

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2018 年 12 月 6 日 (06.12.2018)



(10) 国际公布号
WO 2018/218708 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2017/089139
- (22) 国际申请日: 2017 年 6 月 20 日 (20.06.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201710395849.7 2017 年 5 月 27 日 (27.05.2017) CN
- (71) 申请人: 中国矿业大学 (CHINA UNIVERSITY OF MINING AND TECHNOLOGY) [CN/CN]; 中国江苏省徐州市大学路 1 号, Jiangsu 221116 (CN)。
- (72) 发明人: 周勇 (ZHOU, Yong); 中国江苏省徐州市大学路 1 号, Jiangsu 221116 (CN)。刘兵 (LIU, Bing); 中国江苏省徐州市大学路 1 号, Jiangsu

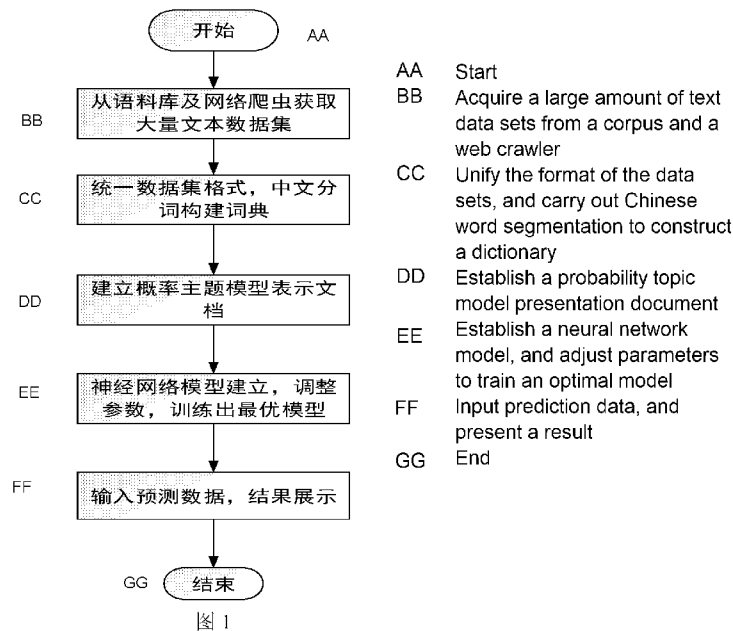
221116 (CN)。刘敬学 (LIU, Jinxue); 中国江苏省徐州市大学路 1 号, Jiangsu 221116 (CN)。王重秋 (WANG, Chongqiu); 中国江苏省徐州市大学路 1 号, Jiangsu 221116 (CN)。

(74) 代理人: 南京瑞弘专利商标事务所 (普通合伙) (NANJING RUIHONG PATENT & TRADEMARK OFFICE (ORDINARY PARTNERSHIP)); 中国江苏省南京市玄武区太平门街 1 号 304 室, Jiangsu 210016 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,

(54) Title: DEEP-LEARNING-BASED PUBLIC OPINION HOTSPOT CATEGORY CLASSIFICATION METHOD

(54) 发明名称: 一种基于深度学习的舆情热点类别划分方法



(57) Abstract: A deep-learning-based public opinion hotspot category classification method, comprising: collecting and pre-processing a training data set, establishing a probability topic presentation model, carrying out two probability distribution presentations, i.e. document-topic and topic-vocabulary, on a text data set, and inputting an obtained topic-vocabulary matrix into a pre-established neural network model for training so as to learn text features; and a network output layer selecting Softmax to carry out normalization processing and classification prediction. The method solves the dimensionality reduction problem of long text public opinion hotspot data, and realises the automatic extraction of deep features of public opinion hotspot information, so that the classification of multiple categories of public opinion hotspots is more accurate.

MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

— 发明人资格 (细则4.17(iv))

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种基于深度学习的舆情热点类别划分方法, 包括: 训练数据集的采集和预处理, 建立概率主题表示模型, 对文本数据集进行文档-主题和主题-词汇两个概率分布表示, 并将获得的主体-词汇矩阵输入到预先搭建好的神经网络模型中训练学习文本特征, 网络输出层选择 Softmax 归一化处理分类预测; 所述方法解决了长文本舆情热点数据的降维问题, 实现了舆情热点信息的深层特征的自动提取, 使得舆情热点多种类划分更加准确。

一种基于深度学习的舆情热点类别划分方法

技术领域

本发明涉及深度学习和自然语言处理技术领域，尤其是一种基于深度学习的舆情热点类别划分方法，方法是一种具体应用于舆情分析的舆情热点科学类型划分的中文文本分类方法。

背景技术

在如今信息技术高度发达的年代，社会上某个事件发生后，广大群众可以迅速通过各种途径了解到事情的来龙去脉，随之产生大量的评论。这就是舆情，对于民众舆情的分析对后续如何处理事件起着至关重要的作用。在分析某一舆情热点之前应先对其进行科学的类型界定。比如热点事件一般可以分为突发自然灾害事件、生产安全事故、群体性事件、公共卫生事件、公权力形象、司法事件、经济民生事件、社会思潮、境外涉华突发事件等。类型界定的正确与否，对后续事件的分析比较，应对政策的制定都极为关键。

现有的相关技术有相近专利 CN201310072137.3 公开了一种网络舆情分析方法，以主题匹配代替简单舆情分析中的词语匹配，分析并判断页面是否属于舆情，属于一个二分类的问题。专利 CN201410023154.2 涉及一种基于情感分析和隐马尔科夫模型融合的方法，通过利用情感倾向性信息，提高了股市预测的准确性。

通过调查和分析发现，现有技术存在以下问题：

(1) 舆情热点信息由于其特殊性，文本的长度不确定，对于较长的文本若采用传统的文本表示方法，对于数据降维处理而言难度较大；

(2) 仅考虑数据的情感倾向性信息，对分类判断的特征提取单一，无法提取更深层次的特征，分类的效果不佳；

(3) 现有舆情热点的分类技术，大部分是舆情的二分类，缺乏对舆情热点种类划分更细致的解决方案。

发明内容

发明目的：本发明针对来源于舆情热点事件的文本数据，为解决现有技术的分类功能不足，准确率不高的问题，提出一种基于深度学习的舆情热点类别划分方法。

技术方案：为实现上述技术效果，本发明的技术方案为：

一种基于深度学习的舆情热点类别划分方法，该方法包括两个阶段，第一个阶段为

训练阶段，第二个阶段为分类阶段；

训练阶段包括步骤：

(1) 采集满足需求量的包含舆情热点的文本数据作为训练数据，并根据舆情热点的种类在训练数据中添加标签；对采集到的训练数据进行预处理，将训练数据保存为统一格式并构建训练数据中文词典；

(2) 根据训练数据中文词典建立训练数据概率主题模型，并采用 Gibbs 抽样方法学习训练数据概率主题模型，得到训练数据的文档-主题分布矩阵和主题-词汇分布矩阵，并依据文档-主题分布矩阵采集训练数据的分类标签，将主题-词汇分布矩阵作为深度学习模型所需训练矩阵；

(3) 建立深度学习模型，所述深度学习模型包括具有多层隐层的 DBN 神经网络和 softmax 层，所述 DBN 神经网络的最后一层隐层的输出数据作为 softmax 层的输入数据，softmax 层作为所述深度学习模型的输出层，对 DBN 神经网络的输出结果进行归一化处理；

(4) 将训练数据、训练数据中文词典、训练数据的分类标签输入深度学习模型进行循环训练，并保存满足预设指标的最优深度学习模型；

分类阶段包括步骤：

(5) 采集预测数据，根据舆情热点的种类在预测数据中添加标签；对采集到的预测数据进行预处理，将预测数据保存为统一格式并构建预测数据中文词典；

(6) 根据预测数据中文词典建立预测数据概率主题模型，并采用 Gibbs 抽样方法学习预测数据概率主题模型，得到预测数据的文档-主题分布矩阵和主题-词汇分布矩阵，并依据文档-主题分布矩阵采集预测数据的分类标签；

(7) 将预测数据，预测数据分类标签和预测数据中文词典输入步骤(4)得到的最优模型，最优模型根据预测数据词典将输入的预测数据映射为输入层大小的矩阵，最优模型的输出层预测数据的多分类的结果进行归一化处理并根据预设阈值选择出预测数据的一种类别或一组类别。

进一步的，所述训练数据和预测数据的采集方法为：从网络公开新闻语料库中直接获取和/或通过网络爬虫从网络新闻中爬取。

进一步的，所述预处理的具体步骤包括：

a、统一数据的格式：将采集到的 xml 文件去掉尖括号，转存为 TXT 格式；将每个

xml 文件命名为：舆情热点种类标签+文件编号；

b、将每类文件存为一个 TXT 文件，TXT 文件中每一行为一条新闻数据，TXT 文件名为舆情热点种类标签；TXT 文件的总数目为舆情热点的种类数目；

c、采用中文分词工具对采集的文本数据进行分词表示和去除停用词处理，得到文本数据的中文词典；中文词典中，每个中文单词对应一个索引号，中文词典的每一行为：一个单词+该单词的索引号。

进一步的，所述构建训练数据概率主题模型的步骤包括：

1) 对采集到的训练数据中的每篇文档中的每个词随机的赋值一个话题编号；

2) 重新遍历训练数据，对每篇文档的每个词，按照 Gibbs 采样公式重新采样话题编号，然后对训练数据中的每个词所属的主题编号进行更新，更新即为重新对每个词进行随机赋值；

3) 重复步骤 2) 所述的对训练数据的采样过程，直到采样结果收敛；

4) 采样收敛后，统计训练数据中话题词汇频率矩阵，即为主题-词汇分布矩阵。

进一步的，所述得到最优深度学习模型的步骤包括：

(5-1) 将训练数据、训练数据中文词典、训练数据的分类标签输入深度学习模型，并根据训练数据中文词典将输入数据映射为深度学习模型输入层形式的矩阵；

(5-2) 将读取的训练数据划分为训练数据集和测试数据集，所述训练数据集参与深度学习模型的训练过程，测试数据集用于测试每次训练的准确率；

(5-3) 设置每次训练的数据集大小和最大迭代次数，初始化深度学习模型；

(5-4) 将训练数据集中的数据分批次送入深度学习模型，计算每次训练的损失函数，所述损失函数用于描述深度学习模型的输出结果与实际期望结果的误差；采用梯度下降的方法来最小化损失函数，每下降 N 步，就用测试数据集测试训练数据的准确率；

(5-5) 循环执行步骤 (5-4) 直至满足预设的迭代次数或者训练数据的准确率达到预设期望值，此时的深度学习模型为最优深度学习模型。

有益效果：与现有技术相比，本发明具有以下优势：

本发明是针对舆情热点数据的特殊性和现有技术的不足之处，将 LDA 概率主题文本表示方法和 DBN 神经网络模型相结合，实现了高准确度多分类的效果。本算法的优点可以分为两个方面：

一是可以对不同长度的文本数据进行规格化表示和主题划分，实现了初步的数据降

维过程，提供了分类算法的效率。

二是 DBN 神经网络相较于其他复杂的网络模型简单，“无监督训练+有监督微调”的训练方式，使得在保证分类准确率的前提下，训练阶段的时间花费较少，也降低训练的复杂度。神经网络输出层选择 Softmax 回归模型，将多分类的结果进行归一化处理并根据预设阈值选择出预测的一种类别或一组类别。实现了舆情热点信息的深层特征的自动提取，使得舆情热点多种类划分更加准确。

附图说明

图 1 是本发明所述基于深度学习的舆情热点类别划分方法的流程图；

图 2 是 LDA 概率模型示意图；

图 3 是 DBN 神经网络模型示意图。

具体实施方式

下面结合附图对本发明作更进一步的说明。

本发明的整个流程划分为两个阶段，前期的模型训练阶段和后期的预测分类阶段。这两个阶段对于数据集的预处理过程和概率主题表示过程相似，也就是说我们读入神经网络中的文本表示在两个阶段是一样，不同的是前期的文本用于模型的训练，后期的文本用于进行结果分类预测。

下面根据附图来进一步的描述本发明的技术方案。

图 1 是本发明实施流程图，在具体实施的过程中可以分为两个阶段。第一阶段，包括训练数据集的采集、预处理过程、概率主题文本表示和深度学习模型的训练部分。第二阶段，包括预测数据的选取和预处理和概率主题文本表示，读取第一阶段的训练好的模型，进行分类预测。具体为：

（一）第一阶段：

1. 训练数据的采集和预处理

（1）数据集采集：

训练数据集的来源有两种，一种是选用网上公布的新闻数据集主要有搜狗实验室的新闻语料和网易新闻语料，根据舆情热点的种类进行人工的划分添加标签。另一种方式采用爬虫程序按照种类从国内各大主流的媒体采集数据，添加标签。两种合一，构成大数据集合。

（2）预处理：

a、统一数据集的格式。将采集到的 xml 文件去尖括号，转存为 TXT 格式。xml 文件文件的名字为舆情热点的种类标签+文件编号；

b、每类文件存为一个 TXT 文件，其中每一行为一条新闻数据；TXT 文件的文件名为为舆情热点的种类标签，TXT 文件的数目为舆情热点的种类数目。

c、采用中文分词工具对采集的中文文本数据进行分词表示，去停用词并构建中文词典。

2. 概率主题模型的建立

依据经预处理建立的词典将文本表示成文档-主题（doc-topic）的分布和主题-词汇（topic-Word）的分布，并依据文档-主题（doc-topic）来进行人工神经网络分类标签采集，主题-词汇（topic-Word）矩阵则为神经网络所需训练矩阵。

3. 深度学习模型的建立

深度学习训练模型选择深度置信网络（DBN），是由两层受限玻尔兹曼机（RBM）组成的有向图连接模型，其中第一层为可视层（v）也叫作输入层，它由 m 个可视节点组成，第二层为隐层（h），也就是特征提取层，由 n 个隐藏节点组成。建立具有三层隐层的 DBN 神经网络。输出层选择 Softmax 层。

4. 输入数据，进行深度学习模型训练，保存最优模型

（1）深度学习模型读取输入数据，输入数据标签和输入数据中文词典，并根据输入数据中文词典将输入数据映射为深度学习模型输入层大小的矩阵。

（2）将读取的训练数据划分为 train 集和 test 集，并设置每次训练的数据集大小，按照每批次迭代的数据集大小计算最大迭代次数。

（3）初始化神经网络模型。

（4）将训练数据分批次输入模型，并计算损失，在一定的迭代次数用 test 数据集测试 accuracy。

（5）在迭代次数内，循环训练。直到 accuracy 达到预期。保存最优模型。

（二）第二阶段：

1. 预测数据的在线采集和预处理

利用爬虫程序，在国内主流的网站采集待分析的文本数据，采取和第一阶段一样的预处理过程。

2. 概率主题模型的建立。

3. 读取第一阶段的最优模型，输入预测数据，做出预测

读取第一阶段也就是模型训练时的最优模型，读取预测数据，预测数据标签和预测数据词典，并根据预测数据词典将输入的预测数据映射为输入层大小的矩阵。分类输出

层 Softmax 回归模型，将多分类的结果进行归一化处理并根据预设阈值选择出预测的一种类别或一组类别。

4. 分类结果的展示。

展示有两种选择：a. 对输入数据，预测模型给出对应的预测所分某一种种类。b. 设置类别数目的阈值，预测模型给出输入数据的一组类别的名称和相应的概率。

图 2 所示为 LDA 模型的训练细节。假设我们收集的待分类的原始文本数据集中有 D 篇文档，其中所有的词汇表示为： $w = (w_1, \dots, w_D)$ 。去词汇对应的主题表示为： $z = (z_1, \dots, z_D)$ 。其中 w_m 表示的是第 m 篇文档中的词汇， z_m 表示为这些词对应的主题的编号。

如图 2 所示，LDA 概率模型主要有两个物理过程： $\vec{\alpha} \rightarrow \vec{\theta}_m \rightarrow z_{m,n}$ ：这个物理过程代表的含义是：在生成第 m 篇文档的时候，先随机生成一个有关文档话题参数 $\vec{\theta}_m$ ，然后通过这个参数生成文档中第 n 个词对应的话题编号 $z_{m,n}$ 。 $\vec{\beta} \rightarrow \vec{\phi} \rightarrow (w_{m,n} | k = z_{m,n})$ ：这个物理过程表示，随机的生成一个有关话题-词汇的参数 $\vec{\phi}_k$ ，在生成的参数 $\vec{\phi}_k$ 中，挑选话题 $k = z_{m,n}$ 的词汇生成第 m 篇文档中的第 n 个词汇。

最后根据图 3 介绍 DBN 模型的训练细节。

DBN 是由两层 RBM 组成的有向图连接模型，在预训练过程中，左侧 x 是输入，右边 y 是输出。当所有层训练结束后，由最上层开始向下有监督的进行微调。如图 3 所示训练一个具有三个隐藏层的 DBN。其中 $W1, W2, W3$ 即为从特征中学习到的特征计算所得到的权值。

DBN 模型是基于受限玻尔兹曼机模型 (RBM) 作为网络的基本建模单元，RBM 网络共有两层，其中第一层为可视层 (v) 也叫作输入层，它由 m 个可视节点组成，第二层为隐层 (h)，也就是特征提取层，由 n 个隐藏节点组成。如图所示的 DBN 模型则是在靠近可视层的部分使用有向图连接，在远离可视层的部分使用 RBM。

DBN 模型的训练包括无监督的预训练和有监督的微调两部分。

在预训练过程中，首先将可视向量值映射给隐藏节点，然后可视节点由隐藏节点重构；这些新的可视节点再次计算隐藏节点的数值，重构激活的隐藏节点，这样就获取新的隐藏节点。隐层和可视层输入之间的重构误差作为权值更新的主要计算依据。

在有监督的微调过程中使用 contrastive wake-sleep 算法进行优化，除了顶层的 RBM，

其他层的 **RBM** 的权重被分为向上的认知权重和向下的生成权重。

监督的微调过程还包括以下两个阶段：

Wake 阶段：认知过程中，通过外界的特征和向上的权重（认知权重）产生每一层的抽象表示，并且使用梯度下降修改层间的下行权重（生成权重）。

Sleep 阶段：生成过程，通过顶层表示和向下权重，生成底层的状态，同时修改层间向上的权重。

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，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

权 利 要 求 书

1、一种基于深度学习的舆情热点类别划分方法，其特征在于，该方法包括两个阶段，第一个阶段为训练阶段，第二个阶段为分类阶段；

训练阶段包括步骤：

(1) 采集满足需求量的包含舆情热点的文本数据作为训练数据，并根据舆情热点的种类在训练数据中添加标签；对采集到的训练数据进行预处理，将训练数据保存为统一格式并构建训练数据中文词典；

(2) 根据训练数据中文词典建立训练数据概率主题模型，并采用 Gibbs 抽样方法学习训练数据概率主题模型，得到训练数据的文档-主题分布矩阵和主题-词汇分布矩阵，并依据文档-主题分布矩阵采集训练数据的分类标签，将主题-词汇分布矩阵作为深度学习模型所需训练矩阵；

(3) 建立深度学习模型，所述深度学习模型包括具有多层隐层的 DBN 神经网络和 softmax 层，所述 DBN 神经网络的最后一层隐层的输出数据作为 softmax 层的输入数据，softmax 层作为所述深度学习模型的输出层，对 DBN 神经网络的输出结果进行归一化处理；

(4) 将训练数据、训练数据中文词典、训练数据的分类标签输入深度学习模型进行循环训练，并保存满足预设指标的最优深度学习模型；

分类阶段包括步骤：

(5) 采集预测数据，根据舆情热点的种类在预测数据中添加标签；对采集到的预测数据进行预处理，将预测数据保存为统一格式并构建预测数据中文词典；

(6) 根据预测数据中文词典建立预测数据概率主题模型，并采用 Gibbs 抽样方法学习预测数据概率主题模型，得到预测数据的文档-主题分布矩阵和主题-词汇分布矩阵，并依据文档-主题分布矩阵采集预测数据的分类标签；

(7) 将预测数据，预测数据分类标签和预测数据中文词典输入步骤(4)得到的最优模型，最优模型根据预测数据词典将输入的预测数据映射为输入层大小的矩阵，最优模型的输出层预测数据的多分类的结果进行归一化处理并根据预设阈值选择出预测数据的一种类别或一组类别。

2、根据权利要求 1 所述的一种基于深度学习的舆情热点类别划分方法，其特征在于，所述训练数据和预测数据的采集方法为：从网络公开新闻语料库中直接获取和/或通过网络爬虫从网络新闻中爬取。

3、根据权利要求 1 所述的一种基于深度学习的舆情热点类别划分方法，其特征在于，所述预处理的具体步骤包括：

a、统一数据的格式：将采集到的 xml 文件去掉尖括号，转存为 TXT 格式；将每个 xml 文件命名为：舆情热点种类标签+文件编号；

b、将每类文件存为一个 TXT 文件，TXT 文件中每一行为一条新闻数据，TXT 文件名为舆情热点种类标签；TXT 文件的总数目为舆情热点的种类数目；

c、采用中文分词工具对采集的文本数据进行分词表示和去除停用词处理，得到文本数据的中文词典；中文词典中，每个中文单词对应一个索引号，中文词典的每一行为：一个单词+该单词的索引号。

4、根据权利要求 3 所述的一种基于深度学习的舆情热点类别划分方法，其特征在于，所述构建训练数据概率主题模型的步骤包括：

1) 对采集到的训练数据中的每篇文档中的每个词随机的赋值一个话题编号；

2) 重新遍历训练数据，对每篇文档的每个词，按照 Gibbs 采样公式重新采样话题编号，然后对训练数据中的每个词所属的主题编号进行更新；

3) 重复步骤 2) 所述的对训练数据的采样过程，直到采样结果收敛；

4) 采样结果收敛后，统计训练数据中话题词汇频率矩阵，即为主题-词汇分布矩阵。

5、根据权利要求 4 所述的一种基于深度学习的舆情热点类别划分方法，其特征在于，所述得到最优深度学习模型的步骤包括：

(5-1) 将训练数据、训练数据中文词典、训练数据的分类标签输入深度学习模型，并根据训练数据中文词典将输入数据映射为深度学习模型输入层形式的矩阵；

(5-2) 将读取的训练数据划分为训练数据集和测试数据集，所述训练数据集参与深度学习模型的训练过程，测试数据集用于测试每次训练的准确率；

(5-3) 设置每次训练的数据集大小和最大迭代次数，初始化深度学习模型；

(5-4) 将训练数据集中的数据分批次送入深度学习模型，计算每次训练的损失函数，所述损失函数用于描述深度学习模型的输出结果与实际期望结果的误差；采用梯度下降的方法来最小化损失函数，每下降 N 步，就用测试数据集测试训练数据的准确率；

(5-5) 循环执行步骤 (5-4) 直至满足预设的迭代次数或者训练数据的准确率达到预设期望值，此时的深度学习模型为最优深度学习模型。

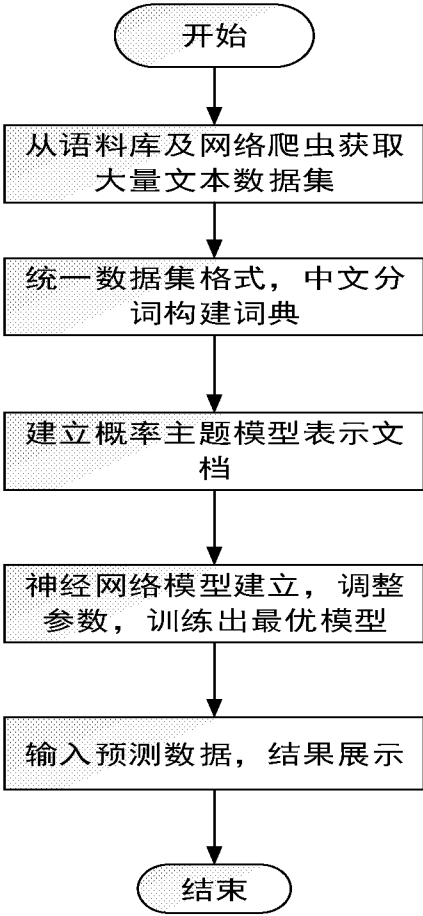


图 1

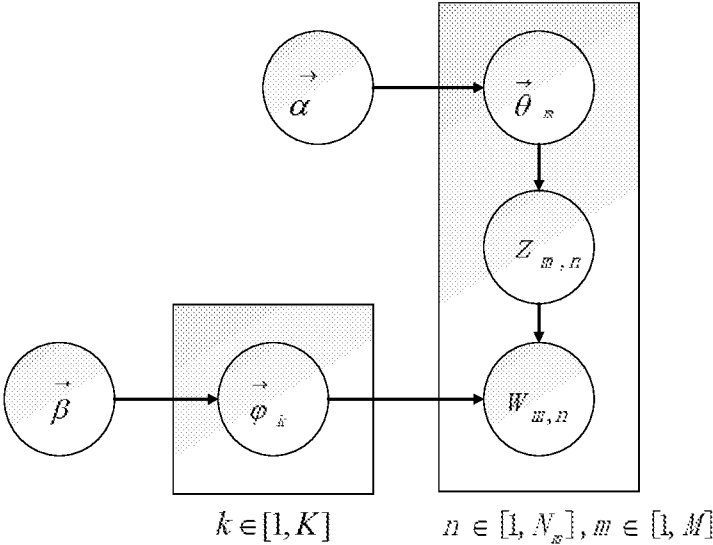


图 2

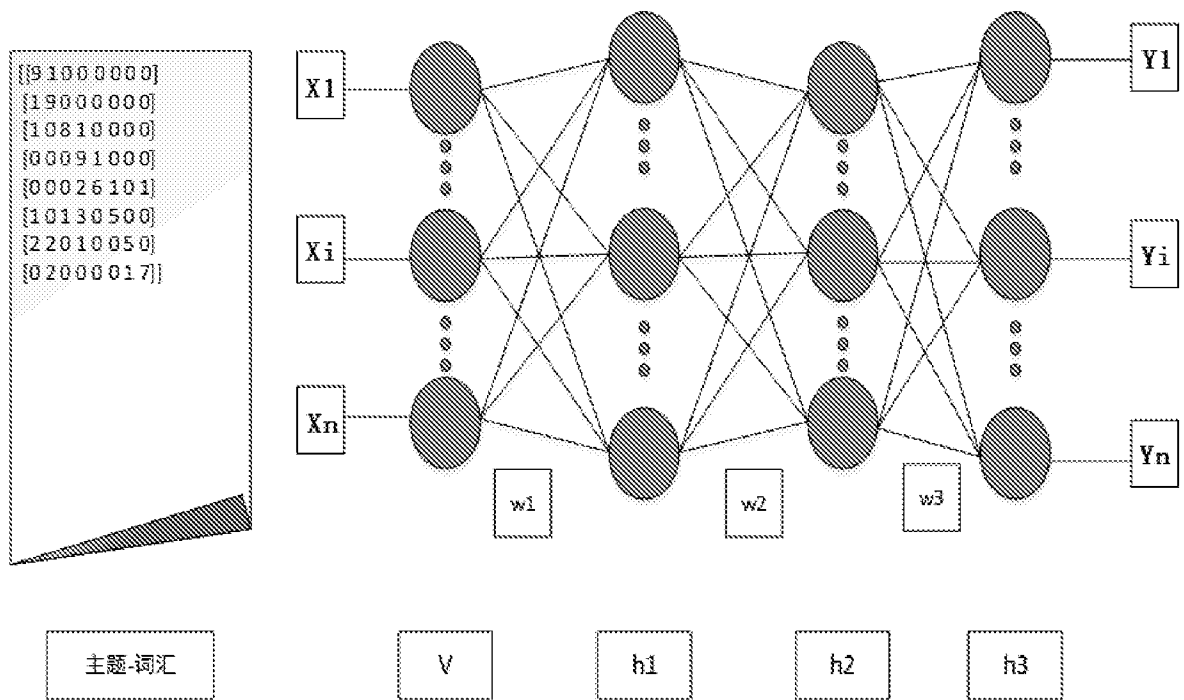


图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/089139

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, IEEE, CNKI: 文本, 文档, 分类, 类别, 舆情, 舆论, 深度, 主题, 词汇, 神经网络, text, document, sort, classification, public sentiment, deep, DBN, LDA, topic model, word, neural network

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	侯思耘. 基于主题模型和深度置信网络的文本分类方法研究. 中国优秀硕士学位论文全文数据库信息科技辑. 15 September 2015 (15.09.2015), no. 9, ISSN: 1674-0246, pp. 1138-1696, text, chapter 2, section 2.1, chapter 3, section 3.2, chapter 5. (Chinese Master's Theses Full-text Database (Information Science & Technology)), non-official translation (HOU, Siyun Research on Text Classification Method of Topic Modeling and Deep Brief Network)	1-5
Y	CN 104881685 A (TSINGHUA UNIVERSITY) 02 September (02.09.2015), description, paragraph [0061], and figure 2	1-5
A	CN 105975478 A (BEIJING JIAOTONG UNIVERSITY et al.) 28 September 2016 (28.09.2016), entire document	1-5
A	US 2007018953 A1 (THE BOEING COMPANY) 25 January 2007 (25.01.2007), entire document	1-5

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 26 January 2018	Date of mailing of the international search report 24 February 2018
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer REN, Xingchao Telephone No. (86-10) 53961375

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/089139

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104881685 A	02 September 2015	None	
CN 105975478 A	28 September 2016	None	
US 2007018953 A1	25 January 2007	US 2007094219 A1	26 April 2007
		EP 1750221 A1	07 February 2007

A. 主题的分类 G06F 17/30 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																	
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNPAT, WPI, EPODOC, IEEE, CNKI: 文本, 文档, 分类, 类别, 舆情, 舆论, 深度, 主题, 词汇, 神经网络, text, document, sort, classification, public sentiment, deep, DBN, LDA, topic model, word, neural network																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>Y</td> <td>侯思耘. "基于主题模型和深度置信网络的文本分类方法研究." 中国优秀硕士学位论文全文数据库信息科技辑., 第9期, 2015年 9月 15日 (2015 - 09 - 15), ISSN: 1674-0246, 第1138-1696页, 正文第二章第2.1节, 第三章第3.2节, 第五章</td> <td>1-5</td> </tr> <tr> <td>Y</td> <td>CN 104881685 A (清华大学) 2015年 9月 2日 (2015 - 09 - 02) 说明书第[0061]段, 附图2</td> <td>1-5</td> </tr> <tr> <td>A</td> <td>CN 105975478 A (北京交通大学 等) 2016年 9月 28日 (2016 - 09 - 28) 全文</td> <td>1-5</td> </tr> <tr> <td>A</td> <td>US 2007018953 A1 (THE BOEING COMPANY) 2007年 1月 25日 (2007 - 01 - 25) 全文</td> <td>1-5</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	Y	侯思耘. "基于主题模型和深度置信网络的文本分类方法研究." 中国优秀硕士学位论文全文数据库信息科技辑., 第9期, 2015年 9月 15日 (2015 - 09 - 15), ISSN: 1674-0246, 第1138-1696页, 正文第二章第2.1节, 第三章第3.2节, 第五章	1-5	Y	CN 104881685 A (清华大学) 2015年 9月 2日 (2015 - 09 - 02) 说明书第[0061]段, 附图2	1-5	A	CN 105975478 A (北京交通大学 等) 2016年 9月 28日 (2016 - 09 - 28) 全文	1-5	A	US 2007018953 A1 (THE BOEING COMPANY) 2007年 1月 25日 (2007 - 01 - 25) 全文	1-5
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
Y	侯思耘. "基于主题模型和深度置信网络的文本分类方法研究." 中国优秀硕士学位论文全文数据库信息科技辑., 第9期, 2015年 9月 15日 (2015 - 09 - 15), ISSN: 1674-0246, 第1138-1696页, 正文第二章第2.1节, 第三章第3.2节, 第五章	1-5															
Y	CN 104881685 A (清华大学) 2015年 9月 2日 (2015 - 09 - 02) 说明书第[0061]段, 附图2	1-5															
A	CN 105975478 A (北京交通大学 等) 2016年 9月 28日 (2016 - 09 - 28) 全文	1-5															
A	US 2007018953 A1 (THE BOEING COMPANY) 2007年 1月 25日 (2007 - 01 - 25) 全文	1-5															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
* 引用文件的具体类型: "A" 认为不特别相关的表示了现有技术一般状态的文件 "E" 在国际申请日的当天或之后公布的在先申请或专利 "L" 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) "O" 涉及口头公开、使用、展览或其他方式公开的文件 "P" 公布日先于国际申请日但迟于所要求的优先权日的文件 "T" 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 "X" 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 "Y" 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 "&" 同族专利的文件																	
国际检索实际完成的日期		国际检索报告邮寄日期															
2018年 1月 26日		2018年 2月 24日															
ISA/CN的名称和邮寄地址		受权官员															
中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		任兴超 电话号码 (86-10)53961375															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/089139

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	104881685	A	2015年 9月 2日	无			
CN	105975478	A	2016年 9月 28日	无			
US	2007018953	A1	2007年 1月 25日	US	2007094219	A1	2007年 4月 26日
				EP	1750221	A1	2007年 2月 7日