(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 105183831 A
(43)申请公布日 2015.12.23

(21)申请号 201510545940.3

(22)申请日 2015.08.31

(71)申请人 上海德唐数据科技有限公司
地址 201600 上海市松江区漕河泾开发区松江高科技园莘砖公路518号11幢404-2室

(72)发明人 罗登 周贤华 万享 张玉志

(74)专利代理机构 深圳市科吉华烽知识产权事务所(普通合伙) 44248
代理人 孙伟

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

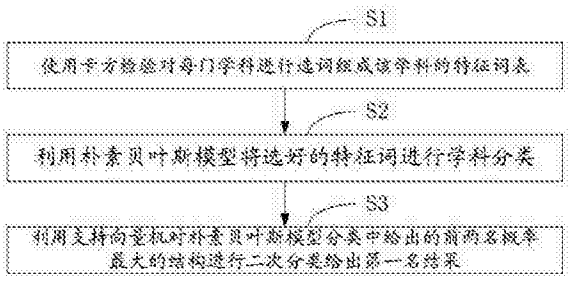
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种针对不同学科题目文本分类的方法

(57)摘要

本发明适用于数据预处理技术领域，提供了一种针对不同学科题目文本分类的方法，所述方法包括以下步骤：A、使用卡方检验对每门学科进行选词组成该学科的特征词表；B、利用朴素贝叶斯模型将选好的特征词进行学科分类；C、利用支持向量机对朴素贝叶斯模型分类中给出的前两名概率最大的结构进行二次分类给出第一名结果。通过两次分类，使得分类平均正确率得到提高，本方法实现简单、操作简单、使用方便准确，对学科间的分类更加准确，有效的提高了邻近学科之间分类的正确率。



CN 105183831 A

1.一种针对不同学科题目文本分类的方法，其特征在于：所述方法包括以下步骤：

A、使用卡方检验对每门学科进行选词组成该学科的特征词表；

B、利用朴素贝叶斯模型将选好的特征词进行学科分类；

C、利用支持向量机对朴素贝叶斯模型分类中给出的前两名概率最大的结果进行二次分类给出第一名结果。

2.根据权利要求1所述的方法，其特征在于：所述步骤A中还包括步骤：

A1、将选出的词按照该词与学科的关联性做排序。

3.根据权利要求2所述的方法，其特征在于：所述步骤A中还包括步骤：

A2、利用词频表对组成的特征词进行词频过滤生成新的特征词表。

4.根据权利要求3所述的方法，其特征在于：所述卡方检验是统计样本的实际值与理论值之间的偏离程度，根据偏离程度大小确定理论值是否正确；其中偏差程度为：$\sum_{i=1}^{n}\frac{(x_i-E)^2}{E}$，E为理论值，$x_1$，$x_2$，...$x_i$，...$x_n$为实际值。

5.根据权利要求4所述的方法，其特征在于：所述步骤B中计算文档d属于某个类别C_i的概率为：$P(C_i|d)=\frac{P(d|C_i)P(C_i)}{P(d)}$，其中$P(d|C_i)=P(w_1|C_i)P(w_2|C_i)...P(w_j|C_i)...P(w_m|C_i)$，m为文档d分词的个数，$P(w_j|C_i)$就代表词汇$w_j$属于类别$C_i$的概率。

6.根据权利要求5所述的方法，其特征在于：对于$P(d|C_i)=P(w_1|C_i)P(w_2|C_i)...P(w_j|C_i)...P(w_m|C_i)$式中$P(C_i)$和P(d)在同一文档中大小值一样。

一种针对不同学科题目文本分类的方法

技术领域

[0001] 本发明涉及数据预处理技术，尤其涉及一种针对不同学科题目文本分类的方法。

背景技术

[0002] 随着网络上文本信息的爆炸式增长，对文本的处理需求越来越迫切，同时要求的精度和准确性也越来越高，尤其是在文档分类和信息检索等领域，经常需要对大批量的文档进行自动分类。

[0003] 目前的文本分类方法主要包含三个环节，即文本表示、特征提取和文本分类，一般来说不同的文本分类方法主要区别在于如何表示文本。在文本表示方面，主要有基于词典向量和基于深度学习两种文本分类法，前者直接将文本按照分词结果表示为向量，向量的每个位表示在文档中有无该分词或者通过某种加权方法后得到的值，而后者一般通过深度学习方法将词表示成向量，向量中每一位没有具体的意义，但整个向量可用来描述该词与其他词之间的联系；在特征提取方面，除了常用的词频、逆向文档频率等指标，还有信息增益以及卡方检验等统计学方法；在文本分类方面，常用的分类法如朴素贝叶斯、k 邻近、支持向量机以及神经网络等方法都可以用于文本分类。

[0004] 目前的文本分类方法在处理特征明显、类别间相互差异较大的文本时有较高的正确率，但在处理有一定相似度的文本时效果会降低，以常见的初、高中九门学科的题目，即数、语、外、物、化、生、政、史、地为例，其中理科与文科之间比较容易分类，但理科或者文科内部的各科间都有一定的相似度。在基于词典向量的方法中，选择特征词时一般都会使用统计学习方法，在统计时一般只考虑了词的信息，而词与词之间的关联则被忽略；而基于深度学习的方法在把词表示成向量后，虽然向量中包含了词与词之间的关联信息，但在用词向量表示整个文本时，由于不同文本的长度变化幅度大，难以找到统一的特征输入分类器，在一些使用深度学习的方案中将文本长度固定，这样的做法不可避免会带来信息的丢失。

发明内容

[0005] 为了解决现有技术中的问题，本发明提供了一种针对不同学科题目文本分类的方法。

[0006] 本发明是这样实现的，一种针对不同学科题目文本分类的方法，所述 方法包括以下步骤：

[0007] A、使用卡方检验对每门学科进行选词组成该学科的特征词表；

[0008] B、利用朴素贝叶斯模型将选好的特征词进行学科分类；

[0009] C、利用支持向量机对朴素贝叶斯模型分类中给出的前两名概率最大的结果进行二次分类给出第一名结果。

[0010] 本发明的进一步技术方案是：所述步骤 A 中还包括步骤：

[0011] A1、将选出的词按照该词与学科的关联性做排序。

[0012] 本发明的进一步技术方案是：所述步骤 A 中还包括步骤：

[0013] A2、利用词频表对组成的特征词进行词频过滤生成新的特征词表。

[0014] 本发明的进一步技术方案是：所述卡方检验是统计样本的实际值与理论值之间的偏离程度，根据偏离程度大小确定理论值是否正确；其中偏差程度为：$\sum_{i=1}^{n}\frac{(x_i-E)^2}{E}$，E为理论值，$x_1$，$x_2$，…$x_i$，…$x_n$为实际值。

[0015] 本发明的进一步技术方案是：所述步骤B中计算文档d属于某个类别C_i的概率为：$P(C_i|d)=\frac{P(d|C_i)P(C_i)}{P(d)}$，其中$P(d|C_i)=P(w_1|C_i)P(w_2|C_i)\cdots P(w_j|C_i)\cdots P(w_m|C_i)$，m为文档d分词的个数，$P(w_j|C_i)$就代表词汇$w_j$属于类别$C_i$的概率。

[0016] 本发明的进一步技术方案是：对于$P(d|C_i)=P(w_1|C_i)P(w_2|C_i)\cdots P(w_j|C_i)\cdots P(w_m|C_i)$式中$P(C_i)$和$P(d)$在同一文档中大小值一样。

[0017] 本发明的有益效果是：通过两次分类，使得分类平均正确率得到提高，本方法实现简单、操作简单、使用方便准确，对学科间的分类更加准确，有效的提高了邻近学科之间分类的正确率。

附图说明

[0018] 图1是本发明实施例提供的针对不同学科题目文本分类的方法的流程图。

图2是卡方检验选词流程图。

具体实施方式

[0019] 针对现有方法的不足，本方案设计了一个新的二次分类处理方法，在选择特征词的基础上根据不同的阶段确定有效的分类策略。为了使词典中的特征词尽可能具有代表性，本方案使用卡方检验选词。卡方检验是统计学中一种专门用于相关分析的假设检验方法，其模型中包含了对相关文档 频率的统计，比仅统计词频要更可靠，而且卡方检验是在每个类别中得到一系列特征词，这比使用信息增益在总体上得到的特征词更有针对性。

[0020] 在使用卡方检验得到特征词后，文档就可以表示成由这些特征词组成的向量，接下来要考虑如何进行分类。由于卡方检验得到的词表是经过相关性排序的，利用这一点，在每个类别的特征词表中依次对每个词赋权值，然后在分类时，根据文档分词后的匹配情况，对每一个类别都得到一个权值之和，最后以该和值大小来判断属于哪个类别。这种方法在对特征词赋予权值时使用了自定义的模型来进行量化，得到的量化值与每个特征词的重要性并不一定相符。本方案使用NBM进行分类，对卡方检验选出来的特征词，经过词频统计得到先验概率，然后在分类时根据贝叶斯公式计算文档属于每个类别的概率。相比于自定义模型，NBM有理论基础且应用广泛，而且其中先验概率的计算考虑了词在文档中重复出现的个数，这在一定程度上弥补了卡方检验的不足之处，即仅考虑词在不同文档中出现次数。

[0021] 在上一节中提到，在文本分类中统计学习方法一般只考虑了单个词的信息，词与词之间的关联往往被忽略，这个问题对于卡方检验和NBM来说都是存在的，再加上NBM需要假设文档中出现的词与词之间相互独立，而这一点在实际中难以满足。为了尽可能弥补这些缺陷，本方案在NBM基础上，添加了SVM进行二次分类。SVM是一种寻找最优分界面的模型，其寻找最优界面的过程隐性地包含了寻找不同词之间的最佳组合，而且SVM并不要

求输入的特征之间满足任何相关性条件。综上，将SVM用于优化分类结果，是一种合适的选择。

[0022] 图1示出了本发明提供的一种针对不同学科题目文本分类的方法的流程图，其详述如下：

[0023] 步骤S1，使用卡方检验对每门学科进行选词组成该学科的特征词表；使用卡方检验对每门学科进行选词，并且对选出的词按照该词与学科的关联性做一个排序，组成该学科的特征词表。卡方检验基本思想是统计样本的实际值与理论值之间的偏离程度，根据偏离程度大小确定理论值是否正确。设理论值为E，实际值为 $x_1, x_2, \cdots x_i, \cdots x_n$，偏差程度的计算公式为：$\sum_{i=1}^{n}\frac{(x_i-E)^2}{E}$，具体到文本分类中，一般假设某个词t与某个类别c不相关，即t不是类c的特征词。这样求到的卡方值越大，与假设偏差越大，说明t与c越相关；卡方值越小，与假设偏差越小，说明t与c越不相关。此时卡方值可用于衡量词t与类别c的相关程度。

[0024] 在得到词t与类别c的卡方值过程中，实际值为下表中的四种文档数：

[0025] 四种文档数

[0026]

	属于类别c	不属于类别c
包含词t	A	B
不包含词t	C	D

[0027] 对所有的文档按表中条件统计出A、B、C和D(即实际值)，以A为例，A表示既包含词t又属于类别c的文档数，其理论值即为属于类别c的文档数(A+C)乘以文档包含词t的概率(A+B)/N，其中N为总文档数：$E_{11}=(A+C)\frac{A+B}{N}$，则A的偏差程度为：$D_{11}=\frac{(A-E_{11})^2}{N}$，同理得到B、C、D的偏差程度 D_{12}、D_{21}、D_{22}，最后词t与类别c的卡方值为：

[0028] $x^2(t, c) = D_{11}+D_{12}+D_{21}+D_{22}$，由于一般只关心每个词对某个类的相关大小顺序，而不关心具体的卡方值，一般将上式可简化为：$\chi^2(t,c)=\frac{(AD-BC)}{(A+B)(C+D)}$，以上为卡方检验的原理和卡方值的计算方法，其流程图乳图2所示。

[0029] 每个词相对于每个类别都会有一个卡方值，选择卡方值大小排前两位的类别作为该词所属的类，这样每个类别都会被分配一定数量的特征词，由于卡方检验没有考虑词在文档中的个数，可能会夸大低频词的作用，所以需要经过词频表过滤一遍。

[0030] 之所以流程中选择每个词分配到两个类别中，是因为不同学科之间的关键词有很多是共有的，不能只分配给一个类别，而分配的类别过多会降低特征词表总体上与对应类别的相关性。

[0031] 步骤S2，利用朴素贝叶斯模型将选好的特征词进行学科分类；NBM(Naive Bayesian Model，NBM，朴素贝叶斯)首先假设给定目标值时属性之间相互条件独立，即文档中出现的词与词之间相互独立，然后根据文档中的每个词属于每个类别的先验概率，以及贝叶

斯公式，得到该文档 d 属于某个类别的概率。计算公式如下：$P(C_i|d)=\frac{P(d|C_i)P(C_i)}{P(d)}$，其中 $P(d|C_i) = P(w_1|C_i)P(w_2|C_i)\cdots P(w_j|C_i)\cdots P(w_m|C_i)$，m 为文档 d 分词的个数，$P(w_j|C_i)$ 就代表词汇 w_j 属于类别 C_i 的概率，可以用该词在该类中的词频来估计。式中的 $P(C_i)$ 和 P(d) 对于同一个文档的计算来说是一样大小的值，可以不用进行计算。

[0032] 步骤 S3，利用支持向量机对朴素贝叶斯模型分类中给出的前两名概率 最大的结构进行二次分类给出第一名结果。经过对 NBM 分类结果（对应表 3.1）的观察，某些类别间比较容易混淆：数学 vs 物理、语文 vs 政治、语文 vs 历史、英语 vs 政治、物理 vs 化学、化学 vs 生物、政治 vs 地理、政治 vs 历史、地理 vs 历史（相关数据见附表 1）。这些科目间部分题目具有较高的相似性，难以分辨。本方案使用 SVM 对以上九种情况进行二次分类，经过二次分类的优化，有效提高正确率。

[0033] 在学科分类中，每门学科选取 10000 道题目，最终得到每科的特征词数如下表：

[0034] 每科的特征词数

[0035]

学科	数学	语文	英语	物理	化学
特征词数	863	12656	9733	1764	1651
学科	生物	政治	地理	历史	
特征词数	1480	3349	1537	2598	

[0036] 对卡方检验选出来的词，经去重处理得到 7333 个词作为总词表；对总词表中的每个词计算它出现在每个类中的概率；在测试时，根据贝叶斯公式计算测试样本属于每个类的概率，以概率最大的类别作为判定结果。通过这一处理，使得分类平均正确率从卡方检验赋值分类方法的 85％提升到 93％。

[0037] 在本方案中对于 SVM 的二次分类，使用方检验选出来的 7333 个词来表示每个题目，作为 SVM 的输入，这样 SVM 的输入就是一个 7333 维的向量，每个位对应一个词，用 -1 表示题目中没有该词，1 表示题目中有该词。对以上九种情况的每个类别各取 2000 个样本训练 SVM 分类器，在测试时，如果 NBM 分类法给出的前两名为九种情况中的一种，则调用相应的 SVM 分类器进行二次分类。

[0038] 通过这一处理，使得分类平均正确率进一步从 NBM 分类方法的 92％提升到 96％。这样就比较可靠地实现了对题目的学科分类。

[0039] 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。

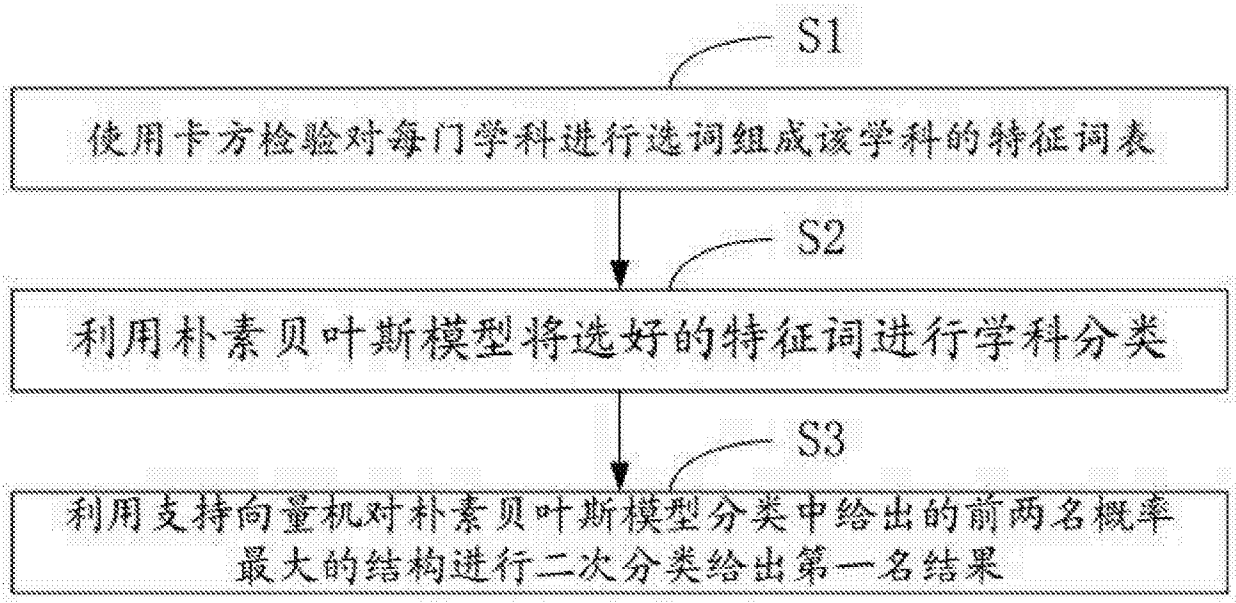

S1
使用卡方检验对每门学科进行选词组成该学科的特征词表
S2
利用朴素贝叶斯模型将选好的特征词进行学科分类
S3
利用支持向量机对朴素贝叶斯模型分类中给出的前两名概率最大的结构进行二次分类给出第一名结果

图 1

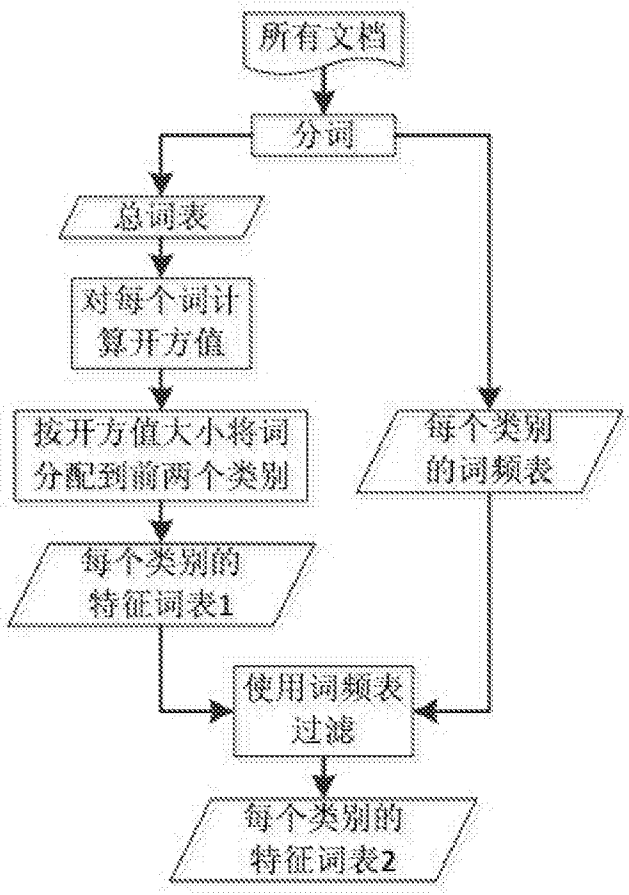

所有文档
分词
总词表
对每个词计算开方值
按开方值大小将词分配到前两个类别
每个类别的词频表
每个类别的特征词表1
使用词频表过滤
每个类别的特征词表2

图 2