文章编号：2095-2163(2020)12-0049-05 中图分类号：TP391 文献标志码：A

基于自注意力机制的用户画像

张 维，陈泽宇
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：传统的词向量模型生成的词向量，存在着难以表达出一词多义和学习到词与词之间的依赖关系的问题。针对于此，本文提出基于自注意力机制的用户画像。首先采用自注意力机制，将所有单词信息编码进每一个单词中，学习查询句中词的语义，理解一词多义、一义多词。然后利用多头注意力机制提升模型能力，全面理解查询句中词与词之间的复杂语义。最后利用支持向量机(SVM)分类算法，得到用户基本属性的分类结果，构建用户画像。实验结果表明，模型分类精度高于使用词向量模型和 LDA 模型方法的分类精度。

关键词：自注意力机制；词向量模型；用户画像；SVM

User portrait based on self-attention mechanism

ZHANG Wei, CHEN Zeyu
(School of Electrical and Electronic Engineering, Shanghai University of Engineering and Technology, Shanghai 201620, China)

【Abstract】The word vector generated by the traditional word vector model is difficult to express the polysemy of a word and to learn the dependence between words. In view of this, this paper proposes a user portrait based on self-attention mechanism. Firstly, the self-attention mechanism is adopted to encode all word information into each word, learn the semantics of the words in the query sentence, and understand polysemy and polysemy. Then the multi-head attention mechanism is used to improve the model's ability to fully understand the complex semantics between words in the query sentence. Finally, support vector machine (SVM) classification algorithm is used to get the classification results of the basic attributes of users, and the user portrait is constructed. The experimental results show that the classification accuracy of the model is higher than that of the word vector model and LDA model.

【Key words】Self-attention mechanism; Word vector; User portrait; SVM

0 引 言

随着互联网技术的快速发展，用户在使用软件或浏览网页时的大量行为数据系统会保存下来，但其中能够精确表达用户喜好的行为数据却很少。利用信息过滤技术，对用户行为数据进行深入的分析和挖掘，能够筛选出表达用户兴趣爱好的显式或隐式的行为数据，帮助企业精准定位客户需求，推动企业发展。用户画像[1-2]能够有效的对用户行为信息过滤，深层次分析用户的基本属性和行为习惯。利用贴标签的方式[3]描述用户的喜好，可以帮助企业精准的服务客户，提高用户的消费体验。

用户画像技术的具有极高的应用价值。文献[4]通过研究 APP 中得到的用户行为数据，构造用户画像，分析出用户流失域的主要问题。文献[5]采用贝叶斯网络来构造用户画像，使用多元线性回归模型追踪到用户随时间不断变化的兴趣爱好。用户画像的构建，面临着各种各样的挑战。例如，在使用传统的向量空间模型构建用户画像时，文本向量存在特征稀疏的问题，不能够准确的表达出词与词之间的语义关系。文献[6]中提出一种在搜索引擎上构建短文本用户画像的方法。在文本向量构建时引入词向量模型，解决了传统方法中的特征稀疏问题，但此方法依然无法表达词与词之间的语义关系。为了解决这个问题，文献[7]提出一种将文本向量与对应的文本主题分布相融合的方法，它能够使词向量具有更加丰富的含义。文献[8]提出基于改进词向量模型的用户画像研究，利用 LDA 模型生成查询词的主题信息，将查询词和其主题信息拼接输入到神经网络模型中，学习包含主题信息的词向量。针对上述的问题，本文提出基于自注意力机制的用户画像，采用自注意力机制，学习查询词中词与词之间的强依赖关系，化解语序不同所带来的语义差异问题，利用多头注意力机制，将关注不同语义方面的各个注意力机制进行拼接，输入神经网络，得到更加全面的语义理解。最后使用 SVM 分类算法对属性标签进行分类。

作者简介：张 维(1995-)，男，硕士研究生，主要研究方向：人工智能、推荐算法；陈泽宇(1995-)，男，硕士研究生，主要研究方向：自然语言处理。

通讯作者：陈泽宇 Email：240791324@qq.com

收稿日期：2020-08-18

1 相关工作

1.1 自注意力机制

生活中存在着很多牵动科学的规律,注意力机制就来源于对客观规律的认知。人类在观察客观事物时,首先会注意到自己最感兴趣的部分,自动忽略了其他的信息,这是人类先天具有的选择性注意意识。注意力机制(Attention mechanism)[9]与这种人类选择性注意方式相似。科学家将这种注意力方式通过数学抽象成一种组合函数,通过计算概率分布的形式,来表明词与词之间的联系。

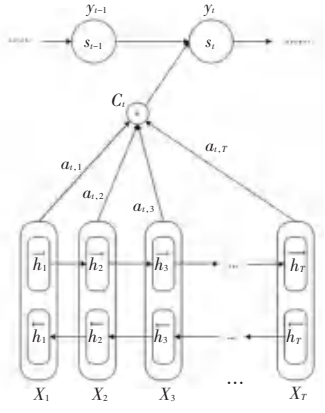

图 1 注意力机制的结构图

Fig. 1 The structure diagram of attention mechanism

针对传统的 Encoder-Decoder 框架对输入序列区分程度不高的问题,Bahdanau 等人[10]引入了注意力机制到 Encode 端和 Decoder 端,其提出的注意力机制的模型结构图如图 1 所示。注意力机制能够提取出稀疏数据的特征,但是没有办法提取出数据之间的相关性。2017 年,Ashish Vaswani 等[11]提出了 Transformer 模型,该模型使用自注意力机制替换掉 Encoder-Decoder 框架的注意力机制,其能够提取出词与词之间的关联度。模型在机器翻译、语音识别等[12]多个领域取得了前所未有的效果。如图 2 所示,这是 Transformer 模型中一个 encoder 的内部结构。编码器先将单词以词向量 x_i 作为模型的输入,一般来说,输入 x_i 是经前馈神经网络训练后的词向量[13]。因此,这种方式解决了 one-hot 编码的词向量过于稀疏的问题。当输入 x_i 进入 Encoder 时,首先会经过 self-attention 层的计算,对单词进行编码得到 z_i。z_i 将输入序列中所有单词信息编码进每一个单词中,通过模型的深入提炼,可以关注到词与词之间的关联程度,然后将 self-attention 层的输出传入一个前馈神经网络中得到 r_i。为了对单词进行非结构化编码,每一个前馈神经网络的参数都是相同的。

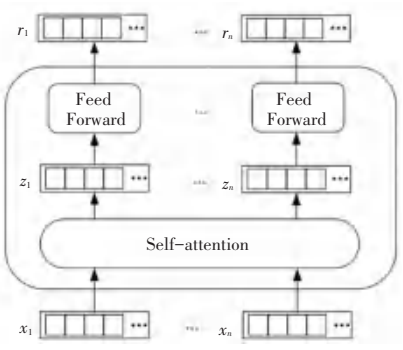

图 2 Encoder 结构图

Fig. 2 Encoder structure diagram

Decoder 的结构与 Encoder 相似,但在 transformer 中 decoder 只是比 Encoder 多了 Encoder-Decoder Attention 层。Transformer 在为模型的输入进行初始化时,还加入了词的绝对位置信息,来解决词序的问题。

$$x' = x + t.$$

其中,t 用来表示词的绝对位置编码,通过人为设定。位置编码是利用 512 维的向量表示,值的大小在-1~1 之间。自注意力机制中的计算过程如下:

Input: $x_i W^Q\ W^K\ W^V d_k$ **Output**: z_i

1: W^Q, W^K, W^V = modelTrain( ) //通过模型生成 3 个权重向量 W^Q , W^K , W^V

2: $q = x_i \cdot W^Q$ //计算 查询词向量 q ,关键词向量 k ,权重值向量 v

3: $k = x_i \cdot W^K$

4: $v = x_i \cdot W^V$

5: $attentionScore = Q \times K^T$ // $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ 是 q , k , v 的矩阵表示

6: $Attention(Q,K,V)$ = softMax($attentionScore, V, d_k$) //归一化并根据公式 1 计算 Attention

7: z_i = sum($Attention(Q,K,V)$) //计算 z_i

z_i 是输入另一种举证的表示形式。每一个单词对应行向量,列是词向量表示的一个维度。self-attention 的计算公式如下。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}(\frac{Q \times \boldsymbol{K}^T}{\sqrt{d_k}}) \cdot \boldsymbol{V} \quad (1)$$

其中,$\sqrt{d_k}$ 为伸缩因子,起调节作用,目的是减小 Q 和 K 的内积值,Q 、K 、V 是 q 、k 、v 向量的矩阵表示。

1.2 多头注意力机制

虽然每个人都具有选择性注意意识,但是每个人的注意点有所差别。因此,将每个人的注意力点

进行整合，就可以挖掘出因不同原因容易忽略掉的信息。语义的表达也类似，词与词之间的语义关系比词的语义要复杂很多，单个注意力机制无法把词与词之间的语义关系描述清楚，所以需要利用多个注意力机制，来帮助模型深入理解文本。这种描述词与词之间语义关系的机制称为多头注意力机制（Multi-head self-attention）[14]。在 Multi-head self-attention 中，每个机制上的注意点都有不同的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 权重矩阵，并且每个注意力机制可以计算出具有不同关联程度的 $\boldsymbol{Z}$ 矩阵。最后将多个 $\boldsymbol{Z}$ 矩阵通过拼接的方式，表达出更全面的关联程度。计算过程如图 3 所示。

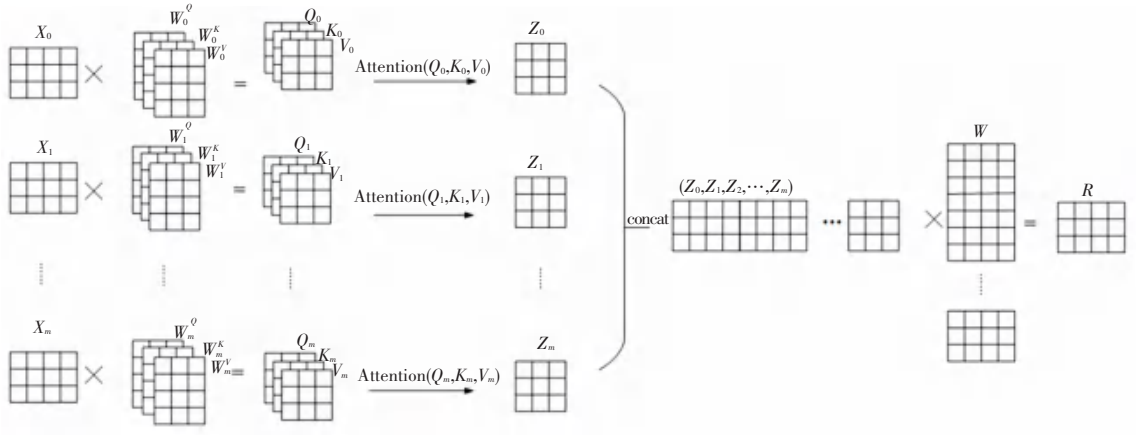

图 3　Multi-head self-attention 计算过程

Fig. 3　Multi-head self-attention calculation process

1.3　SVM 分类模型

SVM 是一种关于统计学习理论的机器学习方法[15]，是一种经典的二分类模型。基本模型是线性分类器，被定义在特征空间上，间隔最大化方式，使其有别于感知机。SVM 所包含的核技巧，使其成为实质上的非线性分类器。

SVM 学习，能够正确分类训练数据集，并且几何间隔最大化的分离超平面。$\vec{\omega} \cdot x + \vec{b} = 0$ 即为分离超平面。由此可知，超平面是受参数 $\vec{w}$ 和 $\vec{b}$ 影响。样本到超平面的距离如公式(2)所示：

$$d = \frac{|\vec{\omega} \cdot x + \vec{b}|}{\| \omega \|}. \tag{2}$$

距离超平面最近的样本被称为支持向量。类别不同的支持向量与超平面的距离表示为：$\gamma = \max \frac{2}{\| \omega \|^2}$。因此，需要最小化参数 $\| \omega \|$ 来得到最大 γ 值。

2　基于自注意力机制的用户画像模型

在本算法的自注意力机制中，矩阵 $\boldsymbol{K}$、$\boldsymbol{Q}$ 和 $\boldsymbol{V}$ 是相同的。通过 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 相乘得到一个 attention score 矩阵。利用不同的矩阵 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$，可以将单词映射到不同空间上，提高 attention score 矩阵的泛化能力。自注意力机制能够将一篇文本中的某个词对这篇文本中其他词的关注程度表示出来。例如：要理解“我不吃苹果是因为其太甜了”这句话中的“其”指代是“苹果”还是“我”，通过计算机来识别是非常困难的。但通过 self-attention 查看“其”与文本中其它词的关注程度，可以发现大部分的注意力都集中在“苹果”上。

模型的输入类似于 Transformer 的输入，将词的原始编码及其位置编码相加作为模型的输入。自注意力机制可以将输入序列中所有单词信息编码进每一个单词中，所以文本向量不用将训练得到的每个词向量相加来表示，只需将输入序列中第一个位置设置成[CLS]，模型训练结束后的输出向量即为文本向量。最后将文本向量作为用户特征放入分类模型中，对用户属性进行分类。基于自注意力机制的用户画像的流程如图 4 所示。

基于自注意力机制的用户画像模型如图 5 所示。其算法实现步骤如下：

(1)对每一个输入序列的起始位置设置一个[CLS]。

(2)用 10 层 Encoder 为输入序列中的单词进行编码。

(3)编码器第一个位置的输出作为用户特征。

(4)使用 SVM 分类算法对用户特征分类。

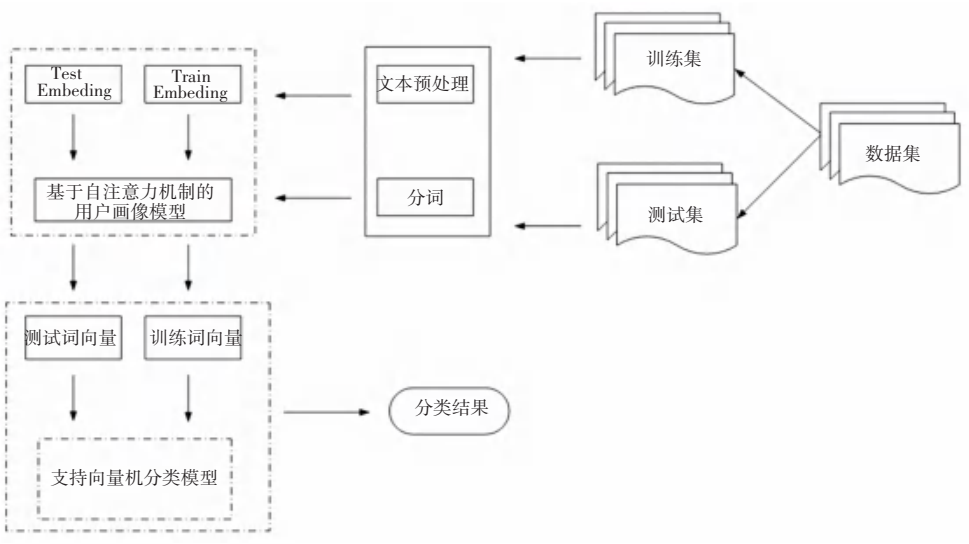

图 4 基于自注意力机制的用户画像流程

Fig. 4 User portrait process based on self-attention mechanism

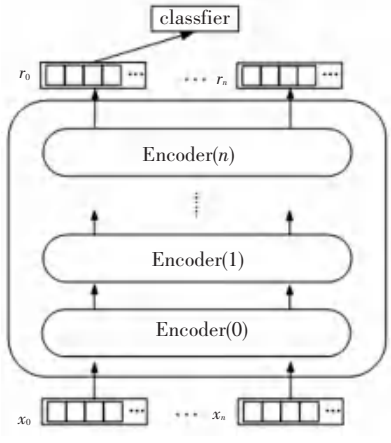

图 5 基于自注意力机制的用户画像模型

Fig. 5 User portrait model diagram based on self - attention mechanism

3 实验及结果分析

3.1 数据集

实验数据集中包含了 10 万条用户数据和搜索文本。其中包括用户的学历、ID 编码、性别、年龄等用户属性标签。数据集取自于中国计算机学会(CCF)数据竞赛。部分特征表示见表 1。

ID 是通过加密唯一表示。在数据集中,人为把年龄属性分为 6 类,由数字 1-6 表示;学历属性分为 6 类,由数字 1-6 表示;性别属性被分为 2 类,由数字 1-2 表示。各类属性均用数字 0 表示未知。本文最终实验结果是将 10 万条数据进行 5 次五折交叉验证得到。将数据集划分成 5 份,其中 4 份作为训练集,1 份作为测试集,一共划分了 5 次数据集来进行实验,最后将 5 次实验的结果取平均作为最终结果。

表 1 实验数据集

Tab. 1 Experimental data set

ID	Gender	Age	Education	Querytext
A3A0E25C1D63	2	1	5	你微笑如花是什么歌　高考体检
8AB7D1A9F687	1	3	3	伊苏树海拉比兔在哪　广州车展门票
C85C8C2A6F00	1	4	4	木耳的功效与作用　集成吊顶颜色怎么搭配
6931EFC26D229	2	4	3	野鸡蛋储存方法　儿童舞蹈视频

3.2 模型评估指标

将 3 种使用不同模型方法的查全率 R 、查准率 P 、和 $F1$ 的值作为评价标准[16]。分别计算用户的年龄、学历、性别属性的分类召回率 R 、精确率 P 和 $F1$ 值。

$$R = \frac{TP}{TP + FN}, \tag{3}$$

$$P = \frac{TP}{TP + FP}, \tag{4}$$

$$F1 = \frac{2 \times P \times R}{P + R}. \tag{5}$$

其中, TP 表示算法对某一类预测正确的个数; FP 表示将其它类的数据错误预测为本类的个数; FN 表示本类数据被错误预测的个数。

3.3　实验结果与分析

本文与使用 LDA 模型和 Word2Vec 模型的方法进行对比分析。每个算法统一使用 SVM 分类器对用户属性进行分类。实验结果见表 2。3 种属性的分类召回率均值、精确率均值和 $F1$ 均值如图 6 所示。

表 2　不同算法的分类性能

Tab. 2　Classification performance of different algorithms

类别	LDA			Word2Vec			本文方法		
	召回率	精准率	F1 值	召回率	精准率	F1 值	召回率	精准率	F1 值
性别	0.792	0.791	0.813	0.831	0.830	0.830	**0.836**	**0.837**	**0.832**
年龄	0.562	0.542	0.556	0.612	0.582	0.601	**0.635**	**0.612**	**0.631**
学历	0.593	0.565	0.578	0.628	0.602	0.625	**0.640**	**0.632**	**0.641**

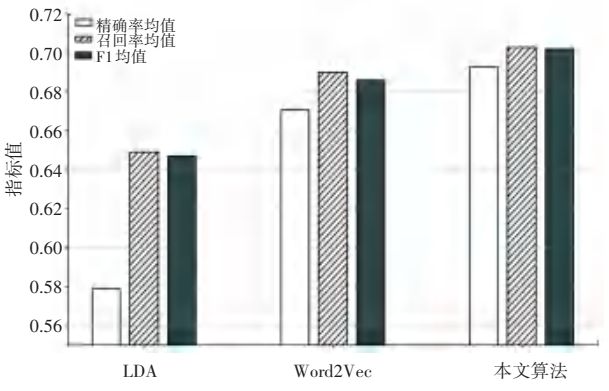

图 6　不同算法的实验结果对比

Fig. 6　The comparison of experimental results in different algorithms

从实验结果中可以看出，本文算法分类效果明显优于其他两种算法。本文方法的分类准确率均值较使用 LDA 模型的方法提高了 9%，较使用 Word2Vec 模型的方法提高了 2.1%；平均召回率较使用 LDA 模型的方法提高了 4.9%，较使用 Word2Vec 模型的方法提高了 1.2%；F1 均值较使用 LDA 模型的方法提高了 1.52%，较使用 Word2Vec 模型的方法提高了 0.9%。因此，基于自注意力机制的方法可以增强词与词之间的依赖关系，能够更好的理解查询句的语义。

4　结束语

本文研究了在搜索引擎上短文本用户画像的相关问题。通过实验将本文所提出的基于自注意力机制模型的方法与 LDA 模型、Word2Vec 模型的方法，在性别、年龄、学历 3 种属性的分类精准率均值、召回率均值和 F1 均值上进行了对比与分析。实验结果表明使用自注意力机制模型的方法对用户搜索文本、进行文本表示时能得到更好的效果。但在数据集中还存在特征分布不平衡等问题。例如，年龄分布不均匀、学历分布也不均匀等。因此，进一步将通过平衡数据的方法来提高分类精度。

参考文献

[1] 费鹏. 用户画像构建技术研究[D]. 大连：大连理工大学，2017.
[2] 黄文彬，徐山川，吴家辉. 移动用户画像构建研究[J]. 现代情报，2016，36(10)：54-61.
[3] 牛温佳. 用户网络行为画像：大数据中的用户网络行为画像分析与内容推荐应用[M]. 北京：电子工业出版社，2016.
[4] 李映坤. 大数据背景下用户画像的统计方法实践研究[D]. 北京：首都经济贸易大学，2016.
[5] 张小可，沈文明，杜翠凤. 贝叶斯网络在用户画像构建中的研究[J]. 移动通信，2016，40(22)：22-25.
[6] 李雅坤. 基于搜索引擎的用户画像构建方法研究[D]. 太原：山西财经大学，2018.
[7] 张小川，余林峰，桑瑞婷，等. 融合 CNN 和 LDA 的短文本分类研究[J]. 软件工程，2018，26(6)：17-20.
[8] 陈泽宇，黄勃. 改进词向量模型的用户画像研究[J]. 计算机工程与应用，2020，56(1)：180-182.
[9] TAMAKA，H，TOKUTANA T，et al. Algorithms for The Maximum Subarray Problem Based on Matrix Multiplication[J]. Interdisciplinary Information Sciences，1998，6(2)：99-104.
[10] 祝元勃. 基于深度学习与自注意力机制的情感分类方法研究[D]. 西安：西安理工大学，2019.
[11] ASHUSH V，NOAM S，Niki Parmar，et al. Attention Is All You Need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach，CA，USA，2017：1-1.
[12] BAHANAU D，CHO K，BENGIO Y. Neural MachineTranslation by Jointly Learning to Align and Translate[J]. Computer ence，2014.
[13] TURIAN J，RATINOV L，BENGIO Y，et al. Word Representations：A Simple and General Method for Semi-supervised Learning[C]//Proceedings of the 48th Annual Meeting of the As-sociation for Computational Linguistics. Uppsala，Sweden，2010：384-390.
[14] 朱张莉，饶元，吴渊，等. 注意力机制在深度学习中的研究进展[J]. 中文信息学报，2019，33(6)：1-11.
[15] MCCALLUM A，NIGAM K. A Comparison of Event Models for Naive Bayes Text Classification[C]// AAAI-98 WORKSHOP N LEARNING FOR TEXT CATEGORIZATION，1998，62(2)：41-47.
[16] 周志华.机器学习[M]. 北京：清华学出版社，2016.