(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114936723 A
(43)申请公布日 2022.08.23

(21)申请号 202210856300.4

(22)申请日 2022.07.21

(71)申请人 中国电子科技集团公司第三十研究所
地址 610000 四川省成都市高新区创业路6号
申请人 国家计算机网络与信息安全管理中心

(72)发明人 丁建伟 陈周国 李欣泽 沈亮 张震 石瑾 杨宇 王鑫 刘志洁 李航

(74)专利代理机构 成都九鼎天元知识产权代理有限公司 51214
专利代理师 黎飞

(51)Int.Cl.
G06Q 10/04(2012.01)
G06Q 50/00(2012.01)
G06N 3/08(2006.01)
G06N 3/04(2006.01)

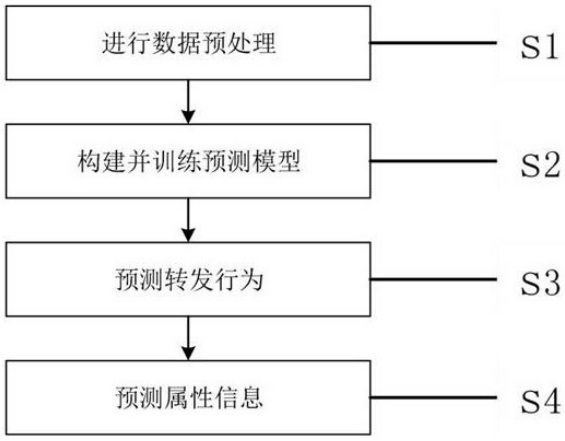
权利要求书4页 说明书11页 附图4页

(54)发明名称

一种基于数据增强的社交网络用户属性预测方法及系统

(57)摘要

本发明涉及数据挖掘技术领域，公开了一种基于数据增强的社交网络用户属性预测方法及系统，该属性预测方法，基于用户的历史行为序列，推断用户未来一段时间的行为序列，通过将历史行为序列与预测得到的行为序列进行拼接，扩大行为序列的长度，对用户的行为数据进行增强。本发明解决了现有技术存在的线网络用户行为序列长度较短时不能为用户属性预测任务提供足够信息、从而导致用户属性预测的预测准确性低的问题。



CN 114936723 A

1.一种基于数据增强的社交网络用户属性预测方法,其特征在于,基于用户的历史行为序列,推断用户未来一段时间的行为序列,通过将历史行为序列与预测得到的行为序列进行拼接,扩大行为序列的长度,对用户的行为数据进行增强。

2.根据权利要求1所述的一种基于数据增强的社交网络用户属性预测方法,其特征在于,包括以下步骤:

S1,进行数据预处理:提取社交网络用户的转发行为数据,对提取的转发行为数据进行预处理,获得预处理后的用户行为数据;

S2,构建并训练预测模型:利用预处理后的用户行为数据,构建用户行为数据预测模型,并训练用户行为数据预测模型;

S3,预测转发行为:利用训练好的用户行为数据预测模型预测用户未来一段时间的转发行为,获得增强后的用户行为数据;

S4,预测属性信息:利用增强后的用户行为数据预测社交网络中用户的属性信息。

3.根据权利要求2所述的一种基于数据增强的社交网络用户属性预测方法,其特征在于,步骤S1包括以下步骤:

S11,提取社交网络用户的行为信息:对于给定用户u_i,首先获取u_i在某一段时间跨度内的转发行为序列$H_{u_i}=\{u_1,u_5,\cdots,u_i,\cdots,u_m\}$,在起始位置添加特殊标志符[SOS],在行为序列的末尾添加特殊标志符[EOS];并将转发行为对应的时间戳$T_{u_i}=\{t_{u_1},t_{u_5},\cdots,t_{u_i},\cdots,t_{u_m}\}$进行记录;其中,i表示用户的编号,$u_i$表示第i个用户的昵称,$t_{u_i}$表示第i个用户的转发行为对应的时间戳;

S12,首先计算当前转发行为与参照时刻之间的时间间隔,然后以事先设定的周期将时间间隔转换为时间ID,最后计算相邻转发行为之间时间ID的差分值。

4.根据权利要求3所述的一种基于数据增强的社交网络用户属性预测方法,其特征在于,步骤S2包括以下步骤:

S21,构建用户转发行为嵌入表示层:用户转发行为嵌入表示层包括行为序列嵌入表示层$\mathbf{B}$、行为先后顺序嵌入表示层$\mathbf{D}$、时间信息嵌入层$\mathbf{T}$;其中,行为序列嵌入表示层$\mathbf{B}$用于将用户转发行为序列中的每一个昵称转换为k维的向量表示,行为先后顺序嵌入表示层$\mathbf{D}$用于将每一个被转发用户的转发顺序编号转换为k维的向量表示,时间信息嵌入层$\mathbf{T}$用于将每一个被转发用户的转发时间信息转换为k维的向量表示,将上述三种向量表示按位相加,得到生成式预训练用户行为数据预测模型的输入$\mathbf{X}'$;k≥2且k为整数;

S22,构建编码器:构建包括多头自注意力模块MHA、基于位置的全连接前馈网络FFN的编码器;其中,多头自注意力模块MHA基于缩放点积自注意力用户行为数据预测模型,将嵌入表示矩阵X'作为输入,输出注意力评分矩阵$\mathbf{A}$;基于位置的全连接前馈网络FFN,将注意力评分矩阵$\mathbf{A}$作为输入,输出解码器的隐层表示H;

S23,构建行为序列输出层:行为序列输出层为全连接神经网络,用于将解码器的隐层表示$\mathbf{H}$作为输入,输出预测结果,并通过Softmax函数计算预测结果与真实值之间的误差值δ,所述预测结果指用户下一时刻转发行为;

步骤S24，通过误差反向传播的训练方式更新用户行为数据预测模型参数，直到误差值δ达到最低为止，保存最终的用户行为数据预测模型参数，获得训练好的用户行为数据预测模型；误差值δ达到最低为止指用户行为数据预测模型收敛的情形。

5.根据权利要求4所述的一种基于数据增强的社交网络用户属性预测方法，其特征在于，步骤S22包括以下步骤：

S221，利用嵌入表示矩阵$\mathbf{X}'$作为输入，先将$\mathbf{X}'$整理为$\mathbf{H}^0=\left[\mathbf{x}_1,\cdots,\mathbf{x}_{|x|}\right]$，再将$\mathbf{H}^0$输入到一个$L$层的Transformer网络中进行编码：

$$\mathbf{H}^l=\text{Transformer}_l\left(\mathbf{H}^{l-1}\right);$$

其中，$l\in[1,L]$，$\mathbf{H}^l=\left[\mathbf{h}_1^l,\cdots,\mathbf{h}_{|x|}^l\right]$，$\mathbf{H}^0$表示用户行为的向量表征，$\mathbf{H}^l$表示第$l$层用户行为数据预测模型的输出结果，$l$表示Transformer网络中的层编号，$\text{Transformer}_l\left(\mathbf{H}^{l-1}\right)$表示第$l$层Transformer网络编码后的结果，$|x|$表示步骤S11中转发行为序列$H_{u_i}$的长度，$\mathbf{h}_1^l,\cdots,\mathbf{h}_{|x|}^l$分别表示第1至第$|x|$个行为在第$l$层的向量表征；在每个Transformer编码器中，有多头注意力机制以聚合前一层的输出向量；第l层的Transformer中的一个自注意力头$\mathbf{A}_l$的计算公式如下：

$$\mathbf{Q}=\mathbf{H}^{l-1}\mathbf{W}_l^Q,$$

$$\mathbf{K}=\mathbf{H}^{l-1}\mathbf{W}_l^K,$$

$$\mathbf{V}=\mathbf{H}^{l-1}\mathbf{W}_l^V,$$

$$\mathbf{M}_{ij}=\begin{cases}0, & \text{可以参照}\\ -\infty, & \text{不能参照}\end{cases},$$

$$\mathbf{A}=\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{k}}+\mathbf{M}\right)\mathbf{V};$$

其中，前层的输出$\mathbf{H}^{l-1}\in\mathbb{R}^{|x|\times d_h}$通过参数为$\mathbf{W}_l^Q,\mathbf{W}_l^K,\mathbf{W}_l^V\in\mathbb{R}^{d_h\times d_k}$的线性变换分别得到查询向量$\mathbf{Q}$、键向量$\mathbf{K}$、值向量$\mathbf{V}$；掩码矩阵$\mathbf{M}\in\mathbb{R}^{|x|\times|x|}$控制着行为之间是否能够被参照，$\mathbf{W}_l^Q$表示第$l$层查询向量对应的线性变换矩阵，$\mathbf{W}_l^K$表示第$l$层键向量对应的线性变换矩阵，$\mathbf{W}_l^V$表示第$l$层值向量对应的线性变换矩阵，i表示掩码矩阵$\mathbf{M}$的行标号，j表示掩码矩阵$\mathbf{M}$的列标号，$\mathbf{M}_{ij}$表示掩码矩阵$\mathbf{M}$的第i行第j列元素，$\mathbf{K}^\top$表示键向量的转置，$d_k$表示隐藏层神经元个数，$d_h$表示用户行为的嵌入维度；

S222，以交叉熵作为损失函数，采用梯度下降法训练行为生成部分的神经网络，将用户的转发行为进行独热编码；

S223:返回步骤S221,循环执行步骤S221~步骤S222,直至训练用户行为数据预测模型收敛为止。

6.根据权利要求5所述的一种基于数据增强的社交网络用户属性预测方法,其特征在于,步骤S222中,独热编码计算公式为:

$$\varepsilon = -\sum_{i=2}^{n}\sum_{j=1}^{N}\left(u_{i,j}log\left(\hat{u}_{i,j}\right) + \left(1-u_{i,j}\right)log\left(1-\hat{u}_{i,j}\right)\right);$$

其中,ε表示损失值,$u_{i,j}$表示真实用户在第i时刻转发用户j的概率,n表示时刻总数,N表示所分析的用户的总数量,$\hat{u}_{i,j}$表示社交网络用户行为数据预测模型预测的用户i在第j时刻转发用户的概率。

7.根据权利要求6所述的一种基于数据增强的社交网络用户属性预测方法,其特征在于,步骤S3包括以下步骤:

S31,利用词嵌入层以及位置嵌入层对行为序列对应的时间ID差分序列进行嵌入表示,然后用户行为数据预测模型通过L个编码层得到时间ID差分序列的隐藏表示,最后利用交叉熵损失函数计算用户行为数据预测模型预测结果与期望值之间的偏差;

S32,将嵌入表示矩阵$\mathbf{X}'$输入到多头自注意力机制ATT中,得到注意力权重矩阵,并将注意力矩阵$\mathbf{A}$输入到基于位置的全连接前馈网络FFN中,得到隐层表示$\mathbf{H}$,并通过步骤S23得到用户下一时刻转发行为的预测结果$\mathbf{y}$,最后将预测结果$\mathbf{y}$拼接到输入序列中;

S33,循环执行步骤S31~S32,直至得到用户未来一段时间的转发行为预测结果$\mathbf{Y}$。

8.根据权利要求7所述的一种基于数据增强的社交网络用户属性预测方法,其特征在于,步骤S4包括以下步骤:

S41,将步骤S33得到的用户行为预测结果$\mathbf{Y}$与用户历史行为序列$\mathbf{X}$进行拼接得到增强后的用户行为数据$\mathbf{Y}'$,并利用步骤S21得到用户行为数据$\mathbf{Y}'$的嵌入矩阵$\overline{\mathbf{Y}}$,对嵌入矩阵$\overline{\mathbf{Y}}$按列求平均得到社交网络用户属性预测用户行为数据预测模型的输入特征$\overline{\mathbf{X}}$;

S42,将特征向量$\overline{\mathbf{X}}$输入到预训练好的用户行为数据预测模型中进行调整,得到用户属性的所属概率值$\mathbf{p}$。

9.根据权利要求8所述的一种基于数据增强的社交网络用户属性预测方法,其特征在于,步骤S42中,对用户行为数据预测模型进行调整的具体方法为:

仅保留用户行为数据预测模型的嵌入层以及编码器模块,并添加Flatten层做维度变换,最后输入包括依次相连的线性层、激活层、线性层的前馈神经网络,将隐藏特征映射到真实标签,具体的计算方式如下式:

$$\mathbf{h}_1 = \mathrm{FlattenLayer}(\mathbf{h}_0),$$

$$\mathrm{FFN}(\mathbf{h}_1, W_1, W_2, b_1, b_2) = \max(0, \mathbf{h}_1 W_1 + b_1)W_2 + b_2;$$

其中，$\mathbf{h}_0$表示用户行为数据预测模型最后一个编码器的输出向量，$\mathbf{h}_1$表示属性预测结果，$\mathrm{FFN}(\mathbf{h}_1, W_1, W_2, b_1, b_2)$表示全连接前馈神经网络的输出，$W_1$、$W_2$表示权重，$b_1$表示、$b_2$表示偏置。

10.一种基于数据增强的社交网络用户属性预测系统，其特征在于，基于权利要求1至9任一项所述的一种基于数据增强的社交网络用户属性预测方法，包括依次电相连的以下模块：

数据预处理模块：用以，提取社交网络用户的转发行为数据，对提取的转发行为数据进行预处理，获得预处理后的用户行为数据；

预测模型构建并训练模块：用以，利用预处理后的用户行为数据，构建用户行为数据预测模型，并训练用户行为数据预测模型；

转发行为预测模块：用以，利用训练好的用户行为数据预测模型预测用户未来一段时间的转发行为，获得增强后的用户行为数据；

属性信息预测模块：用以，利用增强后的用户行为数据预测社交网络中用户的属性信息。

一种基于数据增强的社交网络用户属性预测方法及系统

技术领域

[0001] 本发明涉及数据挖掘技术领域，具体是一种基于数据增强的社交网络用户属性预测方法及系统。

背景技术

[0002] 在过去的二十年间涌现出越来越多的社交平台(例如Twitter、Facebook、Instagram等等)，在这些社交平台上，用户可以第一时间阅读感兴趣的推文并将自己的想法添加到评论区与他人分享。除了评论之外，用户还可以利用更加便捷的转发功能，在原始推文的基础上添加评论后进行转发，便可与粉丝分享你的观点。在推特平台中，用户之间通过转发关系构成错综复杂的网络结构，这种推文传播方式具有传播快、覆盖广的特点，使得某些推文能够在短时间内形成极大的关注和影响。

[0003] 虽然已有工作针对社交网络中的转发行为预测进行了多项研究，但是这些方法均着眼于推文的被转发概率或者用户转发推文的可能性，并未对用户在未来一段时间内的转发对象进行深入研究。除此之外，上述算法的预测精度与特征的构建及选择息息相关，为了对用户的转发行为进行更加精准的预测，数据科学家需要根据业务背景以及专家知识构建大量特征用于机器学习模型的训练，这使得训练好的模型很难迁移到其它数据集或相关业务场景之下。在模型的训练方式上，由于特征工程与模型训练是分开执行的，所以很难选择最佳的特征组合对建立好的模型进行训练，而端到端的训练方式可以将特征构建与预测任务进行有效的结合，从而使得模型能够对不同特征之间的相对重要程度以及特征与预测任务之间的关联性进行全面的考量。

[0004] 随着深度学习技术的快速发展，数据增强技术已经在自然语言处理、语音识别、目标检测等多个领域取得了巨大的进展。如何将数据增强技术应用于行为建模等相关技术是接下来用户属性预测任务的研究重点。

发明内容

[0005] 为克服现有技术的不足，本发明提供了一种基于数据增强的社交网络用户属性预测方法及系统，解决现有技术存在的线网络用户行为序列长度较短时不能为用户属性预测任务提供足够信息、从而导致用户属性预测的预测准确性低的问题。

[0006] 本发明解决上述问题所采用的技术方案是：

一种基于数据增强的社交网络用户属性预测方法，基于用户的历史行为序列，推断用户未来一段时间的行为序列，通过将历史行为序列与预测得到的行为序列进行拼接，扩大行为序列的长度，对用户的行为数据进行增强。

[0007] 作为一种优选的技术方案，包括以下步骤：

S1，进行数据预处理：提取社交网络用户的转发行为数据，对提取的转发行为数据进行预处理，获得预处理后的用户行为数据；

S2，构建并训练预测模型：利用预处理后的用户行为数据，构建用户行为数据预测

模型，并训练用户行为数据预测模型；

S3，预测转发行为：利用训练好的用户行为数据预测模型预测用户未来一段时间的转发行为，获得增强后的用户行为数据；

S4，预测属性信息：利用增强后的用户行为数据预测社交网络中用户的属性信息。

[0008] 作为一种优选的技术方案，步骤S1包括以下步骤：

S11，提取社交网络用户的行为信息：对于给定用户u_i，首先获取u_i在某一段时间跨度内的转发行为序列$H_{u_i}=\{u_1,u_5,\cdots,u_i,\cdots,u_m\}$，在起始位置添加特殊标志符[SOS]，在行为序列的末尾添加特殊标志符[EOS]；并将转发行为对应的时间戳$T_{u_i}=\{t_{u_1},t_{u_5},\cdots,t_{u_i},\cdots,t_{u_m}\}$进行记录；其中，i表示用户的编号，$u_i$表示第i个用户的昵称，$t_{u_i}$表示第i个用户的转发行为对应的时间戳；

S12，首先计算当前转发行为与参照时刻之间的时间间隔，然后以事先设定的周期将时间间隔转换为时间ID，最后计算相邻转发行为之间时间ID的差分值。

[0009] 作为一种优选的技术方案，步骤S2包括以下步骤：

S21，构建用户转发行为嵌入表示层：用户转发行为嵌入表示层包括行为序列嵌入表示层B、行为先后顺序嵌入表示层D、时间信息嵌入层T；其中，行为序列嵌入表示层B用于将用户转发行为序列中的每一个昵称转换为k维的向量表示，行为先后顺序嵌入表示层D用于将每一个被转发用户的转发顺序编号转换为k维的向量表示，时间信息嵌入层T用于将每一个被转发用户的转发时间信息转换为k维的向量表示，将上述三种向量表示按位相加，得到生成式预训练用户行为数据预测模型的输入X'；k≥2且k为整数；

S22，构建编码器：构建包括多头自注意力模块MHA、基于位置的全连接前馈网络FFN的编码器；其中，多头自注意力模块MHA基于缩放点积自注意力用户行为数据预测模型，将嵌入表示矩阵X'作为输入，输出注意力评分矩阵A；基于位置的全连接前馈网络FFN，将注意力评分矩阵A作为输入，输出解码器的隐层表示H；

S23，构建行为序列输出层：行为序列输出层为全连接神经网络，用于将解码器的隐层表示H作为输入，输出预测结果，并通过Softmax函数计算预测结果与真实值之间的误差值δ，所述预测结果指用户下一时刻转发行为；

步骤S24，通过误差反向传播的训练方式更新用户行为数据预测模型参数，直到误差值δ达到最低为止，保存最终的用户行为数据预测模型参数，获得训练好的用户行为数据预测模型；误差值δ达到最低为止指用户行为数据预测模型收敛的情形。

[0010] 作为一种优选的技术方案，步骤S22包括以下步骤：

S221，利用嵌入表示矩阵X'作为输入，先将X'整理为$\mathbf{H}^0=\left[\mathbf{x}_1,\cdots,\mathbf{x}_{|x|}\right]$，再将$\mathbf{H}^0$输入到一个$L$层的Transformer网络中进行编码：

$$\mathbf{H}^l=\text{Transformer}_l\left(\mathbf{H}^{l-1}\right);$$

其中，$l \in [1, L]$，$\mathbf{H}^l = \left[\mathbf{h}_1^l, \cdots, \mathbf{h}_{|x|}^l\right]$，$\mathbf{H}^0$表示用户行为的向量表征，$\mathbf{H}^l$表示第$l$层用户行为数据预测模型的输出结果，$l$表示Transformer网络中的层编号，$\text{Transformer}_l\left(\mathbf{H}^{l-1}\right)$表示第$l$层Transformer网络编码后的结果，$|x|$表示步骤S11中转发行为序列$H_{u_i}$的长度，$\mathbf{h}_1^l, \cdots, \mathbf{h}_{|x|}^l$分别表示第1至第$|x|$个行为在第$l$层的向量表征；在每个Transformer编码器中，有多头注意力机制以聚合前一层的输出向量；第l层的Transformer中的一个自注意力头$\mathbf{A}_l$的计算公式如下：

$$\mathbf{Q} = \mathbf{H}^{l-1}\mathbf{W}_l^Q,$$

$$\mathbf{K} = \mathbf{H}^{l-1}\mathbf{W}_l^K,$$

$$\mathbf{V} = \mathbf{H}^{l-1}\mathbf{W}_l^V,$$

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{可以参照} \\ -\infty, & \text{不能参照} \end{cases},$$

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{k}} + \mathbf{M}\right)\mathbf{V};$$

其中，前层的输出$\mathbf{H}^{l-1} \in \mathbb{R}^{|x| \times d_h}$通过参数为$\mathbf{W}_l^Q, \mathbf{W}_l^K, \mathbf{W}_l^V \in \mathbb{R}^{d_h \times d_k}$的线性变换分别得到查询向量$\mathbf{Q}$、键向量$\mathbf{K}$、值向量$\mathbf{V}$；掩码矩阵$\mathbf{M} \in \mathbb{R}^{|x| \times |x|}$控制着行为之间是否能够被参照，$\mathbf{W}_l^Q$表示第$l$层查询向量对应的线性变换矩阵，$\mathbf{W}_l^K$表示第$l$层键向量对应的线性变换矩阵，$\mathbf{W}_l^V$表示第$l$层值向量对应的线性变换矩阵，i表示掩码矩阵$\mathbf{M}$的行标号，j表示掩码矩阵$\mathbf{M}$的列标号，$\mathbf{M}_{ij}$表示掩码矩阵$\mathbf{M}$的第i行第j列元素，$\mathbf{K}^\top$表示键向量的转置，$d_k$表示隐藏层神经元个数，$d_h$表示用户行为的嵌入维度；

S222，以交叉熵作为损失函数，采用梯度下降法训练行为生成部分的神经网络，将用户的转发行为进行独热编码；

S223：返回步骤S221，循环执行步骤S221~步骤S222，直至训练用户行为数据预测模型收敛为止。

[0011] 作为一种优选的技术方案，步骤S222中，独热编码计算公式为：

$$\varepsilon = -\sum_{i=2}^{n}\sum_{j=1}^{N}\left(u_{i,j}log\left(\hat{u}_{i,j}\right) + \left(1 - u_{i,j}\right)log\left(1 - \hat{u}_{i,j}\right)\right);$$

其中，ε表示损失值，$u_{i,j}$表示真实用户在第i时刻转发用户j的概率，n表示时刻

总数，N表示所分析的用户的总数量，$\hat{u}_{i,j}$表示社交网络用户行为数据预测模型预测的用户i在第j时刻转发用户的概率。

[0012] 作为一种优选的技术方案，步骤S3包括以下步骤：

S31，利用词嵌入层以及位置嵌入层对行为序列对应的时间ID差分序列进行嵌入表示，然后用户行为数据预测模型通过L个编码层得到时间ID差分序列的隐藏表示，最后利用交叉熵损失函数计算用户行为数据预测模型预测结果与期望值之间的偏差；

S32，将嵌入表示矩阵$\mathbf{X}'$输入到多头自注意力机制ATT中，得到注意力权重矩阵，并将注意力矩阵$\mathbf{A}$输入到基于位置的全连接前馈网络FFN中，得到隐层表示$\mathbf{H}$，并通过步骤S23得到用户下一时刻转发行为的预测结果$\mathbf{y}$，最后将预测结果$\mathbf{y}$拼接到输入序列中；

S33，循环执行步骤S31~S32，直至得到用户未来一段时间的转发行为预测结果$\mathbf{Y}$。

[0013] 作为一种优选的技术方案，步骤S4包括以下步骤：

S41，将步骤S33得到的用户行为预测结果$\mathbf{Y}$与用户历史行为序列$\mathbf{X}$进行拼接得到增强后的用户行为数据$\mathbf{Y}'$，并利用步骤S21得到用户行为数据$\mathbf{Y}'$的嵌入矩阵$\overline{\mathbf{Y}}$，对嵌入矩阵$\overline{\mathbf{Y}}$按列求平均得到社交网络用户属性预测用户行为数据预测模型的输入特征$\overline{\mathbf{X}}$；

S42，将特征向量$\overline{\mathbf{X}}$输入到预训练好的用户行为数据预测模型中进行调整，得到用户属性的所属概率值$\mathbf{p}$。

[0014] 作为一种优选的技术方案，步骤S42中，对用户行为数据预测模型进行调整的具体方法为：

仅保留用户行为数据预测模型的嵌入层以及编码器模块，并添加Flatten层做维度变换，最后输入包括依次相连的线性层、激活层、线性层的前馈神经网络，将隐藏特征映射到真实标签，具体的计算方式如下式：

$$\mathbf{h}_1 = \mathrm{FlattenLayer}(\mathbf{h}_0),$$

$$\mathrm{FFN}(\mathbf{h}_1, W_1, W_2, b_1, b_2) = \max(0, \mathbf{h}_1 W_1 + b_1) W_2 + b_2;$$

其中，$\mathbf{h}_0$表示用户行为数据预测模型最后一个编码器的输出向量，$\mathbf{h}_1$表示属性预测结果，$\mathrm{FFN}(\mathbf{h}_1, W_1, W_2, b_1, b_2)$表示全连接前馈神经网络的输出，$W_1$、$W_2$表示权重，$b_1$表示、$b_2$表示偏置。

[0015] 一种基于数据增强的社交网络用户属性预测系统，基于所述的一种基于数据增强的社交网络用户属性预测方法，包括依次电相连的以下模块：

数据预处理模块：用以，提取社交网络用户的转发行为数据，对提取的转发行为数据进行预处理，获得预处理后的用户行为数据；

预测模型构建并训练模块：用以，利用预处理后的用户行为数据，构建用户行为数据预测模型，并训练用户行为数据预测模型；

转发行为预测模块：用以，利用训练好的用户行为数据预测模型预测用户未来一段时间的转发行为，获得增强后的用户行为数据；

属性信息预测模块：用以，利用增强后的用户行为数据预测社交网络中用户的属

性信息。

[0016] 本发明相比于现有技术,具有以下有益效果:

(1)本发明通过采用数据增强的方法,仅基于用户的历史行为序列,便可以合理预测其未来一段时间的行为,通过将历史行为序列与预测得到的行为序列进行拼接,可以有效扩大行为序列的长度,便于为广泛的下游任务提供更加丰富的行为信息;

(2)本发明通过嵌入表示,可以得到用户行为的通用化向量表示方法,基于行为的向量表示方法,可以将其应用于不同的下游任务并均取得非常准确的预测效果;

(3)本发明通过使用微调的方法,以增强后的用户行为数据作为输入,预测该用户的属性信息,从而有效克服特征维度高、筛选有效特征难等问题。

附图说明

[0017] 图1为用户行为数据预测模型的示意图;

图2为本发明所述的一种基于数据增强的社交网络用户属性预测方法的步骤示意图;

图3为数据预处理流程图;

图4为自注意力模块的算法流程图。

具体实施方式

[0018] 下面结合实施例及附图,对本发明作进一步的详细说明,但本发明的实施方式不限于此。

[0019] 实施例1

如图1至图4所示,本发明公开一种基于数据增强的社交网络用户属性预测方法。

[0020] 1)能够充分融合行为信息及其相关的时间信息,并对行为模式识别与行为数据生成两种任务同时进行学习;2)提出一种共享参数的深度学习模型,通过使用精心设计的注意力掩码机制控制行为数据增强过程中所用到的行为以及时间信息;3)选用Transformer的编码器部分进行建模,通过采用多头掩码自注意力机制,可以使模型在预测当前时刻的转发用户昵称时,有侧重地对其所有的历史转发记录进行分析;4)在行为生成模型中引入转发时间信息,帮助模型学习更加复杂的转发行为模式,从而较为准确地预测出用户未来一段时间的转发对象。5)对预训练好的模型针对用户属性预测任务进行微调,输出的结果表示用户属性的所属概率。本发明公开的一种基于数据增强的社交网络用户属性预测方法可应用于在线社交网络中的用户属性补全,从而帮助社交媒体平台建立更加完整的用户画像,所需要的数据在现实网络中易于获取,计算复杂度低,可以在社交网络的用户属性预测任务中获得非常高的准确率。

[0021] 一种基于数据增强的社交网络用户属性预测方法,根据用户的转发行为信息能够快速预测社交网络中用户的属性信息,具体包括以下步骤:

步骤S1:用户行为数据预处理

步骤S11:提取社交网络用户的转发行为信息,用u_i表示第i个用户的昵称,用t_{u_i}表示用户u_i转发行为对应的时间戳。这里用转发对象的昵称表示用户的转发行为,对于给

定用户u_i，首先获取其某一段时间跨度内的转发行为数据$H_{u_i}=\{u_1,u_5,\cdots,u_i,\cdots,u_m\}$，其中在起始位置添加特殊标志符[SOS]，在行为序列的末尾添加特殊标志符[EOS]。并记录用户u_i转发行为对应的时间戳$T_{u_i}=\{t_{u_1},t_{u_5},\cdots,t_{u_i},\cdots,t_{u_m}\}$；

步骤S12：首先计算当前转发行为与参照时刻之间的时间间隔，然后按照事先设定的周期将时间间隔转换为时间ID，最后计算相邻转发行为之间时间ID的差分值；

步骤S2：构建用户行为数据预测模型，模型一共由三个模块组成，包括用户转发行为嵌入表示层，编码器和行为序列输出层；模型的构建具体包括以下步骤：

步骤S21：构建用户转发行为嵌入表示层：用户转发行为嵌入表示层包括行为序列嵌入表示层B、行为先后顺序嵌入表示层D、时间信息嵌入层T；其中，行为序列嵌入表示层B用于将用户转发行为序列中的每一个昵称转换为k维的向量表示，行为先后顺序嵌入表示层D用于将每一个被转发用户的转发顺序编号转换为k维的向量表示，时间信息嵌入层T用于将每一个被转发用户的转发时间信息转换为k维的向量表示，将上述三种向量表示按位相加，得到生成式预训练用户行为数据预测模型的输入X'；k≥2且k为整数；

步骤S22：构建编码器，包括多头自注意力模块(Multi-head Attention, MHA)、基于位置的全连接前馈网络(Feed Forward Neural Network, FFN)。多头自注意力模块MHA基于缩放点积自注意力模型，将嵌入表示矩阵X'作为输入，输出注意力评分矩阵；基于位置的全连接前馈网络FFN，将注意力评分矩阵A作为输入，输出解码器的隐层表示H；

步骤S23：构建行为序列输出层：行为序列输出层为全连接神经网络，用于将解码器的隐层表示H作为输入，输出预测结果，并通过Softmax函数计算预测结果与真实值之间的误差值δ，所述预测结果指用户下一时刻转发行为；

步骤S24：通过误差反向传播的训练方式更新用户行为数据预测模型的参数，直到误差值δ达到最低为止，保存最终的模型参数；

步骤S3：利用训练好的用户行为数据预测模型预测用户未来一段时间的转发行为，具体包括以下步骤：

步骤S31：利用时间嵌入层以及位置嵌入层对行为序列对应的时间ID差分序列进行嵌入表示，然后户行为数据预测模型通过L个编码层得到时间ID差分序列的隐藏表示，最后利用交叉熵损失函数计算户行为数据预测模型预测结果与期望值之间的偏差；

步骤S32：将嵌入表示矩阵X'输入到多头自注意力机制ATT中，得到注意力权重矩阵，并将注意力矩阵A输入到基于位置的全连接前馈网络FFN中，得到隐层表示H，并通过步骤S23得到用户下一时刻转发行为的预测结果y，最后将预测结果y拼接到输入序列中；

步骤S33：循环执行步骤S31、S32得到用户未来一段时间的转发行为预测结果Y；

步骤S4：利用增强后的用户行为数据推理社交网络中用户的属性信息，具体包括以下步骤：

步骤S41：将步骤S33得到的用户行为预测结果Y与用户历史行为序列X进行拼接得到增强后的用户行为数据Y'，并利用步骤S21得到用户行为数据Y'的嵌入矩阵$\overline{\mathrm{Y}}$，对嵌入矩阵$\overline{\mathrm{Y}}$按列求平均得到社交网络用户属性预测用户行为数据预测模型的输入特征$\overline{\mathrm{X}}$；

步骤S42：根据步骤S41，将特征向量$\overline{\mathbf{X}}$输入到预训练好的用户行为数据预测模型中进行调整，得到用户属性的所属概率值$\mathbf{p}$；

2.根据权力要求1所述的基于数据增强的社交网络用户行为生成模型，其特征在于所述步骤S22中的编码器构建方式具体包括以下步骤：

步骤S1：利用嵌入表示矩阵作为输入，先将嵌入表示矩阵整理为。再将其输入到一个层的Transformer网络中对输入进行编码：

S221，利用嵌入表示矩阵$\mathbf{X}'$作为输入，先将$\mathbf{X}'$整理为$\mathbf{H}^0=\left[\mathbf{x}_1,\cdots,\mathbf{x}_{|x|}\right]$，再将$\mathbf{H}^0$输入到一个$L$层的Transformer网络中进行编码：

$$\mathbf{H}^l=\text{Transformer}_l\left(\mathbf{H}^{l-1}\right);$$

其中，$l\in[1,L]$，$\mathbf{H}^l=\left[\mathbf{h}_1^l,\cdots,\mathbf{h}_{|x|}^l\right]$，$\mathbf{H}^0$表示用户行为的向量表征，$\mathbf{H}^l$表示第$l$层用户行为数据预测模型的输出结果，$l$表示Transformer网络中的层编号，$\text{Transformer}_l\left(\mathbf{H}^{l-1}\right)$表示第$l$层Transformer网络编码后的结果，$|x|$表示步骤S11中转发行为序列$H_{u_i}$的长度，$\mathbf{h}_1^l,\cdots,\mathbf{h}_{|x|}^l$分别表示第1至第$|x|$个行为在第$l$层的向量表征；在每个Transformer编码器中，有多头注意力机制以聚合前一层的输出向量；第l层的Transformer中的一个自注意力头$\mathbf{A}_l$的计算公式如下：

$$\mathbf{Q}=\mathbf{H}^{l-1}\mathbf{W}_l^Q,$$

$$\mathbf{K}=\mathbf{H}^{l-1}\mathbf{W}_l^K,$$

$$\mathbf{V}=\mathbf{H}^{l-1}\mathbf{W}_l^V,$$

$$\mathbf{M}_{ij}=\begin{cases}0, & \text{可以参照}\\ -\infty, & \text{不能参照}\end{cases},$$

$$\mathbf{A}=\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{k}}+\mathbf{M}\right)\mathbf{V};$$

其中，前层的输出$\mathbf{H}^{l-1}\in\mathbb{R}^{|x|\times d_h}$通过参数为$\mathbf{W}_l^Q,\mathbf{W}_l^K,\mathbf{W}_l^V\in\mathbb{R}^{d_h\times d_k}$的线性变换分别得到查询向量$\mathbf{Q}$、键向量$\mathbf{K}$、值向量$\mathbf{V}$；掩码矩阵$\mathbf{M}\in\mathbb{R}^{|x|\times|x|}$控制着行为之间是否能够被参照，$\mathbf{W}_l^Q$表示第$l$层查询向量对应的线性变换矩阵，$\mathbf{W}_l^K$表示第$l$层键向量对应的线性变换矩阵，$\mathbf{W}_l^V$表示第$l$层值向量对应的线性变换矩阵，i表示掩码矩阵$\mathbf{M}$的行标号，j表示掩码矩阵$\mathbf{M}$的列标号，$\mathbf{M}_{ij}$表示掩码矩阵$\mathbf{M}$的第i行第j列元素，$\mathbf{K}^\top$表示键向量的转置，$d_k$表示隐藏层神经元个数，$d_h$表示用户行为的嵌入维度；

不同的掩码矩阵$\mathbf{M}$控制关注到不同的上下文信息，例如在双向掩码矩阵中，掩码

矩阵的值为0,表示所有的行为之间都能够相互注意到。

[0022] S222,以交叉熵作为损失函数,采用梯度下降法训练行为生成部分的神经网络,将用户的转发行为进行独热编码(One-Hot Encoding);

独热编码计算公式为:

$$\varepsilon = -\sum_{i=2}^{n}\sum_{j=1}^{N}\left(u_{i,j}log\left(\hat{u}_{i,j}\right) + \left(1-u_{i,j}\right)log\left(1-\hat{u}_{i,j}\right)\right);$$

其中,ε表示损失值,$u_{i,j}$表示真实用户在第i时刻转发用户j的概率,n表示时刻总数,N表示所分析的用户的总数量,$\hat{u}_{i,j}$表示社交网络用户行为数据预测模型预测的用户i在第j时刻转发用户的概率。

[0023] S223:返回步骤S221,循环执行步骤S221~步骤S222,直至训练用户行为数据预测模型收敛为止。

[0024] 本发明的目的在于针对在线网络用户行为序列长度较短,不能为用户属性预测任务提供足够信息的问题,提出一种基于数据增强的社交网络用户属性预测方法,能够对用户的行为数据进行有效增强,从而提高社交网络用户属性预测的预测准确性。

[0025] 本发明具有以下有益效果:

(1)本发明通过采用数据增强的方法,仅基于用户的历史行为序列,便可以合理预测其未来一段时间的行为,通过将历史行为序列与预测得到的行为序列进行拼接,可以有效扩大行为序列的长度,便于为广泛的下游任务提供更加丰富的行为信息;

(2)本发明通过嵌入表示,可以得到用户行为的通用化向量表示方法,基于行为的向量表示方法,可以将其应用于不同的下游任务并均取得非常准确的预测效果;

(3)本发明通过使用微调的方法,以增强后的用户行为数据作为输入,预测该用户的属性信息,从而有效克服特征维度高、筛选有效特征难等问题。

[0026] 实施例2

如图1至图4所示,作为实施例1的进一步优化,在实施例1的基础上,本实施例还包括以下技术特征:

本发明为一种基于数据增强的社交网络用户属性预测方法,包括以下步骤:

步骤S1:用户行为数据预处理;

为了验证本文所提算法的有效性,本实施例中的用户转发行为数据集来自于Internet Archive网站。从该网站获取了2019年9月1日至2019年9月30日之间的推特用户数据,这些数据包含来自5,971,242个用户的50,560,219条推文信息。首先,从原始数据中提取用户昵称、被转发用户昵称、转发时间戳作为用户的转发行为数据。由于不同转发次数下的用户数量服从幂律分布,为了确保用户有足够多的历史转发行为供模型分析,仅考虑转发次数大于10的用户进行分析。在样本选择方面,考虑到计算资源的有限性,对每一类转发次数下的样本进行了随机采样。行为推理模型预训练的一个要素就是如何对每一类转发次数下的样本进行采样,这个选择是一个零和博弈,如果转发行为较多的样本采样频率过

高，模型可能过拟合；如果转发行为较少的样本训练次数不够，模型就会欠拟合。因此，采用XLM中使用的方法，假设有M种转发次数，每一种转发次数下对应的样本记为$\{S_i\}_{i=1,\cdots,M}$，而每一种转发次数下的样本数目记为n_i。然后，将每一种转发次数下的样本随机打乱后按照概率$\{q_i\}_{i=1,\cdots,M}$进行随机采样，其中q_i的计算公式如下：

$$q_i = \frac{p_i^{\alpha}}{\sum_{j=1}^{N} p_j^{\alpha}} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}$$

不难发现，α值越大，对于样本数较多的类别采样频率越高，惩罚力度越小，本发明中选择$\alpha = 0.5$。通过上述采样方式，共得到2,038个用户的219,341条历史转发记录，采样数据中共包含74,936种用户昵称。

[0027] 然后将用户的转发对象昵称映射为0至N之间的一个整数，其中N表示数据集中所有出现过的昵称所构成的集合的大小。通过上述步骤，一个用户的转发行为序列便可表示为由多个数字构成的列表。这里，取该列表的前n项作为模型的输入：$u_1, u_2, u_3, u_4, \ldots\ldots, u_n$，而后项作为模型的期望输出：$u_{m-n}, \ldots\ldots, u_m$。数据预处理过程如图3所示。

[0028] 步骤S2：构建用户行为数据预测模型，如图2所示模型一共由四个模块组成，包括用户转发行为嵌入表示层，编码器、行为序列输出层；

构建用户转发行为嵌入表示层：用户转发行为嵌入表示层包括行为序列嵌入表示层B、行为先后顺序嵌入表示层D、时间信息嵌入层T；行为序列嵌入表示层B用于将用户转发行为序列中的每一个昵称转换为512维的向量表示，行为先后顺序嵌入表示层D用于将每一个被转发用户的转发顺序编号转换为512维的向量表示，时间信息嵌入层T用于将每一个被转发用户的转发时间信息转换为k维的向量表示，将上述三种向量表示按位相加，得到生成式预训练模型的输入$\mathrm{X}' \in R^{n\times 512}$；

解码器包括多头自注意力模块MHA、基于位置的全连接前馈网络FFN。多头自注意力模块MHA基于缩放点积自注意力模型，将嵌入表示矩阵X'作为输入，输出注意力权重矩阵$\mathrm{A} \in R^{16\times 16}$，具体计算方式如图3所示；基于位置的全连接前馈网络FFN，将注意力权重矩阵A作为输入，输出解码器的隐层表示$\mathrm{H} \in R^{n\times 512}$。

[0029] 行为序列输出层为全连接神经网络，将解码器的隐层表示H作为输入，输出用户下一时刻转发行为的预测结果，这里计算得到的是对每一个推特用户的转发概率，概率值越大，表示用户在下一时刻更有可能转发该用户的推文，并通过Softmax函数计算预测结果与真实值之间的误差值。通过使用梯度下降法反复更迭模型参数，使得误差值δ最小化，并将此时的模型参数保存下来。

[0030] 步骤S3：利用训练好的用户行为数据预测模型预测用户未来一段时间的转发行为；

基于训练好的用户行为数据预测模型，通过反复将模型的预测值加入到输入序列

中进一步解码，就可以得到用户在未来一段时间的转发行为序列。将生成的转发行为序列与用户的历史行为序列进行合并便可得到增强后的用户行为数据。

[0031] 步骤S4：利用增强后的用户行为数据推断用户的属性信息；

采用微调的方式训练用户属性推断模型。在微调过程中，为了使模型能够对下游任务进行端到端的训练，仅保留模型的嵌入层以及编码器模块，并在此基础上添加Flatten层做维度变换，最后采用“线性层-激活层-线性层”的结构将隐藏特征映射到真实标签，具体的计算方式如下式：

$$\mathbf{h}_1 = \mathrm{FlattenLayer}(\mathbf{h}_0),$$

$$\mathrm{FFN}(\mathbf{h}_1, W_1, W_2, b_1, b_2) = \max(0, \mathbf{h}_1 W_1 + b_1) W_2 + b_2;$$

其中，$\mathbf{h}_0$表示用户行为数据预测模型最后一个编码器的输出向量，$\mathbf{h}_1$表示属性预测结果，$\mathrm{FFN}(\mathbf{h}_1, W_1, W_2, b_1, b_2)$表示全连接前馈神经网络的输出，$W_1$、$W_2$表示权重，$b_1$表示、$b_2$表示偏置。

[0032] 其中表示用户行为数据预测模型最后一个编码器的输出向量。微调模型中新增线性层权重参数以正态分布初始化，偏置参数初始化为常数。

[0033] 为了检验在本实施例中本发明所提出的基于数据增强的用户属性推断方法的效果，在整个数据集上进行了测试，整个数据集包含2,038名Twitter用户以及这些用户的219,341条转发行为。选择该数据集中80%左右用户的行为信息以及属性信息作为训练集，10％左右的用户行为以及属性信息作为验证集，10％左右的用户行为以及属性信息作为测试集。在测试集上计算BLEU-4和Accuracy两个值分别作为行为生成和属性推断的评价指标。

[0034] BLEU-4指标的计算方法为：c_i表示模型生成的行为序列，$s_i = s_{i1}, \cdots, s_{im}$表示m个参考结果，$h_k(c_i)$表示元素$w_k$在行为序列$c_i$中出现的次数，$h_k(s_{ij})$表示元素$w_k$在参考结果$s_{ij}$中出现的次数，$w_k$表示序列中的第k个n-gram词组，$\max_{j \in m} h_k(s_{ij})$表示元素$w_k$在各条参考结果中的最大出现次数。基于上述定义，我们给出各阶n-gram的精度计算公式：

$$P_n = \frac{\sum_i \sum_k \min\left(h_k(c_i), \max_{j \in m} h_k(s_{ij})\right)}{\sum_i \sum_k \min\left(h_k(c_i)\right)}$$

Accuracy指标计算的是所有预测正确样本占所有实验样本的比例。

[0035] 本次实施例的实验结果如下：

行为序列生成的BLEU-4稳定在5.98，测试集Accuracy值稳定在0.81。

[0036] 该实验结果表明本发明所提出的基于数据增强的用户属性推断方法可以在在线社交网络中实现对用户属性推断并取得很好的效果。

[0037] 如上所述，可较好地实现本发明。

[0038] 本说明书中所有实施例公开的所有特征，或隐含公开的所有方法或过程中的步骤，除了互相排斥的特征和/或步骤以外，均可以以任何方式组合和/或扩展、替换。

[0039] 以上所述，仅是本发明的较佳实施例而已，并非对本发明作任何形式上的限制，依

据本发明的技术实质,在本发明的精神和原则之内,对以上实施例所作的任何简单的修改、等同替换与改进等,均仍属于本发明技术方案的保护范围之内。

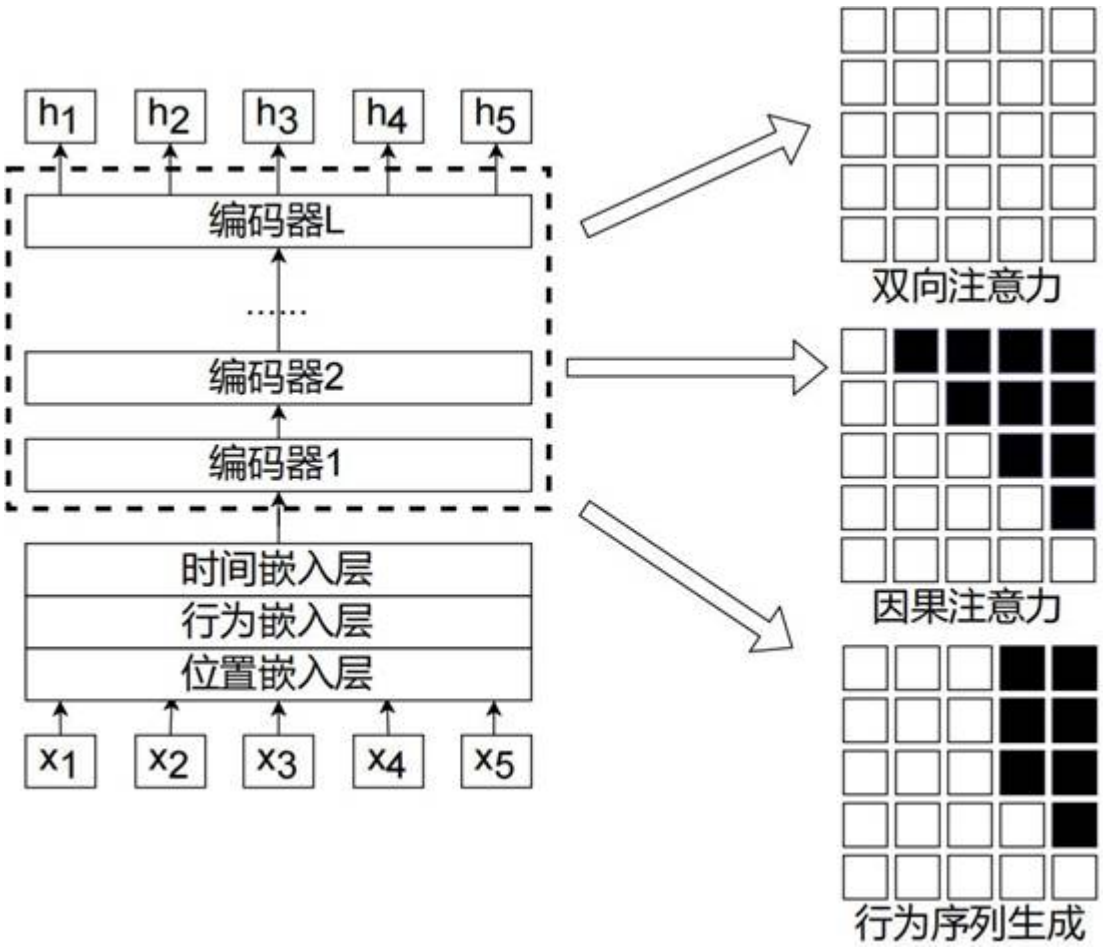

h1
h2
h3
h4
h5
编码器L
.....
编码器2
编码器1
时间嵌入层
行为嵌入层
位置嵌入层
x1
x2
x3
x4
x5
双向注意力
因果注意力
行为序列生成

图1

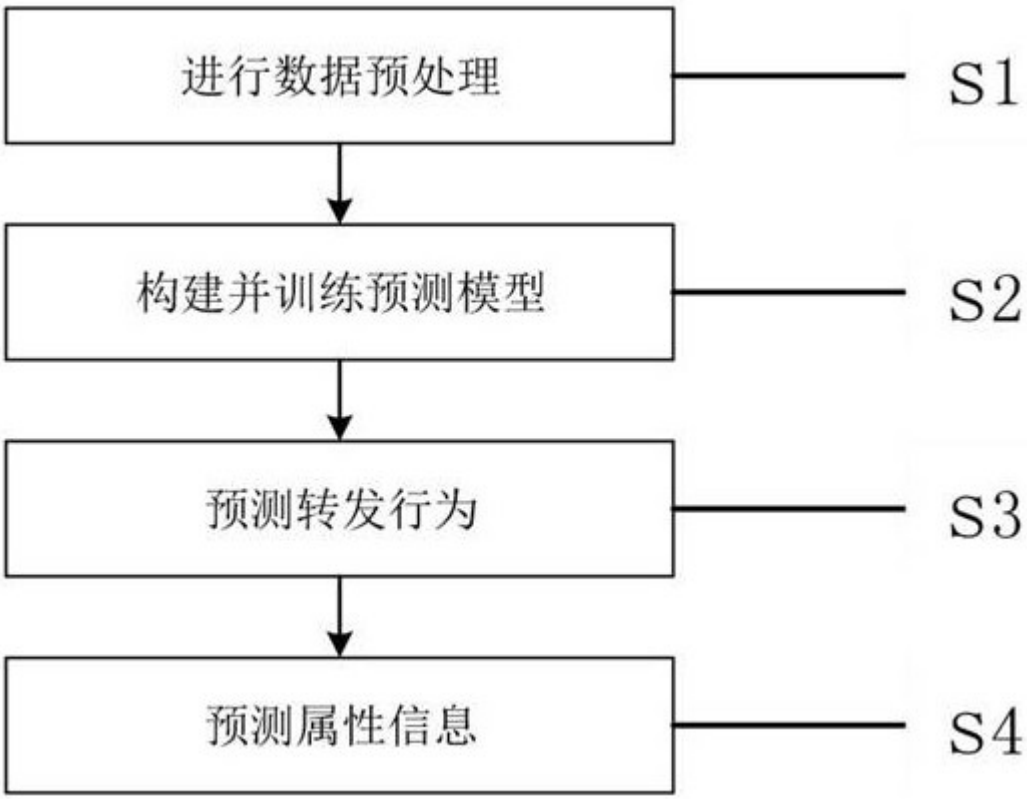


图2

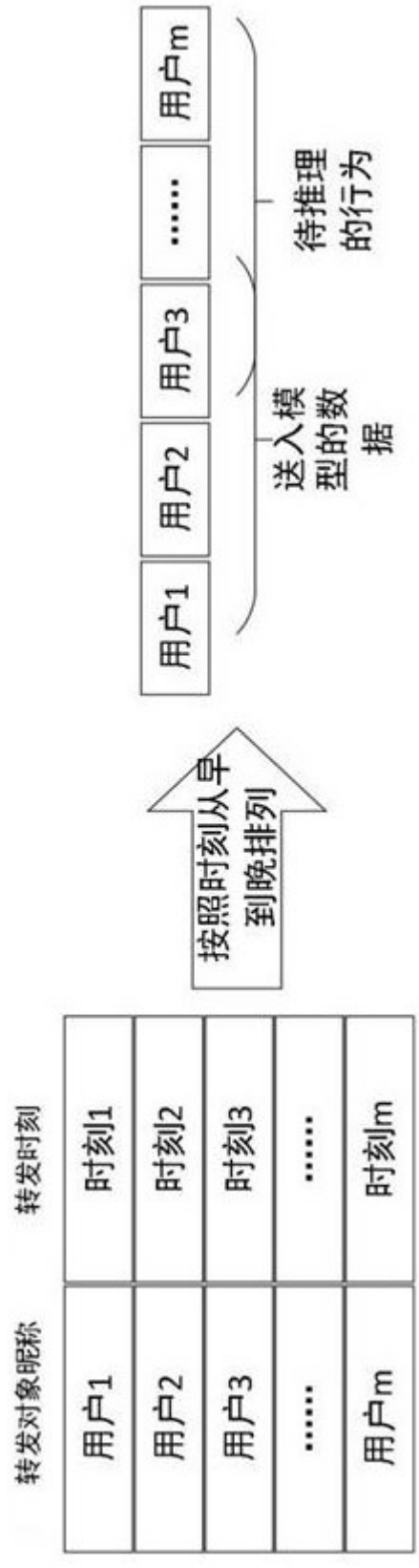
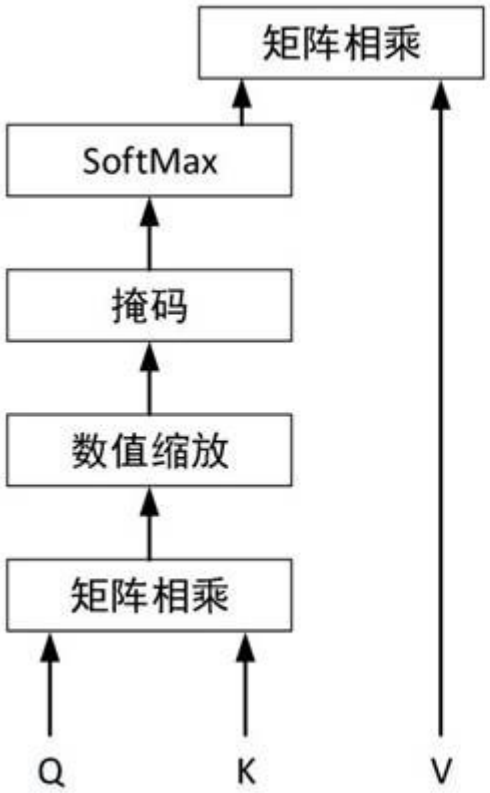

转发对象昵称
转发时刻
用户1
时刻1
用户2
时刻2
用户3
时刻3
……
……
用户m
时刻m
按照时刻从早到晚排列
用户1
用户2
用户3
……
用户m
送入模型的数据
待推理的行为

图3


矩阵相乘
SoftMax
掩码
数值缩放
矩阵相乘
Q
K
V

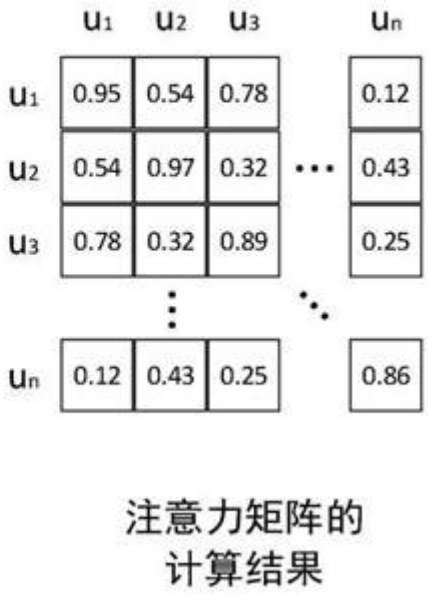

u1 u2 u3 un
u1 0.95 0.54 0.78 0.12
u2 0.54 0.97 0.32 0.43
u3 0.78 0.32 0.89 0.25
un 0.12 0.43 0.25 0.86
注意力矩阵的
计算结果

图4