(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115409135 B
(45)授权公告日 2023.02.03

(21)申请号 202211365338.8

(22)申请日 2022.11.03

(65)同一申请的已公布的文献号
申请公布号 CN 115409135 A

(43)申请公布日 2022.11.29

(73)专利权人 南昌惠联网络技术有限公司
地址 330000 江西省南昌市红谷滩区红角洲学府大道899号江西慧谷-红谷创意产业园1号楼A栋六楼A6-04室

(72)发明人 洪葵 胡盛利 钟天生 黄隆辉 龚晖 周涛 熊新宇 薛萌

(74)专利代理机构 南昌明佳知识产权代理事务所(普通合伙) 36132
专利代理师 熊赣荣

(51)Int.Cl.
G06F 18/2415(2023.01)
G06N 3/0464(2023.01)
G06N 3/0442(2023.01)
G06N 3/084(2023.01)
G06F 40/30(2020.01)
G06F 16/35(2019.01)

(56)对比文件
CN 110134786 A,2019.08.16
CN 111209402 A,2020.05.29
CN 112949713 A,2021.06.11

审查员 曹宁

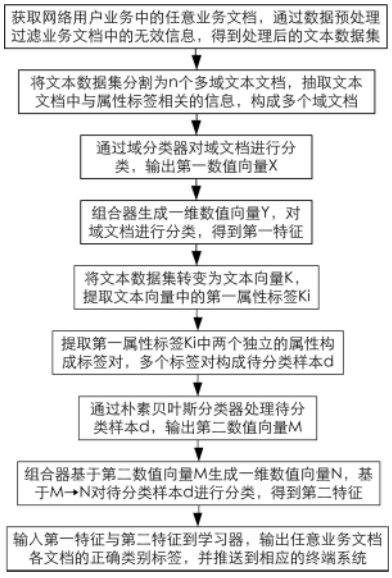
权利要求书2页 说明书6页 附图4页

(54)发明名称

一种网络业务文档的分类管理方法

(57)摘要

本发明公开了一种网络业务文档的分类管理方法。该分类管理方法通过多域分类与加权朴素贝叶斯分类并行的方式对网络业务文档进行特征提取与分类。首先,将业务文档进行数据预处理得到为文本数据集,对文本数据集进行分割处理成域文档后,通过域分类器得到第一特征。其次,将文本数据集通过空间向量模型转变为文本向量,获取属性标签,以属性相似的标签构成标签对,若干个标签对组成待分类样本,并按照文本数据集的特点与词频进行属性加权。最后,通过朴素贝叶斯分类器得到第二特征,第一特征与第二特征共同执行分类决策,并将分类决策结果推送至相应的终端系统。



CN 115409135 B

1.一种网络业务文档的分类管理方法,其特征在于,包括以下步骤:

步骤1:获取网络用户业务中的任意业务文档,通过数据预处理过滤业务文档中的无效信息,得到处理后的文本数据集;

步骤2:将文本数据集分割为n个多域文本文档,抽取文本文档中与属性标签相关的信息,构成多个域文档;

步骤3:通过域分类器对域文档进行处理,输出基于该域文档的第一数值向量$X=(SE_1,SE_2,\cdots,SE_n)$,$X\in R_n$;

步骤4:组合器基于第一数值向量X生成一维数值向量Y=(SE),基于X→Y对域文档进行分类,得到第一特征,

域分类器只处理唯一一个的域文档,域分类的域分类模型抽取域文档中的属性特征,域分类模型为各个域文档计算其置信度J,J∈R,置信度J能够作为域文档属于噪声标签的似然程度,置信度J在数值上与第一数值向量X的一维数值向量Y=(SE)相等,每个域分类模型对应唯一一个域文档;

步骤5:将文本数据集通过向量空间模型转变为文本向量K,提取文本向量K中的第一属性标签K_i,i=1,2,…,n;

步骤6:提取第一属性标签K_i任意两个独立的属性构成标签对,所述标签对构成待分类样本d;

步骤7:通过朴素贝叶斯分类器对待分类样本d进行处理,输出结果处理后输出基于该文档的第二数值向量$M=(SR_1,SR_2,\cdots,SR_n)$,$M\in R_n$;

步骤8:组合器基于第二数值向量M生成一维数值向量N=(SR),基于M→N对待分类样本d进行分类,得到第二特征;

步骤9:输入第一特征与第二特征到学习器,输出任意业务文档各文档的正确类别标签,并推送到相应的终端系统。

2.根据权利要求1所述的网络业务文档的分类管理方法,其特征在于,所述数据预处理方法剔除任意业务文档中的冠词、连词、空格字符、人称代词、形容词,得到文本数据集,并通过特征提取,分离噪声标签。

3.根据权利要求2所述的网络业务文档的分类管理方法,其特征在于,特征提取根据包含空间复杂度、时间复杂度与提取准确率的约束条件获取文本数据集的特征值,按照网络用户业务的文本类别统计特征值,根据特征值的大小构建特征词集合,其中,文本类别q中词c的特征值v=FF(c)*DF(c)*[1/QF(c)],其中,FF(c)为词c在文本类别q中最大出现频率,DF(c)为文本类别q中出现词c的文档总数量,QF(c)代表文本数据集中出现词c的类别总数量。

4.根据权利要求1所述的网络业务文档的分类管理方法,其特征在于,通过选定的属性标签将文本数据集分割为多域文本文档,所述属性标签是文本数据集的分类标准。

5.根据权利要求1所述的网络业务文档的分类管理方法,其特征在于,第一特征为组合器对第一数值向量X经过处理得到的二值结果,该第一特征包含属性标签与噪声标签。

6.根据权利要求1所述的网络业务文档的分类管理方法,其特征在于,第一属性标签包含有文本数据集中的全部属性特征,第一属性标签为属性特征中的词在高维空间内映射所产生的集合。

7.根据权利要求1所述的网络业务文档的分类管理方法,其特征在于,第二特征为组合器对第二数值向量X经过处理得到的二值结果,该第二特征包含语义标签与噪声标签。

一种网络业务文档的分类管理方法

技术领域

[0001] 本发明涉及文档处理技术，尤其涉及一种网络业务文档的分类管理方法。

背景技术

[0002] 提取业务文档的文本内容，根据文本属性与特征进行分类，是网络平台自动处理用户文件的有效手段。现有技术下，文本分类技术大部分采用朴素贝叶斯分类器进行集中处理，以提取关键词的方式对文本进行分类管理。例如，文献《面向互联网文本的大规模层次分类技术研究》(何力，博士学位论文，2014)中提到的基于贪心策略的文本数据特征提取的方法，对文本信息分为多个阶段进行层次化处理，提高了分类的精度，大大减少了噪声标签。再例如，CN106897428B提到的构建特征词集合，评估特征词与标准相关度的加权朴素贝叶斯分类学习方法，都属于典型的集中分类方式。网络业务文档具有多样化的特点，文本形式受到用户学历、表述方式、文本属性等多方面的限制。现有技术希望利用更加高效的文本分类管理方法，提取网络用户业务文档中包含的有效文本信息，实现数据的精准推送。

发明内容

[0003] 本发明提出了一种网络业务文档的分类管理方法，通过多域属性分类与加权朴素贝叶斯分类并行的方法进行文本特征提取。本发明对网络用户业务文档信息进行多重维度的分割，以分割得到的域文档进行横向分类，获取第一特征，以属性标签进行纵向向分类，获取第二特征。第一特征与第二特征通过学习器为任意业务文档进行分类，并反馈至对应的终端系统。

[0004] 本申请的发明目的可通过以下技术方案实现：

[0005] 一种网络业务文档的分类管理方法，包括以下步骤：

[0006] 步骤1：获取网络用户业务中的任意业务文档，通过数据预处理过滤业务文档中的无效信息，得到处理后的文本数据集；

[0007] 步骤2：将文本数据集分割为n个多域文本文档，抽取文本文档中与属性标签相关的信息，构成多个域文档；

[0008] 步骤3：通过域分类器对域文档进行处理，输出基于该域文档的第一数值向量$X=(SE_1,SE_2,\cdots,SE_n)$，$X\in R_n$；

[0009] 步骤4：组合器基于第一数值向量X生成一维数值向量Y=(SE)，基于X→Y对域文档进行分类，得到第一特征；

[0010] 步骤5：将文本数据集通过向量空间模型转变为文本向量K，提取文本向量K中的第一属性标签K_i，(i=1,2,…,n)；

[0011] 步骤6：提取第一属性标签K_i任意两个独立的属性构成标签对，所述标签对构成待分类样本d；

[0012] 步骤7：通过朴素贝叶斯分类器对待分类样本d进行处理，输出结果处理后输出基于该文档的第二数值向量$M=(SR_1,SR_2,\cdots,SR_n)$，$M\in R_n$；

[0013] 步骤8:组合器基于第二数值向量M生成一维数值向量N=(SR),基于M→N对待分类样本d进行分类,得到第二特征;

[0014] 步骤9:输入第一特征与第二特征到学习器,输出任意业务文档各文档的正确类别标签,并推送到相应的终端系统。

[0015] 在本发明中,所述数据预处理方法剔除任意业务文档中的冠词、连词、空格字符、人称代词、形容词,得到文本数据集,并通过特征提取,分离噪声标签。

[0016] 在本发明中,特征提取根据包含空间复杂度、时间复杂度与提取准确率的约束条件获取文本数据集的特征值,按照网络用户业务的文本类别统计特征值,根据特征值的大小构建特征词集合,其中,文本类别q中词c的特征值v=FF(c)*DF(c)*[1/QF(c)],其中,FF(c)为词c在文本类别q中最大出现频率,DF(c)为文本类别q中出现词c的文档总数量,QF(c)代表文本数据集中出现词c的类别总数量。

[0017] 在本发明中,通过选定的属性标签将文本数据集分割为多域文本文档,所述属性标签是文本数据集的分类标准。

[0018] 在本发明中,域分类器只处理唯一一个的域文档,域分类的域分类模型抽取域文档中的属性特征,域分类模型为各个域文档计算其置信度J,J∈R,置信度J能够作为域文档属于噪声标签的似然程度,每个域分类模型对应唯一一个域文档。

[0019] 在本发明中,第一特征为组合器对第一数值向量X经过处理得到的二值结果,该第一特征包含属性标签与噪声标签。

[0020] 在本发明中,第一属性标签包含有文本数据集中的全部属性特征,第一属性标签为属性特征中的词在高维空间内映射所产生的集合。

[0021] 在本发明中,待分类样本d为多个相似的标签对,通过对平台中文本数据集的各个属性特征进行权重提取,包括词形、词距、词长以及词序进行超参数预设,分别得到α_1、α_2、α_3、α_4,计算各个属性标签的综合相似度,构成待分类样本d。

[0022] 在本发明中,第二特征为组合器对第二数值向量X经过处理得到的二值结果,该第二特征包含语义标签与噪声标签。

[0023] 实施本发明的这种网络业务文档的分类管理方法,具有以下有益效果:本发明采用了横向域分类与纵向加权朴素贝叶斯分类相结合的文本分类与学习方法,对于不同区域组织记录标签数据命名规则不同、表述形式不一致、语法或非法字符错误的情况有着显著的优化效果。除此之外,传统的集中化文本分类方法以神经网络或机器学习为基础,面对网络业务文档这类复杂度较高、属性较多的文本内容,实际文本分类的效果较差,而本发明提供了域分类器作为限制,加权朴素贝叶斯分类中提取的待分类样本为属性标签中的综合相似度较高的标签对,减少了重复且不必要的分类过程,有效提高了文本属性分类的效率。这种面向网络业务文档的横向属性提取方法,获取多个属性特征为分类器提供支持,有效提高了整体分类模型的鲁棒性。

附图说明

[0024] 图1为本发明的这种网络业务文档的分类管理方法的原理图;

[0025] 图2为本发明的这种网络业务文档的分类管理方法的流程图;

[0026] 图3为本发明的域分类器的文本处理过程示意图;

[0027] 图4为第一数值向量在域分类器中分类原理的示意图；

[0028] 图5为本发明的贝叶斯分类器的文本处理过程示意图；

[0029] 图6为本发明的属性加权朴素贝叶斯算法对待分类样本d进行处理的流程图。

具体实施方式

[0030] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述。

[0031] 网络业务文档是指各类业务平台中用户输入的文本内容，各个网站上用户反馈文本内容被集成到网络平台的后台数据库中。大型网络平台需要短时间内处理处理大量的业务文本内容，由于业务文本内容具有较多的属性，涉及的相关属性信息是实时变化的，传统的机器分类与神经网络分类方法容易受到用户表述方式与表述水平的影响，提取用户反馈信息的文本特征较为困难。参照图1，在具体实施过程中，将文本属性作为主观影响因素，将用户表述作为客观影响因素，利用本发明的多域文本属性分类学习的方法，多维度提取网络业务文档信息中的特征标签，利用域分类器与朴素贝叶斯分类器进行多重分类。参照图2，本发明实施的一种网络业务文档的分类管理方法具体包括以下几个步骤。

[0032] 步骤1：通过API接口获取网络业务中的业务文档，业务文档经过数据预处理过滤业务文档中的无效信息，得到处理后的文本数据集。在进行数据预处理的过程中，具体分为个4个步骤进行，分别是文本标记筛除、文本分词、文本词根提取、文本稀有词与冠词筛除。在本实施例中，文本标记筛除包括标点符号、数字、大小写统一；文本分词的目的在于确定特征提取的基本处理单位，通过ICTCLAS系统接口可以进行自动化的文本分词处理；文本词根提取过程主要将词根相同，词形态不同的词语进行拟合，构成相同的语义单位；文本稀有词与冠词包括连词、代词、副词、助词，排除不具有参考意义的词语内容。进一步的，由于业务文档中包含有不具有参考价值的噪声标签，在数据预处理阶段还需要通过向量空间模型分离噪声标签。

[0033] 步骤2：将文本数据集分割为n个多域文本文档，抽取文本文档中与属性标签相关的信息，构成多个域文档。所述属性标签为业务文档的信息分类。

[0034] 本实施例优选的中文文本分类数据集THUCNews中以新闻文档为源数据，按照属性特征提取进行分类的工具包，选取政治、社会两个标签作为属性标签的人为定义的分类标准。属性标签的抽取采用正则式抽取原则对多域文本文档进行抽取，构建多个域文档。

[0035] 本实施例中，特征提取根据包含空间复杂度、时间复杂度与提取准确率的约束条件获取文本数据集的特征值，按照网络用户业务的文本类别统计特征值，根据特征值的大小构建特征词集合，其中，文本类别q中词c的特征值v=FF(c)*DF(c)*[1/QF(c)]，其中，FF(c)为词c在文本类别q中最大出现频率，DF(c)为文本类别q中出现词c的文档总数量，QF(c)代表文本数据集中出现词c的类别总数量。

[0036] 步骤3：参照图3，域分类器对域文档进行处理，输出基于该域文档的第一数值向量$X=(SE_1,SE_2,\cdots,SE_n)$，$X\in R_n$。所述域分类器只处理唯一一个的域文档，域分类中的分类原则为域分类模型，域分类模型对域文档中的属性特征进行抽取，并自动进行训练与更新，域分类模型为各个域文档计算其置信度J，$J\in R$，置信度J能够作为域文档属于噪声标签的似然程度，置信度J在数值上与第一数值向量X的一维数值向量Y=(SE)相等，每个域分类模型对

应唯一一个域文档。

[0037] 步骤4:组合器基于第一数值向量X生成一维数值向量Y=(SE),基于X→Y对域文档进行分类,得到第一特征。参照图4,第一数值向量X为多个域分类器中输出的多个数值向量构成的集合,第一数值向量X通过二值类别标签,即属性标签与噪声标签组合的形式生成以Y=(SE)的分类结果,当类别二值类别标签L为噪声标签时,输出Y=1;当类别二值标签L为非噪声标签时,输出Y=0。当类别标签位置,第一数值向量进行分类预测时,当二值类别标签L为噪声标签时,输出Y={Y丨0.5＜Y≤1};若二值类别标签L为非噪声标签时,输出{Y丨0≤Y≤0.5}。支持向量模型的分类通过支持向量机构建,SVM通过核函数的映射方法合理解决非线性的分类问题,尤其是对于数值类的向量分类问题,单独的域分类器均可以生成对应的一维数值向量,第一特征为组合器对第一数值向量X经过处理得到的二值结果,包含噪声标签与非噪声标签。

[0038] 步骤5:将文本数据集通过向量空间模型转变为文本向量K,提取文本向量K中的第一属性标签K_i,(i=1,2,…,n)。对于特定的文本数据集,赋予一特征识别的属性序列W=(W_1,W_2,…,W_n),在分类网络中,通过对文本上下文特征与局部特征的提取,将所取特征中的各个词语映射到高维空间,通过语言模型提取词,从而得到文本向量。在本实施例中,所述语言模型采用优选的BERT系列语言模型。

[0039] 在本实施例中,所述分类网络包括词嵌入层、特征提取层、注意力层以及全连接层,第一属性标签的提取需要在特征提取层中进行,通过提取文本属性以及描述的上下文相关特征。优选的CNN模块中的多卷积核对文本向量中的局部特征进行提取,局部特征的集合为第一属性标签。本实施例中,局部特征ci=f(w_cg+b_c),其中,f(•)为非线性激活函数,w_c为CNN模块的卷积核,b_c为偏置项,g表示在文本向量K中,某一词向量在特定位置上所构成的向量矩阵。K_i={c_1,c_2,…,c_i},(i=1,2,…,n)。

[0040] 步骤6:提取第一属性标签K_i任意两个独立的属性构成标签对,所述标签对构成待分类样本d,d={w_1,w_2,…,w_p}。其中,待分类样本d为多个相似的标签对,通过对文本数据集的各个属性特征进行权重提取,包括词形、词距、词长以及词序进行超参数预设,分别得到α_1、α_2、α_3、α_4,按照权重计算各个属性标签的综合相似度,综合相似度高的单独成对。

[0041] 在本实施例中,第一属性标签K_i中任取两个标签K_1与K_2,计算K_1与K_2的词形、词距、词长以及词序。其中,词形相似度$S_s(K_1,K_2)=\frac{2|K_1\cap K_2|}{|K_1|+|K_2|}$,词序相似度$S_o(K_1,K_2)=1-\frac{re(K_1,K_2)}{|K_1+K_2|-1}$,词长相似度$S_l(K_1,K_2)=1-abs\frac{|K_1|-|K_2|}{|K_1|+|K_2|}$,词距相似度$S_d(K_1,K_2)=1-\frac{\mathrm{ed}(K_1,K_2)}{max(|K_1|,|K_2|)}$。其中,$|K_1\cap K_2|$代表标签K1与K2中包含共同字词的总数量,$|K_1|$与$|K_2|$分别表示$K_1$与$K_2$中包含字词数量,$re(K_1,K_2)$代表标签$K_1$对于$K_2$的逆序总字词量,$\mathrm{ed}(K_1,K_2)$代表 K_1转化为K_2的操作最少次数。根据设定的超参数值α_1、α_2、α_3、α_4,各个属性标签的综合相似度为:S=$S_s(K_1,K_2)*\alpha_1+S_o(K_1,K_2)*\alpha_2+S_l(K_1,K_2)*\alpha_3+S_d(K_1,K_2)*\alpha_4$。

[0042] 为了减少重复的多次计算,获得更加集中的属性标签综合相似度比较结果,本实施例优选的标签相似度矩阵S_L可有效提高综合相似度比较效率,$S_L=\begin{bmatrix}S_{11}&\cdots&S_{1n}\\ \vdots&\ddots&\vdots\\ S_{n1}&\cdots&S_{nn}\end{bmatrix}$。遍历$S_L$矩阵中的各个元素值,按照设定的标签标注阈值,评估是否为近似标签对。

[0043] 步骤7:参照图5,通过朴素贝叶斯分类器对待分类样本d进行处理,输出结果处理后输出基于该文档的第二数值向量M=(SR$_1$,SR$_2$,…,SR$_n$),M∈R$_n$。由于待分类样本d={w$_1$,w$_2$,…,w$_p$}中包含多个标签对,其先验概率计算效率较高,故本实施例优选的改进后的属性加权朴素贝叶斯算法通过对待分类样本d进行处理,参照图6,具体分为以下几个步骤:

[0044] 步骤71:获取待分类样本d的类标签u(d),计算待分类样本d个各个属性w$_p$与不同类标签u之间的距离相关系数,并计算出w$_p$属性的距离相关系数总和;

[0045] 步骤72:根据属性w$_p$的权值大小、先验概率、条件概率,并对待分类样本d的类标签u(d)进行分类;

[0046] 步骤73:类标签u(d)返回到待分类样本d。其中,改进后的属性加权朴素贝叶斯算法表达式:$U(d) = \text{argmax}[\log P(u) + \sum_{i=1}^{p} \zeta f_i \log P(w_p|u)]$,其中,ζ为属性加权值,f$_i$为选取的词在整个待分类样本d中的出现频率,P(u)为先验概率,$P(w_p|u)$为条件概率。

[0047] 在本实施例中,为了提高本方法对文本数据集处理的敏感程度,待分类样本d通过属性加权的方式进行朴素贝叶斯分类。任取一随机变量A={a$_1$,a$_2$,…,a$_n$},构成n个独立的条件属性,则随机变量A取值为a$_i$,(i=1,2,…,s;s为随机变量A的属性值个数)。对于任意的一随机变量B,构成m个独立的决策属性,随机变量B的取值为b$_j$,(j=1,2,…,t;t为随机变量B的属性标签个数)。通过A与B两个序列之间的相关系数作为加权值ζ,$\zeta = \left|\frac{Cov(A,B)}{\sqrt{D(A)*D(B)}}\right|$, $\zeta \in (0,1)$。其中,D(A)与D(B)分别为随机变量A与随机变量B的方差,Cov(A,B)为随机变量A与随机变量B的协方差。

[0048] 步骤8:组合器基于第二数值向量M生成一维数值向量N=(SR),基于M→N对待分类样本d进行分类,得到第二特征。第二特征为组合器对第二数值向量X经过处理得到的二值结果,包含语义标签与噪声标签。

[0049] 本实施例优选的提取语义标签的方法基于LSTM模块的双向结构实现,语句的双向语义特征需要建立在全局语义关系的基础上,将第二数值向量进行转换,获取文本中包含属性的上下文特征,获得更加丰富的语义局部特征。

[0050] 在神经网络的全连接层中,softmax层中通过将语义局部特征与文本上下文特征向量,通过拼接后转换为语义标签。进一步的,为了加强模型的自我学习能力,在LSTM提取语义标签的过程中,本实施例通过优选的传播计算交叉熵分类损失的算法为组合器的语义特征提取实现模型参数的更新,所述模型参数更新通过典型的反向传播算法实现,在经过多次迭代,且分类损失不再产生下降的情况下,选择收敛后的参数作为组合器使用的文本分类感知学习模型。

[0051] 步骤9:输入第一特征与第二特征到学习器,输出任意业务文档各文档的正确类别标签,并推送到相应的终端系统。其中,学习器能够接收多个方面的信息内容,并不仅限于特征值,组合器将第一特征与第二特征输入学习器后,学习器会请求用户为该业务文档进行反馈,对应的反馈传输到各个域分类器与朴素贝叶斯分类器。其中,第一特征与第二特征的反馈是独立的,第一特征反馈至域分类器,第二特征反馈至朴素贝叶斯分类器。在本实施例中,任意业务文档的标签至少包括诉求内容、所属区域、归口类型,按照标签推送至相应的终端系统。

[0052] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改，等同替换和改进等，均应包含在本发明的保护范围之内。

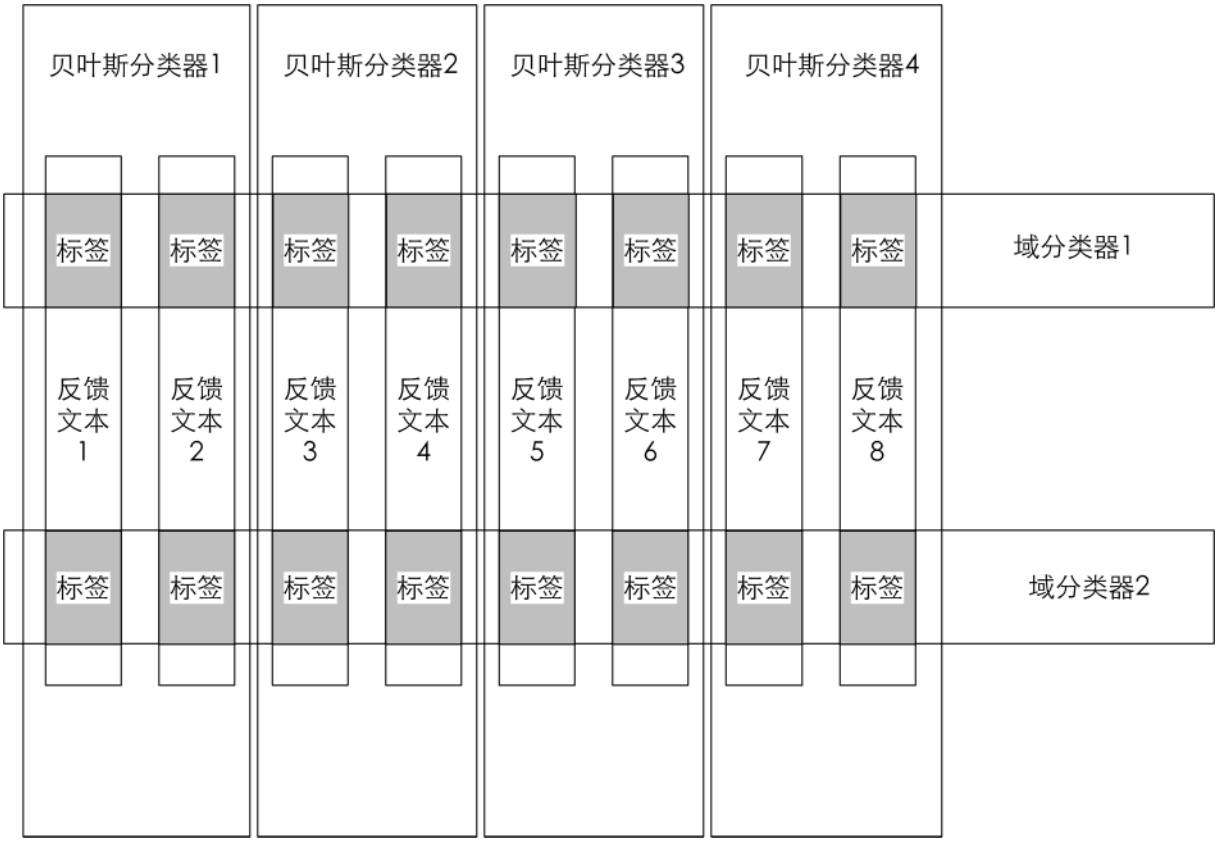

贝叶斯分类器1
贝叶斯分类器2
贝叶斯分类器3
贝叶斯分类器4
标签
域分类器1
反馈文本1
反馈文本2
反馈文本3
反馈文本4
反馈文本5
反馈文本6
反馈文本7
反馈文本8
域分类器2

图1

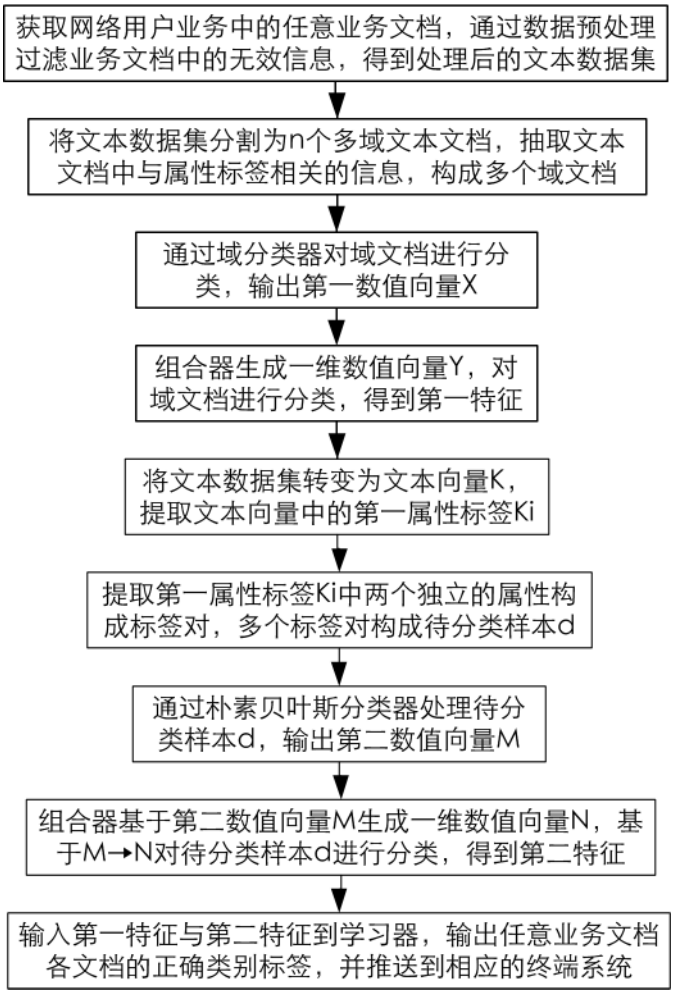

获取网络用户业务中的任意业务文档，通过数据预处理过滤业务文档中的无效信息，得到处理后的文本数据集
将文本数据集分割为n个多域文本文档，抽取文本文档中与属性标签相关的信息，构成多个域文档
通过域分类器对域文档进行分类，输出第一数值向量X
组合器生成一维数值向量Y，对域文档进行分类，得到第一特征
将文本数据集转变为文本向量K，提取文本向量中的第一属性标签Ki
提取第一属性标签Ki中两个独立的属性构成标签对，多个标签对构成待分类样本d
通过朴素贝叶斯分类器处理待分类样本d，输出第二数值向量M
组合器基于第二数值向量M生成一维数值向量N，基于M→N对待分类样本d进行分类，得到第二特征
输入第一特征与第二特征到学习器，输出任意业务文档各文档的正确类别标签，并推送到相应的终端系统

图2

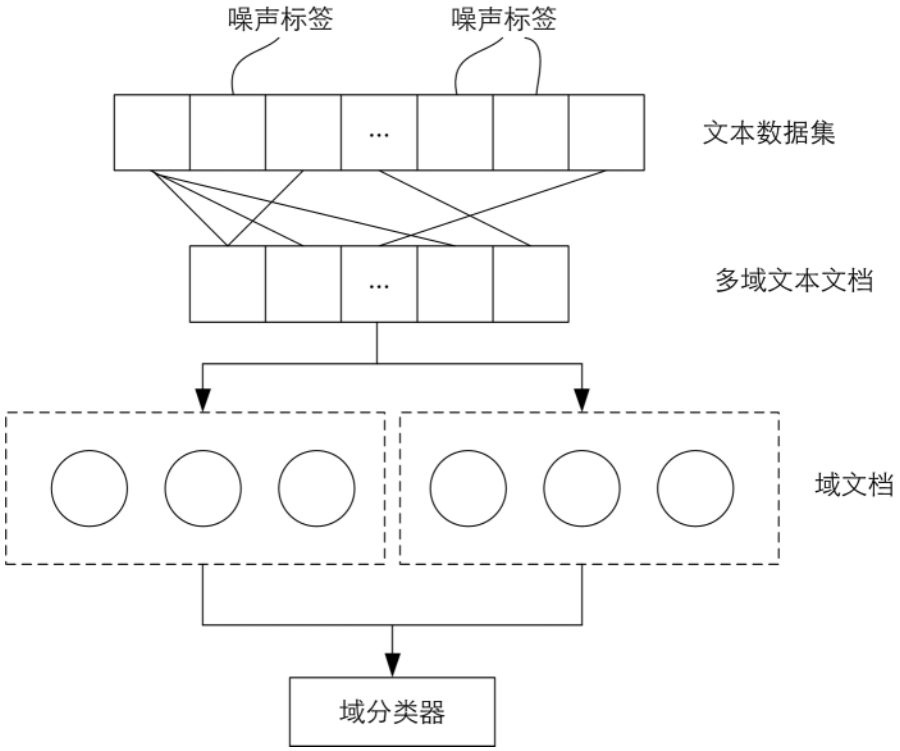

噪声标签
噪声标签
...
文本数据集
...
多域文本文档
域文档
域分类器

图3

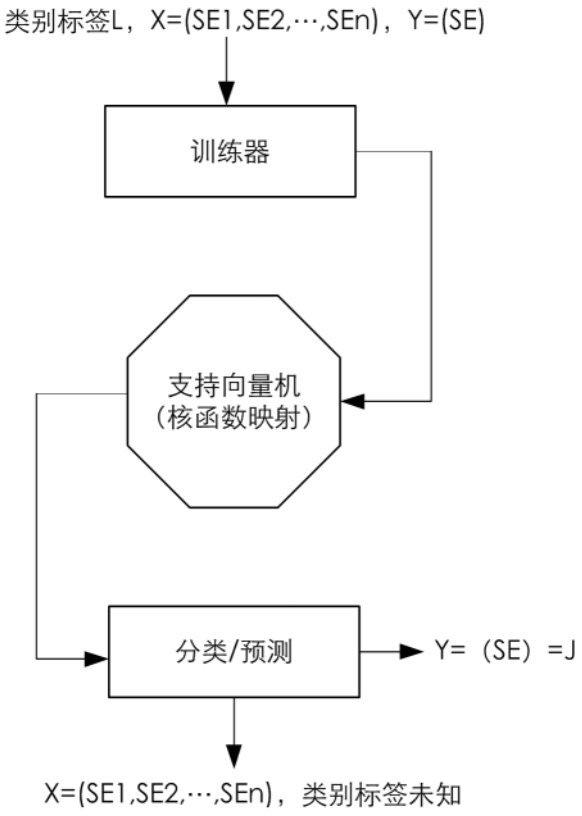

类别标签L，X=(SE1,SE2,…,SEn)，Y=(SE)
训练器
支持向量机
（核函数映射）
分类/预测
Y=（SE）=J
X=(SE1,SE2,…,SEn)，类别标签未知

图4

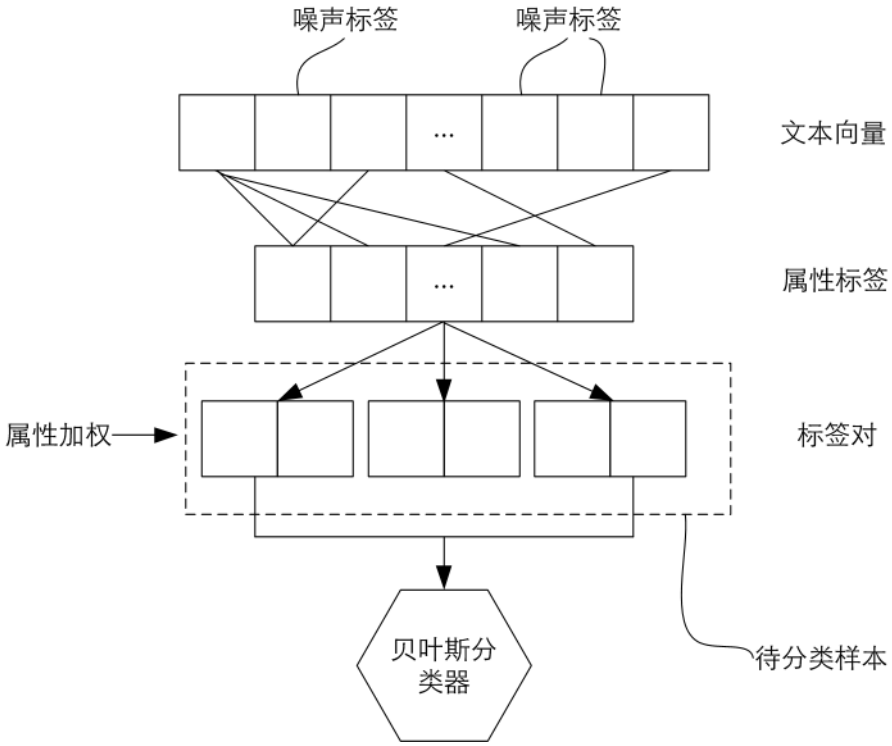

噪声标签
噪声标签
文本向量
属性标签
属性加权
标签对
贝叶斯分类器
待分类样本

图5

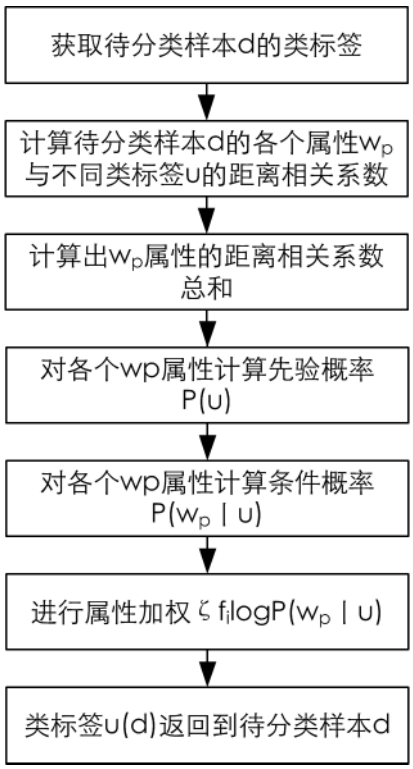

获取待分类样本d的类标签
计算待分类样本d的各个属性w_p与不同类标签u的距离相关系数
计算出w_p属性的距离相关系数总和
对各个wp属性计算先验概率P(u)
对各个wp属性计算条件概率$P(w_p \mid u)$
进行属性加权 $\zeta f_i \log P(w_p \mid u)$
类标签u(d)返回到待分类样本d

图6