(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 112347262 B
(45)授权公告日 2021.04.13

(21)申请号 202110028487.4
(22)申请日 2021.01.11
(65)同一申请的已公布的文献号
申请公布号 CN 112347262 A
(43)申请公布日 2021.02.09
(73)专利权人 北京江融信科技有限公司
地址 100080 北京市海淀区苏州街55号3层01-A441
(72)发明人 衣得平 李海滨 翁国海
(74)专利代理机构 北京天盾知识产权代理有限公司 11421
代理人 张彩珍
(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/332(2019.01)
G06F 40/35(2020.01)
(56)对比文件
US 2015199593 A1,2015.07.16
CN 107766853 A,2018.03.06
审查员 刘长勇

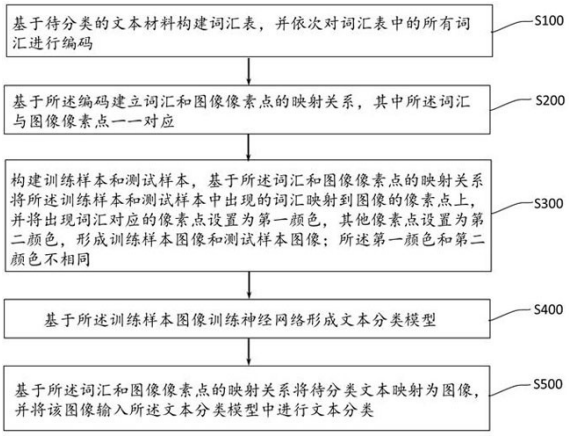
权利要求书2页 说明书7页 附图2页

(54)发明名称

一种文本分类方法及系统、意图分类系统和机器人

(57)摘要

本发明公开了一种文本分类方法及系统、意图分类系统和机器人，包括构建词汇表，依次对词汇表中的所有词汇进行编码；基于编码建立词汇和图像像素点的映射关系，词汇与图像像素点一一对应；构建训练样本和测试样本，基于词汇和图像像素点的映射关系将训练样本和测试样本中出现的词汇映射到图像的像素点上，并将出现词汇对应的像素点设置为第一颜色，其他像素点设置为第二颜色，形成训练样本图像和测试样本图像；第一颜色和第二颜色不相同；基于训练样本图像训练神经网络形成文本分类模型；将待分类文本映射为图像，并将其输入文本分类模型中进行文本分类；该方法把文本映射到一个图像上，用简单的神经网络或卷积神经网络进行文本分类，准确率高。



CN 112347262 B

1.一种文本分类方法,其特征在于,包括以下步骤:

S100:基于待分类的文本材料构建词汇表,并依次对词汇表中的所有词汇进行编码;

S200:基于所述编码建立词汇和图像像素点的映射关系,其中所述词汇与图像像素点一一对应,一个词汇编码固定对应一个像素点;

S300:构建训练样本和测试样本,基于所述词汇和图像像素点的映射关系将所述训练样本和测试样本中出现的词汇映射到图像的像素点上,并将出现词汇对应的像素点设置为第一颜色,其他像素点设置为第二颜色,形成训练样本图像和测试样本图像;所述第一颜色和第二颜色不相同;

S400:基于所述训练样本图像训练神经网络形成文本分类模型;

S500:基于所述词汇和图像像素点的映射关系将待分类文本映射为图像,并将该图像输入所述文本分类模型中进行文本分类。

2.根据权利要求1所述的文本分类方法,其特征在于,所述步骤S100中还包括:统计词汇表中各词汇的词频,基于词频的大小依次对词汇表中的所有词汇进行编码,并删除词汇表中词频低于设定阈值的词汇。

3.根据权利要求1所述的文本分类方法,其特征在于,所述步骤S100中还包括:将待分类的文本材料进行分词,基于分词后得到的词汇构建词汇表。

4.根据权利要求1所述的文本分类方法,其特征在于,所述词汇表中的各词汇包括单词汇、第一组合词汇和第二组合词汇;所述第一组合词汇由相邻的单词汇组合构成,所述第二组合词汇由不相邻的单词汇组合构成。

5.根据权利要求1所述的文本分类方法,其特征在于,所述S100中还包括:根据语义信息或语法信息增加集合词。

6.一种文本分类系统,其特征在于,包括词汇表构建模块、编码模块、映射模块、样本图像生成模块、文本分类模型构建模块和文本分类模块;

所述词汇表构建模块用于基于待分类的文本材料构建词汇表;

所述编码模块用于依次对所述词汇表中的所有词汇进行编码;

所述映射模块用于基于所述编码建立词汇和图像像素点的映射关系,其中所述词汇与图像像素点一一对应,一个词汇编码固定对应一个像素点;

所述样本图像生成模块用于构建训练样本和测试样本,基于所述词汇和图像像素点的映射关系将所述训练样本和测试样本中出现的词汇映射到图像的像素点上,并将出现词汇对应的像素点设置为第一颜色,其他像素点设置为第二颜色,形成训练样本图像和测试样本图像;所述第一颜色和第二颜色不相同;

所述文本分类模型构建模块用于基于所述训练样本图像训练神经网络形成文本分类模型;

所述文本分类模块用于基于所述词汇和图像像素点的映射关系将待分类文本映射为图像,并将该图像输入所述文本分类模型中进行文本分类。

7.根据权利要求6所述的文本分类系统,其特征在于,所述词汇表中的各词汇包括单词汇、第一组合词汇和第二组合词汇;所述第一组合词汇由相邻的单词汇组合构成,所述第二组合词汇由不相邻的单词汇组合构成。

8.一种意图分类系统,其特征在于,包括意图分类树和如权利要求6或7所述的文本分

类系统，通过所述文本分类系统将文本进行分类得到分类结果，将所述分类结果输入所述意图分类树中匹配得到意图。

9.根据权利要求8所述的意图分类系统，其特征在于，所述意图分类树包括一级类别意图和N-1个级别的类别子意图，第N级类别子意图下只含有一个意图。

10.一种机器人，包括如权利要求8或9所述的意图分类系统。

一种文本分类方法及系统、意图分类系统和机器人

技术领域

[0001] 本发明涉及文本分类技术领域，具体涉及一种文本分类方法及系统、意图分类系统和机器人。

背景技术

[0002] 文本分类应用于很多领域，比如垃圾信息的检测、新闻的自动分类等；现有的神经网络文本分类大部分需要词向量来实现，通过词向量把高维的信息压缩到较小的维度，然后输入到神经网络里进行分类；然而词向量需要进行大量文本统计，存在语法、语义以及人工信息均难以融合进去等缺陷；而且词表中所有的词向量构成的向量空间维度还是相对较大，例如词向量是500的一个文章，如果文章长度是1万个词汇那么至少要500*10000这么大的矩阵来存储，而对于100个词汇的小文章只用到前面一部分矩阵空间500*100，即词向量构成的图像并非固定宽度的图像，图片之间有大有小，甚至相差100倍，用普通的CNN等网络不能识别较大的图片，而这种“图片”不能缩放，如果缩放的话就丢失词向量信息，因此必须用到较为复杂的神经网络或卷积神经网络才能进行文本分类。

发明内容

[0003] 针对上述问题，本发明的一个目的是提供一种文本分类方法，该方法无需使用词向量，直接把待分类的文本映射到一个图像上，然后用简单的神经网络或卷积神经网络进行文本分类，准确率高。

[0004] 本发明的第二个目的是提供一种文本分类系统。

[0005] 本发明的第三个目的是提供一种意图分类系统。

[0006] 本发明的第四个目的是提供一种机器人。

[0007] 本发明所采用的第一个技术方案是：一种文本分类方法，包括以下步骤：

[0008] S100：基于待分类的文本材料构建词汇表，并依次对词汇表中的所有词汇进行编码；

[0009] S200：基于所述编码建立词汇和图像像素点的映射关系，其中所述词汇与图像像素点一一对应；

[0010] S300：构建训练样本和测试样本，基于所述词汇和图像像素点的映射关系将所述训练样本和测试样本中出现的词汇映射到图像的像素点上，并将出现词汇对应的像素点设置为第一颜色，其他像素点设置为第二颜色，形成训练样本图像和测试样本图像；所述第一颜色和第二颜色不相同；

[0011] S400：基于所述训练样本图像训练神经网络形成文本分类模型；

[0012] S500：基于所述词汇和图像像素点的映射关系将待分类文本映射为图像，并将该图像输入所述文本分类模型中进行文本分类。优选地，所述步骤S100中还包括：统计词汇表中各词汇的词频，基于词频的大小依次对词汇表中的所有词汇进行编码，并删除词汇表中词频低于设定阈值的词汇。

[0013] 优选地，所述步骤S100中还包括：将待分类的文本材料进行分词，基于分词后得到的词汇构建词汇表。

[0014] 优选地，所述词汇表中的各词汇包括单词汇、第一组合词汇和第二组合词汇；所述第一组合词汇由相邻的单词汇组合构成，所述第二组合词汇由不相邻的单词汇组合构成。

[0015] 优选地，所述S100中还包括：根据语义信息或语法信息增加集合词。本发明所采用的第二个技术方案是：一种文本分类系统，包括词汇表构建模块、编码模块、映射模块、样本图像生成模块、文本分类模型构建模块和文本分类模块；

[0016] 所述词汇表构建模块用于基于待分类的文本材料构建词汇表；

[0017] 所述编码模块用于依次对所述词汇表中的所有词汇进行编码；

[0018] 所述映射模块用于基于所述编码建立词汇和图像像素点的映射关系，其中所述词汇与图像像素点一一对应；

[0019] 所述样本图像生成模块用于构建训练样本和测试样本，基于所述词汇和图像像素点的映射关系将所述训练样本和测试样本中出现的词汇映射到图像的像素点上，并将出现词汇对应的像素点设置为第一颜色，其他像素点设置为第二颜色，形成训练样本图像和测试样本图像；所述第一颜色和第二颜色不相同；

[0020] 所述文本分类模型构建模块用于基于所述训练样本图像训练神经网络形成文本分类模型；

[0021] 所述文本分类模块用于基于所述词汇和图像像素点的映射关系将待分类文本映射为图像，并将该图像输入所述文本分类模型中进行文本分类。

[0022] 优选地，所述词汇表中的各词汇包括单词汇、第一组合词汇和第二组合词汇；所述第一组合词汇由相邻的单词汇组合构成，所述第二组合词汇由不相邻的单词汇组合构成。本发明所采用的第三个技术方案是：一种意图分类系统，包括意图分类树和如第二技术方案中所述的文本分类系统，通过所述文本分类系统将文本进行分类得到分类结果，将所述分类结果输入所述意图分类树中匹配得到意图。

[0023] 优选地，所述意图分类树包括一级类别意图和N-1个级别的类别子意图，第N级类别子意图下只含有一个意图。

[0024] 本发明所采用的第四个技术方案是：一种机器人，包括如第三技术方案中所述的意图分类系统。

[0025] 上述技术方案的有益效果：

[0026] (1)本发明公开的文本分类方法简单，不需要训练词向量，简单的将分词信息映射到像素点上就可以实现文本的分类，而且能达到较高的准确性。

[0027] (2)本发明公开的文本分类方法能将语法和语义信息融合进去，通过添加适量的语法、语义信息来增加文本分类的准确性。

[0028] (3)本发明公开的文本分类方法中通过将词汇一一映射到图像的像素点上形成词汇和像素点的映射关系，使其空间维度非常小，而且不同的文本可以映射形成固定宽度的图像，无需使用复杂的神经网络或卷积神经网络来进行文本分类，只需用最简单的神经网络就能实现较高准确性的分本分类。

[0029] (4)本发明公开的文本分类方法能用于机器人的意图识别上，比如先对机器人的意图做一个简单的分类，然后在每个小的类别上再做分类，一直分类细化下去最终每个类

别只有一个意图;机器人能用该文本分类方法计算获取到的句子的意图,最后在机器人的意图类别中匹配到具体的意图;机器人例如为银行的聊天机器人、各种客服机器人等。

附图说明

[0030] 图1为本发明的一个实施例提供的一种文本分类方法的流程示意图;

[0031] 图2为本发明的一个实施例提供的一种文本分类系统结构示意图;

[0032] 图3为本发明的一个实施例提供的意图分类树的结构示意图。

具体实施方式

[0033] 下面结合附图和实施例对本发明的实施方式作进一步详细描述。以下实施例的详细描述和附图用于示例性地说明本发明的原理,但不能用来限制本发明的范围,即本发明不限于所描述的优选实施例,本发明的范围由权利要求书限定。

[0034] 在本发明的描述中,需要说明的是,除非另有说明,“多个”的含义是两个或两个以上;术语“第一”“第二”等仅用于描述目的,而不能理解为指示或暗示相对重要性;对于本领域的普通技术人员而言,可视具体情况理解上述术语在本发明中的具体含义。

[0035] 实施例一

[0036] 图1为本发明的一个实施例提供的一种文本分类方法,包括以下步骤:

[0037] S100:基于待分类的文本材料构建词汇表,并依次对词汇表中的所有词汇进行编码;

[0038] 搜集待分类的文本材料,待分类的文本材料包括搜狗的新闻、清华新闻分类语料和今日头条中文新闻(短文本);其中,搜狗的新闻约有5万文本数据;清华新闻分类语料根据新浪新闻RSS订阅频道2005~2011年间的历史数据筛选过滤生成,约有78万的文本数据;今日头条中文新闻(短文本)约有38万的文本数据;

[0039] 将待分类的文本材料进行分词,分词后得到各单词汇,由相邻的单词汇组合形成第一组合词汇;基于单词汇和第一组合词汇构建词汇表;

[0040] 将词汇表中的各单词汇以及第一组合词汇依次进行编码,词汇表中的每个词汇均具有其独有的编码;编码会根据词汇表中词汇的变动实时更新。

[0041] 进一步的,在一个实施例中,步骤S100中还包括:统计词汇表中各词汇的词频,将词频低于设定阈值的词汇去掉;

[0042] 对词汇表中的各单词汇以及第一组合词汇的出现频率进行统计得到各单词汇以及第一组合词汇的词频,将词频低于设定阈值的词汇去掉;设定的阈值例如为10,词频太低就不会有泛化作用,去掉低频词汇形成较小的词汇表,在保证分类准确率的情况下减小运算量,提高处理效率;词汇表中的词汇量约为6万。

[0043] 进一步的,在一个实施例中,将不相邻的单词汇与单词汇进行组合形成第二组合词汇,统计第二组合词汇的词频,并将高频的第二组合词汇作为词汇表的补充数据;补充数据在原先的词汇表后面加入,原先词汇表中的词汇不需要重新排序。

[0044] 进一步的,在一个实施例中,通过在词汇表中添加语义信息来增加文本分类的准确性,例如中国、美国和日本属于国家集合,本发明通过定义一个集合词“{国家s}”,当出现若干国家相关的词汇,则在词汇表中增加“{国家s}”这个集合词,以此提高文本分类的准确

性;也可以通过添加适当的语法信息来增加文本分类的准确性,例如加入搭配结构集合词,{主语s}+{谓语s}+{宾语s},这些集合词实际上就是一些新的特殊词汇,可以扩大原先的词汇表,在原先的词汇表后面加入,原先词汇表中的词汇不需要重新排序,通过词汇表中高频词汇的扩增增加文本分类的准确性。

[0045] 上述这类语义信息和语法信息需要统计并人工矫正,在词汇表中加入语义信息或/和语法信息后准确率可以提高0.43%～5%;可见,本发明公开的文本分类方法能将语法和语义信息融合进去,通过添加适量的语法、语义信息来增加文本分类的准确性。

[0046] 进一步的,在一个实施例中,基于词频的大小对依次对词汇表中的所有词汇进行编码;

[0047] 将词汇表中的各单词汇以及第一组合词汇按其词频的高低依次进行排序编码,例如将出现频率最高的词汇编码为1,其次的编码为2,…(但并不限于此);词汇表中的每个词汇均具有其独有的编码;编码会根据词汇表中词汇的变动实时更新。

[0048] S200:基于编码的大小将词汇表中的所有词汇一一映射到图像的像素点上形成词汇和像素点的映射关系;一个编码固定对应一个像素点;

[0049] 例如300*300维度的图像是由300行像素点和300列像素点构成的,图像上共有90000个像素点;

[0050] 将编码为1的词汇映射到图像中位于第1行、第1列的像素点;

[0051] 将编码为2的词汇映射到图像中位于第1行、第2列的像素点;

[0052] ...

[0053] 将编码为301的词汇映射到图像中位于第2行、第1列的像素点;

[0054] 将编码为302的词汇映射到图像中位于第2行,第2列的像素点;

[0055] ...;

[0056] (但并不限于此)。

[0057] 本发明公开的文本分类方法简单,不需要训练词向量,简单的将分词信息映射到像素点上就可以实现文本的分类,而且能达到较高的准确性。

[0058] S300:构建训练样本和测试样本,基于词汇和像素点的映射关系将训练样本和测试样本中的词汇映射到图像中,将训练样本和测试样本中词汇对应的像素点设置为第一颜色,其他像素点设置为第二颜色,第一颜色和第二颜色不相同;将各训练样本和测试样本保存为不同的训练样本图像和测试样本图像;

[0059] 本发明的训练样本是10个类别,包括体育、娱乐、家居、房产、教育、时尚、时政、社会、科技和财经,每个类别为5000个文本,从每个类别中抽取1/10的样本作为测试样本,这样训练样本就是每个类别4500个文本,测试样本每个类别500个样本;

[0060] 进一步的,在一个实施例中将第一颜色设置为白色、第二颜色设置为黑色或将第一颜色设置为黑色、第二颜色设置为白色;即将训练样本和测试样本中的出现的词汇根据词汇与像素点的映射关系映射到图像中,如果训练样本和测试样本的文本中有这个词汇就把该词汇对应的像素点变成白色,如果没有这个词汇,则该词汇对应的像素点为黑色;把图片保存为一个黑白图片,通过采用黑白图片的形式来减少运算量。

[0061] 例如300*300维度的图像上共有90000个像素点,

[0062] 如果有编码为1的词汇,那么就把第1行,第1列设置为白色,

[0063] 如果有编码为2的词汇,那么就把第1行,第2列设置为白色,

[0064] ...

[0065] 如果有编码为301的词汇,那么就把第2行,第1列设置为白色,

[0066] 如果有编码为302的词汇,那么就把第2行,第2列设置为白色,

[0067] ..

[0068] 其他点设置为黑色(但并不限于此)。

[0069] S400:基于训练样本图像训练神经网络形成文本分类模型;基于测试样本图像测试文本分类模型的准确率;

[0070] 将上述训练样本图像输入至神经网络进行训练形成文本分类模型,将测试样本图像输入文本分类模型进行测试得到准确率;测试样本准确率高达94%。

[0071] 本发明无需采用复杂的神经网络或卷积神经网络来进行文本分类,只需用最简单的神经网络,例如只有一层sofomax层就可以进行文本分类;将词汇映射图像输入至sofomax层采用以下代码:

[0072] model = Sequential()

[0073] model.add(Input(shape=(图片宽度,图片高度,1)))

[0074] model.add(Flatten())

[0075] model.add(Dense(分类个数n))

[0076] model.add(Activation('softmax'))。

[0077] 本发明公开的文本分类方法中通过将词汇一一映射到图像的像素点上形成词汇和像素点的映射关系,使其空间维度非常小,而且不同的文本可以映射形成固定宽度的图像,无需使用复杂的神经网络或卷积神经网络来进行文本分类,只需用最简单的神经网络就能实现较高准确性的分本分类。

[0078] S500:基于词汇和图像像素点的映射关系将待分类文本映射为图像,并将该图像输入文本分类模型中进行文本分类。

[0079] 下面结合具体算例分析本发明的实际效果:

[0080] 将约为78万个的清华新闻分类语料文本数据进行分类,分别采用本发明公开的分类方法和THUCTC工具包对上述文本数据进行分类,得到相应的分类准确率;基于本发明公开的分类方法进行分类的准确率为94%,而采用THUCTC工具包进行分类的准确率为88.6%;采用词向量的方法进行分类的准确率为91%;由此可见,使用本发明公开的分类方法进行文本分类得到的准确率明显高于基于THUCTC工具包进行的分类和基于词向量进行分类的准确率。

[0081] 实施例二

[0082] 图2为本发明的一个实施例提供的一种文本分类系统,包括词汇表构建模块、编码模块、映射模块、样本图像生成模块、文本分类模型构建模块、测试模块和文本分类模块;

[0083] 所述词汇表构建模块用于基于待分类的文本材料构建词汇表;

[0084] 所述编码模块用于依次对所述词汇表中的所有词汇进行编码;

[0085] 所述映射模块用于基于编码建立词汇和图像像素点的映射关系,其中词汇与图像像素点一一对应;

[0086] 所述样本图像生成模块用于构建训练样本和测试样本,基于所述词汇和图像像素

点的映射关系将所述训练样本和测试样本中出现的词汇映射到图像的像素点上，并将出现词汇对应的像素点设置为第一颜色，其他像素点设置为第二颜色，形成训练样本图像和测试样本图像；所述第一颜色和第二颜色不相同；

[0087] 所述文本分类模型构建模块用于基于所述训练样本图像训练神经网络形成文本分类模型；

[0088] 所述测试模块用于基于所述测试样本图像测试所述文本分类模型的准确率；

[0089] 所述文本分类模块用于基于词汇和图像像素点的映射关系将待分类文本映射为图像，并将该图像输入文本分类模型中进行文本分类。

[0090] 实施例三

[0091] 发明的一个实施例提供了一种意图分类系统，包括意图分类树和实施例二中的文本分类系统，通过文本分类系统将文本进行分类得到分类结果，将分类结果输入意图分类树中匹配得到意图。

[0092] 本发明公开的文本分类系统能应用于机器人的意图分类，例如聊天机器人、客服机器人的意图分类(但并不限于此)；本发明通过分析分类错误的句子，在词汇表中加入相关的语法和语义信息来提高句子分类的准确性，进而提高意图匹配的准确性。

[0093] 首先构建意图分类树，如图3所示，将意图构建为分类树，一级类别包括m个分支，a_1，a_2，...a_m；如果一级类别的某个分支下的意图大于1个，则对该一级类别分支再进行细分构建二级类别；如果二级类别的某个分支下的意图大于1个，则对该二级类别分支再进行细分构建三级类别，...以此类推，直至N级类别下每个分支只有一个意图。

[0094] 例如一级类别包括10个分支，a_1，a_2，... a_10，当a_1这个分支下包括10个子意图，则再在a_1这个分支下构建二级类别；这样不断的分类下去，最终每个分支类别只有一个意图，这样就可以将通过文本分类系统分类后的结果输入至预设的意图分类树中，最终匹配到准确的意图。

[0095] 实施例四

[0096] 发明的一个实施例提供了一种机器人，包括实施例三中的意图分类系统。

[0097] 本发明公开的文本分类方法能用于机器人的意图识别上，比如先对机器人的意图做一个简单的分类，然后在每个小的类别上再做分类，一直分类细化下去最终每个类别只有一个意图；机器人能用该文本分类方法计算获取到的句子的意图，最后在机器人的意图类别中匹配到具体的意图；机器人例如为银行的聊天机器人、各种客服机器人等。

[0098] 本领域普通技术人员可以意识到，结合本发明实施例中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

[0099] 在本申请所提供的实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连

接,可以是电性,机械或其它的形式。

[0100] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0101] 另外,在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。

[0102] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案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一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0103] 以上所述,仅为本发明的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应以权利要求的保护范围为准。

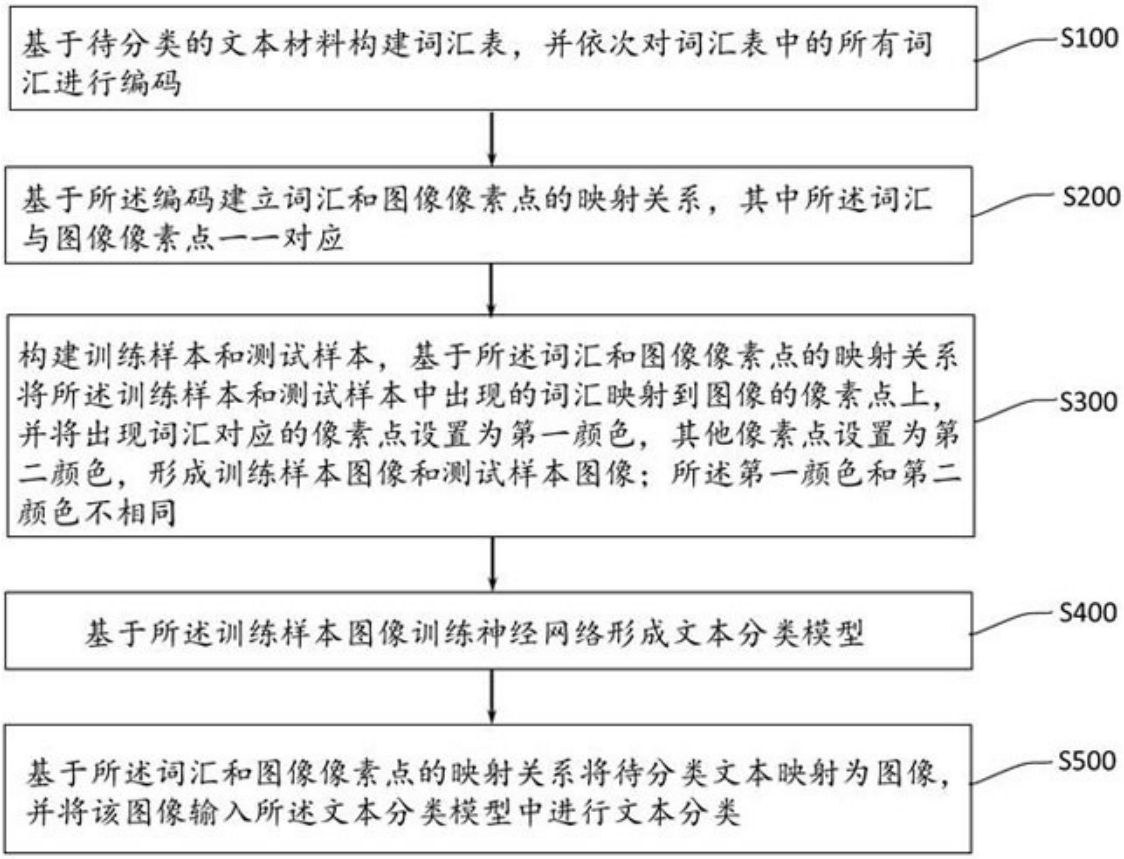

基于待分类的文本材料构建词汇表，并依次对词汇表中的所有词汇进行编码
S100
基于所述编码建立词汇和图像像素点的映射关系，其中所述词汇与图像像素点一一对应
S200
构建训练样本和测试样本，基于所述词汇和图像像素点的映射关系将所述训练样本和测试样本中出现的词汇映射到图像的像素点上，并将出现词汇对应的像素点设置为第一颜色，其他像素点设置为第二颜色，形成训练样本图像和测试样本图像；所述第一颜色和第二颜色不相同
S300
基于所述训练样本图像训练神经网络形成文本分类模型
S400
基于所述词汇和图像像素点的映射关系将待分类文本映射为图像，并将该图像输入所述文本分类模型中进行文本分类
S500

图1

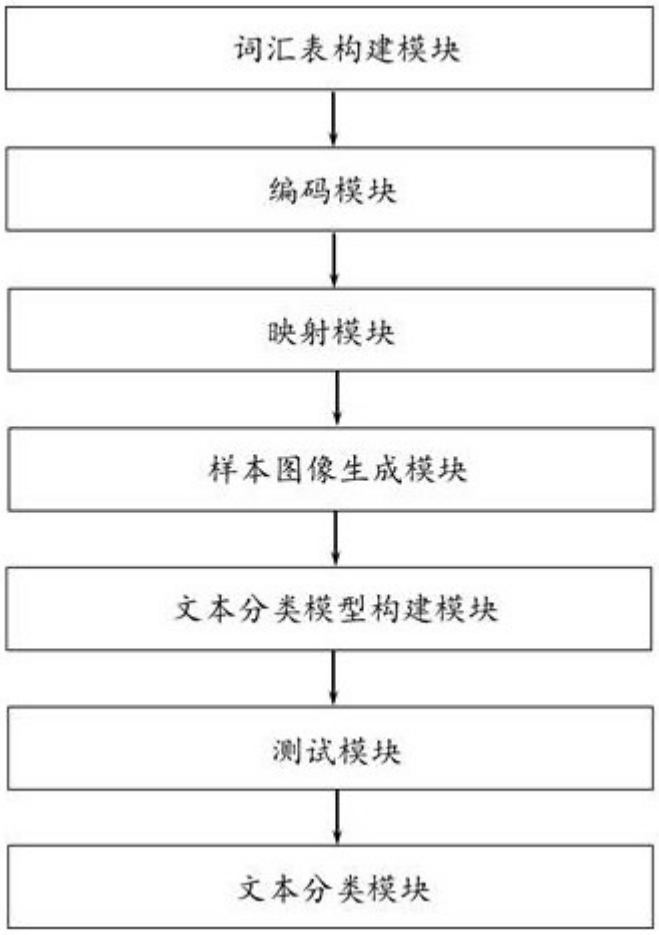

词汇表构建模块
编码模块
映射模块
样本图像生成模块
文本分类模型构建模块
测试模块
文本分类模块

图2

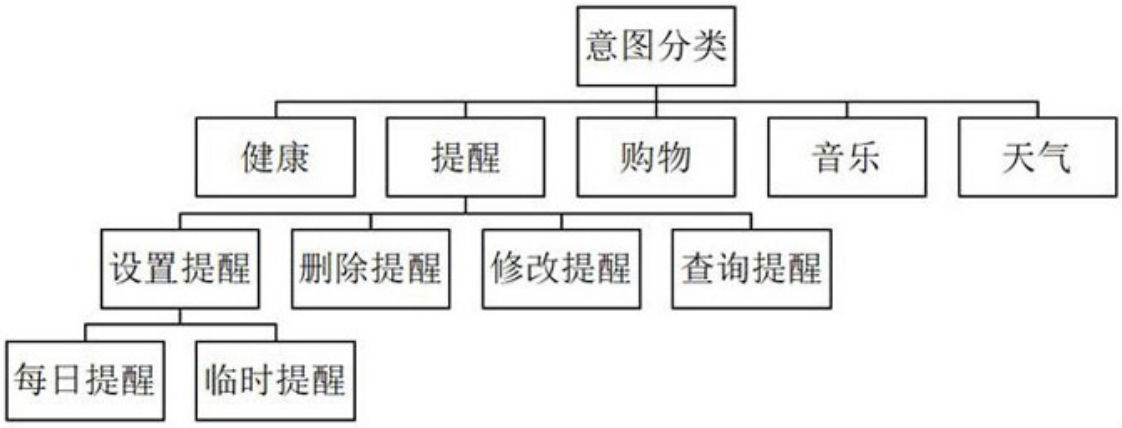

意图分类
健康
提醒
购物
音乐
天气
设置提醒
删除提醒
修改提醒
查询提醒
每日提醒
临时提醒

图3