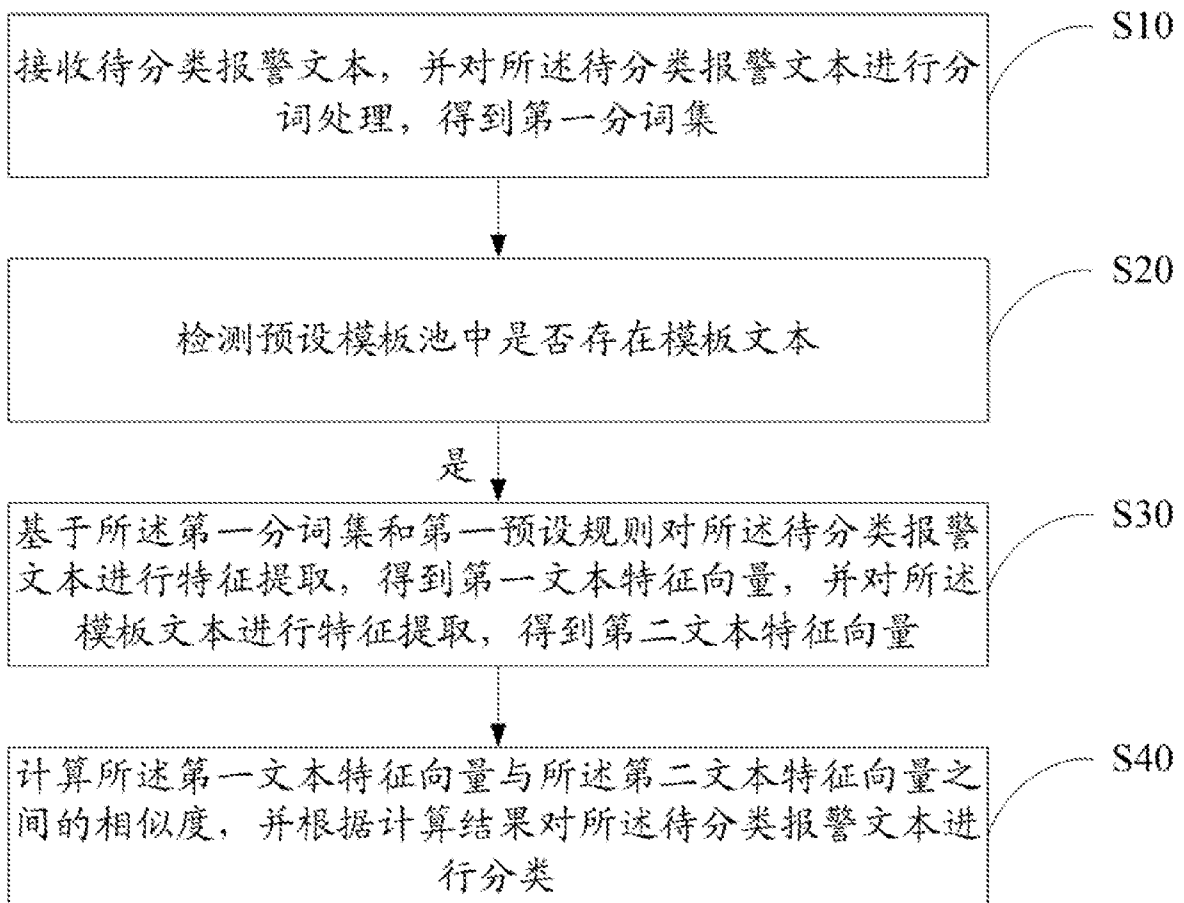


图 2

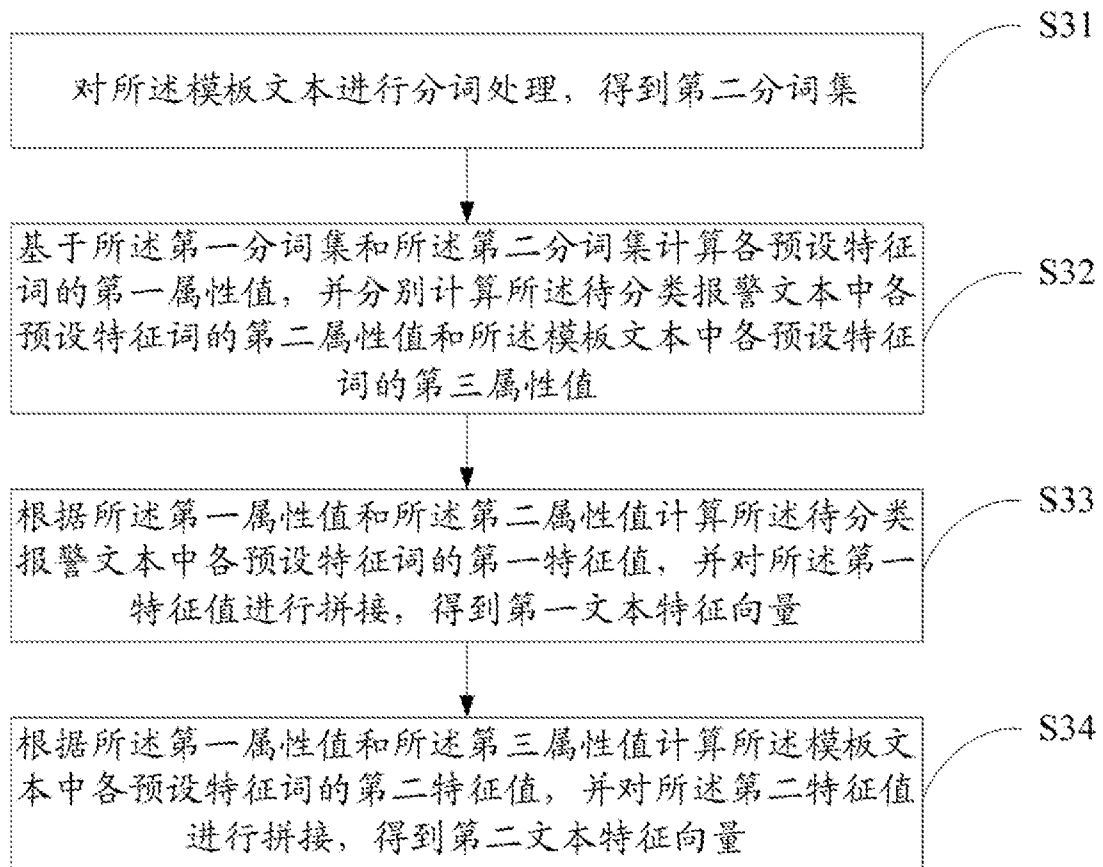


图 3

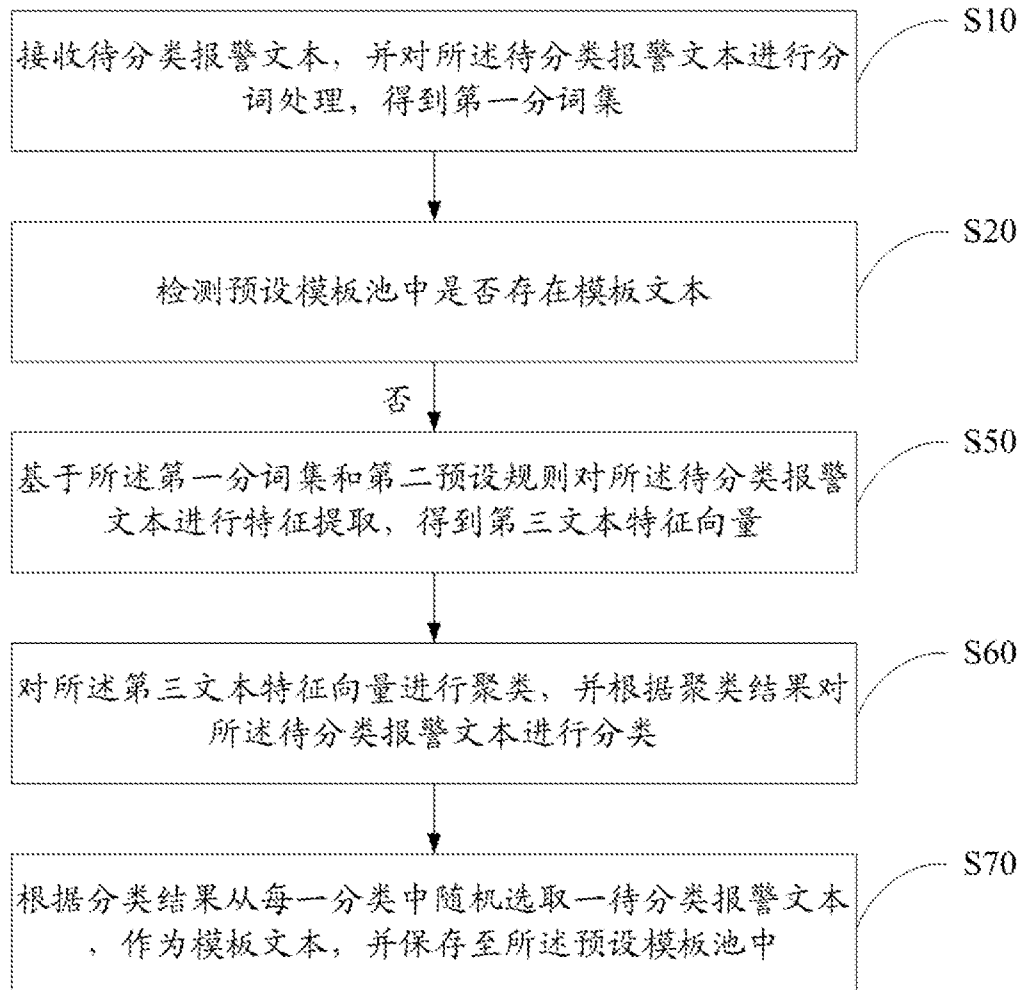


图 4

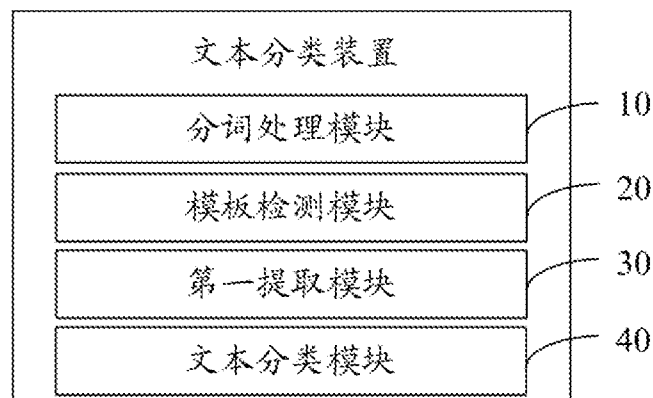


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/078389

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/35(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, EPODOC, WPI, CNKI, ISI: 文本, 文档, 分类, 类别, 分词, 词频, 次数, 频率, 词, 相似度, 相似性, 特征, 值, 向量,
document, classify, word, frequency, similar, feature, vector**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110008343 A (SHENZHEN QIANHAI WEIZHONG BANK CO., LTD.) 12 July 2019 (2019-07-12) claims 1-10	1-14
X	CN 105045812 A (SHANGHAI GAOXIN COMPUTER SYSTEMS CO., LTD.) 11 November 2015 (2015-11-11) description, paragraphs [0054]-[0088]	1-3, 5, 8-10, 12, 15, 16, 18, and 19
Y	CN 105045812 A (SHANGHAI GAOXIN COMPUTER SYSTEMS CO., LTD.) 11 November 2015 (2015-11-11) description, paragraphs [0054]-[0088]	4, 6, 7, 11, 13, 14, 17, 20
Y	CN 103257957 A (SHENZHEN TENCENT COMPUTER SYSTEMS COMPANY LIMITED) 21 August 2013 (2013-08-21) description, paragraphs [0073]-[0079]	4, 6, 7, 11, 13, 14, 17, 20
A	CN 104112026 A (CHINA UNITED NETWORK COMMUNICATIONS LIMITED.) 22 October 2014 (2014-10-22) entire document	1-20
A	US 2014052728 A1 (NEC CORPORATION) 20 February 2014 (2014-02-20) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

13 May 2020

Date of mailing of the international search report

28 May 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/078389

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110008343	A	12 July 2019	None			
CN	105045812	A	11 November 2015	CN	105045812	B	29 January 2019
CN	103257957	A	21 August 2013	CN	103257957	B	08 September 2017
CN	104112026	A	22 October 2014	CN	104112026	B	08 September 2017
US	2014052728	A1	20 February 2014	JP	5534280	B2	25 June 2014
				WO	2012147428	A1	01 November 2012

国际检索报告

国际申请号

PCT/CN2020/078389

A. 主题的分类

G06F 16/35 (2019.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT, EPDOC, WPI, CNKI, ISI: 文本, 文档, 分类, 类别, 分词, 词频, 次数, 频率, 词, 相似度, 相似性, 特征, 值, 向量, document, classify, word, frequency, similar, feature, vector

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 110008343 A (深圳前海微众银行股份有限公司) 2019年 7月 12日 (2019 - 07 - 12) 权利要求1-10	1-14
X	CN 105045812 A (上海高欣计算机系统有限公司) 2015年 11月 11日 (2015 - 11 - 11) 说明书第[0054]-[0088]段	1-3, 5, 8-10, 12, 15, 16, 18, 19
Y	CN 105045812 A (上海高欣计算机系统有限公司) 2015年 11月 11日 (2015 - 11 - 11) 说明书第[0054]-[0088]段	4, 6, 7, 11, 13, 14, 17, 20
Y	CN 103257957 A (深圳市腾讯计算机系统有限公司) 2013年 8月 21日 (2013 - 08 - 21) 说明书第[0073]-[0079]段	4, 6, 7, 11, 13, 14, 17, 20
A	CN 104112026 A (中国联合网络通信集团有限公司) 2014年 10月 22日 (2014 - 10 - 22) 全文	1-20
A	US 2014052728 A1 (NEC CORPORATION) 2014年 2月 20日 (2014 - 02 - 20) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 5月 13日

国际检索报告邮寄日期

2020年 5月 28日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

李娜

电话号码 86-(10)-53961403

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/078389

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	110008343	A	2019年 7月 12日	无	
CN	105045812	A	2015年 11月 11日	CN	105045812 B 2019年 1月 29日
CN	103257957	A	2013年 8月 21日	CN	103257957 B 2017年 9月 8日
CN	104112026	A	2014年 10月 22日	CN	104112026 B 2017年 9月 8日
US	2014052728	A1	2014年 2月 20日	JP	5534280 B2 2014年 6月 25日
				WO	2012147428 A1 2012年 11月 1日