(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 104199857 A
(43)申请公布日 2014.12.10

(21)申请号 201410400522.0

(22)申请日 2014.08.14

(71)申请人 西安交通大学
地址 710049 陕西省西安市咸宁西路 28 号

(72)发明人 刘均 马健 郑庆华 张未展 吴蓓

(74)专利代理机构 西安通大专利代理有限责任公司 61200
代理人 陆万寿

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

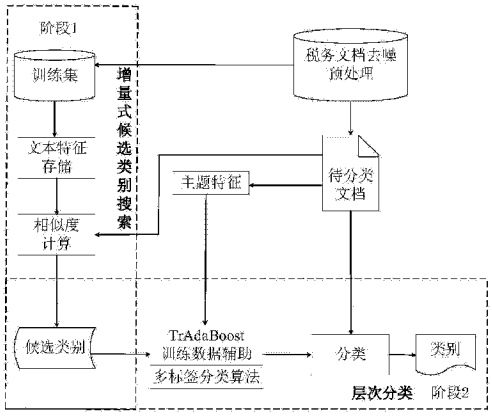
权利要求书3页 说明书7页 附图2页

(54)发明名称

一种基于多标签分类的税务文档层次分类方法

(57)摘要

一种基于多标签分类的税务文档层次分类方法，先从隐含狄利克雷分布模型中抽取生成的主题分布，构建税务文档的隐含狄利克雷分布主题特征。然后构建训练数据对应的 tf·idf 特征向量，计算包含训练数据和待分类文档的 tf·idf 特征向量，计算相似度获取候选类别标签。最后对候选类别标签节点的源数据补充辅助数据，用迁移学习算法 TrAdaBoost 构建基于迁移学习的多标签分类模型，对待分类文档进行分类。该方法将层次分类问题转换成“搜索－分类”两阶段过程，使用增量式候选类别搜索大大减少计算量，降低计算复杂度，用基于迁移学习的多标签分类模型将税务文档映射到税种层次类别上，有效利用了辅助数据，提升了分类性能。



CN 104199857 A

1. 一种基于多标签分类的税务文档层次分类方法，其特征在于，包括以下步骤：

1) 税务文档主题特征构建：

1-1) 对待分类的税务文档进行去噪预处理，得到待分类文档；

1-2) 指定待分类文档的主题个数，从隐含狄利克雷分布模型中抽取生成的主题分布，构建待分类文档的隐含狄利克雷分布主题特征，得到待分类文档的主题分布以及每个主题对应词的分布；

2) 增量式候选类别搜索：

2-1) 将若干篇已经标过分类标签的税务文档作为训练数据，去除训练数据中的停用词，构建训练数据的 tf•idf 特征向量，将 tf 矩阵中不为 0 的词表进行存储，并将生成的词汇列表、tf 列表、idf 值列表按序存储；

2-2) 对于待分类文档 r，根据保存的词汇列表计算 tf_r 值，tf_r 值是待分文档 r 的 tf 向量，将待分类文档 r 中出现但在当前词汇列表中未出现的词汇添加到词汇列表后面，然后根据 tf_r 值重新计算 idf 值，重新计算 tf•idf 特征向量，得到包含训练数据和待分类文档的 tf•idf 特征向量；

2-3) 计算待分类文档 r 和训练数据的相似度，获取候选类别标签；

3) 基于迁移学习的训练数据构建及多标签分类：

3-1) 对于每个候选类别标签节点，其本身对应的训练数据为源数据，借助其祖先节点和孩子节点对应的训练数据对源数据进行补充，补充的训练数据为辅助数据；

3-2) 利用迁移学习算法 TrAdaBoost 从辅助数据中选择出适合用于构建分类模型的数据，并构建基于迁移学习的多标签分类模型；

3-3) 利用基于迁移学习的多标签分类模型，结合隐含狄利克雷分布主题特征，对待分类文档进行分类，得到待分类文档所属的税种层次类别。

2. 如权利要求 1 所述的基于多标签分类的税务文档层次分类方法，其特征在于：所述的步骤 1-1) 中对待分类的税务文档进行去噪预处理的具体步骤为：先将待分类的税务文档转换成文本格式，对转换后的税务文档进行数据清洗，删除由于转换导致的乱码文档，去除重复文档，同时去除元数据信息，其中元数据信息包括文档标题和作者。

3. 如权利要求 1 所述的基于多标签分类的税务文档层次分类方法，其特征在于：所述的步骤 1-2) 中指定待分类文档的主题个数为 10 ～ 20 个。

4. 如权利要求 1 所述的基于多标签分类的税务文档层次分类方法，其特征在于：所述的步骤 1-2) 中隐含狄利克雷分布模型中所有隐藏变量和可见变量的联合分布如下：

$$P(w_i,z_i,\theta_i,\phi|\alpha,\eta)=\prod_{j=1}^{N}P(\theta_i|\alpha)P(z_{i,j}|\theta_i)P(\phi|\eta)P(w_{i,j}|\phi_{z_{i,j}})$$

其中 α、η 表示狄利克雷分布，i 表示第 i 篇税务文档，j 表示第 j 个词，N 表示文档的总词数，P() 表示多项式分布的共轭先验概率，w_i 表示从税务文档 i 中抽取生成的词语，z_i 表示从税务文档 i 中抽取生成的主题，θ_i 表示税务文档 i 的主题多项式分布，ϕ 表示词语分布，$z_{i,j}$ 表示从主题多项式分布 θ_i 中抽取生成税务文档 i 第 j 个词的主题，$\phi_{z_{i,j}}$ 表示从

狄利克雷分布 η 中抽取生成主题 $z_{i,j}$ 的词语多项式分布，$w_{i,j}$ 表示从词语多项式分布 $\phi_{z_{i,j}}$ 中抽样生成的词语。

5. 如权利要求 1 所述的基于多标签分类的税务文档层次分类方法，其特征在于：所述的步骤 2-3) 中使用余弦相似度计算待分类文档 r 和训练数据的相似度，选择并得到最相似的 k 个候选类别标签，k = 5 ～ 20。

6. 如权利要求 5 所述的基于多标签分类的税务文档层次分类方法，其特征在于：对于向量 γ 和向量 λ，余弦相似度计算公式为：

$$\cos(\gamma,\lambda)=\frac{\sum_{s=1}^{S}\gamma_s\lambda_s}{\sqrt{\sum_{s=1}^{S}\gamma_s^{\,2}}\sqrt{\sum_{s=1}^{S}\lambda_s^{\,2}}}$$

其中：s 表示向量分量的下标，即该分量位于向量中的位置，S 表示向量的维度，γ_s 表示向量 γ 的第 s 个分量，λ_s 表示向量 λ 的第 s 个分量。

7. 如权利要求 1 所述的基于多标签分类的税务文档层次分类方法，其特征在于：所述的步骤 3-1) 的具体操作为：对于候选类别标签节点 C_a 和其他任意的候选类别标签节点 C_b，对于 C_a 的任意祖先节点 P_a，在满足 $P_a \notin P(C_b)$ 时，将 P_a 对应的训练数据补充到 C_a 的源数据中，并将 P_a 的其他非候选类别标签节点的孩子节点对应的训练数据补充到 C_a 的源数据中，同时将 C_a 的孩子节点对应的训练数据补充到 C_a 的源数据中，其中 $P(C_b)$ 表示 C_b 的祖先节点的集合。

8. 如权利要求 1 所述的基于多标签分类的税务文档层次分类方法，其特征在于：所述的步骤 3-2) 的具体操作为：

①输入辅助数据 $T_a(sy_k)$、源数据 $T_b(sy_k)$、待分类文档、多标签 k- 近邻算法、迭代总次数 N 和训练数据集 T，$T = T_a(sy_k) \cup T_b(sy_k)$；

②初始化：设置初始权重向量 $w^1=(w_1^1,w_2^1,\cdots,w_{n+m}^1)$，其中 w_j^1 为初始权重向量中的第 j 个向量，其值为 0 ～ 1 的随机数，n 为 $T_a(sy_k)$ 中数据的个数，m 为 $T_b(sy_k)$ 中数据的个数；并设置 $\beta=1/(1+\sqrt{2\ln n/N})$；

③迭代计算：

i 设置迭代次数 t = 1,…,N；

ii 设置权重分布 p^t，使其满足

$$p^t=\frac{w^t}{\sum_{j=1}^{n+m}w_j^t}$$

其中 w^t 是第 t 次迭代后的权重向量，w_j^t 是 w^t 的第 j 个向量；

iii 调用多标签 k- 近邻算法，依据训练数据集 T 以及 T 上的权重分布 p^t 和待分类文档，得到分类器 h_t；

iv 计算 h_t 在 $T_b(sy_k)$ 上的错误率 ε_t，

$$\varepsilon_t = \sum_{j=n+1}^{n+m} \frac{w_j^t hloss_{T_b}(h_t)}{\sum_{j=n+1}^{n+m} w_j^t}$$

其中 $hloss_{T_b}(h_t)$ 是分类器 h_t 在 $T_b(sy_k)$ 上的汉明损失；

v 设置 $\beta_t = \varepsilon_t/(1-\varepsilon_t)$；

vi 设置迭代后的权重向量 w_j^{t+1} 为：

$$w_j^{t+1} = \begin{cases} w_j^t \beta^{hloss_{T_b}(h_t)} & \text{当} j=1,\cdots,n \\ w_j^t \beta^{-hloss_{T_b}(h_t)} & \text{当} j=n+1,\cdots,n+m \end{cases}$$

④输出基于迁移学习的多标签分类模型：

$$h_f(x) = \left\{ y \mid \sum_{t=N/2}^{N} \ln(1/\beta_t) f_t(x,y) \geq \frac{1}{2} \sum_{t=N/2}^{N} \ln(1/\beta_t) \right\}$$

其中 $f_t(x,y)$ 是分类器 h_t 在标签类别 y 上的预测值，x 是待分类文档的主题特征，$h_f(x)$ 是标签分类器。

9. 如权利要求 8 所述的基于多标签分类的税务文档层次分类方法，其特征在于：所述的迭代总次数 N = 50 ～ 100 次。

一种基于多标签分类的税务文档层次分类方法

技术领域

[0001] 本发明属于数据挖掘领域，具体涉及一种基于多标签分类的税务文档层次分类方法。

背景技术

[0002] 随着互联网的迅猛发展，各种资源以指数形式迅速增长，税务文档也大量涌现在网络中，造成了人们获取过程中的信息过载问题。如何有效的对税务文档进行组织和管理是解决税务资源获取过程中信息过载问题的关键，是对税收有重要意义的一项工作。

[0003] 为了有效组织和管理互联网上的海量税务文档，通常按照一个主题类别层次或大规模的概念或对税务文档进行分类，以更好地访问和搜索这些税务文档。税收分类是按一定标准对各种税收进行的分类，一个国家的税收体系通常是由许多不同的税种构成的。将税务文档按照税收进行分类是一种有效的组织方式。

[0004] 申请人经过查新，没有找到有关对税务文档进行层次分类的专利，因而检索了一篇与本专利相关的已授权专利：一种使用本体进行文本文档自动分类的方法［专利号：ZL201010210107.0］；在该专利中，发明人使用带权重的关键词集合表示文本文档的特征信息，通过计算文本文档和分类目录之间的相似值对文本文档进行自动分类。该发明所述方法使用简单的相似度来进行分类，且只能分到一种类别里，无法对有多个主题的文档进行分类，而且难以分到层次类别里面。

发明内容

[0005] 本发明的目的在于提供一种基于多标签分类的税务文档层次分类方法，能够有效的对税务文档进行组织和管理，解决税务资源获取过程中信息过载的问题。

[0006] 为达到上述目的，本发明采用的技术方案为：

[0007] 一种基于多标签分类的税务文档层次分类方法，包括以下步骤：

[0008] 1）税务文档主题特征构建：

[0009] 1-1）对待分类的税务文档进行去噪预处理，得到待分类文档；

[0010] 1-2）指定待分类文档的主题个数，从隐含狄利克雷分布模型中抽取生成的主题分布，构建待分类文档的隐含狄利克雷分布主题特征，得到待分类文档的主题分布以及每个主题对应词的分布；

[0011] 2）增量式候选类别搜索：

[0012] 2-1）将若干篇已经标过分类标签的税务文档作为训练数据，去除训练数据中的停用词，构建训练数据的 tf·idf 特征向量，将 tf 矩阵中不为 0 的词表进行存储，并将生成的词汇列表、tf 列表、idf 值列表按序存储；

[0013] 2-2）对于待分类文档 r，根据保存的词汇列表计算 tf_r 值，tf_r 值是待分文档 r 的 tf 向量，将待分类文档 r 中出现但在当前词汇列表中未出现的词汇添加到词汇列表后面，然后根据 tf_r 值重新计算 idf 值，重新计算 tf·idf 特征向量，得到包含训练数据和待分类

文档的 tf·idf 特征向量；

[0014] 2-3) 计算待分类文档 r 和训练数据的相似度，获取候选类别标签；

[0015] 3) 基于迁移学习的训练数据构建及多标签分类：

[0016] 3-1) 对于每个候选类别标签节点，其本身对应的训练数据为源数据，借助其祖先节点和孩子节点对应的训练数据对源数据进行补充，补充的训练数据为辅助数据；

[0017] 3-2) 利用迁移学习算法 TrAdaBoost 从辅助数据中选择出适合用于构建分类模型的数据，并构建基于迁移学习的多标签分类模型；

[0018] 3-3) 利用基于迁移学习的多标签分类模型，结合隐含狄利克雷分布主题特征，对待分类文档进行分类，得到待分类文档所属的税种层次类别。

[0019] 所述的步骤 1-1) 中对待分类的税务文档进行去噪预处理的具体步骤为：先将待分类的税务文档转换成文本格式，对转换后的税务文档进行数据清洗，删除由于转换导致的乱码文档，去除重复文档，同时去除元数据信息，其中元数据信息包括文档标题和作者。

[0020] 所述的步骤 1-2) 中指定待分类文档的主题个数为 10 ～ 20 个。

[0021] 所述的步骤 1-2) 中隐含狄利克雷分布模型中所有隐藏变量和可见变量的联合分布如下：

[0022] $$P(w_i,z_i,\theta_i,\phi|\alpha,\eta)=\prod_{j=1}^{N}P(\theta_i|\alpha)P(z_{i,j}|\theta_i)P(\phi|\eta)P(w_{i,j}|\phi_{z_{i,j}})$$

[0023] 其中 α、η 表示狄利克雷分布，i 表示第 i 篇税务文档，j 表示第 j 个词，N 表示文档的总词数，P() 表示多项式分布的共轭先验概率，w_i 表示从税务文档 i 中抽取生成的词语，z_i 表示从税务文档 i 中抽取生成的主题，θ_i 表示税务文档 i 的主题多项式分布，ϕ 表示词语分布，$z_{i,j}$ 表示从主题多项式分布 θ_i 中抽取生成税务文档 i 第 j 个词的主题，$\phi_{z_{i,j}}$ 表示从狄利克雷分布 η 中抽取生成主题 $z_{i,j}$ 的词语多项式分布，$w_{i,j}$ 表示从词语多项式分布 $\phi_{z_{i,j}}$ 中抽样生成的词语。

[0024] 所述的步骤 2-3) 中使用余弦相似度计算待分类文档 r 和训练数据的相似度，选择并得到最相似的 k 个候选类别标签，k ＝ 5 ～ 20。

[0025] 对于向量 γ 和向量 λ，余弦相似度计算公式为：

[0026] $$\cos(\gamma,\lambda)=\frac{\sum_{s=1}^{S}\gamma_s\lambda_s}{\sqrt{\sum_{s=1}^{S}\gamma_s^{\,2}}\sqrt{\sum_{s=1}^{S}\lambda_s^{\,2}}}$$

[0027] 其中：s 表示向量分量的下标，即该分量位于向量中的位置，S 表示向量分量的总个数，即向量的维度，γ_s 表示向量 γ 的第 s 个分量，λ_s 表示向量 λ 的第 s 个分量。

[0028] 所述的步骤 3-1) 的具体操作为：对于候选类别标签节点 C_a 和其他任意的候选类别标签节点 C_b，对于 C_a 的任意祖先节点 P_a，在满足 $P_a \notin P(C_b)$ 时，将 P_a 对应的训练数据补充到 C_a 的源数据中，并将 P_a 的其他非候选类别标签节点的孩子节点对应的训练数据补充到 C_a 的源数据中，同时将 C_a 的孩子节点对应的训练数据补充到 C_a 的源数据中，其中 $P(C_b)$ 表示

C_b 的祖先节点的集合。

[0029] 所述的步骤 3-2) 的具体操作为：

[0030] ①输入辅助数据 $T_a(sy_k)$、源数据 $T_b(sy_k)$、待分类文档、多标签 k- 近邻算法、迭代总次数 N 和训练数据集 T，$T = T_a(sy_k) \cup T_b(sy_k)$；

[0031] ②初始化：设置初始权重向量 $w^1 = (w_1^1, w_2^1, \cdots, w_{n+m}^1)$，其中 w_j^1 为初始权重向量中的第 j 个向量，其值为 0 ～ 1 的随机数，n 为 $T_a(sy_k)$ 中数据的个数，m 为 $T_b(sy_k)$ 中数据的个数；并设置 $\beta = 1/(1+\sqrt{2\ln n / N})$；

[0032] ③迭代计算：

[0033] i 设置迭代次数 $t = 1, \cdots, N$；

[0034] ii 设置权重分布 p^t，使其满足

[0035] $$p^t = \frac{w^t}{\sum_{j=1}^{n+m} w_j^t}$$

[0036] 其中 w^t 是第 t 次迭代后的权重向量，w_j^t 是 w^t 的第 j 个向量；

[0037] iii调用多标签 k- 近邻算法，依据训练数据集 T 以及 T 上的权重分布 p^t 和待分类文档，得到分类器 h_t；

[0038] iv计算 h_t 在 $T_b(sy_k)$ 上的错误率 ε_t，

[0039] $$\varepsilon_t = \sum_{j=n+1}^{n+m} \frac{w_j^t hloss_{T_b}(h_t)}{\sum_{j=n+1}^{n+m} w_j^t}$$

[0040] 其中 $hloss_{T_b}(h_t)$ 是分类器 h_t 在 $T_b(sy_k)$ 上的汉明损失；

[0041] v 设置 $\beta_t = \varepsilon_t/(1-\varepsilon_t)$；

[0042] vi设置迭代后的权重向量 w_j^{t+1} 为：

[0043]

$$w_j^{t+1} = \begin{cases} w_j^t \beta^{hloss_{T_b}(h_t)} & \text{当} j = 1, \cdots, n \\ w_j^t \beta^{-hloss_{T_b}(h_t)} & \text{当} j = n+1, \cdots, n+m \end{cases}$$

[0044] ④输出基于迁移学习的多标签分类模型：

[0045] $$h_f(x) = \left\{ y \mid \sum_{t=N/2}^{N} \ln(1/\beta_t) f_t(x,y) \geq \frac{1}{2} \sum_{t=N/2}^{N} \ln(1/\beta_t) \right\}$$

[0046] 其中 $f_t(x, y)$ 是分类器 h_t 在标签类别 y 上的预测值，x 是待分类文档的主题特征，$h_f(x)$ 是标签分类器。

[0047] 所述的迭代总次数 N = 50 ～ 100 次。

[0048] 相对于现有技术，本发明的有益效果为：

[0049] 本发明提供的基于多标签分类的税务文档层次分类方法，主要包括税务文档主题特征构建、增量式候选类别搜索、基于迁移学习的训练数据构建及多标签分类这三个部分，通过构建税务文档的主题特征，将层次分类问题转换成“搜索 - 分类”两阶段过程，使用增量式候选类别搜索，根据构建的待分类文档的隐含狄利克雷分布主题特征，针对每个需要

进行税务分类的待分类文档，只计算该待分类文档的文本特征，采用 tf•idf 向量来表示文档，并基于此采用余弦相似度来计算待分类文档 r 和训练数据的相似度，获取候选类别标签，大大减少计算量，降低计算复杂度，基于改进传统的 Adaboost 算法，使用迁移学习算法 TrAdaBoost(Transfer AdaBoost) 构建基于迁移学习的多标签分类模型，依据待分类文档的数据逐步调整辅助数据和源数据的权重，利用不同权重的数据进行训练，充分利用辅助数据帮助待分类文档进行分类，达到了有效迁移知识的目的，大大提升了分类性能，并且能够有效的对税务文档进行组织和管理，解决税务资源获取过程中信息过载的问题。

附图说明

[0050] 图 1 是基于多标签分类的税务文档层次分类方法流程图；

[0051] 图 2 是增量式候选类别搜索流程图；

[0052] 图 3 是税种类别层次树状样例图；

[0053] 图 4 是训练数据辅助选择示例图。

具体实施方式

[0054] 下面结合附图对本发明作进一步的详细说明。

[0055] 税务文档是指在税务领域中对税务进行描述、分析和研究的资料和文章。税种层次类别是指按照一定标准对各种税种进行分类所构成的税收体系。

[0056] 本发明提供的基于多标签分类的税务文档层次分类方法，包括如下 3 个过程：

[0057] 1) 税务文档主题特征构建，包括 2 个步骤：

[0058] 1-1) 对待分类的税务文档进行去噪预处理，即将不同类型的待分类的税务文档全部转换成文本类型，对文档进行数据清洗，删除由于转换导致的乱码文档，去除重复文档，同时去除文档标题、作者等元数据信息，得到待分类文档；

[0059] 1-2) 由若干待分类文档构成文档集合，针对文档集合，指定主题的个数（一般为 10 ～ 20 个），从隐含狄利克雷分布模型中抽取生成的主题分布，构建每篇待分类文档的隐含狄利克雷分布主题特征，即给出每篇文档的主题分布以及每个主题对应词的分布。其中主题是指文档的一个概念、一个方面，它表现为一系列相关的词语。隐含狄利克雷分布简称 LDA(Latent Dirichlet allocation)，LDA 首先由 Blei,David M.、吴恩达和 Jordan,Michael I 于 2003 年提出。隐含狄利克雷分布模型中所有隐藏变量和可见变量的联合分布如下：

[0060] $$P(w_i,z_i,\theta_i,\phi|\alpha,\eta)=\prod_{j=1}^{N}P(\theta_i|\alpha)P(z_{i,j}|\theta_i)P(\phi|\eta)P(w_{i,j}|\phi_{z_{i,j}})$$

[0061] 其中 α、η 表示狄利克雷分布，i 表示第 i 篇税务文档，j 表示第 j 个词，N 表示文档的字数，P() 表示多项式分布的共轭先验概率，w_i 表示从税务文档 i 中抽取生成的词语，z_i 表示从税务文档 i 中抽取生成的主题，θ_i 表示税务文档 i 的主题多项式分布，ϕ 表示词语分布，$z_{i,j}$ 表示从主题多项式分布 θ_i 中抽取生成税务文档 i 第 j 个词的主题，$\phi_{z_{i,j}}$ 表示从狄利克雷分布 η 中抽取生成主题 $z_{i,j}$ 的词语多项式分布，$w_{i,j}$ 表示从词语多项式分布 $\phi_{z_{i,j}}$

中抽样生成的词语。

[0062] 2) 增量式候选类别搜索，包括 3 个步骤：

[0063] 2-1) 将若干篇已经标过分类标签的税务文档作为训练数据，去除训练数据中的停用词（在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后回自动过滤掉某些字或词，即停用词。在这里，停用词是指出现频率很高但实际意义不大的词，比如我你他等等。），构建训练数据的 tf·idf 特征向量，tf 矩阵是一个稀疏矩阵，为了存储简约，只将 tf 矩阵中不为 0 的词表进行存储，格式为【词编号 :词频】，将生成的词汇列表、tf 列表、idf 列表按序存储，如果列表较小，则将其直接保存到内存中，最终全部分类完成后再更新到硬盘上面，而这些文件在磁盘上所占空间开销很小，如果列表较大，则直接存到磁盘；

[0064] 2-2) 对于待分类文档 r，进行主体部分的抽取，根据保存的词汇列表计算 tf_r 值（tf 值从小到大排序，前 r 个 tf 值构成 tf_r 值），tf_r 值是待分类文档 r 的 tf 向量，并将待分类文档 r 中出现但在当前词汇列表中未出现的词汇添加到词汇列表后面，然后根据 tf_r 值重新计算 idf 值，再读取保存的 tf 列表，重新计算 tf·idf 特征向量，得到包含训练数据和待分类文档的 tf·idf 特征向量；

[0065] 2-3) 采用步骤 2-2) 计算得到的 tf·idf 特征向量来表示文档（训练数据和待分类文档），使用余弦相似度计算待分类文档 r 和训练数据的相似度，选择并得到最相似的 k 个候选类别标签，k = 5 ～ 20。

[0066] 对于向量 γ 和向量 λ，余弦相似度计算公式为：

[0067] $$\cos(\gamma,\lambda)=\frac{\sum_{s=1}^{S}\gamma_s\lambda_s}{\sqrt{\sum_{s=1}^{S}\gamma_s^2}\sqrt{\sum_{s=1}^{S}\lambda_s^2}}$$

[0068] 其中 :s 表示向量分量的下标，即该分量位于向量中的位置，S 表示向量分量的总个数，即向量的维度，γ_s 表示向量 γ 的第 s 个分量，λ_s 表示向量 λ 的第 s 个分量。

[0069] 3) 基于迁移学习的训练数据构建及多标签分类，包括 3 个步骤：

[0070] 3-1) 层次分类是指将一篇税务文档挂载到税种层次类别中的一个或多个节点上面，从而得到一个或多个分类标签。对于候选类别标签节点 C_a 和其他任意的候选类别标签节点 C_b，C_a 本身对应的训练数据（未补充的）为源数据，对于 C_a 的任意祖先节点 P_a，只要 $P_a \notin P(C_b)$，其中 $P(C_b)$ 表示 C_b 的祖先节点集合，那么 P_a 的对应训练数据就可以补充到 C_a 的源数据中，并且将 P_a 的其他非候选类别标签节点的孩子节点对应的训练数据补充到 C_a 的源数据中，同时将候选类别标签节点 C_a 的孩子节点对应的训练数据补充到 C_a 的源数据中。如附图 4 所示，候选类别标签节点 Q 向上补充两层到节点 O 和 M，然后添加相应的孩子节点 R、W、Y、Z。该节点能够补充的所有祖先节点和孩子节点为该节点的辅助节点，所有补充的训练数据为辅助数据，所以候选类别标签节点 Q 对应的辅助节点为 $f_a(Q) = \{M,O,R,W,Y,Z\}$，其中 $f_a(Q)$ 表示节点 Q 的辅助节点；

[0071] 3-2) 使用迁移学习算法 TrAdaboost 去除掉辅助数据中与源数据不相关的数据，对于预测正确的辅助数据，将其权重增加，而对于预测错误的辅助数据，将其权重减小，充分利用辅助数据帮助待分类文档进行分类。

[0072] 迁移学习算法 TrAdaBoost 的具体步骤为：

[0073] ①输入辅助数据 $T_a(sy_k)$、源数据 $T_b(sy_k)$、待分类文档、多标签 k-近邻算法、迭代总次数 N = 50 ～ 100 和训练数据集 T，$T = T_a(sy_k) \cup T_b(sy_k)$；

[0074] ②初始化：设置初始权重向量 $w^1 = (w_1^1, w_2^1, \cdots, w_{n+m}^1)$，其中 w_j^1 为初始权重向量中的第 j 个向量，其值为 0 ～ 1 的随机数，n 为 $T_a(sy_k)$ 中数据的个数，m 为 $T_b(sy_k)$ 中数据的个数；并设置 $\beta = 1/(1+\sqrt{2\ln n / N})$；

[0075] ③迭代计算：

[0076] i 设置迭代次数 t = 1,…,N；

[0077] ii 设置权重分布 p^t，使其满足

[0078] $$p^t = \frac{w^t}{\sum_{j=1}^{n+m} w_j^t}$$

[0079] 其中 w^t 是第 t 次迭代后的权重向量，w_j^t 是 w^t 的第 j 个向量；

[0080] iii调用多标签 k- 近邻算法，依据训练数据集 T 以及 T 上的权重分布 p^t 和待分类文档，得到分类器 h_t；

[0081] iv计算 h_t 在 $T_b(sy_k)$ 上的错误率 ε_t，

[0082] $$\varepsilon_t = \sum_{j=n+1}^{n+m} \frac{w_j^t hloss_{T_b}(h_t)}{\sum_{j=n+1}^{n+m} w_j^t}$$

[0083] 其中 $hloss_{T_b}(h_t)$ 是分类器 h_t 在 $T_b(sy_k)$ 上的汉明损失；

[0084] v 设置 $\beta_t = \varepsilon_t/(1-\varepsilon_t)$；

[0085] vi设置迭代后的权重向量 w_j^{t+1} 为：

[0086]

$$w_j^{t+1} = \begin{cases} w_j^t \beta^{hloss_{T_b}(h_t)} & 当 j = 1,\cdots,n \\ w_j^t \beta^{-hloss_{T_b}(h_t)} & 当 j = n+1,\cdots,n+m \end{cases}$$

[0087] ④输出基于迁移学习的多标签分类模型：

[0088] $$h_f(x) = \left\{ y \mid \sum_{t=N/2}^{N} \ln(1/\beta_t) f_t(x,y) \geq \frac{1}{2} \sum_{t=N/2}^{N} \ln(1/\beta_t) \right\}$$

[0089] 其中 $f_t(x,y)$ 是分类器 h_t 在标签类别 y 上的预测值，x 是待分类文档的主题特征，$h_f(x)$ 是标签分类器。

[0090] 表 1 给出了多标签 TrAdaBoost 算法（迁移学习算法 TrAdaboost）的程序。

[0091] 表 1 多标签 TrAdaBoost 算法

[0092]

算法：多标签 TrAdaBoost 算法

输入：辅助数据 $T_a(sy_k)$和源数据 $T_b(sy_k)$，训练数据集 T，$T=T_a(sy_k)\cup T_b(sy_k)$，待分类文档，多标签 k-近邻算法，迭代总次数 N=50~100 次。

初始化：

1. 设置初始权重向量 $w^1=(w_1^1,w_2^1,\cdots,w_{n+m}^1)$

其中，$w_j^1=\begin{cases}1/n & \text{当}j=1,\cdots,n\\ 1/m & \text{当}j=n+1,\cdots,n+m\end{cases}$；

2. 设置 $\beta=1/(1+\sqrt{2\ln n/N})$。

For 迭代次数 $t=1,\cdots,N$

1. 设置权重分布 p^t 满足 $p^t=\dfrac{w^t}{\sum_{j-1}^{n+m}w_j^t}$；
2. 调用多标签 k-近邻算法，依据训练数据集 T 以及 T 上的权重分布 p^t 和待分类文档，得到分类器 h_t。
3. 计算 h_t 在 $T_b(sy_k)$上的错误率 ε_t：

$$\varepsilon_t=\sum_{j=n+1}^{n+m}\frac{w_j^t hloss_{T_b}(h_t)}{\sum_{j=n+1}^{n+m}w_j^t}$$

其中 $hloss_{T_b}(h_t)$是分类器 h_t 在 $T_b(sy_k)$上的汉明损失。

4. 设置 $\beta_t=\varepsilon_t/(1-\varepsilon_t)$；
5. 设置迭代后的权重向量为：

$$w_j^{t+1}=\begin{cases}w_j^t\beta^{hloss_{T_b}(h_t)} & \text{当}j=1,\cdots,n\\ w_j^t\beta^{-hloss_{T_b}(h_t)} & \text{当}j=n+1,\cdots,n+m\end{cases}$$

输出最终的标签分类器 $h_f(x)$，即基于迁移学习的多标签分类模型：

$$h_f(x)=\left\{y\,|\sum_{t=N/2}^{N}\ln(1/\beta_t)f_t(x,y)\geq\frac{1}{2}\sum_{t=N/2}^{N}\ln(1/\beta_t)\right\}$$

其中 $f_t(x,y)$是分类器 h_t 在标签类别 y 上的预测值，x 是待分类文档的主题特征。

[0093] 3-3）利用步骤 3-2）构建的基于迁移学习的多标签分类模型，结合步骤 1-2）得到的隐含狄利克雷分布主题特征，对待分类文档进行分类预测，得到待分类文档所属的税种层次类别。

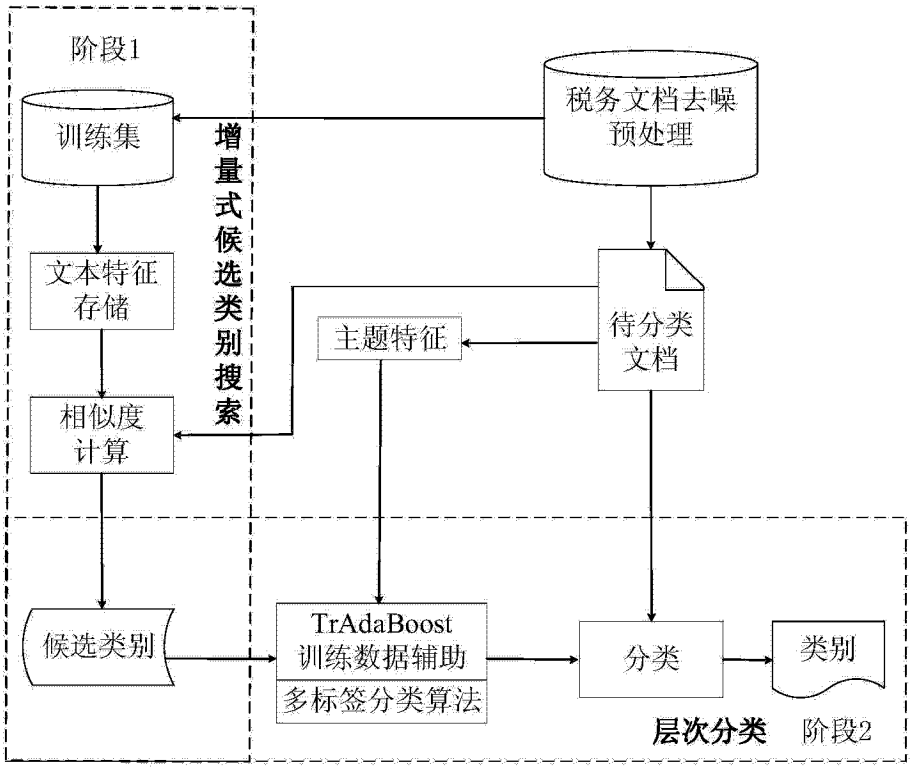

阶段1
训练集
税务文档去噪
预处理
增量式候选类别搜索
文本特征
存储
主题特征
待分类
文档
相似度
计算
候选类别
TrAdaBoost
训练数据辅助
多标签分类算法
分类
类别
层次分类
阶段2

图 1

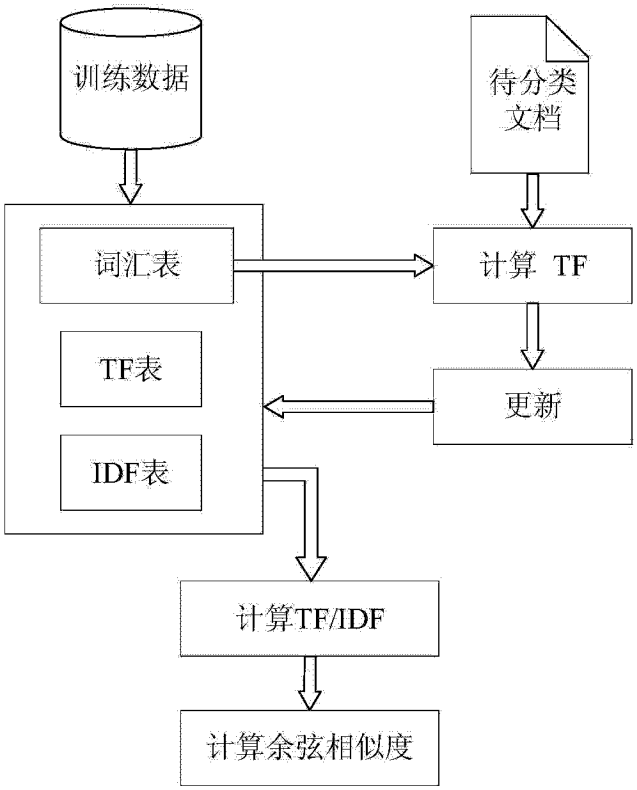

训练数据
待分类
文档
词汇表
计算 TF
TF表
更新
IDF表
计算TF/IDF
计算余弦相似度

图 2

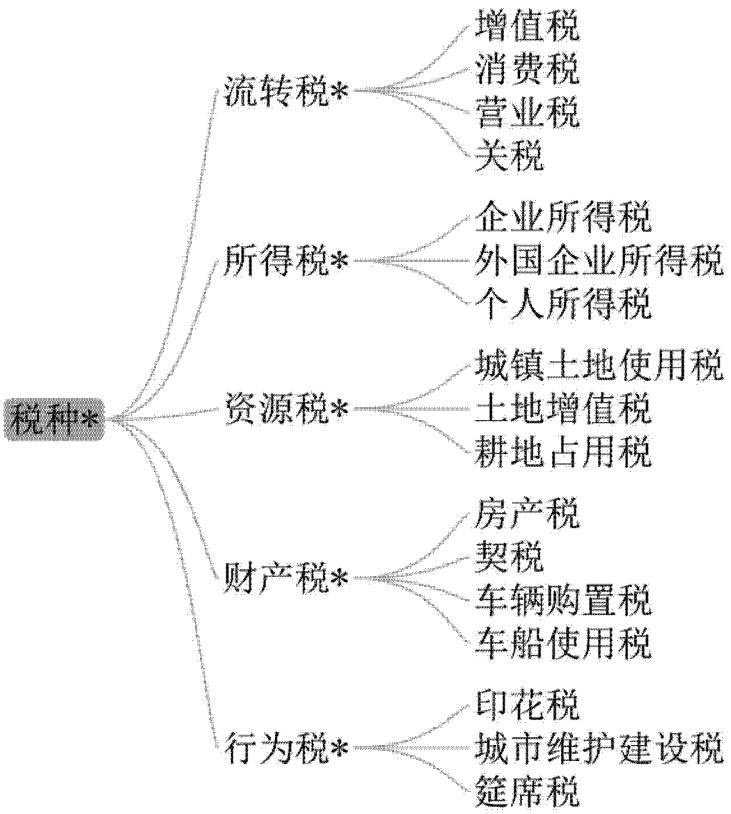

税种*
流转税*
增值税
消费税
营业税
关税
所得税*
企业所得税
外国企业所得税
个人所得税
资源税*
城镇土地使用税
土地增值税
耕地占用税
财产税*
房产税
契税
车辆购置税
车船使用税
行为税*
印花税
城市维护建设税
筵席税

图 3

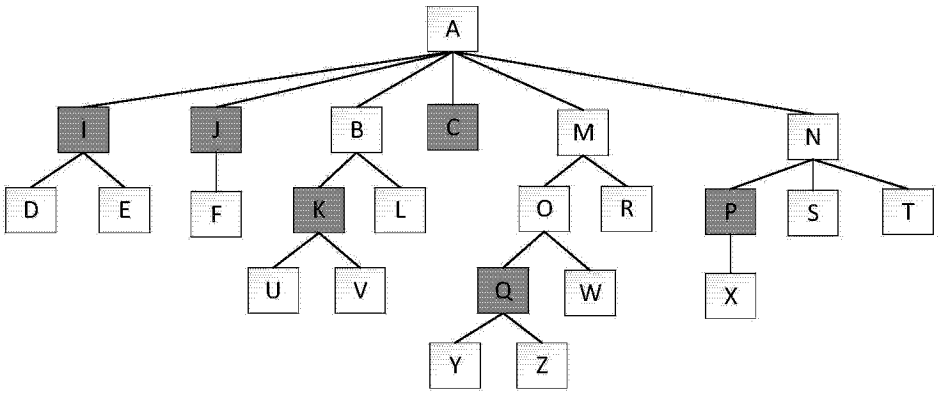

A
I
J
B
C
M
N
D
E
F
K
L
O
R
P
S
T
U
V
Q
W
X
Y
Z

图 4