(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 107220320 A
(43)申请公布日 2017.09.29

(21)申请号 201710356745.5

(22)申请日 2017.05.19

(71)申请人 湘潭大学
地址 411105 湖南省湘潭市雨湖区湘潭大学

(72)发明人 程戈 欧阳建权 周金海 何春辉

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

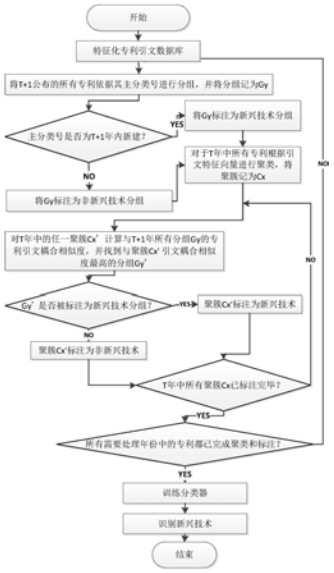
权利要求书1页 说明书5页 附图2页

(54)发明名称

一种基于专利引文的新兴技术识别方法

(57)摘要

本发明涉及数据挖掘领域，具体涉及一种基于专利引文的新兴技术识别方法。步骤如下：S1特征化专利引文；S2将T+1年专利依据其主分类号分组记为Gy；S3如果该主分类号是T+1年新建的则标注Gy为新技术分组；S4对于T年专利根据专利引文特征向量聚类，聚簇记为Cx；S5计算T年的任一C'_x与T+1年Cy的专利同引的耦合度；S6找到与C'_x耦合度最高的G'_y；S7如果G'_y为新兴技术分组则标记为新兴技术；S8循环S4直至T年的Cx都被标记；S9循环S1直至除最大年份外的专利都完成聚类与标注；S10使用标注数据训练分类器；S11使用该分类器判定新兴技术。本发明提出的方法具有较高的新兴技术识别正确率和适用性。



CN 107220320 A

1.基于专利引文的新兴技术识别方法，所述方法包括以下步骤：

S1特征化引文数据库；

S2将在T+1年公布的每一项专利依据其主分类号进行分组，将分组记为Gy；

S3如果该主分类号是T+1年新建立的，将Gy标注为新技术分组，否则记为非新技术分组；

S4对于T年中所有专利根据专利引文特征向量进行聚类，将聚簇记为Cx；

S5对于T年的任一C_x'计算与T+1年所有分组Cy的专利同引的耦合度；

S6找到与C_x'专利同引的耦合度最高的分组G_y'；

S7如果G_y'为新兴技术分组，将聚簇C_x'标记为新兴技术，否则标记为非新型技术；

S8循环步骤4，直至T年所有的聚簇Cx被标记完毕；

S9循环步骤1，直至专利数据除了年份最大的其他专利都完成聚类与标注；

S10 采用标注数据训练分类器；

S11 使用该分类器判定基于专利引文特征向量的聚簇是否为新兴技术。

2.根据权利要求1的方法，其中在所述步骤S1中，特征化引文数据库是指引文数据表达(或者特征)的选择，既抽取引文或专利文件的部分指标数据作为特征数据，多个特征数据构成特征向量，例如选取权利要求项数、引文总数量、非专利文献引文数量、专利分类号、技术生命周期、被引技术的相似性指数、被引技术所有者平均相似性指数等作为特征向量。

3.根据权利要求1-2中任何一项的方法，其中在所述步骤S5中，专利同引的耦合度是指聚簇Cx和Gy的文献耦合相似度(BCS)，计算公式为：

$$BCS_{XY} = \frac{n(C_x \cap G_y)}{n(C_x \cup G_y)}$$ 。

一种基于专利引文的新兴技术识别方法

技术领域

[0001] 本发明涉及计算机数据挖掘领域，具体涉及一种基于专利引文的新兴技术识别方法。

背景技术

[0002] 当今世界，科技的发展已经进入到了一个前所未有的时代。新兴技术发展势头强劲，进步速度迅猛，技术类型层出不穷。新兴技术是新技术的一部分，反过来，新技术就不一定属于新兴技术，正因如此，在所有新技术中对新兴技术进行有效识别就显得至关重要，它将直接关乎到我们的经济、科技的发展速度。随着社会发展与科技进步，各领域里大量的新兴技术如雨后春笋般涌现出来。但是真正能够进入市场并产生较大社会影响的却是寥寥无几，因而，谁能率先识别并应用这些技术指导生产实践，谁就能在竞争中脱颖而出，从而引领群雄。随着社会的发展，新兴技术识别的手段和方法越来越多，复杂性也越来越高，识别难度也在逐步增大。

[0003] 识别方法主要分为主观识别方法和基于文献的识别方法。最早的新兴技术识别方法主要采用专家讨论的形式来实现，此方法比较便捷，主观方法取决于专家的个人经验和能力，存在追随权威和随众现象，以及缺乏客观评价标准等弊端。随着计算机技术的发展，人们收集处理数据能力越来越强。基于文献的新兴技术识别方法成为主要的研究趋势。依据文献来源分为基于非专利文献与专利文献测新兴技术识别方法。主要采用文本聚类技术、主题提取、共词分析、网络演化等方法对新兴技术的识别进行实证研究，利用这些方法来识别新兴技术。通过从这些文献中抽取特征词来构成实体，然后在构建识别模型，在一定程度上降低了主观性的影响，但是特征词抽取的难度较大，而且会造成信息损失。

[0004] 在新兴技术识别中，目标技术和新兴技术的依赖性起到了关键的作用，并且技术发展越快，新兴技术的作用就越突出。正因如此，在所有新技术中对新兴技术进行有效识别就显得至关重要，它将直接关乎到中国的经济、科技的发展速度。随着社会的全面发展，各大领域里的新兴技术快速的涌现出来。但是真正能够进入市场并产生较大社会影响的却是寥寥无几，因而，谁能率先识别并应用这些技术指导生产实践，谁就能在竞争中脱颖而出，从而引领群雄。

发明内容

[0005] 本发明通过对特征化处理的引文数据进行新兴技术标注与识别。采用聚类方法对特征化的引文信息进行聚类，将相似特征信息的专利数据划分到同一个聚族，再利用往年的新兴技术与专利分类号得关系对聚族进行新兴技术标注，利用标注的数据训练分类器，将新兴技术的识别问题转化为一个分类问题。

[0006] 基于专利引文的新兴技术识别方法，所述方法包括以下步骤：

S1特征化用于训练的引文数据库；

S2将在T+1年公布的每一项专利依据其主分类号进行分组，将分组记为Gy；

S3如果该主分类号是T+1年新建立的，将Gy标注为新技术分组，否则记为非新技术分组；

S4对于T年中所有专利根据专利引文特征向量进行聚类，将聚簇记为Cx；

S5对于T年的任一C_x'计算与T+1年所有分组Cy的专利同引的耦合度；

S6找到与C_x'专利同引的耦合度最高的分组G_y'；

S7如果G_y'为新兴技术分组，将聚簇C_x'标记为新兴技术，否则标记为非新型技术；

S8循环步骤4，直至T年所有的聚簇Cx被标记完毕；

S9循环步骤1，直至专利数据除了年份最大的其他专利都完成聚类与标注；

S10 采用标注数据训练分类器；

S11 使用该分类器判定基于专利引文特征向量的聚簇是否为新兴技术。

[0007] 所述步骤S1中，特征化引文数据库是指引文数据表达(或者特征)的选择，既抽取引文或专利文件的部分指标数据作为特征数据，多个特征数据构成特征向量，例如选取权利要求项数、引文总数量、非专利文献引文数量、专利分类号、技术生命周期、被引技术的相似性指数、被引技术所有者平均相似性指数等作为特征向量。

[0008] 所述步骤S5中，专利同引的耦合度是指聚簇Cx和Gy的文献耦合相似度(BCS)，计算公式为：

$$BCS_{XY}=\frac{n(C_x\cap G_y)}{n(C_x\cup G_y)}$$

本发明的技术效果或优点：

相比现有的技术方案，本发明提出的基于专利引文分析的新兴技术识别方法可以降低现有识别方法的主观性，简化了特征提取的复杂度，可以客观快速的对专利数据进行新兴技术标注，这些标注数据可以用于训练各种分类器，因此该方法具有良好的可扩展性，可以高效迅速准确的预测新兴技术。

附图说明

[0009] 图1是基于专利引文的新兴技术识别方法流程图。

[0010] 图2是深度神经网络分类器的系统结构图

具体实施方式

[0011] 下面结合附图和实施例，对本发明的具体实施方式做进一步描述。

[0012] 基于专利引文的新兴技术识别方法，如图1所示，所述方法包括以下步骤：

S1特征化用于训练的引文数据库；

S2将在T+1年公布的每一项专利依据其主分类号进行分组，将分组记为Gy；

S3如果该主分类号是T+1年新建立的，将Gy标注为新技术分组，否则记为非新技术分组；

S4对于T年中所有专利根据专利引文特征向量进行聚类，将聚簇记为Cx；

S5对于T年的任一C_x'计算与T+1年所有分组Cy的专利同引的耦合度；

S6找到与C_x'专利同引的耦合度最高的分组G_y'；

S7如果G_y'为新兴技术分组，将聚簇C_x'标记为新兴技术，否则标记为非新型技术；

S8循环步骤4，直至T年所有的聚簇Cx被标记完毕；

S9循环步骤1，直至专利数据除了年份最大的其他专利都完成聚类与标注；

S10 采用标注数据训练分类器；

S11 使用该分类器判定基于专利引文特征向量的聚簇是否为新兴技术。

[0013] 在步骤S1中，特征化引文数据库是指引文数据表达(或者特征)的选择，既抽取引文或专利文件的部分指标数据作为特征数据，多个特征数据构成特征向量。在本实施例中采用如下特征数据：

1)权利要求项数；2)引文总数量；3)非专利文献引文数量；4)专利分类号；5)技术生命周期，本实施例中采用如下计算公式：$TCT_i = \text{median}_j\{|T_i - T_j|\}$

其中T_i是第i篇专利申请日期，T_j是第i篇专利引用的第j篇专利的申请日期；

6)被引技术的相似性指数(CTSI)专利分类系统对不同领域的技术进行了划分。大类只是限定了大概的领域，而小类才会给出更具体的领域，在实际中往往是采用大类和小类相结合来共同构成专利的分类号。本实施例采用如下的计算公式：

下面给出用于两个主分类号之间相似性计算的公式：

$$CS_{ij} = \begin{cases} 0, & \text{如果 i 和 j 大类和小类都不相同} \\ 0.5, & \text{如果 i 和 j 大类相同，小类不同} \\ 1, & \text{如果 i 和 j 大类和小类都相同} \end{cases}$$

如果一项专利往往拥有几项分类号，因此需要求出两项专利分类号之间的平均相似度(PCS_{pq})，以下是PCS_{pq}的表达式：

$$PCS_{pq} = \frac{\sum_{i=1}^{N_p}\sum_{j=1}^{N_q} CS_{ij}}{N_p N_q}$$

这里N_p和N_q表示专利p和专利q各自所拥有的分类号的数量。

[0014] 最后，再来计算第x篇专利的被引技术相似性指数，指标的计算公式如下：

$$CTSI(x) = \frac{\sum_{n=1}^{N} PCS_{xn}}{N}$$

此处，N是x引用的专利总数，n是被x引用的第n项专利；

7)被引技术所有者平均相似性指数(CASI)。一项专利通常情况下有一个或多个专利权人，采取下面的公式计算两项技术的专利权人相似性指标：

$$CASI(x) = \frac{\sum_{n=1}^{N} AS_{xn}}{N}$$

其中

$$AS_{pq} = \frac{\sum_{i=1}^{N_p}\sum_{j=1}^{N_q} A_i A_j}{N_p N_q}$$

这里N_p和N_q是专利p和专利q各自的专利权人的数量，

$$A_iA_j=\begin{cases}1，\text{如果专利i和专利j的专利权人是相同的}\\0，\text{否则}\end{cases}$$。

[0015] 在步骤S4中，对专利引文数据进行特征化后进行聚类操作，在本实施例中结合两种聚类算法和美国专利分类体系的优势来设计聚类步骤。首先，使用DBSCAN聚类算法按不同的年份对引文数据进行聚类，得到该数据集的聚簇类别数K1，然后考虑美国专利分类体系中的大类数目为450，这样就可以得到两个聚类的数目，为了减少误差，最终取这两个类别数的平均值，即K=(K1+450)/2，并向上取整。这样得到的这个K就更加接近真实的类别数，然后将K值带入K-means聚类算法，按不同年份对引文数据进行聚类。

[0016] 在步骤S5中专利同引的耦合度是指聚簇Cx和Gy的文献耦合相似度(BCS)，在本实施例中采用以下计算公式：

$$BCS_{XY}=\frac{n(C_x\cap G_y)}{n(C_x\cup G_y)}$$

在步骤S10采用标注数据训练分类器，在本实施例中采用深度神经网络作为分类器。如图2所示，该分类器可分为四层，如图所示第一层是输入层，本层需要对输入数据进行预处理，形成统一格式的数据矩阵；然后就是深度神经网络层，该层由3层RBM堆叠而成，主要功能就是对数据进行重构，自动提取出合适的特征；接下来是分类器所在的决策层，该层使用Logistic Regression算法来设计分类器，然后再对分类结果应用Softmax算法进行概率转换。将结果中概率较大的所对应的下标作为分类结果，因为原分类结果只有两个维度，因此最终的分类结果只有0或者1，0代表非新兴技术，1代表新兴技术。

[0017] 本实施例中选取RBM算法作为深度信念网络各层之间的重构算法。信念网络里面各层之间RBM调节的主要通过多个隐含层的相互转化，从而为RBM内部的参数调节提供训练目标，通过降低重构矩阵与原矩阵的差异来达到调节RBM参数的最终目标。对于RBM的参数学习采用对数似然度极大化的思想来获取RBM算法中参数θ，θ的表达式定义如下：

$$\theta^*=\arg\max_\theta\sum_{k=1}^{K}\log P(v^{(k)}|\theta)$$

为了获得最优参数，可以使用随机梯度上升法，其中关键步骤是计算关于各个模型参数的偏导数。由式2.1可以求出求关于分布P的均值。

[0018] 深度模型的反馈微调主要通过三个过程来实现：加载参数、构造数据矩阵、循环调节。其中前两个过程主要是在完成整个深度模型前期的准备工作，而循环调节过程才是整个深度模型反馈调节机制的核心。随层次增加，深度表示的维度也在逐渐变化，在反馈微调阶段，先通过识别模型自底向上进行转换，到了最上层之后，再进行自顶向下的生成模型的转换，从而生成对各个层次的重构展现。最后通过对原始表示和重构表示的不断优化调节，从而来实现两者的误差最小化。

[0019] 本实施例中采用BP算法对自底向上的识别模型和自顶向下的生成模型相结合的方式来进行微调。经过网络的识别模型，本文可以近似得到深度模型对输入数据最初的各个层次上的表示形式，并得到一个深度模型对样本最高层次的抽象表示形式，通过该生成

模型，本文可以从模型的最高层次表示形式出发，重构展示深度模型对样本数据的各个层次的表示，这样就可以为原来的每个层级的训练提供优化目标。经过各个层次的不断调节，生成模型就可以重构出具有较低误差的训练样本，通过以上步骤模型可以自动学习出原样本的数据特征，即最高层次的抽象表示形式。

[0020] 上面是本发明提供的基于专利引文的新兴技术识别方法优选实施方式，并不构成对本发明的保护权限，任何在本发明上的改进，只要原理相同，都包含在本发明的权利要求保护范围之内。

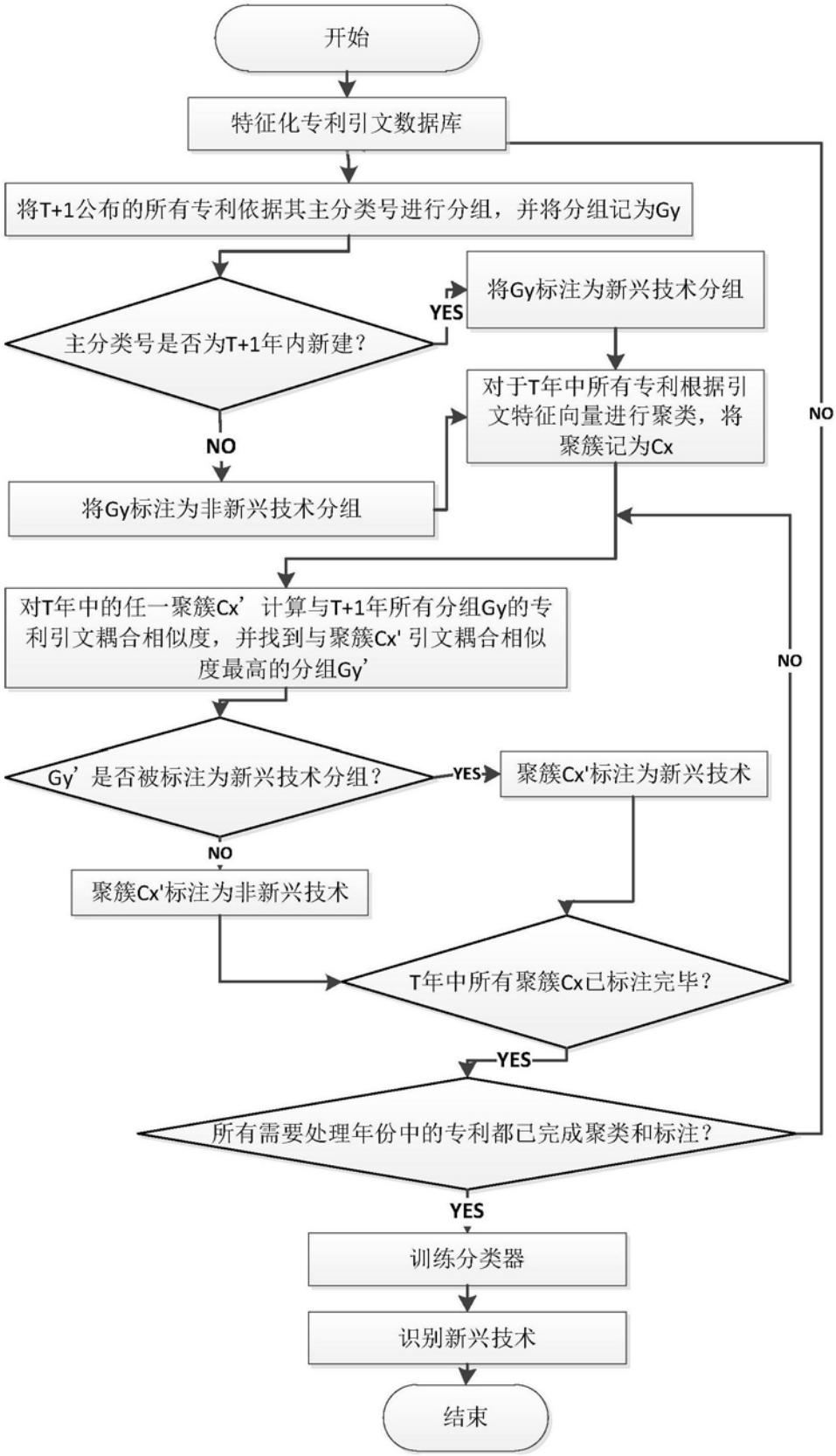

开始
特征化专利引文数据库
将T+1公布的所有专利依据其主分类号进行分组，并将分组记为Gy
主分类号是否为T+1年内新建？
YES
将Gy标注为新兴技术分组
NO
将Gy标注为非新兴技术分组
对于T年中所有专利根据引文特征向量进行聚类，将聚簇记为Cx
NO
对T年中的任一聚簇Cx’计算与T+1年所有分组Gy的专利引文耦合相似度，并找到与聚簇Cx' 引文耦合相似度最高的分组Gy’
Gy’是否被标注为新兴技术分组？
YES
聚簇Cx'标注为新兴技术
NO
聚簇Cx'标注为非新兴技术
T年中所有聚簇Cx已标注完毕？
NO
YES
所有需要处理年份中的专利都已完成聚类和标注？
NO
YES
训练分类器
识别新兴技术
结束

图1

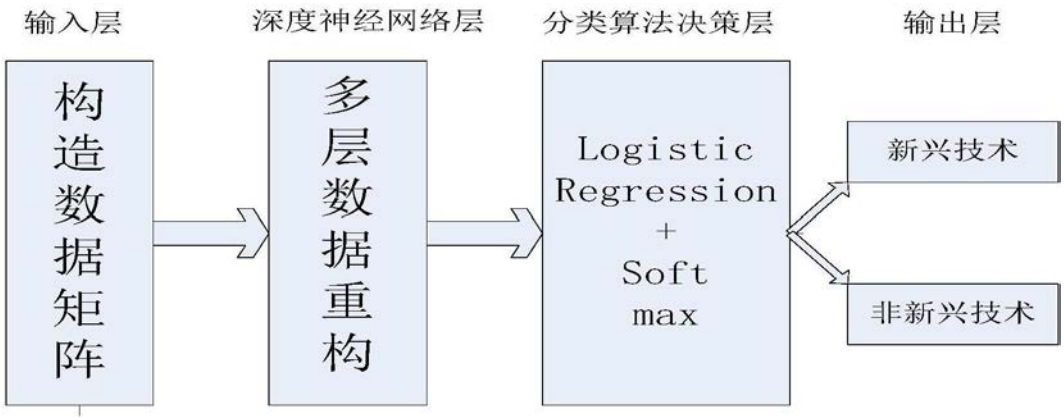

输入层
深度神经网络层
分类算法决策层
输出层
构造数据矩阵
多层数据重构
Logistic Regression + Soft max
新兴技术
非新兴技术

图2