(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 103336766 B
(45)授权公告日 2016.12.28

(21)申请号 201310278012.6

(22)申请日 2013.07.04

(65)同一申请的已公布的文献号
申请公布号 CN 103336766 A

(43)申请公布日 2013.10.02

(73)专利权人 微梦创科网络科技(中国)有限公司
地址 100080 北京市海淀区海淀北二街10号701室

(72)发明人 姜贵彬

(74)专利代理机构 北京市京大律师事务所 11321
代理人 张璐 方晓明

(51)Int.Cl.
G06F 17/27(2006.01)
G06F 17/30(2006.01)

(56)对比文件
CN 101784022 A,2010.07.21,
CN 101784022 A,2010.07.21,
CN 101996241 A,2011.03.30,
龚垒.基于支持向量机的垃圾短信过滤方法研究.《中国优秀硕士学位论文全文数据库信息科技辑》.2011,(第09期),I136－964.
侯旭东.基于内容的短消息智能分析系统研究.《中国优秀硕士学位论文全文数据库信息科技辑》.2011,(第05期),I136-260.

审查员 孙国辉

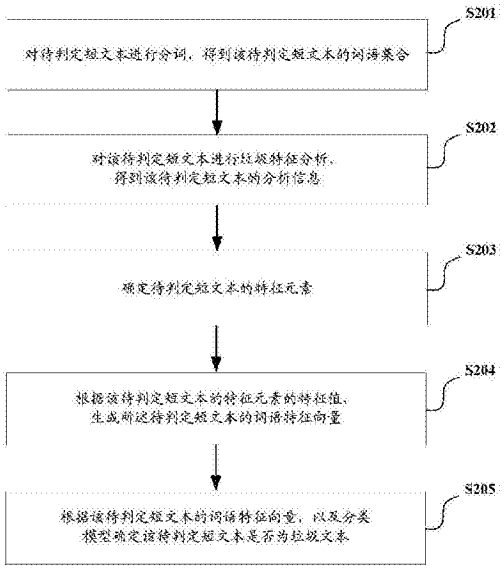
权利要求书4页 说明书11页 附图3页

(54)发明名称

短文本垃圾识别以及建模方法和装置

(57)摘要

本发明公开了一种短文本垃圾识别以及建模方法和装置，所述方法包括：对待判定短文本进行分词得到词语集合，并对所述待判定短文本进行垃圾特征分析得到分析信息；将所述待判定短文本的分析信息以及词语集合中每个词语分别与预先确定的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成所述待判定短文本的词语特征向量；根据所述待判定短文本的词语特征向量，以及分类模型，确定所述待判定短文本是否为垃圾文本；其中分类模型是结合训练集中的样本数，选择合适的分类算法预先训练出的。由于采用扩充了分析信息的特征值的词语特征向量进行垃圾识别，从而提高了识别垃圾文本的识别准确率。



CN 103336766 B

1.一种短文本垃圾识别方法，其特征在于，包括：

对待判定短文本进行分词得到词语集合，并对所述待判定短文本进行垃圾特征分析得到分析信息；

将所述待判定短文本的分析信息以及词语集合中每个词语分别与预先确定的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成所述待判定短文本的词语特征向量；

根据所述待判定短文本的词语特征向量，以及预先训练出的分类模型，确定所述待判定短文本是否为垃圾文本；其中，

所述分析信息包括如下任一信息，或如下信息的任意组合：

是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息、二元词的占比信息、不同词性词汇搭配比例、标点符号与名词的数量比例信息；以及

所述分析信息的特征值具体包括：

对于所述是否包含联系方式特征的信息，其特征值为二值的0或1；

对于所述干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息，其特征值为0～1之间的数值。

2.如权利要求1所述的方法，其特征在于，在所述生成所述待判定短文本的词语特征向量之前，还包括：

对与所述特征元素集合中的特征元素相匹配的分析信息的特征值进行归一化：

将其中是否包含联系方式特征的信息的特征值归一化为二值的0或100；

将其中干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息的特征值乘以100，得到0～100之间的归一化数值。

3.如权利要求1或2所述的方法，其特征在于，所述词语的特征值根据如下方法得到：

计算该词语的TF、IDF值，并根据如下公式1计算出该词语的特征值：

log(TF+1.0)×IDF （公式1）。

4.如权利要求1或2所述的方法，其特征在于，所述分类模型的训练方法，以及所述特征元素集合的确定方法包括：

对于训练集中已区分为垃圾文本，或非垃圾文本的每个短文本，进行分词后得到该短文本的词语集合，并对该短文本进行垃圾特征分析后得到该短文本的分析信息；

针对所述训练集中的每个短文本，计算该短文本的词语集合中每个词语的特征值，并计算该短文本的分析信息的特征值后，对计算出的特征值求取类别区分度；将类别区分度大于设定阈值的词语，以及分析信息作为所述特征元素集合中的特征元素；

针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与所述特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相

匹配的词语或分析信息的特征值，生成该短文本的词语特征向量；

根据所述训练集中各短文本的词语特征向量训练出所述分类模型。

5.如权利要求4所述的方法，其特征在于，所述根据所述训练集中各短文本的词语特征向量训练出所述分类模型具体为：

运用SVM分类算法、或贝叶斯分类算法、或决策树分类算法、或最大熵分类算法，根据所述训练集中各短文本的词语特征向量训练出所述分类模型。

6.一种建模方法，其特征在于，包括：

对于训练集中已区分为垃圾文本，或非垃圾文本的每个短文本，进行分词后得到该短文本的词语集合，并对该短文本进行垃圾特征分析后得到该短文本的分析信息；

针对所述训练集中的每个短文本，计算该短文本的词语集合中每个词语的特征值，并计算该短文本的分析信息的特征值后，对计算出的特征值求取类别区分度；将类别区分度大于设定阈值的词语，以及分析信息作为特征元素集合中的特征元素；

针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与所述特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量；

根据所述训练集中各短文本的词语特征向量训练出分类模型；其中，

所述分析信息包括如下任一信息，或如下信息的任意组合：

是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息、二元词的占比信息、不同词性词汇搭配比例、标点符号与名词的数量比例信息；以及

所述分析信息的特征值具体包括：

对于所述是否包含联系方式特征的信息，其特征值为二值的0或1；

对于所述干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息，其特征值为0～1之间的数值。

7.如权利要求6所述的方法，其特征在于，在所述计算该短文本的分析信息的特征值后，以及所述根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量之前，还包括：

对该短文本的分析信息的特征值进行归一化：

将所述是否包含联系方式特征的信息的特征值归一化为二值的0或100；

将所述干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息的特征值乘以100，得到0～100之间的归一化数值；以及

所述根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量具体为：

根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的归一化后的特征值，生成该短文本的词语特征向量。

8.如权利要求6或7所述的方法，其特征在于，所述根据所述训练集中各短文本的词语特征向量训练出所述分类模型具体为：

运用SVM分类算法、或贝叶斯分类算法、或决策树分类算法、或最大熵分类算法，根据所述训练集中各短文本的词语特征向量训练出所述分类模型。

9.一种建模装置，其特征在于，包括：

特征提取模块，用于对于训练集中已区分为垃圾文本，或非垃圾文本的每个短文本，进行分词后得到该短文本的词语集合，并对该短文本进行垃圾特征分析得到该短文本的分析信息；

特征元素集合确定模块，用于针对所述训练集中的每个短文本，计算该短文本的词语集合中每个词语的特征值，并计算该短文本的分析信息的特征值后，对计算出的特征值求取类别区分度；将类别区分度大于设定阈值的词语，以及分析信息作为特征元素集合中的特征元素；

特征向量确定模块，用于针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与所述特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量；

分类模型构建模块，用于根据所述特征向量确定模块确定出的所述训练集中各短文本的词语特征向量，构建分类模型；

其中，所述分析信息包括如下任一信息，或如下信息的任意组合：

是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息、二元词的占比信息、不同词性词汇搭配比例、标点符号与名词的数量比例信息；以及

所述分析信息的特征值具体包括：

对于是否包含联系方式特征的信息，其特征值为二值的0或1；

对于干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息，其特征值为0～1之间的数值。

10.如权利要求9所述的装置，其特征在于，

所述特征向量确定模块具体用于针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与所述特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的归一化后的特征值，生成该短文本的词语特征向量。

11.一种短文本垃圾识别装置，其特征在于，包括：

特征提取模块，用于对于待判定短文本进行分词后得到词语集合，并对所述待判定短文本进行垃圾特征分析得到分析信息；

特征向量确定模块，用于将所述待判定短文本的分析信息以及词语集合中每个词语分别与预先确定的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成所述待判定短文本的词语特征向量；

垃圾识别模块，用于从所述特征向量确定模块获取所述待判定短文本的词语特征向量后，根据所述待判定短文本的词语特征向量，以及预先训练出的分类模型，确定所述待判定短文本是否为垃圾文本；

其中，所述分析信息包括如下任一信息，或如下信息的任意组合：

是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息、二元词的占比信息、不同词性词汇搭配比例、标点符号与名词的数量比例信息；以及

所述分析信息的特征值具体包括：

对于是否包含联系方式特征的信息，其特征值为二值的0或1；

对于干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息，其特征值为0～1之间的数值。

12.如权利要求11所述的装置，其特征在于，

所述特征向量确定模块具体用于将所述待判定短文本的分析信息以及词语集合中每个词语分别与预先确定的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的归一化后的特征值，生成所述待判定短文本的词语特征向量。

短文本垃圾识别以及建模方法和装置

技术领域

[0001] 本发明涉及互联网领域，尤其涉及一种短文本垃圾识别以及建模方法和装置。

背景技术

[0002] 互联网技术迅猛发展，网上信息爆炸式增长；随着生活、工作节奏的加快，人们越来越倾向于用简短的文字来沟通交流。以twitter(推特)和新浪微博为代表的以较小的短文本来生产、组织和传播信息的SNS(Social Network Service，社会性网络服务)网站，获得网友的青睐。

[0003] 目前，对互联网上的短文本内容进行自动垃圾识别的主要方法是，采用基于分类模型的方法，对于某个短文本内容将其分类为垃圾文本，或非垃圾文本；该方法包括：训练阶段和分类阶段。

[0004] 在训练阶段，根据训练集中大量的短文本进行建模：对于训练集中已区分为垃圾文本，或非垃圾文本的各个短文本，进行分词得到每个短文本的词语集合，根据每个短文本的词语集合计算得到每个短文本的词语特征向量；基于训练集中每个短文本的词语特征向量训练出分类模型。例如，运用SVM(Support Vector Machine，支持向量机)分类算法、或贝叶斯分类算法、或决策树分类算法、或最大熵分类算法，根据所述训练集中各短文本的词语特征向量训练出分类模型。

[0005] 在分类阶段，对于待判定短文本，进行分词得到该待判定短文本的词语集合后，根据该待判定短文本的词语集合计算出该待判定短文本的词语特征向量；根据该待判定短文本的词语特征向量与之前训练出的分类模型，判定该待判定短文本是否为垃圾文本。如何根据该待判定短文本的词语特征向量和分类模型进行垃圾文本的判定有多种算法，为本领域技术人员所熟知，此处不再赘述。

[0006] 但是，在实际应用中，本发明的发明人发现，SNS网站由于其社交属性，在SNS网站上的短文本通常内容简短，基于如此简短内容而提取的词语集合中的词语很少，由此得到的词语特征向量中的有效的特征值非常稀疏，有时得到的短文本的词语特征向量中可能仅有1、2个有效的特征值；基于如此少的特征值进行垃圾文本集和非垃圾文本集的归属判断的准确性大大降低；亦即，目前现有技术的短文本内容的垃圾识别方法识别准确率不高。

发明内容

[0007] 针对上述现有技术存在的缺陷，本发明提供了一种短文本垃圾识别以及建模方法和装置，用以提高对短文本的内容进行垃圾识别的准确性。

[0008] 根据本发明的一个方面，提供了一种短文本垃圾识别方法，包括：

[0009] 对待判定短文本进行分词得到词语集合，并对所述待判定短文本进行垃圾特征分析得到分析信息；

[0010] 将所述待判定短文本的分析信息以及词语集合中每个词语分别与预先确定的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语

或分析信息的特征值，生成所述待判定短文本的词语特征向量；

[0011] 根据所述待判定短文本的词语特征向量，以及预先训练出的分类模型，确定所述待判定短文本是否为垃圾文本。

[0012] 较佳地，所述分析信息包括如下任一信息，或如下信息的任意组合：

[0013] 是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息、二元词的占比信息、不同词性词汇搭配比例、标点符号与名词的数量比例信息。

[0014] 较佳地，所述分析信息的特征值具体包括：

[0015] 对于所述是否包含联系方式特征的信息，其特征值为二值的0或1；

[0016] 对于所述干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息，其特征值为0～1之间的数值。

[0017] 进一步的，在所述生成所述待判定短文本的词语特征向量之前，还包括：

[0018] 对与所述特征元素集合中的特征元素相匹配的分析信息的特征值进行归一化：

[0019] 将其中是否包含联系方式特征的信息的特征值归一化为二值的0或100；

[0020] 将其中干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息的特征值乘以100，得到0～100之间的归一化数值。

[0021] 较佳地，所述词语的特征值根据如下方法得到：

[0022] 计算该词语的TF、IDF值，并根据如下公式1计算出该词语的特征值：

[0023] $\log(TF+1.0)\times IDF$ （公式1）

[0024] 较佳地，所述分类模型的训练方法，以及所述特征元素集合的确定方法包括：

[0025] 对于训练集中已区分为垃圾文本，或非垃圾文本的每个短文本，进行分词后得到该短文本的词语集合，并对该短文本进行垃圾特征分析后得到该短文本的分析信息；

[0026] 针对所述训练集中的每个短文本，计算该短文本的词语集合中每个词语的特征值，并计算该短文本的分析信息的特征值后，对计算出的特征值求取类别区分度；将类别区分度大于设定阈值的词语，以及分析信息作为所述特征元素集合中的特征元素；

[0027] 针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与所述特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量；

[0028] 根据所述训练集中各短文本的词语特征向量训练出所述分类模型。

[0029] 较佳地，所述根据所述训练集中各短文本的词语特征向量训练出所述分类模型具体为：

[0030] 运用SVM分类算法、或贝叶斯分类算法、或决策树分类算法、或最大熵分类算法，根据所述训练集中各短文本的词语特征向量训练出所述分类模型。

[0031] 根据本发明的另一个方面，还提供了一种建模方法，包括：

[0032] 对于训练集中已区分为垃圾文本，或非垃圾文本的每个短文本，进行分词后得到该短文本的词语集合，并对该短文本进行垃圾特征分析后得到该短文本的分析信息；

[0033] 针对所述训练集中的每个短文本，计算该短文本的词语集合中每个词语的特征值，并计算该短文本的分析信息的特征值后，对计算出的特征值求取类别区分度；将类别区分度大于设定阈值的词语，以及分析信息作为特征元素集合中的特征元素；

[0034] 针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与所述特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量；

[0035] 根据所述训练集中各短文本的词语特征向量训练出分类模型。

[0036] 较佳地，所述分析信息包括如下任一信息，或如下信息的任意组合：

[0037] 是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息、二元词的占比信息、不同词性词汇搭配比例、标点符号与名词的数量比例信息。

[0038] 较佳地，所述分析信息的特征值具体包括：

[0039] 对于所述是否包含联系方式特征的信息，其特征值为二值的0或1；

[0040] 对于所述干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息，其特征值为0～1之间的数值。

[0041] 较佳地，在所述计算该短文本的分析信息的特征值后，以及所述根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量之前，还包括：

[0042] 对该短文本的分析信息的特征值进行归一化：

[0043] 将所述是否包含联系方式特征的信息的特征值归一化为二值的0或100；

[0044] 将所述干扰性符号的占比信息、或生僻字的占比信息、或繁体字符的占比信息、或词语间的转移概率、或前后词的词性间转移概率、或名词的占比信息、或动词的占比信息、或标点符号的占比信息、或一元词的占比信息、或二元词的占比信息、或不同词性词汇搭配比例、或标点符号与名词的数量比例信息的特征值乘以100，得到0～100之间的归一化数值；以及

[0045] 所述根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量具体为：

[0046] 根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的归一化后的特征值，生成该短文本的词语特征向量。

[0047] 较佳地，所述根据所述训练集中各短文本的词语特征向量训练出所述分类模型具体为：

[0048] 运用SVM分类算法、或贝叶斯分类算法、或决策树分类算法、或最大熵分类算法，根据所述训练集中各短文本的词语特征向量训练出所述分类模型。

[0049] 根据本发明的另一个方面，还提供了一种建模装置，包括：

[0050] 特征提取模块，用于对于训练集中已区分为垃圾文本，或非垃圾文本的每个短文本，进行分词后得到该短文本的词语集合，并对该短文本进行垃圾特征分析得到该短文本的分析信息；

[0051] 特征元素集合确定模块，用于针对所述训练集中的每个短文本，计算该短文本的词语集合中每个词语的特征值，并计算该短文本的分析信息的特征值后，对计算出的特征值求取类别区分度；将类别区分度大于设定阈值的词语，以及分析信息作为特征元素集合中的特征元素；

[0052] 特征向量确定模块，用于针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与所述特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量；

[0053] 分类模型构建模块，用于根据所述特征向量确定模块确定出的所述训练集中各短文本的词语特征向量，构建分类模型。

[0054] 较佳地，所述特征向量确定模块具体用于针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与所述特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的归一化后的特征值，生成该短文本的词语特征向量。

[0055] 较佳地，所述分析信息包括如下任一信息，或如下信息的任意组合：

[0056] 是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息、二元词的占比信息、不同词性词汇搭配比例、标点符号与名词的数量比例信息。

[0057] 根据本发明的另一个方面，还提供了一种短文本垃圾识别装置，包括：

[0058] 特征提取模块，用于对于待判定短文本进行分词后得到词语集合，并对所述待判定短文本进行垃圾特征分析得到分析信息；

[0059] 特征向量确定模块，用于将所述待判定短文本的分析信息以及词语集合中每个词语分别与预先确定的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成所述待判定短文本的词语特征向量；

[0060] 垃圾识别模块，用于从所述特征向量确定模块获取所述待判定短文本的词语特征向量后，根据所述待判定短文本的词语特征向量，以及预先训练出的分类模型，确定所述待判定短文本是否为垃圾文本。

[0061] 较佳地，所述特征向量确定模块具体用于将所述待判定短文本的分析信息以及词语集合中每个词语分别与预先确定的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的归一化后的特征值，生成所述待判定短文本的词语特征向量。

[0062] 较佳地，所述分析信息包括如下任一信息，或如下信息的任意组合：

[0063] 是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的

占比信息、标点符号的占比信息、一元词的占比信息、二元词的占比信息、不同词性词汇搭配比例、标点符号与名词的数量比例信息。

[0064] 本发明的技术方案中，构成垃圾文本超平面和非垃圾文本超平面的训练集中的各短文本的词语特征向量，和待判定短文本的词语特征向量，都包含了扩充的分析信息的特征值，根据包含了扩充的分析信息的特征值的词语特征向量，对待判定短文本进行垃圾识别，提高了识别垃圾文本的识别率和识别准确率。

附图说明

[0065] 图1为本发明实施例的构建垃圾文本超平面和非垃圾文本超平面流程图；

[0066] 图2为本发明实施例的对待判定的短文本进行垃圾识别的流程图；

[0067] 图3为本发明实施例的建模装置和短文本垃圾识别装置的内部结构框图。

具体实施方式

[0068] 为使本发明的目的、技术方案及优点更加清楚明白，以下参照附图并举出优选实施例，对本发明进一步详细说明。然而，需要说明的是，说明书中列出的许多细节仅仅是为了使读者对本发明的一个或多个方面有一个透彻的理解，即便没有这些特定的细节也可以实现本发明的这些方面。

[0069] 本申请使用的“模块”、“系统”等术语旨在包括与计算机相关的实体，例如但不限于硬件、固件、软硬件组合、软件或者执行中的软件。例如，模块可以是，但并不仅限于：处理器上运行的进程、处理器、对象、可执行程序、执行的线程、程序和/或计算机。举例来说，计算设备上运行的应用程序和此计算设备都可以是模块。一个或多个模块可以位于执行中的一个进程和/或线程内，一个模块也可以位于一台计算机上和/或分布于两台或更多台计算机之间。

[0070] 本发明的发明人考虑到，可以对基于现有技术方法得到的词语特征向量进行扩充：除了包括词语的特征值外，还可包括对短文本进行垃圾特征分析后得到的分析信息的特征值。例如，对短文本进行垃圾特征分析后得到的分析信息可以包括：是否包含联系方式特征，干扰性符号的占比、名词的占比、或动词的占比等。根据该扩充了的词语特征向量，判定其所属的待判定短文本是否为垃圾文本，比现有技术的方法提高了判定的准确率，即提高了垃圾短文本的识别准确率。

[0071] 基于上述考虑，本发明的实施例提供了一种基于分类模型的短文本垃圾识别方法；在分类模型的训练阶段，先进行建模：建模过程中，根据训练集中的各短文本，构建分类模型中的垃圾文本超平面和非垃圾平面超平面；在识别阶段，则可利用构建的分类模型中的垃圾文本超平面和非垃圾平面超平面，进行垃圾短文本的判定。

[0072] 建模过程中，根据训练集中的各短文本进行建模的方法，即构建分类模型中的垃圾文本超平面和非垃圾平面超平面的方法，流程如图1所示，具体步骤包括：

[0073] S101：对训练集中的每个短文本进行分词，得到每个短文本的词语集合。

[0074] 具体地，对于训练集中已区分为垃圾文本，或非垃圾文本的每个短文本，进行分词：将该短文本中连续的字序列划分为一个个词语；在划分出的词语中，去除掉没有实际意义的虚词（如标点、组动词、语气词、叹词、拟声词等）；剩余的词语构成该短文本的词语集

合。

[0075] S102:对训练集中的每个短文本进行垃圾特征分析,得到每个短文本的分析信息。

[0076] 具体地,对于训练集中已区分为垃圾内部,或非垃圾文本的每个短文本,进行垃圾特征分析,得到该短文本的分析信息,具体包括如下任一信息,或如下信息的任意组合:是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息,二元词的占比信息、不同词性词汇搭配比例(比如名词与动词的数量比例信息)、和标点符号与名词的数据比例信息等。

[0077] 其中,针对训练集中每个短文本,该短文本的分析信息可以是在得到该短文本的词语集合之前的预处理过程中提取得到的,也可以是在得到该短文本的词语集合之后得到的。

[0078] 上述的联系方式特征具体可以是一串具有联系意义的数字或字符,例如,电话号码、QQ号码、或URL(Uniform Resource Locator,统一资源定位符)等;通常而言,有些垃圾文本的目的是为了获取私利,而要留下联系方式;因此,短文本中是否有联系方式可以作为判定是否为垃圾文本的一个重要的判定特征。

[0079] 干扰性符号具体可以是不常用的符号,例如,“¥”等;有的垃圾文本为了避免关键词的过滤,采用一些不常用的符号来进行关键词的分隔;因此,统计短文本中干扰性符号出现的比例可以作为判定是否为垃圾文本的一个判定特征。

[0080] 词语间的转移概率指的是相邻两个词语的搭配概率、和相邻两个词语的类型的搭配概率,例如,“垃圾识别”短文本中“垃圾”与“识别”是正常搭配,对应存在一个搭配概率,“垃圾”为名词,“识别”为动词,名词与动词搭配的概率较大;

[0081] 一元词具体可以是单个词语;

[0082] 二元词具体可以是2个词语组成的惯用语、俚语或成语。

[0083] S103:对于训练集中每个短文本,确定该短文本的分析信息的特征值,以及该短文本的词语集合中每个词语的特征值。

[0084] 本步骤中,根据得到的训练集中每个短文本的词语集合,针对每个短文本的词语集合中的每个词语,计算该词语在该短文本中的TF(Term Frequency,词频)值;计算该词语在训练集中的IDF(Inverse Document Frequency,逆向文件频率)值;并根据如下公式1计算出该词语的特征值:

[0085] $\log(TF+1.0)\times IDF$ (公式1)

[0086] 计算得到的词语的特征值通常为0～100之间的数值。

[0087] 本步骤中,针对该训练集中每个短文本,根据得到的该短文本的分析信息,判断该短文本是否包含联系方式特征的信息,若是,则设定是否包含联系方式特征的信息的特征值为1(或0);否则,设定为0(或1);

[0088] 将统计出的干扰性符号占该短文本字符的比例作为该干扰性符号的占比信息的特征值;

[0089] 将统计出的生僻字占该短文本字符的比例作为该生僻字的占比信息的特征值;

[0090] 将统计出的繁体字符占该短文本字符的比例作为该繁体字符的占比信息的特征值;

[0091] 将得到的词语间的转移概率作为该词语间的转移概率的特征值；

[0092] 将统计出的名词占该短文本字符的比例作为该名词的占比信息的特征值；

[0093] 将统计出的动词占该短文本字符的比例作为该动词的占比信息的特征值；

[0094] 将统计出的标点符号占该短文本字符的比例作为该标点符号的占比信息的特征值；

[0095] 将统计出的一元词占该短文本字符的比例作为该一元词的占比信息的特征值；

[0096] 将统计出的二元词占该短文本字符的比例作为该二元词的占比信息的特征值；

[0097] 将统计出的该短文本中的名词与动词的数量比例作为该名词与动词的数量比例信息的特征值；

[0098] 将统计出的该短文本中的标点符号与名词的数量比例作为该符号与名词的数量比例信息的特征值；

[0099] 考虑到上述的各类型占比信息、数量比例信息和转移概率的特征值通常是0～1之间的数值，为了使计算方便，作为一种更优的实施方式，还可对确定出的该训练集中每个短文本的分析信息的特征值进行归一化，得到所述特征值的归一化数值：针对训练集中的每个短文本，将计算得到的是否包含联系方式特征的信息的特征值乘以100，得到是否包含联系方式特征的信息的特征值的归一化数值：0或者100；

[0100] 将统计得到的干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息，二元词的占比信息、名词与动词的数量比例信息、和标点符号与名词的数量比例信息的特征值分别乘以100，分别得到该短文本的各类型的占比、数量比例和转移概率的特征值的归一化数值：0～100之间的数值。

[0101] S104：对于训练集中的每个短文本，对该短文本的词语集合中每个词语的特征值，以及该短文本的分析信息的特征值，求取类别区分度后，选取特征元素集合中的特征元素。

[0102] 具体地，对于训练集中的每个短文本，对该短文本的词语集合中每个词语的特征值，可以采用AUC算法求取每个词语的类别区分度；词语的类别区分度可以反映出该词语对短文本进行垃圾文本或非垃圾文本的类别区分的贡献程度。

[0103] 对于训练集中的每个短文本，根据该短文本的分析信息的特征值或归一化后的特征值求取分析信息的类别区分度；分析信息的类别区分度可以反映出该分析信息对短文本进行垃圾文本或非垃圾文本的类别区分的贡献程度。对于特征值为离散数值的分析信息，可采用卡方检验算法求取类别区分度；对于特征值为连续数值的分析信息，可以采用AUC(Area Under Curve，曲线下面积)算法求取类别区分度。

[0104] 对于训练集中的每个短文本，在计算出该短文本的分析信息的类别区分度，和该短文本的词语集合中每个词语的类别区分度后，将类别区分度大于设定阈值的词语，以及分析信息作为特征元素集合中的特征元素。上述的设定阈值可由技术人员根据经验设置，对于特征值为离散数值和连续数值的不同情况，设置的设定阈值可以不同：例如，对于特征值为离散数值的情况可以设置设定阈值为10，对于特征值为连续数值的连续数值的可以设置设定阈值为0.7。

[0105] S105：针对训练集中的每个短文本，生成该短文本的词语特征向量。

[0106] 本步骤中，针对训练集中的每个短文本，生成该短文本的词语特征向量将该短文

本的分析信息以及词语集合中每个词语分别与特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量。

[0107] 更优地，也可以是针对训练集中的每个短文本，生成该短文本的词语特征向量将该短文本的分析信息以及词语集合中每个词语分别与特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的归一化后的特征值，生成该短文本的词语特征向量。

[0108] 具体地，针对训练集中的每个短文本，该短文本的词语特征向量中各维向量元素分别为特征元素集合中的各特征元素，其中，有的向量元素为该短文本的分析信息，则将该分析信息的特征值或归一化后的特征值作为该向量元素的值；有的向量元素为该短文本的词语集合中的词语，则将该词语的特征值或归一化后的特征值作为该向量元素的值；其它的向量元素值则为空，或0。

[0109] S106：根据得到的训练集中每个短文本的词语特征向量，构建分类模型。

[0110] 本步骤中，可以运用SVM分类算法、或贝叶斯分类算法、或决策树分类算法、或最大熵分类算法，根据训练集中各短文本的词语特征向量训练出分类模型。具体地，可以结合训练集中短文本的数量（即样本数量），选择运用一个合适的算法，根据训练集中各短文本的词语特征向量训练出分类模型。

[0111] 如何根据训练集中各短文本的词语特征向量训练出分类模型的具体方法为本领域技术人员所熟知，此处不再赘述。

[0112] 事实上，上述步骤S101与S102之间没有严格的先后顺序，可以并行执行或先执行步骤S102再执行步骤S101。

[0113] 在训练阶段构建出分类模型后，可以在识别阶段根据构建出的分类模型，对待判定的短文本进行垃圾识别；本发明实施例提供的短文本垃圾识别方法的流程图如图2所示，具体步骤包括：

[0114] S201：对待判定短文本进行分词，得到该待判定短文本的词语集合。

[0115] 具体地，对于待判定短文本进行分词：将该短文本中连续的字序列划分为一个个词语；在划分出的词语中，去除掉没有实际意义的虚词（如标点、组动词、语气词、叹词、拟声词等）；剩余的词语构成该短文本的词语集合。

[0116] S202：对该待判定短文本进行垃圾特征分析，得到该待判定短文本的分析信息。

[0117] 具体地，对于该待判定短文本，进行垃圾特征分析，得到该短文本的分析信息，具体包括如下任一信息，或如下信息的任意组合：是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息，二元词的占比信息、不同词性词汇搭配比例、和标点符号与名词的数据比例信息等。

[0118] 其中，该待判定短文本的分析信息可以是在得到该待判读短文本的词语集合之前的预处理过程中提取得到的，也可以是在得到该待判读短文本的词语集合之后得到的。

[0119] S203：确定待判定短文本的特征元素。

[0120] 具体地，将待判定短文本的分析信息以及词语集合中每个词语分别与上述的特征元素集合中的特征元素进行比较，将与所述特征元素集合中的特征元素相匹配的词语或分

析信息作为该待判定短文本的特征元素。

[0121] S204:根据该待判定短文本的特征元素的特征值,生成所述待判定短文本的词语特征向量。

[0122] 本步骤中,对于该待判定短文本的作为特征元素的词语,计算特征值:计算该词语在该短文本中的TF值,计算该词语在训练集中的IDF值;并根据上述公式1计算出该词语的特征值。

[0123] 本步骤中,对于该待判定短文本的作为特征元素的分析信息,计算特征值:

[0124] 判断作为特征元素的分析信息中是否包含联系方式特征的信息,若是,则设定是否包含联系方式特征的信息的特征值为1(或0);否则,则设定为0(或1);

[0125] 判断作为特征元素的分析信息中是否包含干扰性符号的占比信息;若是,则将统计出的干扰性符号占该短文本字符的比例作为该干扰性符号的占比信息的特征值;

[0126] 判断作为特征元素的分析信息中是否包含生僻字的占比信息;若是,则将统计出的生僻字占该短文本字符的比例作为该生僻字的占比信息的特征值;

[0127] 判断作为特征元素的分析信息中是否包含繁体字符的占比信息;若是,则将统计出的繁体字符占该短文本字符的比例作为该繁体字符的占比信息的特征值;

[0128] 判断作为特征元素的分析信息中是否包含词语间的转移概率;若是,则将得到的词语间的转移概率作为该词语间的转移概率的特征值;

[0129] 判断作为特征元素的分析信息中是否包含名词的占比信息;若是,则将统计出的名词占该短文本字符的比例作为该名词的占比信息的特征值;

[0130] 判断作为特征元素的分析信息中是否包含动词的占比信息;若是,则将统计出的动词占该短文本字符的比例作为该动词的占比信息的特征值;

[0131] 判断作为特征元素的分析信息中是否包含标点符号的占比信息;若是,则将统计出的标点符号占该短文本字符的比例作为该标点符号的占比信息的特征值;

[0132] 判断作为特征元素的分析信息中是否包含一元词的占比信息;若是,则将统计出的一元词占该短文本字符的比例作为该一元词的占比信息的特征值;

[0133] 判断作为特征元素的分析信息中是否包含二元词的占比信息;若是,则将统计出的二元词占该短文本字符的比例作为该二元词的占比信息的特征值;

[0134] 判断作为特征元素的分析信息中是否包含名词与动词的数量比例信息;若是,则将统计出的该短文本中的名词与动词的数量比例作为该名词与动词的数量比例信息的特征值;

[0135] 判断作为特征元素的分析信息中是否包含符号与名词的数量比例信息;若是,则将统计出的该短文本中的标点符号与名词的数量比例作为该符号与名词的数量比例信息的特征值。

[0136] 考虑到该待判定短文本的各类型占比信息、数量比例信息和转移概率的特征值通常为0～1之间的数值,为了使计算方便,作为一种更优的实施方式,还可对该待判定短文本的分析信息的特征值进行归一化,得到所述特征值的归一化数值:针对该待判定短文本,将计算得到的是否包含联系方式特征的信息的特征值乘以100,得到是否包含联系方式特征的信息的特征值的归一化数值:0或者100;

[0137] 将统计得到的干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、

词语间的转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息，二元词的占比信息、名词与动词的数量比例信息、和标点符号与名词的数量比例信息的特征值分别乘以100，分别得到该待判定短文本的各类型的占比、数量比例和转移概率的特征值的归一化数值：0～100之间的数值。

[0138] 本步骤中，根据该待判定短文本的特征元素的特征值或归一化后的特征值，生成所述待判定短文本的词语特征向量；待判定短文本的词语特征向量中各维向量元素分别为特征元素集合中的各特征元素，其中，有的向量元素为该待判定短文本的分析信息，则将该分析信息的特征值或归一化后的特征值作为该向量元素的值；有的向量元素为该待判定短文本的词语集合中的词语，则将该词语的特征值或归一化后的特征值作为该向量元素的值；其它的向量元素值则为空，或0。

[0139] S205：根据该待判定短文本的词语特征向量，以及分类模型确定该待判定短文本是否为垃圾文本。

[0140] 如何根据待判定短文本的词语特征向量，以及分类模型确定该待判定短文本是否为垃圾文本为本领域技术人员所熟知的技术，此处不再赘述。

[0141] 事实上，上述步骤S201与S202之间没有严格的先后顺序，可以并行执行或先执行步骤S202再执行步骤S201。

[0142] 基于上述的建模方法，本发明的实施例提供了一种建模装置，内部结构框图如图3所示，具体包括：特征提取模块301、特征向量确定模块302和分类模型构建模块303、特征元素集合确定模块304。

[0143] 特征提取模块301用于对于训练集中已区分为垃圾文本，或非垃圾文本的每个短文本，进行分词后得到该短文本的词语集合，并对该短文本进行垃圾特征分析得到该短文本的分析信息；其中，短文本的分析信息可以包括如下任一信息，或如下信息的任意组合：

[0144] 是否包含联系方式特征的信息、干扰性符号的占比信息、生僻字的占比信息、繁体字符的占比信息、词语间的转移概率、前后词的词性间转移概率、名词的占比信息、动词的占比信息、标点符号的占比信息、一元词的占比信息、二元词的占比信息、不同词性词汇搭配比例、标点符号与名词的数量比例信息。

[0145] 特征元素集合确定模块304用于针对所述训练集中的每个短文本，从特征提取模块301获取该短文本的词语集合和分析信息，并计算该短文本的词语集合中每个词语的特征值，计算该短文本的分析信息的特征值后，对计算出的特征值求取类别区分度；将类别区分度大于设定阈值的词语，以及分析信息作为特征元素集合中的特征元素；

[0146] 具体地，特征元素集合确定模块304可以针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与所述特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的归一化后的特征值，生成该短文本的词语特征向量。

[0147] 特征向量确定模块302用于针对所述训练集中的每个短文本，将该短文本的分析信息以及词语集合中每个词语分别与特征元素集合确定模块304所得到的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成该短文本的词语特征向量；

[0148] 分类模型构建模块303用于根据所述特征向量确定模块302确定出的所述训练集

中各短文本的词语特征向量，构建分类模型。具体地，分类模型构建模块303运用SVM分类算法、或贝叶斯分类算法、或决策树分类算法、或最大熵分类算法，根据所述训练集中各短文本的词语特征向量训练出所述分类模型。

[0149] 基于上述的短文本垃圾识别方法，本发明实施例提供了一种短文本垃圾识别装置，内部结构框图如图3所示，具体包括：特征提取模块401、特征向量确定模块402、和垃圾识别模块403。

[0150] 其中，特征提取模块401用于对于待判定短文本进行分词后得到词语集合，并对所述待判定短文本进行垃圾特征分析得到分析信息。短文本的分析信息的具体内容前述已经介绍，此处不再赘述。

[0151] 特征向量确定模块402用于从特征提取模块401获取待判定短文本的分析信息以及词语集合，将待判定短文本的分析信息以及词语集合中每个词语分别与预先确定的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的特征值，生成所述待判定短文本的词语特征向量；

[0152] 具体地，特征向量确定模块402可以将所述待判定短文本的分析信息以及词语集合中每个词语分别与预先确定的特征元素集合中的特征元素进行比较，根据与所述特征元素集合中的特征元素相匹配的词语或分析信息的归一化后的特征值，生成所述待判定短文本的词语特征向量。

[0153] 垃圾识别模块403用于从特征向量确定模块402获取所述待判定短文本的词语特征向量后，根据所述待判定短文本的词语特征向量，以及预先训练出的分类模型，确定所述待判定短文本是否为垃圾文本。

[0154] 本发明的技术方案中，训练集中的各短文本的词语特征向量，和待判定短文本的词语特征向量，都包含了扩充的分析信息的特征值，根据包含了扩充的分析信息的特征值的词语特征向量，对待判定短文本进行垃圾识别，提高了识别垃圾文本的识别率和识别准确率。

[0155] 以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以作出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

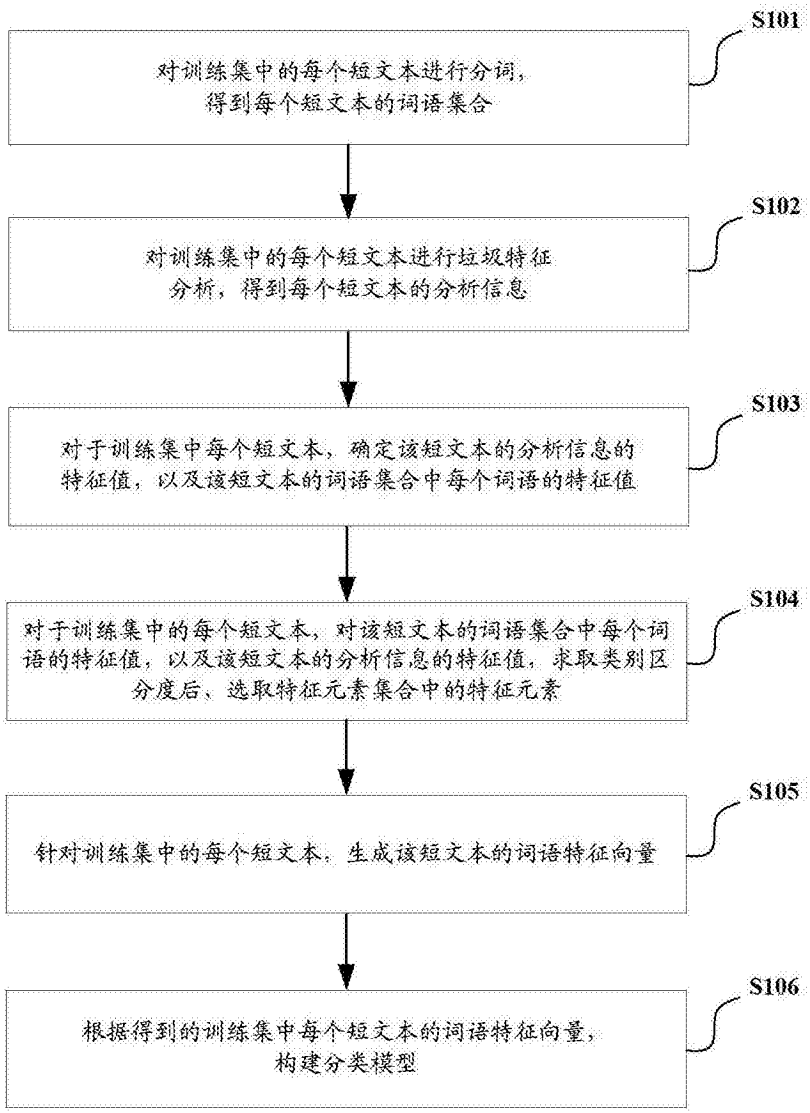

S101
对训练集中的每个短文本进行分词，得到每个短文本的词语集合
S102
对训练集中的每个短文本进行垃圾特征分析，得到每个短文本的分析信息
S103
对于训练集中每个短文本，确定该短文本的分析信息的特征值，以及该短文本的词语集合中每个词语的特征值
S104
对于训练集中的每个短文本，对该短文本的词语集合中每个词语的特征值，以及该短文本的分析信息的特征值，求取类别区分度后、选取特征元素集合中的特征元素
S105
针对训练集中的每个短文本，生成该短文本的词语特征向量
S106
根据得到的训练集中每个短文本的词语特征向量，构建分类模型

图1

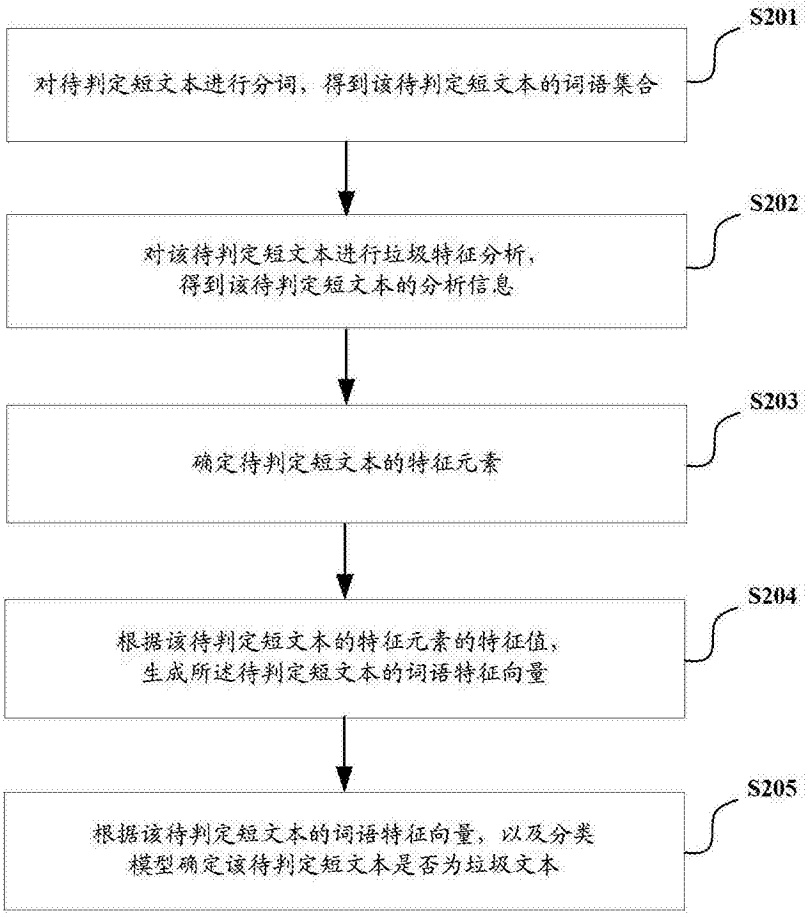

S201
对待判定短文本进行分词，得到该待判定短文本的词语集合
S202
对该待判定短文本进行垃圾特征分析，
得到该待判定短文本的分析信息
S203
确定待判定短文本的特征元素
S204
根据该待判定短文本的特征元素的特征值，
生成所述待判定短文本的词语特征向量
S205
根据该待判定短文本的词语特征向量，以及分类
模型确定该待判定短文本是否为垃圾文本

图2

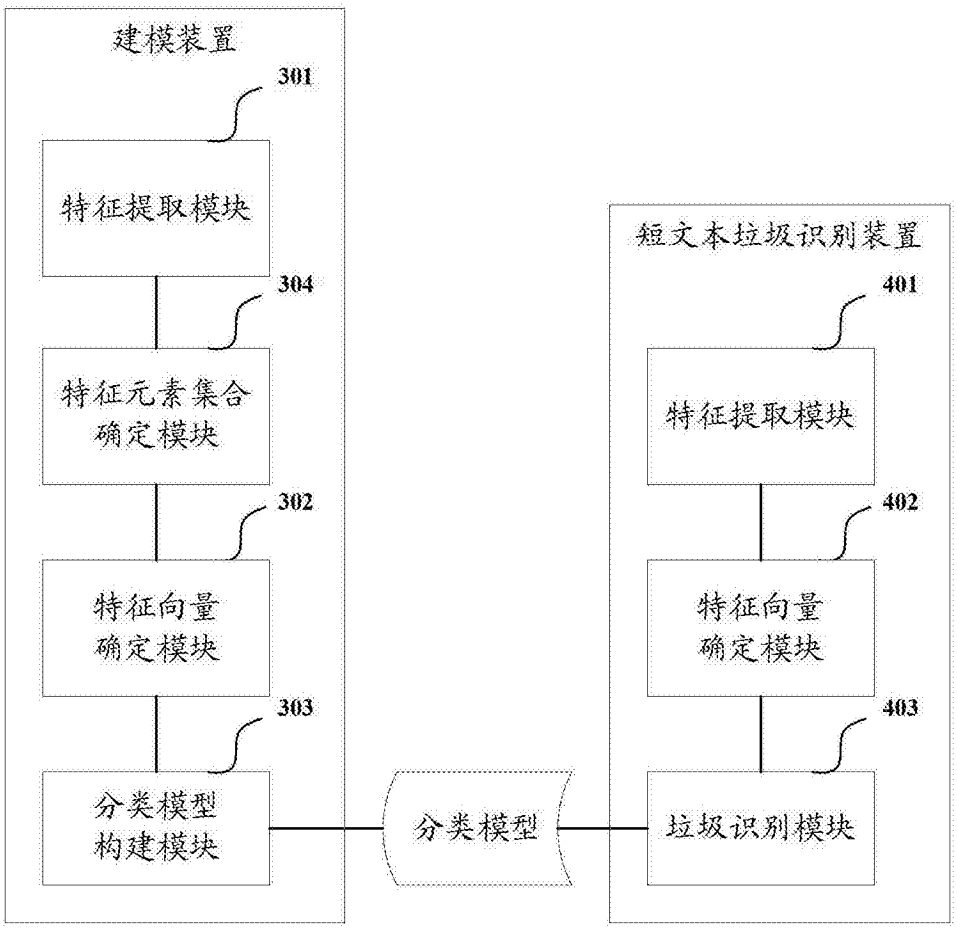

建模装置
301
特征提取模块
304
特征元素集合确定模块
302
特征向量确定模块
303
分类模型构建模块
分类模型
短文本垃圾识别装置
401
特征提取模块
402
特征向量确定模块
403
垃圾识别模块

图3