(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 110134786 A
(43)申请公布日 2019.08.16

(21)申请号 201910397064.2

(22)申请日 2019.05.14

(71)申请人 南京大学
地址 210093 江苏省南京市鼓楼区汉口路22号

(72)发明人 张雷 李博 许磊 顾溢 谢俊元

(74)专利代理机构 南京瑞弘专利商标事务所（普通合伙）32249
代理人 刘珊珊

(51)Int.Cl.
G06F 16/35(2019.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

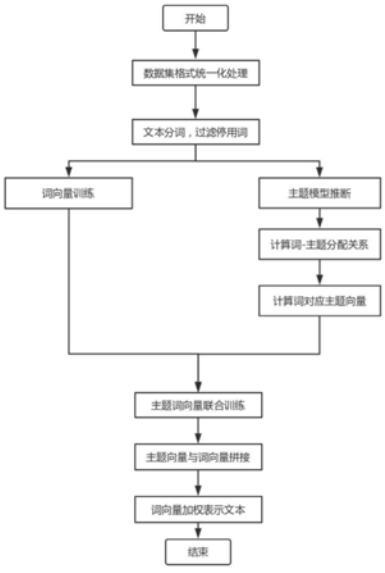
权利要求书3页 说明书7页 附图4页

(54)发明名称

一种基于主题词向量与卷积神经网络的短文本分类方法

(57)摘要

本发明公开了一种基于主题词向量与卷积神经网络的短文本分类方法，包括如下步骤：1)数据采集阶段：根据需求采集短文本数据，对其进行标签标注，作为训练集；2)数据预处理阶段：对文本进行分词，去停用词，无用文本过滤等；3)短文本特征表示，主题层面与词向量层面分别进行表征；4)主题词向量联合训练；5)卷积神经网络分类模型参数优化，迭代；6)新样本进行类别预测。本发明结合短文本数据特点，在特征表示阶段利用主题向量与词向量结合表示，对短文本自身数据特点进行语义特征扩展，在分类模型训练阶段利用卷积神经网络对局部敏感信息抽取能力进一步对文本语义信息进行挖掘，能够对短文本分类任务类别预测准确率等指标进行提高。



CN 110134786 A

1.一种基于主题词向量与卷积神经网络的短文本分类方法,其特征在于,包括如下步骤:

步骤1,采集短文本,对每个短文本进行标签标注,将标注好的短文本作为训练样本;

步骤2,对作为训练样本的短文本进行预处理,统一训练样本的格式,将预处理后的所有训练样本的集合作为语料库D;

步骤3,对语料库D中的每个短文本进行特征表示,包括:

步骤3a)对短文本进行主题级别的特征表示:

步骤3a1)初始化词网络主题模型参数先验参数文档-主题分布参数α与主题-词分布参数β,以及主题数量K;

步骤3a2)通过词网络主题模型挖掘短文本中的潜在词语簇,对于每一个潜在词语簇Z,从Z上的单词的多项式分布中抽取:$\Phi_z \sim Dir(\beta)$;其中,Φ_z表示词语属于词语簇Z的概率分布,Dir(β)表示参数β的狄利克雷分布;

步骤3a3)遍历短文本中的每一个词语w_i,对于w_i的相邻词语列表L_i,在其潜在词语簇中抽取:$\theta_i \sim Dir(\beta)$;其中,θ_i表示文档对应的主题分布;

步骤3a4)对于L_i中的每一个词语w_j:选择潜在词语簇$z_j \sim \Theta_i$,选择相邻词语$\boldsymbol{w_i \sim \Phi_{z_j}}$;其中,$\Theta_i$表示文档-主题分布参数矩阵,$\boldsymbol{\Phi_{z_j}}$表示第j个潜在词语簇的概率分布值;

步骤3a5)至此,得到文档形式表示的短文本,对短文本进行文档主题推断,将文档生成词语的主题分布的期望作为文档生成主题的分布,即:

$$P(z|d) = \sum_{w_i \in W_d} P(z|w_i)P(w_i|d)$$

其中,P(z|d)表示文档生成词语的概率,W_d表示文档集,P(z|w_i)表示词语生成主题的概率,P(w_i|d)表示文档d中单词w_i的经验分布;$\boldsymbol{P(w_i|d) = \frac{f_d(w_i)}{Len(d)}}$,$f_d(w_i)$为文档d中词$w_i$的词频,Len(d)是指文档d的长度;

步骤3a6)进行文档-主题推断,得到文档-主题分布:

$$\theta^D_{d,z} = p(z|d) = \sum_{w_i \in W_d} \theta^{PD}_{i,z} \frac{f_d(w_i)}{Len(d)}$$

其中,$\boldsymbol{\theta^D_{d,z}}$表示主题-词语参数矩阵,$\boldsymbol{\theta^{PD}_{i,z}}$表示词网络主题模型得到的伪文档参数;

步骤3a7)对主题模型参数矩阵进行Gibbs采样估计,得到文档语料库的主题-词分布;

步骤3a8)根据步骤3a6)得到的每个文档的文档-主题分布构建文档-主题分布矩阵,根据步骤3a7)得到的主题-词分布数据构建主题-词分布矩阵;

步骤3a9)由文档-主题分布矩阵与主题-词语分布矩阵得到词-主题分配映射关系;

步骤3b)对短文本进行词语级别的特征表示:

采用Skip-gram模型作为词向量训练模型,根据输入的每一个词向量及相应词的上下文词向量训练词向量训练模型,通过训练好的词向量训练模型可以提取得到短文本中的所有词向量;

步骤4,对每一个词语w_i,对w_i最相关主题中的所有词向量求平均值,将求得的平均值作为w_i的主题向量z_i;以词对<w_i,z_i>为输入,以词对<w_i,z_i>的上下文词对为输出,训练Skip-

gram模型;将w_i和z_i进行向量连接得到主题词向量w^z;

步骤5,对语料库进行字级别预训练得到文本的字向量表示,以字粒度信息对文本进行表示;

步骤6,将主题词向量w^z、字向量送入卷积神经网络进行分类模型训练:

步骤6a)将卷积神经网络的嵌入层设置为主题词向量层与字向量层两个卷积网络,将主题词向量与字向量作为对应卷积层的输入:

步骤6b)卷积层中对连续h个词或字向量进行宽卷积操作,其中宽卷积核窗口宽度为向量维度d,高度为h,$X_{i:i+h-1}$表示卷积核窗口从第i个单元起,作用于文本中的h个词语或字,卷积层通过过滤器提取新的特征;

步骤6c)卷积操作获取了词语的n-gram信息,池化层对卷积层的特征信息进行提取,其中引入注意力机制;输入为卷积层的特征向量,池化层中输入为卷积层特征向量[C_1,C_2,...,C_l],对于不同卷积核提取的特征向量进行权重attention机制自学习:将卷积特征C_i输入tanh层计算C_i的隐藏表示v_i,并通过softmax函数确定卷积特征的注意力权重α_i;

最终通过计算基于注意力权重与卷积特征加权求和输出向量C_α:

$v_i=\tanh(W_cC_i+b_c)$

$\alpha_i=\mathrm{softmax}(W_\alpha v_i)$

$$C_\alpha=\sum_{i=1}^{l}\alpha_iC_i\in R^m$$

其中,W_c表示卷积核的参数矩阵,W_α表示隐藏单元参数,R^m表示向量维度为卷积核数目m;

步骤6d)全连接层将词语级别的特征C_α与字级别的特征C_β进行拼接得到短文本的语义表示S:

$$S=[C_\alpha\oplus C_\beta]$$

步骤6e)分类层的输入是连接层对于文本向量的综合特征表示,分类层由线性变换层与softmax层组成,线性变换层将文本向量转换为与一个维度与类别相同的实数值向量,softmax函数将每一维度的实数值映射为类别的条件概率,其中类别为概率最大的维度,计算公式如下:

$P(y|T)=\mathrm{softmax}(W_sS+b_s)$

$$y=\arg\max_y p(y|T)$$

其中,y表示文本标签类别,T表示类别属性,W_s为卷积网络隐藏单元参数矩阵;b_s为偏置项;

步骤6f)构建计算最小化真实类标$y_j^g(S)$与预测类标y_j的交叉熵损失函数:

$$Loss=-\sum_{i=1}^{N_t}\sum_{j=1}^{N_c}y_j^g(s_i)\cdot\log(y_j(s_i))$$

以最小化损失函数Loss为目标训练神经网络;

步骤7,获取待预测的新的短文本,对新的短文本依次进行预处理、特征表示,通过步骤4得到新的短文本中的主题词向量,将新的短文本的主题词向量和子向量送入训练好的神

经网络进行预测,得到新的短文本的标签。

2.根据权利要求1所述基于主题词向量与卷积神经网络的短文本分类方法,其特征在于,所述步骤3b)中词向量训练模型的目标函数为:

$$\mathcal{L}(U)=\frac{1}{M}\sum_{i=1}^{M}\sum_{k\le c\le k,c\neq 0}\log \Pr(w_{i+c}|w_i)$$

其中,U表示输入的词语序列,U=$\{w_1,\ldots,w_M\}$,词向量训练模型训练时以最大化目标函数为目的进行训练。

3.根据权利要求1所述基于主题词向量与卷积神经网络的短文本分类方法,其特征在于:

所述步骤4中训练Skip-gram模型的目标函数为:

$$\mathcal{L}(U)=\frac{1}{|M|}\sum_{i=1}^{M}\sum_{-k\le c\le k,c\neq 0}\log \Pr(w_{i+c}|w_i)+\log \Pr(z_{i+c}|z_i)$$

训练时以最大化目标函数为目的进行训练。

一种基于主题词向量与卷积神经网络的短文本分类方法

技术领域

[0001] 本发明涉及文本分类领域，尤其是一种基于主题词向量与卷积神经网络的短文本分类方法。

背景技术

[0002] 随着互联网的大规模文本信息的产生，对海量的文本信息进行有效的挖掘与利用需要投入更多的人力物力，文本分类任务已成为重要的处理文本数据的方法，是管理文本语料的重要手段。文本分类是自然语言处理(NLP)的主要研究领域之一。文本分类任务可以理解为通过分析文本的结构特征、语义信息，将文章映射提取到设定好的标签集合中的过程。

[0003] 随着在线通信、新闻快讯、电子商务、社交媒体、在线问答等实时的新型应用的流行以及爆炸式增长，其中这类应用中传播与使用的数据最主要的特点是文字长度短，文本语义信息不足。面对短文本的数据特点，传统的文本表示方法与分类模型算法在长文本中能够取得不错的效果，但直接应用于短文本却往往不能够达到理想的结果指标，其中主要的原因在于，一方面是因为短文本具有稀疏性的特点，长度短，造成了短文本所包含语义短信息不够丰富，无法提供足够的单词共现或上下文信息，很难识别语义信息的元素；另一方面原因在于相较于长文本，在有限的文本长度中，短文本的词语语义具有较强的上下文语境依赖性，提取其有效的语义信息存在难度。基于短文本的数据特点，分类任务的效果主要依赖于对文本的特征表示效果，以及分类模型对于特征向量的学习区别能力。

发明内容

[0004] 发明目的：本发明主要解决的技术问题是，针对短文本文本长度短，语义信息不足，单词共现稀疏的数据特点，造成分类效果不佳的问题。本发明在文本的特征表示方面，基于主题模型与词向量模型对短文本进行语义表示；在分类模型方面，本发明基于深度学习的卷积神经网络模型对短文本进行进一步特征抽取，并最后使用Softmax分类器进行分类。

[0005] 技术方案：为实现上述目的，本发明采用的技术方案为：

[0006] 一种基于主题词向量与卷积神经网络的短文本分类方法，包括如下步骤：

[0007] 步骤1，采集短文本，对每个短文本进行标签标注，将标注好的短文本作为训练样本；

[0008] 步骤2，对作为训练样本的短文本进行预处理，统一训练样本的格式，将预处理后的所有训练样本的集合作为语料库D；

[0009] 步骤3，对语料库D中的每个短文本进行特征表示，包括：

[0010] 步骤3a)对短文本进行主题级别的特征表示：

[0011] 步骤3a1)初始化词网络主题模型参数先验参数文档-主题分布参数α与主题-词分布参数β，以及主题数量K；

[0012] 步骤3a2)通过词网络主题模型挖掘短文本中的潜在词语簇,对于每一个潜在词语簇Z,从Z上的单词的多项式分布中抽取:$\Phi_z \sim Dir(\beta)$;其中,Φ_z表示词语属于词语簇Z的概率分布,Dir(β)表示参数β的狄利克雷分布;

[0013] 步骤3a3)遍历短文本中的每一个词语w_i,对于w_i的相邻词语列表L_i,在其潜在词语簇中抽取:$\theta_i \sim Dir(\beta)$;其中,θ_i表示文档对应的主题分布;

[0014] 步骤3a4)对于L_i中的每一个词语w_j:选择潜在词语簇$z_j \sim \Theta_i$,选择相邻词语$w_i \sim \Phi_{z_j}$;其中,Θ_i表示文档-主题分布参数矩阵,Φ_{z_j}表示第j个潜在词语簇的概率分布值;

[0015] 步骤3a5)至此,得到文档形式表示的短文本,对短文本进行文档主题推断,将文档生成词语的主题分布的期望作为文档生成主题的分布,即:

[0016] $$P(z|d) = \sum_{w_i \in W_d} P(z|w_i)P(w_i|d)$$

[0017] 其中,P(z|d)表示文档生成词语的概率,W_d表示文档集,P(z|w_i)表示词语生成主题的概率,P(w_i|d)表示文档d中单词w_i的经验分布;$P(w_i|d) = \frac{f_d(w_i)}{Len(d)}$,$f_d(w_i)$为文档d中词$w_i$的词频,Len(d)是指文档d的长度;

[0018] 步骤3a6)进行文档-主题推断,得到文档-主题分布:

[0019] $$\theta_{d,z}^{D} = p(z|d) = \sum_{w_i \in W_d} \theta_{i,z}^{PD} \frac{f_d(w_i)}{Len(d)}$$

[0020] 其中,$\theta_{d,z}^{D}$表示主题-词语参数矩阵,$\theta_{i,z}^{PD}$表示词网络主题模型得到的伪文档参数;

[0021] 步骤3a7)对主题模型参数矩阵进行Gibbs采样估计,得到文档语料库的主题-词分布;

[0022] 步骤3a8)根据步骤3a6)得到的每个文档的文档-主题分布构建文档-主题分布矩阵,根据步骤3a7)得到的主题-词分布数据构建主题-词分布矩阵;

[0023] 步骤3a9)由文档-主题分布矩阵与主题-词语分布矩阵得到词-主题分配映射关系;

[0024] 步骤3b)对短文本进行词语级别的特征表示:

[0025] 采用Skip-gram模型作为词向量训练模型,根据输入的每一个词向量及相应词的上下文词向量训练词向量训练模型,通过训练好的词向量训练模型可以提取得到短文本中的所有词向量;

[0026] 步骤4,对每一个词语w_i,对w_i最相关主题中的所有词向量求平均值,将求得的平均值作为w_i的主题向量z_i;以词对<w_i,z_i>为输入,以词对<w_i,z_i>的上下文词对为输出,训练Skip-gram模型;将w_i和z_i进行向量连接得到主题词向量w^z;

[0027] 步骤5,对语料库进行字级别预训练得到文本的字向量表示,以字粒度信息对文本进行表示;

[0028] 步骤6,将主题词向量w^z、字向量送入卷积神经网络进行分类模型训练:

[0029] 步骤6a)将卷积神经网络的嵌入层设置为主题词向量层与字向量层两个卷积网络,将主题词向量与字向量作为对应卷积层的输入:

[0030] 步骤6b)卷积层中对连续h个词或字向量进行宽卷积操作,其中宽卷积核窗口宽度

为向量维度d，高度为h，$X_{i:i+h-1}$表示卷积核窗口从第i个单元起，作用于文本中的h个词语或字，卷积层通过过滤器提取新的特征；

[0031] 步骤6c)卷积操作获取了词语的n-gram信息，池化层对卷积层的特征信息进行提取，其中引入注意力机制；输入为卷积层的特征向量，池化层中输入为卷积层特征向量[C_1，C_2，...，C_l]，对于不同卷积核提取的特征向量进行权重attention机制自学习：将卷积特征C_i输入tanh层计算C_i的隐藏表示v_i，并通过softmax函数确定卷积特征的注意力权重α_i；最终通过计算基于注意力权重与卷积特征加权求和输出向量C_α：

[0032] $v_i = \tanh(W_c C_i + b_c)$

[0033] $\alpha_i = \mathrm{softmax}(W_\alpha v_i)$

[0034] $$C_\alpha = \sum_{i=1}^{l} \alpha_i C_i \in R^m$$

[0035] 其中，W_c表示卷积核的参数矩阵，W_α表示隐藏单元参数，R^m表示向量维度为卷积核数目m；

[0036] 步骤6d)全连接层将词语级别的特征C_α与字级别的特征C_β进行拼接得到短文本的语义表示S：

[0037] $$S = [C_\alpha \oplus C_\beta]$$

[0038] 步骤6e)分类层的输入是连接层对于文本向量的综合特征表示，分类层由线性变换层与softmax层组成，线性变换层将文本向量转换为与一个维度与类别相同的实数值向量，softmax函数将每一维度的实数值映射为类别的条件概率，其中类别为概率最大的维度，计算公式如下：

[0039] $P(y|T) = \mathrm{softmax}(W_s S + b_s)$

[0040] $$y = \arg\max_y p(y|T)$$

[0041] 其中，y表示文本标签类别，T表示类别属性，W_s为卷积网络隐藏单元参数矩阵；b_s为偏置项；

[0042] 步骤6f)构建计算最小化真实类标$y_j^g(S)$与预测类标y_j的交叉熵损失函数：

[0043] $$Loss = -\sum_{i=1}^{N_t} \sum_{j=1}^{N_c} y_j^g(s_i) \cdot \log(y_j(s_i))$$

[0044] 以最小化损失函数Loss为目标训练神经网络；

[0045] 步骤7，获取待预测的新的短文本，对新的短文本依次进行预处理、特征表示，通过步骤4得到新的短文本中的主题词向量，将新的短文本的主题词向量和子向量送入训练好的神经网络进行预测，得到新的短文本的标签。

[0046] 进一步的，所述步骤3b)中词向量训练模型的目标函数为：

[0047] $$\mathcal{L}(U) = \frac{1}{M} \sum_{i=1}^{M} \sum_{k \le c \le k, c \ne 0} \log \Pr(w_{i+c} | w_i)$$

[0048] 其中，U表示输入的词语序列，U＝{w_1，...，w_M}，词向量训练模型训练时以最大化目标函数为目的进行训练。

[0049] 进一步的,所述步骤4中训练Skip-gram模型的目标函数为:

[0050] $$\mathcal{L}(U)=\frac{1}{|M|}\sum_{i=1}^{M}\sum_{-k\le c\le k,c\neq 0}\log Pr(w_{i+c}|w_i)+\log Pr(z_{i+c}|z_i)$$

[0051] 训练时以最大化目标函数为目的进行训练。

[0052] 有益效果:本发明相比现有技术,具有以下有益效果:

[0053] 本发明一种基于主题词向量与卷积神经网络的短文本分类方法,基于神经网络语言模型的词向量与短文本主题模型相结合的表示方法,利用主题模型的全局主题信息以及词向量的局部语义信息对短文本的特征表示进行扩展,在词向量模型中引入WNTM短文本主题模型对词向量进行优化。在词向量的训练过程中将词语的主题向量作为新词引入到词向量训练过程,并考虑到词向量与主题向量之间的差异性,在词向量的构建中分别对其进行训练,增强短文本的局部与全局语义信息表示的准确性。

[0054] 分别将主题词向量信息与字向量信息作为卷积神经网络的输入,从不同层次学习短文本的语义特征。在卷积神经网络的结构中,对传统的卷积神经网络的池化层过程中丢失过多特征信息,面对短文本的数据特点时考虑到对特征向量的保存,在池化层中引入attention机制对不同卷积核提取的特征量进行权重计算,保留了有用的特征信息并进行权重自学习以提升分类任务。

附图说明

[0055] 图1为本发明的特征表示流程图;

[0056] 图2为本发明的WNTWE模型训练框架图;

[0057] 图3为本发明的卷积神经网络流程图;

[0058] 图4为本发明的卷积神经网络架构图。

具体实施方式

[0059] 下面结合附图和具体实施例,进一步阐明本发明,应理解这些实例仅用于说明本发明而不用于限制本发明的范围,在阅读了本发明之后,本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。

[0060] 一种基于主题词向量与卷积神经网络的短文本分类方法,包括如下步骤:

[0061] 如图1所示,

[0062] 步骤1,数据集预处理:将原始文本数据按照统一格式处理,对统一处理好的样本数据进行去噪声;

[0063] 步骤2,文本分词,定制化停用词过滤,构建语料库D:

[0064] 步骤3a)对短文本进行主题级别的特征表示:

[0065] 步骤3a1)初始化词网络主题模型参数先验参数文档-主题分布参数α与主题-词分布参数β,以及主题数量K;

[0066] 步骤3a2)通过词网络主题模型挖掘短文本中的潜在词语簇,对于每一个潜在词语簇Z,从Z上的单词的多项式分布中抽取:$\Phi_z \sim Dir(\beta)$;其中,Φ_z表示词语属于词语簇Z的概率分布,Dir(β)表示参数β的狄利克雷分布;

[0067] 步骤3a3)遍历短文本中的每一个词语w_i,对于w_i的相邻词语列表L_i,在其潜在词语簇中抽取:$\theta_i \sim Dir(\beta)$;其中,θ_i表示文档对应的主题分布;

[0068] 步骤3a4)对于L_i中的每一个词语w_j:选择潜在词语簇$z_j \sim \Theta_i$,选择相邻词语$w_i \sim \Phi_{z_j}$;其中,Θ_i表示文档-主题分布参数矩阵,Φ_{z_j}表示第j个潜在词语簇的概率分布值。

[0069] 步骤3a5)至此,得到文档形式表示的短文本,对短文本进行文档主题推断,将文档生成词语的主题分布的期望作为文档生成主题的分布,即:

[0070] $$P(z|d) = \sum_{w_i \in W_d} P(z|w_i)P(w_i|d)$$

[0071] 其中,$P(z|d)$表示文档生成词语的概率,W_d表示文档集,$P(z|w_i)$表示词语生成主题的概率,$P(w_i|d)$表示文档d中单词w_i的经验分布。$P(w_i|d) = \frac{f_d(w_i)}{Len(d)}$,$f_d(w_i)$为文档d中词$w_i$的词频,Len(d)是指文档d的长度;

[0072] 步骤3a6)进行文档-主题推断,得到文档-主题分布:

[0073] $$\theta_{d,z}^{D} = p(z|d) = \sum_{w_i \in W_d} \theta_{i,z}^{PD} \frac{f_d(w_i)}{Len(d)}$$

[0074] 其中,$\theta_{d,z}^{D}$表示主题-词语参数矩阵,$\theta_{i,z}^{PD}$表示词网络主题模型得到的伪文档参数;

[0075] 步骤3a7)对主题模型参数矩阵进行Gibbs采样估计,得到文档语料库的主题-词分布;

[0076] 步骤3a8)根据步骤3a6)得到的每个文档的文档-主题分布构建文档-主题分布矩阵,根据步骤3a7)得到的主题-词分布数据构建主题-词分布矩阵;

[0077] 步骤3a9)由文档-主题分布矩阵与主题-词语分布矩阵得到词-主题分配映射关系;

[0078] 步骤3b)对短文本进行词语级别的特征表示:

[0079] 采用Skip-gram模型作为词向量训练模型,根据输入的每一个词向量及相应词的上下文词向量训练词向量训练模型,通过训练好的词向量训练模型可以提取得到短文本中的所有词向量;

[0080] 步骤3b1)对短文本进行词向量训练表示;

[0081] 步骤3b2)初始化词向量模型,采用Skip-gram模型进行中心词对上下文窗口进行预测训练,优化模式为负采样;

[0082] 步骤3b3)对于输入层一个词语序列$D=\{w_1,...,w_M\}$,最大化模型的平均对数概率目标函数:

[0083] $$\mathcal{L}(D) = \frac{1}{M}\sum_{i=1}^{M} \sum_{k \le c \le k, c \neq 0} \log Pr(w_{i+c}|w_i)$$

[0084] 步骤3b4)训练目标函数得到词语词向量表示;

[0085] 步骤4,对词语级别的特征表示与主题级别的特征表示进行联合训练,得到主题词向量:

[0086] 步骤4a)通过步骤3a得到的词-主题映射关系,以及步骤3b中得到的训练文本的词向量表示,对词语最相关主题中的词向量求和平均值作为词语w_i的主题向量z_i作为初始化

训练向量；

[0087] 步骤4b)训练模型结合主题向量与词向量的特征表示，并考虑到两种表示的层次关系，词向量的目标函数定义为：

[0088] $$\mathcal{L}(D)=\frac{1}{|M|}\sum_{i=1}^{M}\sum_{-k\le c\le k,c\neq 0}\log \Pr(w_{i+c}|w_i)+\log \Pr(z_{i+c}|z_i)$$

[0089] 模型架构如图2所示，其中模型将词的主题分布作为新词，构成<w_i,z_i>词语与主题向量的独立单元，损失函数中模型对于当前主题-词语的词对分别预测上下文窗口的主题-词语词对；

[0090] 步骤4c)模型训练结束；

[0091] 步骤4d)将模型训练得到的主题向量与词向量进行拼接得到主题词向量w^z；

[0092] 步骤4e)对短文本进行主题向量的求和平均进行文本特征表示；

[0093] 步骤5，如图3所示，将主题词向量与字向量结合送入卷积神经网络进行分类模型训练。

[0094] 步骤5a1嵌入层包含词语与字两个卷积网络，分别使用预训练的主题词向量与字向量作为对应卷积层的输入：

[0095] $$x_{1:l}=x_1\oplus x_2\oplus \ldots\oplus x_l$$

[0096] 步骤5b)卷积层中对连续h个词或字向量进行宽卷积操作，其中宽卷积核窗口宽度为向量维度d，高度为h，$X_{i:i+h-1}$表示卷积核窗口从第i个单元起，作用于文本中的h个词语或字，卷积层通过过滤器提取新的特征，卷积操作公式：

[0097] $c_i=\mathrm{ReLU}(W_c\cdot x_{i:i+h-1}+b)\in R^m$

[0098] 卷积操作中对文本边界进行补齐，设置卷积层的输出长度等于输入长度。每个卷积窗口的卷积核特征使用m个不同的滤波器来执行卷积运算，并将每个窗口的卷积结果特征集表示为C

[0099] 步骤5c)卷积操作获取了词语的n-gram信息，池化层对卷积层的特征信息进行提取，其中引入注意力机制。输入为卷积层的特征向量，池化层中输入为卷积层特征向量[C_1,C_2,...,C_l]，对于不同卷积核提取的特征向量进行权重attention机制自学习其中，W_c表示卷积核的参数矩阵，W_α表示隐藏单元参数，R^m表示向量维度为卷积核数目m：

[0100] $v_i=\tanh(W_cC_i+b_c)$

[0101] $\alpha_i=\mathrm{softmax}(W_\alpha v_i)$

[0102] $$C_\alpha=\sum_{i=1}^{l}\alpha_iC_i\in R^m$$

[0103] 将卷积特征C_i输入tanh层计算C_i的隐藏表示v_i，并通过softmax函数确定卷积特征的注意力权重α_i。最终通过计算基于注意力权重与卷积特征加权求和输出向量C。

[0104] 步骤5d)全连接层将词语级别的特征C_α与字级别的特征C_β进行拼接得到短文本的语义表示S：

[0105] $$S=[C_\alpha\oplus C_\beta]$$

[0106] 步骤5e)分类层的输入是连接层对于文本向量的综合特征表示，分类层由线性变换层与softmax层组成。线性变换层将文本向量转换为与一个维度与类别相同的实数值向

量，softmax函数将每一维度的实数值映射为类别的条件概率，其中类别为概率最大的维度，计算公式如下，其中y表示文本标签类别，T表示类别属性，W_s为卷积网络隐藏单元参数矩阵。b_s为偏置项：

[0107] $P(y|T)=\mathrm{softmax}(W_sS+b_s)$

[0108] $$y=\arg\max_y p(y|T)$$

[0109] 步骤5f)进行模型训练，通过计算最小化真实类标$y_j^g(S)$与预测类标y_j的交叉熵损失函数：

[0110] $$Loss=-\sum_{i=1}^{N_t}\sum_{j=1}^{N_c}y_j^g(s_i)\cdot\log(y_j(s_i))$$

[0111] 其中N_t为训练数据集文本数量，N_c为类别数目，$y_j^g(S)$维度为类别K，对应类标为1，其余维度为0。在模型训练中最小化损失函数，通过反向传播对模型中各层的参数进行迭代更新。模型架构如图4所示。

[0112] 步骤5e)模型训练结束。

[0113] 步骤6，对新样本短文数据进行类标预测。

[0114] 综上所述，本发明结合短文本数据特点，在特征表示阶段利用主题向量与词向量结合表示，对短文本自身数据特点进行语义特征扩展，在分类模型训练阶段利用卷积神经网络对局部敏感信息抽取能力进一步对文本语义信息进行挖掘，能够对短文本分类任务类别预测准确率等指标进行提高。本发明特征表示总体结构如附图1所示。训练主题词向量模型架构如附图2所示。卷积神经网络分类模型流程如附图3所示。神经网络框架图如附图4所示。

[0115] 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。

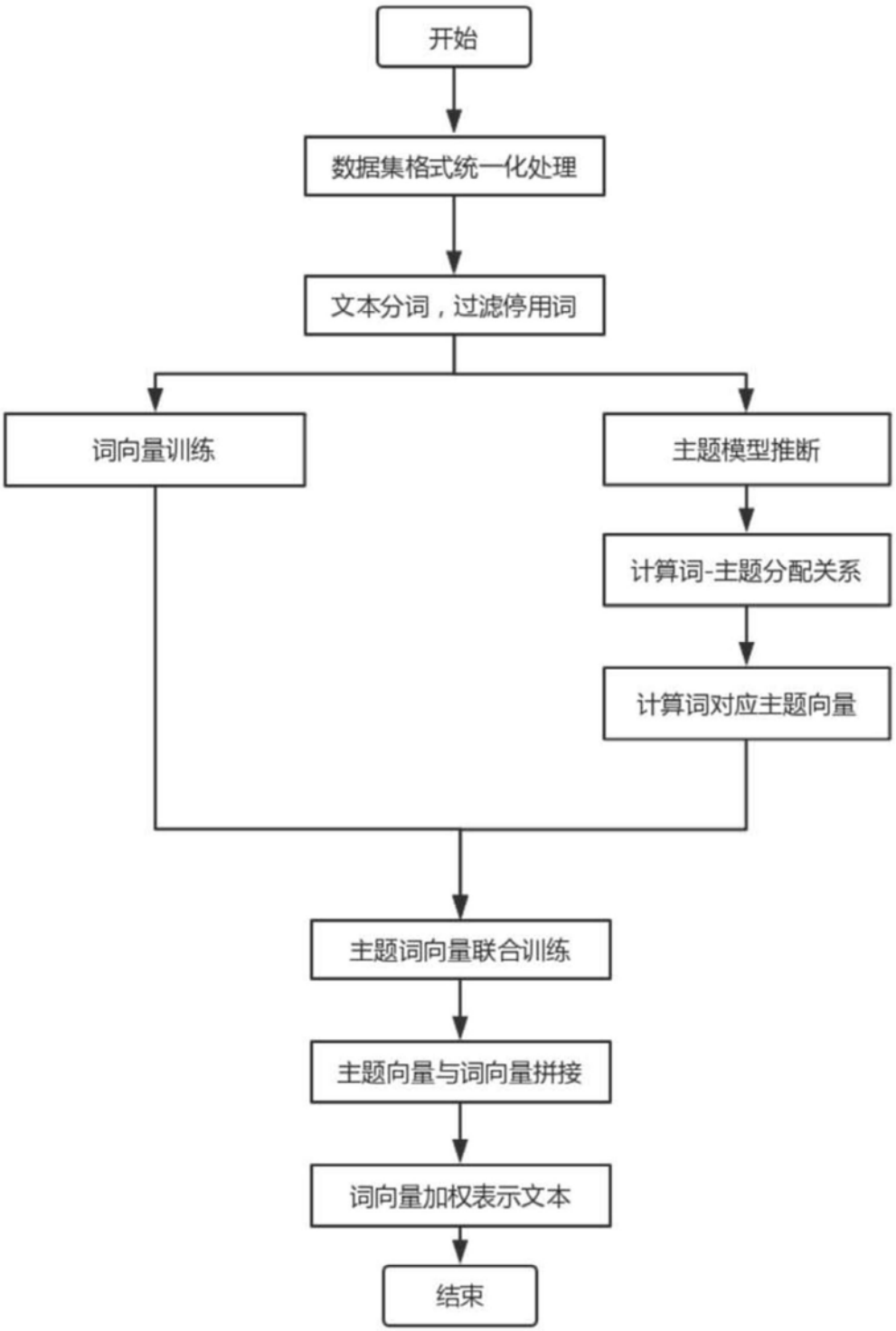

开始
数据集格式统一化处理
文本分词，过滤停用词
词向量训练
主题模型推断
计算词-主题分配关系
计算词对应主题向量
主题词向量联合训练
主题向量与词向量拼接
词向量加权表示文本
结束

图1

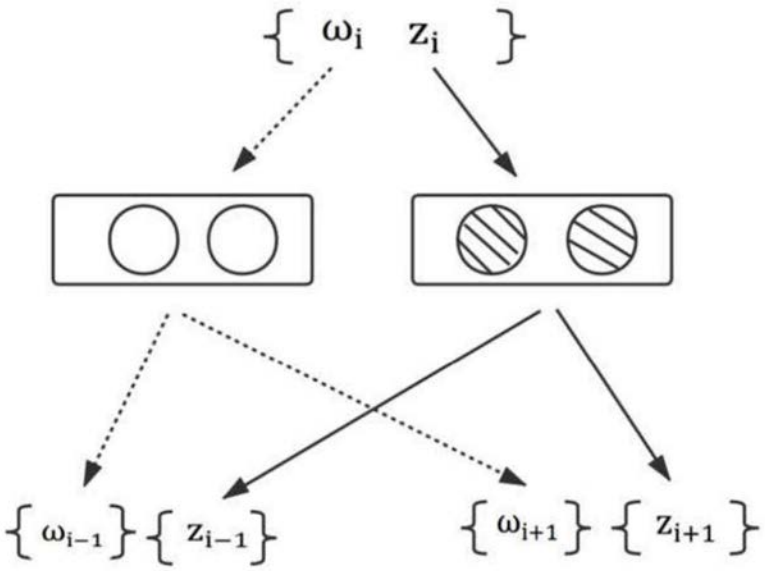


图2

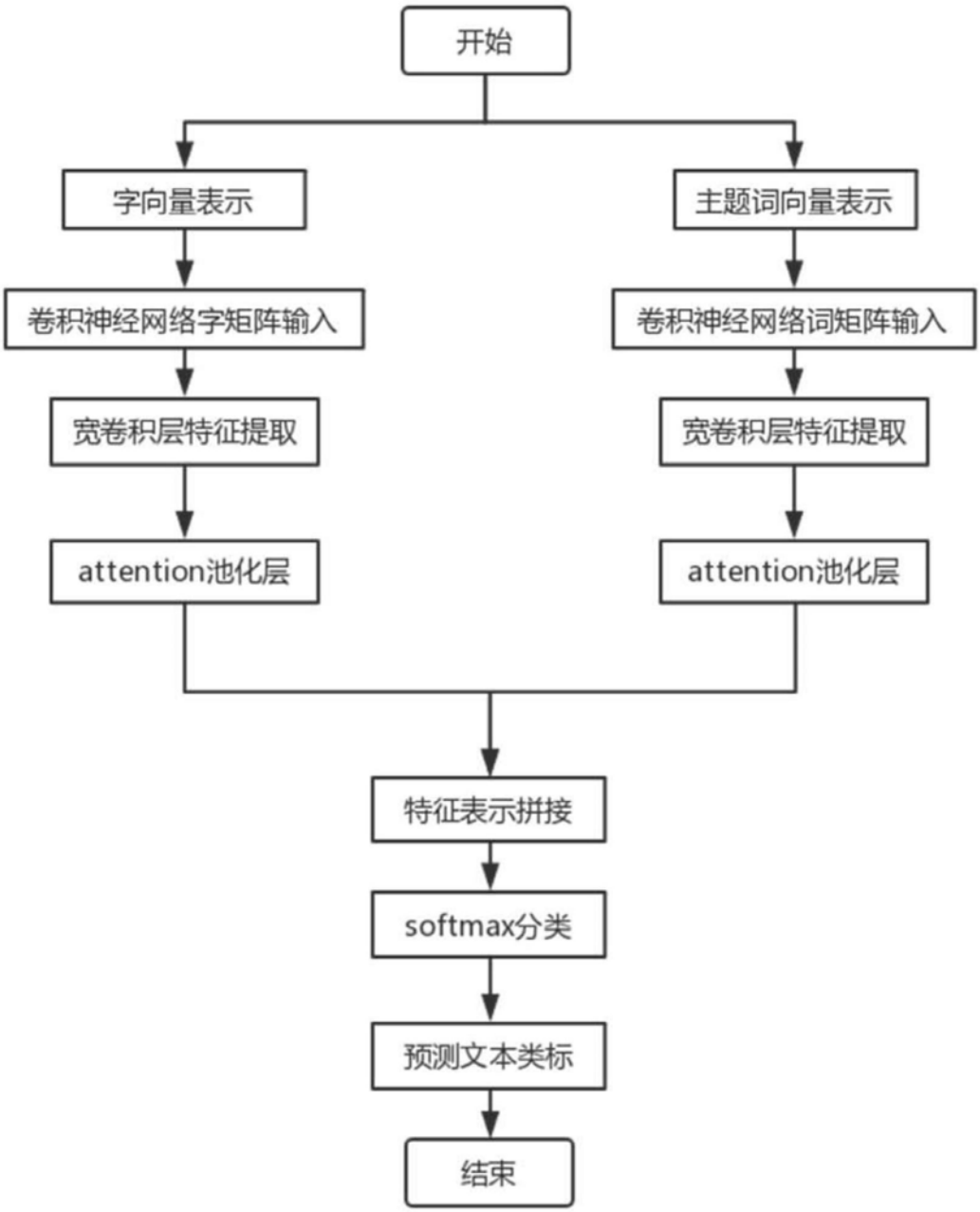

开始
字向量表示
主题词向量表示
卷积神经网络字矩阵输入
卷积神经网络词矩阵输入
宽卷积层特征提取
宽卷积层特征提取
attention池化层
attention池化层
特征表示拼接
softmax分类
预测文本类标
结束

图3

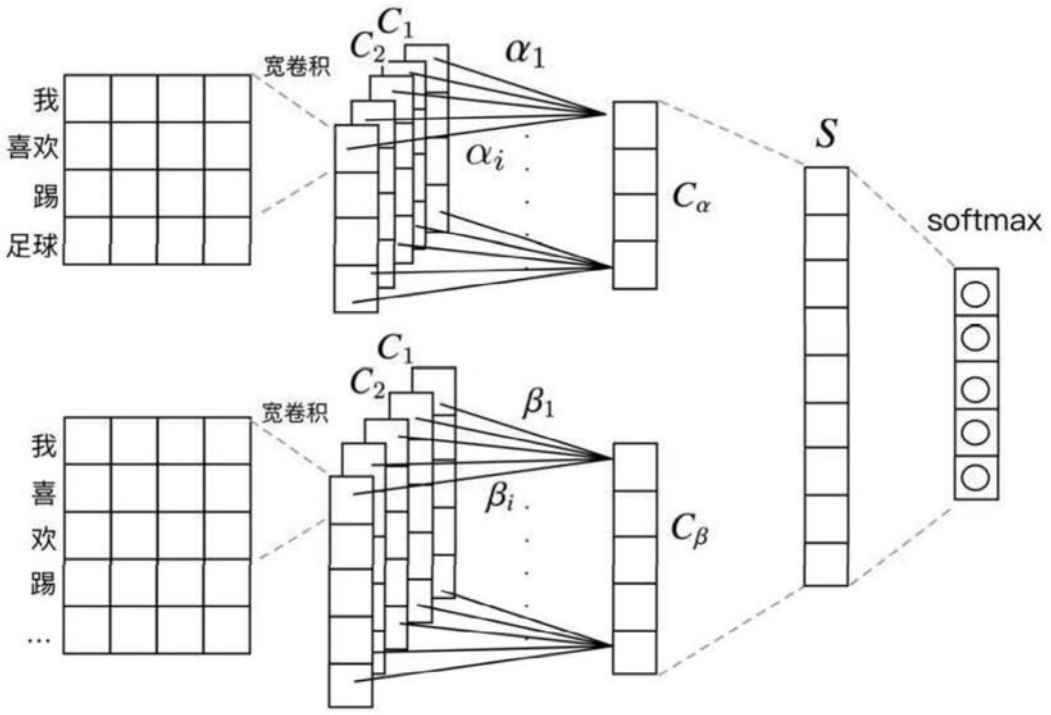

我
喜欢
踢
足球
宽卷积
C_1
C_2
α_1
α_i
C_α
S
softmax
我
喜
欢
踢
...
宽卷积
C_1
C_2
β_1
β_i
C_β

图4