计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900194

多源异构用户生成内容的融合向量化表示学习

纪南巡　孙晓燕　李祯其
中国矿业大学信息与控制工程学院　江苏 徐州 221008
(jinanxun@cumt.edu.cn)

摘　要　随着移动网络和 APPs 的发展,包含用户评价、标记、打分、图像和视频等多源异构数据的用户生成内容(User Generated Contents,UGC)成为提高个性化服务质量的重要依据,对这些数据的融合和表示学习成为其应用的关键。对此,提出一种面向多源文本和图像的融合表示学习。采用 Doc2vec 和 LDA 模型,给出多源文本的向量化表示,采用深度卷积网络获取与评价文本相关的图像特征;给出多源文本向量化表示的多策略融合机制,以及文本和图像卷积融合的表示学习。将所提算法应用于亚马逊含 UGC 内容的商品数据集上,基于 UGC 向量化表示物品的分类准确率说明了该算法的可行性和有效性。

关键词:用户生成内容;表示学习;多源异构;融合;短文本
中图法分类号　TP391

Fusion Vectorized Representation Learning of Multi-source Heterogeneous User-generated Contents

JI Nan-xun,SUN Xiao-yan and LI Zhen-qi
School of Information and Control Engineering,China University of Mining and Technology,Xuzhou,Jiangsu 221008,China

Abstract　With the development of mobile networks and APPs,user generated contents (UGC) containing multi-source heterogeneous data such as evaluations,markings,scoring,images and videos are greatly valuable information for improving the quality of personalized services. The representation learning of fusion and vectorization on the multi-source heterogeneous UGC is the most critical issue for the successful application. Motivated by this,we propose a representation learning method for effectively fusing and vectorizing the comments and image data. We utilize the Doc2vec and LDA models to sufficiently extract the features of the multi-source comments. The images correlated with the comments are represented with deep convolutional network. A hybrid vectorized representation learning for fusing comments and a convolution strategy for integrating images and comments are presented. The feasibility and effectiveness of the proposed method is demonstrated by applying it to typical Amazon public data sets with heterogeneous UGC,in which the vectorized multi-source heterogeneous UGC is taken as the representation of each product and the classification accuracy of the products are compared.

Keywords　User generated contents,Representation learning,Multi-source heterogeneous,Fusion,Short text

1　引言

在个性化服务领域,各类应用 APPs 已成为不可或缺的生活甚至是工作的工具。用户在使用 APPs 的过程中,不仅会产生大量的即时交流信息,还同时提供动态变化的文本评论、打分、标签、图片和视频等海量数据[1]。显然,在当前网络技术下,用户已经变成了数据的主动创造者,提供了大量用户生成内容(User Generated Contents,UGC),其成为个性化服务领域具有极其重要应用价值的数据组成部分。体现用户个性化信息的 UGC 数据具有明显的多源异构特性,其多源化体现在对于同一个对象的描述和评价由不同的人从不同的角度以多种不同的数据形式(如文本、图像或视频等)给出[2]。如何在大量用户生成数据环境中,实现高效的个性化搜索和推荐,在当前个性化服务领域已引起广泛关注。而要实现上述任务,UGC 多源异构数据的向量化表示学习至关重要[3]。

表示学习在文本和图像分类等领域已取得了诸多研究成果[4-8]。而对于多源异构 UGC 数据,由于数据来自不同用户的评价文本、物品的标签描述以及物品的图片信息等(其中,评价文本多为口语化、碎片化、含噪强的短文本[9-10],标签具有不唯一性,图片质量良莠不齐等),从而使得该类多源异构数据的融合和表示学习具有较大的难度。针对短文本分类处理,Zhang 等[11]利用 Word2vec 模型训练词向量后引入 TF-IDF 对训练好的词向量进行加权处理,以实现加权的分类模型对微博短文本进行分类。Zhang 等[12]使用相关词语和隐性话题来增加稀疏文本的话题相关度,同时融合了最大熵分类器和朴素贝叶斯分类器进行文本情感分类。Lai 等[13]基于词嵌入模型,提出了采用深度学习捕获文本的语义表示,并在文本分类方面取得了良好的效果。Chen 等[14]提出了一种基

到稿日期:2020-09-27　返修日期:2021-01-08　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61876184)
This work was supported by the National Natural Science Foundation of China(61876184).
通信作者:孙晓燕(xysun78@126.com)

于潜在狄利克雷分布主题模型和 KNN 的改进的短文本分类方法，使用的 LDA 模型有助于使文本更加注重语义并减少数据的稀疏性。但是，如何将多用户评价的短文本信息，与商品标签描述以及图像描述等进行融合，进而给出更精准的向量化表示，从而使其更好地服务于个性化平台或用户搜索，尚未见到相关研究。

基于此，本文提出了融合多源异构 UGC 数据的物品向量化表示学习策略。如图 1 所示，商品“Chair”既有商家提供的文本描述和图像，又有买家提供的评价，将这些信息进行融合表示学习，进而形成一个含有该物品不同角度信息特征的向量。为此需重点解决如下问题：1）UGC 数据中单条短文本评论的精确向量化表示；2）基于物品描述和用户评价文本的融合；3）对文本向量和图像特征表示等异构信息进行融合与向量化表示。

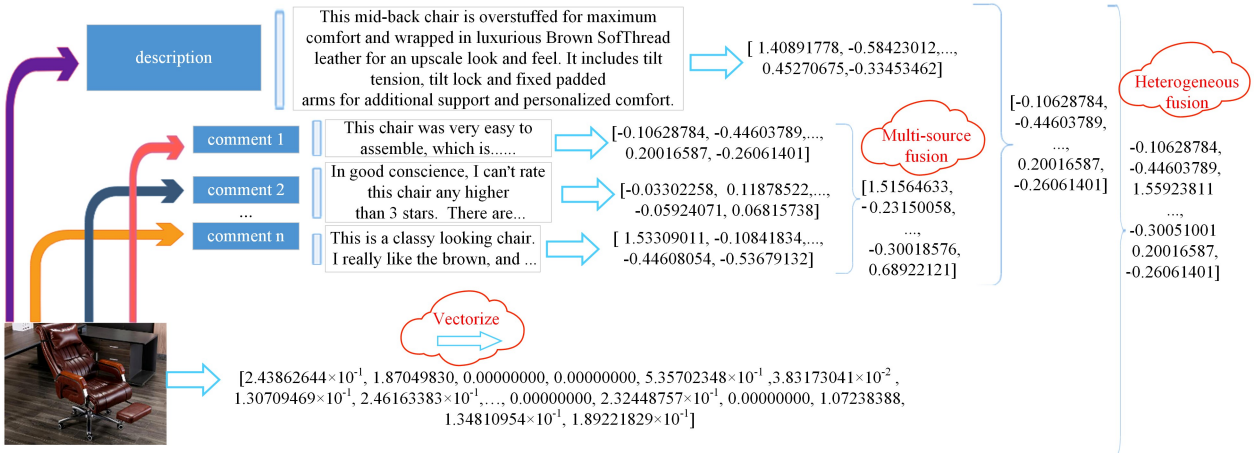


图 1 用户生成多源异构数据向量化表示的示例

Fig. 1 Example of user-generated multi-source heterogeneous data vectorized representation

本文研究思路如下：1）基于 Doc2vec 和 LDA 主题模型实现对用户短文本的精准向量化特征表示。利用 Doc2vec 模型获得短文本上下文的语义关联信息，采用 LDA 主题模型对短文本进行扩展表示学习，对上述两种方法获得的向量采用拼接、相加和张量操作进行融合，从而得到用户生成的短文本评价的向量形式的精准表示。2）在进行多源文本特征融合时，以皮尔逊相似系数衡量物品描述和评论文本之间的相关程度，确定参与融合的评论文本向量并进行融合。3）利用 ResNet 模型获取商品的图像特征，然后基于前 2 个思路，采用卷积融合策略对同一物品的文本和图像特征向量进行融合，得到商品 UGC 数据的向量化表示。本文将所提方法应用于典型亚马逊公开数据集中，实验结果表明，本文提出的短文本表示学习向量化方法更精准，并且基于卷积的多源异构数据的融合策略能够获取 UGC 表示学习更全面的信息，进一步提高对商品描述的精确性，尤其在商品文本数据量少时，融入图像特征对于物品表示精度的改善更为显著。

2 所提算法框架

本文所提算法的框架如图 2 所示。

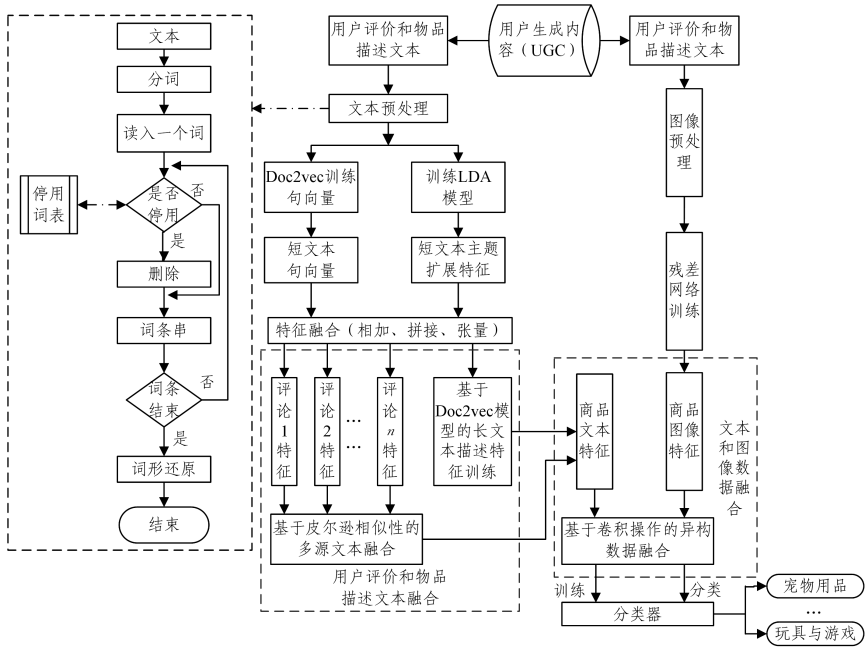


图 2 本文算法框架图（电子版为彩色）

Fig. 2 Framework diagram of proposed algorithm

本文算法主要包括如下 5 部分内容：1）融合 Doc2vec 和 LDA 双模型的短文本向量化表示学习；2）基于相关性计算的用户评价文本和类别文本的融合策略；3）基于 ResNet 迁移机制的物品图像描述的向量化表示；4）文本与图像的卷积融合机制，获得物品包含文本以及图像描述的综合特征表示；5）基于物品向量化表示的分类器设计，利用物品最终融合向量以及类别标签训练分类器，通过分类精度说明多源异构 UGC 数据融合向量化表示的有效性。图 2 中蓝色虚线框部分为本文主要贡献。各模块的具体实现将在第 3 节给出。

3 基于多模型和多融合策略的用户生成内容向量化表示

3.1 用户生成内容的符号表示

用户生成内容主要包括用户评价文本、物品类别文本，以及物品描述的图像信息。用户生成内容的相关符号及说明如表 1 所列。

表 1 用户生成内容符号说明

Table 1 Symbol descriptions of user-generated contents

符号	定义与说明
m	共有 m 类物品
n_k	第 k 类共有 n_k 个物品
$Y^k=\{y_1^k,y_2^k,\cdots,y_{n_k}^k\}$	第 k 类物品集合
d_j	第 k 类第 j 个物品的用户评价文本数量
$T(y_j^k)=\{T_{j1},T_{j2},\cdots,T_{jd_j}\}$	第 k 类第 j 个物品的用户评价文本集合
$D(y_j^k)=\{D_j\}$	第 k 类第 j 个物品的文本描述集合
p_j	第 k 类第 j 个物品图像描述数量
$P(y_j^k)=\{P_{j1},P_{j2},\cdots,P_{jp_j}\}$	第 k 类第 j 个物品图像描述集合
l_c	文本特征向量长度
$\boldsymbol{V}_c(T_{ji})=[v_{i1},v_{i2},\cdots,v_{il_c}]^{\mathrm{T}}$	第 j 个物品的第 i 个用户评价文本特征向量
$\boldsymbol{V}_c(T_j)=[\boldsymbol{V}_c(T_{ji})\mid i=1,2,\cdots,d_j]^{\mathrm{T}}$	第 j 个物品的所有用户评价向量化集合
$\boldsymbol{V}_d(D_j)=[v_1,v_2,\cdots,v_{l_c}]^{\mathrm{T}}$	第 k 类第 j 个物品的文本描述特征向量
l_p	图像特征向量长度
$\boldsymbol{V}_p(P_{ji})=[v_{i1},v_{i2},\cdots,v_{il_p}]^{\mathrm{T}}$	第 j 个物品的第 i 个图像特征向量
$\boldsymbol{V}_p(P_j)=[\boldsymbol{V}_p(P_{ji})\mid i=1,2,\cdots,p_j]^{\mathrm{T}}$	第 j 个物品的所有用户评价向量化集合

3.2 基于 Doc2vec 和 LDA 的文本向量化表示学习

用户生成内容中的文本评价往往相对较短，既有上下文关系，同时又侧重于对物品某些属性的主题表达。例如，对手机的评论中，代表手机特征的主题词“分辨率”“续航待机”“性价比”“拍照像素”等会在评论中反复出现，此时若应用 LDA 模型，则可在短文本评价中更好地发现用户的偏好特征，从而更精准地表达用户评价的聚焦点以及用户的潜在需求。为此，本文采用 Doc2vec 和 LDA 模型共同表示学习用户评价文本和物品描述文本的向量特征，然后将二者进行集成，以精准获得综合反映用户偏好的物品向量化表示。

设 Doc2vec 输出的向量维度为 l_c，基于训练好的 PV-DM 文本向量化表示模型，第 k 类第 j 个物品的第 i 个用户评价文本为 T_{ji}，则该物品所有的文本评价向量为 $\boldsymbol{V}_c^D(T_{ji})=[\boldsymbol{v}_{i1}^D,\boldsymbol{v}_{i2}^D,\cdots,\boldsymbol{v}_{il_c}^D]^{\mathrm{T}}$。同理，输入该物品的描述短语文本 $D(y_j^k)$，也可获得其向量表示为 $\boldsymbol{V}_d^D(D_j)=[\boldsymbol{v}_1^D,\boldsymbol{v}_2^D,\cdots,\boldsymbol{v}_{l_c}^D]^{\mathrm{T}}$。那么，对于第 k 类中第 j 个物品的所有用户评价，经过 Doc2vec 处理后的文本信息构成向量空间 $\boldsymbol{V}_c^D(T_j)=[\boldsymbol{V}_c^D(T_{ji})\mid i=1,2,\cdots,d_j]^{\mathrm{T}}$。

对于 LDA 模型，为便于将其与 Doc2vec 融合，这里设定其特征提取模块的维度也为 l_c，即选取 LDA 的主题个数为 l_c。那么对于评价文本 T_{ji}，经 LDA 主题模型学习后可获得其主题向量表示为 $\boldsymbol{V}_c^L(T_{ji})=[\boldsymbol{v}_{i1}^L,\boldsymbol{v}_{i2}^L,\cdots,\boldsymbol{v}_{il_c}^L]^{\mathrm{T}}$，描述短语对应的主题词向量为 $\boldsymbol{V}_d^L(D_j)=[\boldsymbol{v}_1^L,\boldsymbol{v}_2^L,\cdots,\boldsymbol{v}_{l_c}^L]^{\mathrm{T}}$，那么，对于第 k 类中第 j 个物品的所有用户评价的主题词向量为 $\boldsymbol{V}_c^L(T_j)=[\boldsymbol{V}_c^L(T_{ji})\mid i=1,2,\cdots,d_j]^{\mathrm{T}}$。

将基于 Doc2vec 提取的评价文本语义信息，以及 LDA 模型所得主题词分布信息进行融合，则可获得信息丰富的用户评价文本向量化特征表示。这里考虑了 3 种不同的融合方式，具体如下。

（1）拼接融合：将同一物品同一用户评价的 $\boldsymbol{V}_c^D(T_{ji})=[\boldsymbol{v}_{i1}^D,\boldsymbol{v}_{i2}^D,\cdots,\boldsymbol{v}_{il_c}^D]^{\mathrm{T}}$ 和 $\boldsymbol{V}_c^L(T_{ji})=[\boldsymbol{v}_{i1}^L,\boldsymbol{v}_{i2}^L,\cdots,\boldsymbol{v}_{il_c}^L]^{\mathrm{T}}$ 串接组成一个向量，如式（1）所示：

$$\boldsymbol{V}_c(T_{ji})=[\boldsymbol{v}_{i1}^D,\boldsymbol{v}_{i2}^D,\cdots,\boldsymbol{v}_{il_c}^D,\boldsymbol{v}_{i1}^L,\boldsymbol{v}_{i2}^L,\cdots,\boldsymbol{v}_{il_c}^L]^{\mathrm{T}} \tag{1}$$

显然，此时特征向量维度由单一模型的 l_c 变为 $2l_c$，可视为特征扩充。拼接融合简单易操作，但是可能会带来特征冗余或者噪声干扰。

（2）相加融合：将 $\boldsymbol{V}_c^D(T_{ji})=[\boldsymbol{v}_{i1}^D,\boldsymbol{v}_{i2}^D,\cdots,\boldsymbol{v}_{il_c}^D]^{\mathrm{T}}$ 和 $\boldsymbol{V}_c^L(T_{ji})=[\boldsymbol{V}_{i1}^L,\boldsymbol{v}_{i2}^L,\cdots,\boldsymbol{v}_{il_c}^L]^{\mathrm{T}}$ 对应元素相加，构成维度为 l_c 的新向量，如式（2）所示：

$$\begin{aligned}\boldsymbol{V}_c(T_{ji})&=[\boldsymbol{v}_{i1}^D,\boldsymbol{v}_{i2}^D,\cdots,\boldsymbol{v}_{il_c}^D]^{\mathrm{T}}+[\boldsymbol{v}_{i1}^L,\boldsymbol{v}_{i2}^L,\cdots,\boldsymbol{v}_{il_c}^L]^{\mathrm{T}}\\&=[\boldsymbol{v}_{ij}^D+\boldsymbol{v}_{ij}^L\mid j=1,2,\cdots,l_c]^{\mathrm{T}}\end{aligned} \tag{2}$$

相加融合不改变特征维度，直接把两类模型的输出向量进行累加，简单易操作，但是也极有可能带来噪声扰动。

（3）张量融合：Zadeh 等[15]指出当两个向量通过笛卡儿积方式进行融合时，可获得待融合向量各维度之间的交互信息，这里采用张量融合方法来实现 Doc2vec 文本向量和 LDA 主题向量的融合，如式（3）所示：

$$\begin{aligned}\boldsymbol{V}&=\boldsymbol{V}_c^D(T_{ji})\otimes\boldsymbol{V}_c^L(T_{ji})\\&=[\boldsymbol{v}_{i1}^D,\boldsymbol{v}_{i2}^D,\cdots,\boldsymbol{v}_{il_c}^D]^{\mathrm{T}}\times[\boldsymbol{v}_{i1}^L,\boldsymbol{v}_{i2}^L,\cdots,\boldsymbol{v}_{il_c}^L]\\&=\begin{bmatrix}\boldsymbol{v}_{i1}^D*\boldsymbol{v}_{i1}^L & \boldsymbol{v}_{i1}^D*\boldsymbol{v}_{i2}^L & \cdots & \boldsymbol{v}_{i1}^D*\boldsymbol{v}_{il_c}^L\\ \boldsymbol{v}_{i2}^D*\boldsymbol{v}_{i1}^L & \boldsymbol{v}_{i2}^D*\boldsymbol{v}_{i2}^L & \cdots & \boldsymbol{v}_{i2}^D*\boldsymbol{v}_{il_c}^L\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{v}_{il_c}^D*\boldsymbol{v}_{i1}^L & \boldsymbol{v}_{il_c}^D*\boldsymbol{v}_{i2}^L & \cdots & \boldsymbol{v}_{il_c}^D*\boldsymbol{v}_{il_c}^L\end{bmatrix}\end{aligned} \tag{3}$$

对其按行进行展开，则可得到融合后的特征向量 $\boldsymbol{V}_c(T_{ji})$。显然，3 种融合策略中，张量融合的维度大大增加，为 $l_c\times l_c$，并且该融合方法可涵盖更全面的特征以及特征交互间的信息。

同理，采用上述张量融合策略，可得到关于物品描述基于 Doc2vec 以及 LDA 的融合向量表示 $\boldsymbol{V}_d(D_j)=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{l_c}]^{\mathrm{T}}$。

对于物品 j，若将获得的所有用户评价文本向量$\boldsymbol{V}_c(T_{ji})$ $(i=1,2,\cdots,n_j)$和物品描述文本向量$\boldsymbol{V}_d(D_j)$直接进行融合，则评价内容和描述内容相关性较小的评价文本可能会带来噪声和扰动，并增加计算量。因此，这里采用式(4)进一步计算$\boldsymbol{V}_c(T_{ji})(i=1,2,\cdots,n_j)$和$\boldsymbol{V}_d(D_j)$的皮尔逊相关性，选择相关性较大的用户评价文本和物品描述文本进行融合。

$$\rho_{X_i,Y}=\frac{\mathrm{cov}(X_i,Y)}{\sigma_{X_i}\sigma_Y}=\frac{E[(X_i-\mu_{X_i})(Y-\mu_Y)]}{\sigma_{X_i}\sigma_Y} \tag{4}$$

其中，$X=\boldsymbol{V}_c(T_{ji})$，$i=1,2,\cdots,n_j$；$Y=\boldsymbol{V}_d(D_j)$；$\mu_{X_i}$ 和μ_Y 分别为 X_i 和 Y 的均值；σ_{X_i} 和σ_Y 分别为 X_i 和 Y 的方差。选择$\rho_{X_i,Y}\geqslant\alpha$ 的$\boldsymbol{V}_c(T_{ji})$作为被评价物品的代表评价向量，对这些向量直接相加，以保证其维度不变，然后将其与$\boldsymbol{V}_d(D_j)$进行融合，可采用拼接、相加或张量融合等策略，从而获得对第 k 类中物品 j 的文本信息的向量化表示，记为$\boldsymbol{V}_T(y_j^k)$。

3.3 基于 ResNet 模型的图像特征提取

电商平台对于商品信息的表达，除了利用文字对物品进行描述和评价外，往往还提供大量图片信息。图片附带的信息量更大，在表达情感、态度、揭示状态、描述事物方面具有更加直观的优势。多源文本数据的融合能够较为准确地表示用户对物品的文本描述，若能进一步融入图像等数据，则能进一步丰富 UGC 数据对物品的描述，使其更为准确和全面地表征用户关注物品的全局信息。

采用 ResNet 模型提取与带有用户评价的商品相关的图片特征信息。首先，使用 ImageNet 数据集预训练残差网络，然后，将该网络迁移到本文的物品图像特征提取中进行微调。将与物品 y_j^k 相关的图像集合 $P(y_j^k)=\{P_{j1},P_{j2},\cdots,P_{jp_j}\}$输入微调后的 ResNet 网络，则网络输出的向量作为各图片特征，即获得$\boldsymbol{V}_p(P_{ji})=[\boldsymbol{v}_{i1},\boldsymbol{v}_{i2},\cdots,\boldsymbol{v}_{il_p}]^{\mathrm{T}}$。对于该物品所有图像集的特征向量集合$\boldsymbol{V}_p(P_j)=[\boldsymbol{V}_p(P_{ji})\mid i=1,2,\cdots,p_j]^{\mathrm{T}}$，采取平均加权操作，获得综合所有图片信息后的向量，式(5)为该物品的图像特征表示。

$$\boldsymbol{V}_p(y_j^k)=\frac{\sum_{i=1}^{p_j}\boldsymbol{V}_p(P_{ji})}{p_j} \tag{5}$$

3.4 UGC 文本图像异构数据的融合

文本和图像具有明显的异构特性，其代表的信息既具有差异性，又具有互补性，而各自的特征提取模型完全不同，两者的特征维度也有明显的差异性。为此，设计异构特征数据融合策略至关重要。受卷积神经网络的启发，本文提出基于卷积操作[16-17]的异构数据特征融合方法。

对于物品 y_j^k 的文本综合特征 $\boldsymbol{V}_T(y_j^k)=\{\boldsymbol{v}_{Tj}(i)\mid i=1,2,\cdots,l_c\}^{\mathrm{T}}$ 和图像特征 $\boldsymbol{V}_p(y_j^k)=\{\boldsymbol{v}_{pj}(i)\mid i=1,2,\cdots,l_p\}^{\mathrm{T}}$，两类特征基于卷积的融合如式(6)所示：

$$\boldsymbol{V}(y_j^k)=(\boldsymbol{V}_T*\boldsymbol{V}_p)[n]=\sum_{\tau=1}^{l_c}\boldsymbol{V}_T(\tau)\boldsymbol{V}_p(n-\tau) \tag{6}$$

该操作不需要考虑待融合特征的维度，且融合后的特征维度远远低于张量融合的维度。与拼接融合相比，卷积融合的方法不仅能够充分考虑异构数据特征在各个维度上的交互，还能够增强数据中的重要特征，并对噪声特征进行滤波。

4 实例应用与分析

4.1 实验背景与设置

4.1.1 实验背景

亚马逊提供了含商品评论(评级、文本、投票信息)和元数据(描述、类别信息、价格、品牌和链接等)等 UGC，因此，本文选择相关数据集验证所提算法的性能。实验分别针对评价短文本＋类别文本融合表示学习，以及文本＋图像融合表示学习进行实际应用。在短文本评论特征表示实验中，选取汽车、户外运动物品、办公用品、玩具和游戏商品、Android 软件 5 类商品的评论数据，抽取出评论数据中的文本评论数据并标记上相应标签以构建均衡实验数据集，再以 3∶1 的比例将其分为训练集和测试集进行实例验证；在确定表征商品所需最少评论数以及文本图像融合表示实验中，由于数据集中的商品图像数据存在部分类别缺失以及损坏的情况，选取亚马逊数据集中类别为宠物用品、软件、办公用品、玩具和游戏的评论数据集中的 Reviewtext 字段和元数据集中的 Description 字段，再按照商品编号 asin 字段进行对应合并，然后过滤掉同一商品评论数小于 20 条的所有评论数据，并按照元数据集中所使用的文本描述爬取出相应图像数据。经整合与过滤后的数据分布如图 3 和图 4 所示。

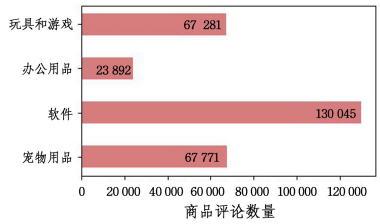


图 3 各类物品评论数据的分布

Fig. 3 Product review data distribution

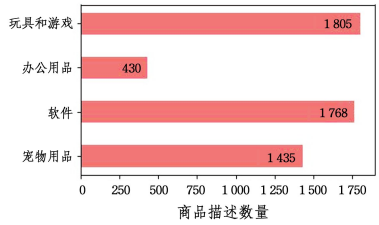


图 4 各类物品描述数据的分布

Fig. 4 Product description data distribution

4.1.2 实验设置

实验分为 4 个部分：1)针对单条短文本特征的融合与表示的精准性。通过对 LDA 主题模型和 Doc2vec 模型从不同角度提取出来的特征进行相加、拼接、张量，对比这 3 种融合方法在逻辑回归、SVM、随机森林、朴素贝叶斯 4 种分类器上的分类效果，从而验证选取 LDA 和 Doc2vec 模型融合的有效性以及确定短文本数据的一致性向量表示方法。2)准确表示商品文本特征所需文本的数量。本实验在准确表示商品短文本评论数据的基础上加入商品的文本描述数据，并采用与实验 1)同样的方法进行验证分析，从而确定充分表达商品信息所需融合商品评论的数量。3)基于长短文本相似度计算评价

文本选择的合理性。实验2）所选取的文本评论数据均为随机选取，没有考虑评价间的相关性，这里将继续验证本文提出的基于文本间皮尔逊相似度选择用户评论文本的合理性。4）含文本和图像多源异构UGC的融合表示学习的有效性。基于上述实验，通过分类准确率的提升验证本文提出的算法在文本和图像多源异构融合表示学习上的有效性。

本文实验环境为：Intel Core i7-7700HQ 处理器、主频2.8 GHz、内存8 GB、1 TB 硬盘的 PC 机。操作系统为 Windows10，编程语言为 Python，编译环境为 JetBrains PyCharm 2017。

4.2 算法参数确定

4.2.1 分类器模型

为了尽可能减少分类器对 UGC 向量化后分类精度的影响，采用多种分类器，比较其分类精度的变化，若绝大多数分类器的分类精度均较高，则说明本文提出的 UGC 向量化表示学习有效。为此，选择 SVM[18]、随机森林、逻辑回归、朴素贝叶斯4种分类器对融合与表示的商品特征进行分类。实验采用经验法结合网格搜索确定各分类器的参数：SVM 分类器的核函数为高斯核函数，惩罚参数设置为1；随机森林分类器[19-20]选用1 200棵决策树[21]，决策树支持的标准选为“Gini”（“Gini 系数”），树的最大深度为25，叶子节点最少样本数为2；逻辑回归分类器[22]中设置正则化系数为0.01，优化算法选择适合数据量大的“sag”优化算法，即随机平均梯度下降优化算法，最大迭代次数为1 000；朴素贝叶斯分类器选用对于离散数据具有较好分类效果的 BernouliNB 分类器[23]。

为了衡量分类器对向量化后物品的分类性能，采用准确率和 F1_score[24]来评估各个模型的分类效果，各指标定义可参考文献[21]，不再赘述。

4.2.2 Doc2vec 参数选择

本文利用 Gensim 库中的 Doc2vec 模型对预处理之后的商品评论进行训练，然后通过交叉验证以及网格搜索确定了分类效果最佳的 Doc2vec 的基本参数，如表2所列。

表2 Doc2vec 的基本参数取值及说明

Table 2 Parameters and description of Doc2vec

参数	参数取值	参数说明
min_count	1	丢弃词频少于1的单词
window	2	当前词与预测词的最大距离
vector_size	200	特征向量的维度
hs	0	调用负采样
negative	5	噪声词频率
alpha	0.25	初始学习率
min_alpha	0.000 25	学习率最小值
dm	1	使用 PV-DM 算法

利用 Doc2vec 提取文本特征向量，向量维度对于后续融合具有重要影响，因此，这里考虑不同维度的文本特征在各分类器上的分类精度，以确定 Doc2vec 输出的向量维度。

若 Doc2vec 模型的特征向量维度过小，则不能充分表示语料库的文本信息，而特征向量维度过大则会增加训练以及运算的时间复杂度。因此，设置特征向量维度从50维增加到300维，每隔50维记录一次分类准确率，结果如图5所示。

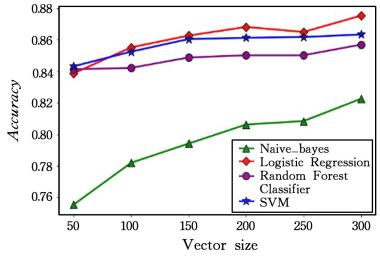


图5 不同维度句向量在各分类器上的分类表现

Fig. 5 Performance of different dimensions for sentence vectors on each classifier

由图5可知，特征向量维度在50～200维时，4种分类器的分类精度呈现比较明显的上升趋势，特征向量在200～300维时，分类精度的上升趋势相对趋于平缓。考虑到运行时间、空间复杂度和语义表达的充分程度，这里选取200维作为 Doc2vec 训练语料库中每条评论的特征向量维度。

4.2.3 LDA 参数选择

经过 LDA 主题模型的训练，可以得到每条评论相对于各个主题的权重矩阵，主题数的多少反映了主题划分粒度的粗细程度。在选取 LDA 主题数时，综合考虑 LDA 模型随主题数变化的分类精度来确定 LDA 模型的主题数。LDA 模型主题数在不同分类器上的表现如图6所示。

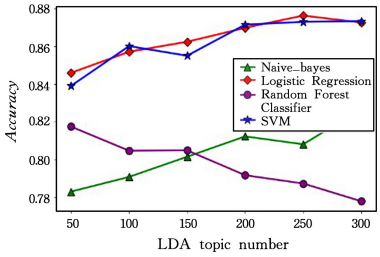


图6 不同主题数在各分类器上的分类表现

Fig. 6 Performance of different topics on each classifier

从图6可知，对于随机森林分类器，随着主题数的增加，分类精度呈现逐渐下降的趋势，但其他3种分类器均呈现出上升的趋势；对于逻辑回归分类器、SVM 分类器、朴素贝叶斯分类器，当 LDA 模型主题数为50～200时，均呈现比较明显的上升趋势，在主题数在200～300时上升的趋势趋于平缓。再次考虑 LDA 模型与 Doc2vec 特征向量的匹配度，其主题数选为200。

4.3 实验结果与分析

4.3.1 单条短文本特征融合与表示的精准性

基于 Doc2vec 和 LDA 文本特征向量化模型，对于各物品可得到其所有用户生成短文本评论的句向量和主题特征，选用3.2节所述单条短文本特征的3种融合表示方法获得单个物品多条用户评价短文本向量，即每一物品的每条短评价文本视为一个样本，利用4种分类器进行物品分类。不同融合方法在各分类器上的分类准确率和 F1_score 值分别如表3和表4所列。其中，仅利用 Doc2vec 向量的分类结果为比较

基准;3 种融合方法的提升率如表中括号内数值所示;明显优于其他融合方法的结果在表中以“粗体+*”的形式给出。

表 3 不同的融合方法在各分类器上的分类准确率

Table 3 Classification accuracy of different fusion methods on each classifier

分类算法	融合方法			
	Doc2vec 向量	拼接	相加	张量
Logistic Regression	0.869	**0.917 (+4.8%)***	0.869 (—)	0.913 (+4.4%)
Random Forest	0.849	**0.893 (+4.4%)***	0.841 (−0.8%)	0.807 (−4.2%)
SVM	0.871	**0.915 (+4.4%)***	0.872 (−0.1%)	0.905 (+3.4%)
Naïve_bayes	0.798	**0.851 (+5.3%)***	0.796 (−0.2%)	0.835 (+3.7%)

表 4 不同的融合方法在各分类器上的 *F1_score* 值

Table 4 *F1_score* value of different fusion methods on each classifier

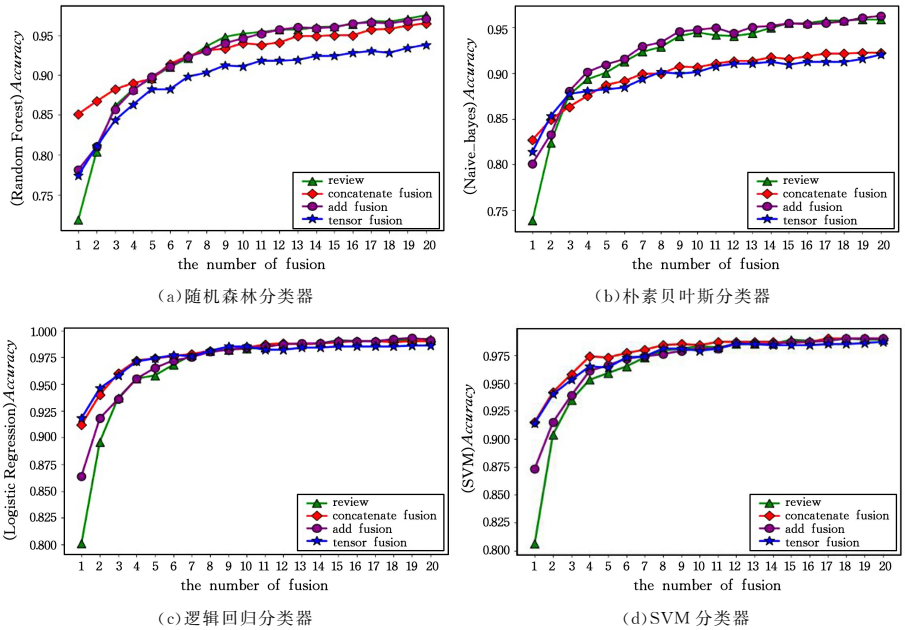
分类算法	融合方法			
	Doc2vec 向量	拼接	相加	张量
Logistic Regression	0.867	0.904(+3.7%)	0.857(−1%)	**0.91(+4.3%)**
Random Forest	0.854	**0.87(+1.6%)**	0.865(+1.1%)	0.832(−2.2%)
SVM	0.866	0.902(+3.6%)	0.851(−1.5%)	**0.907(+4.1%)**
Naïve_bayes	0.799	**0.844(+4.5%)**	0.796 (−0.3%)	0.841(+4.3%)

由表 3 和表 4 可以看出:1)相对于只使用 Doc2vec 训练出的特征直接分类,拼接融合后的分类结果在 4 种分类器上的准确率均有 4%～6%的提升;2)采用相加融合的方法融合句向量和主题特征得到的新特征,在分类结果上基本与融合前的分类结果保持一致;3)对于张量融合后的分类准确率,除随机森林算法略有下降外,其他 3 种分类算法均有 3%～5%的提升,但是张量融合方法得到的特征向量维度远远大于其他两种方法,从而造成计算复杂度和时间复杂度较大。综上,针对 LDA 主题权重特征与 Doc2vec 训练得到的句向量的融合,拼接融合的方法在短评论文本的分类中表现效果最佳。这也表明采用两种文本向量化方法可在一定程度上丰富文本特征,从而提高基于 LDA 和 Doc2vec 模型融合向量的分类精度。

4.3.2 准确表示文本特征所需文本的数量

每个物品都会有大量用户评价和类别信息,融合这些信息可从不同角度更准确地刻画物品,那么,需要选择既影响 UGC 向量化表示的准确性,又影响算法的计算复杂度的用户评价来参与商品特征的融合表示。为此,本文考虑随机选择不同数量的评价文本参与商品文本特征的融合表示对分类精度的影响,以折衷选择参与融合文本的数量。融合文本规模与分类精度的变化关系如图 7 所示。

由图 7 可知:1)随着参与融合的评论数量的增加,4 种分类器的分类精度均呈现上升趋势,且在评论数目达到 8～10 条时,各种融合方法的分类精度变化较小;2)对于逻辑回归和支持向量机分类器,实验所用的 3 种融合方法能达到精度一致的效果,而对于随机森林分类器和朴素贝叶斯分类器,不同的融合方法在分类时的表现略有偏差,但均能达到在某一阶段采用各融合方法得到的商品特征分类准确率不再发生明显变化的效果。



(a)随机森林分类器 (b)朴素贝叶斯分类器

(c)逻辑回归分类器 (d)SVM 分类器

图 7 融合文本数量对融合表示的影响

Fig. 7 Influence of number of fusion texts on fusion representation

4.3.3 基于文本相似度计算的用户评价文本选择的合理性

上述确定参与融合文本的数量时,随机选择用户评价文本,没有考虑文本间以及文本和类别标签间的相关性。3.2节给出了基于皮尔逊相关系数的评价文本选择策略,为了说明该选择的合理性,本文采用SVM对融合后的特征进行分类,实验结果如图8所示。

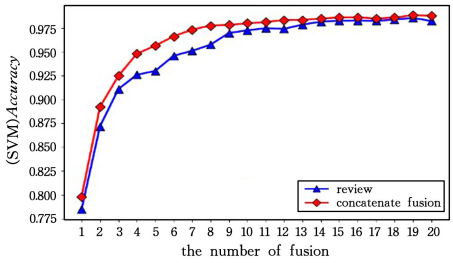


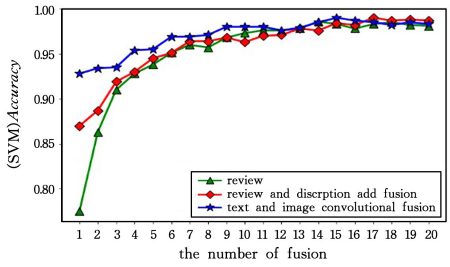
图8 基于相似度计算选择文本融合的分类精度

Fig. 8 Accuracy of texts selection with similarity

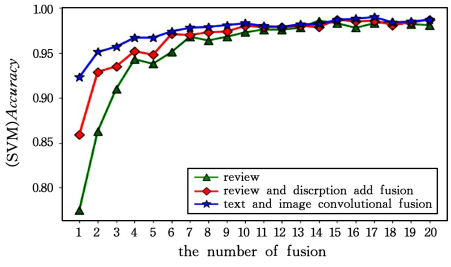
对比图7和图8可知:1)随着融合的评论数据数量的增加,根据皮尔逊相关性衡量文本描述和文本评价的相似程度进而确定所要融合的评论数据在分类精度上均比随机选择用户评论融合表示商品特征后的分类精度有不同程度的提升;2)在选取与商品描述相关程度较高的4~8条评论时,精度的提升较为明显,这也证明了根据相似性来确定融合评论数据的有效性;3)随着融合评论条数的增加,采用随机选择和基于文本相似性选择用户评价实现商品特征的融合表示均能达到提高分类精度的效果,且在评论数目大于9时,分类精度不再发生明显变化。也就是说,在电商平台中商品特征的文本融合表示这一场景中,每种商品的评论条数只需要选择9条左右即可较为准确地表达该商品的文本评价特征,这一结果与4.3.2节中的结论是一致的。通过文本间的相似性合理选择评价文本实现商品特征的融合表示能大大减少融合所需数据量,提高融合表征学习的效率。

4.3.4 含文本和图像的多源异构UGC的融合表示学习的有效性

根据亚马逊数据集中商品的文本描述链接,按照商品编号爬取出对应的商品图像数据,将爬取出的图像数据利用ResNet模型训练学习得到对应的图像特征,并在4.3.3节的基础上采用卷积融合策略加入商品的图像特征。同样以分类准确率衡量加入图像特征后能否更加准确地表示商品的全局信息,实验结果如图9所示,其中蓝色五角星为图像和文本融合的结果。从图9可知,对商品的文本特征和图像特征采用卷积的方法融合表示,无论文本特征融合时选取多少用户评论均能够有效提升商品表示的准确度。尤其是在文本评论数量很少时,精度的提升更加明显,从85%左右提升到97%左右,也就是说,对具有极少评论的商品进行表示时,加入商品的图像数据能更加有效地起到特征扩展以及补充的作用,能够有效改善商品的全局信息表示。该实验结果进一步证明了本文提出的多源异构UGC向量化融合表示学习的有效性。



(a)文本相加后和图像卷积融合效果图



(b)文本拼接后和图像卷积融合效果图

图9 文本和图像卷积融合后的分类表现(电子版为彩色)

Fig. 9 Classification performance after text and image convolution fusion

结束语 为了使电商平台中的用户生成多源异构数据,更好地服务于个性化搜索和推荐系统,本文提出一种多源异构数据融合表示的算法。首先,对于单条短文本评价,给出了一种基于Doc2vec模型和LDA模型融合表示的方法;为了准确表达商品所涵盖的文字信息,选用皮尔逊相关系数作为衡量标准,合理地选取与商品密切相关的数据参与融合;最后,考虑到文本特征的不足,在文本准确表示后采用卷积的方法将文本与商品的图像特征进行融合。本文算法在亚马逊公开数据集上的应用表现说明了该融合的有效性和表示的准确性。

本文实现了从特征级对多源异构数据的融合,实验所选用的数据集为公开大型数据集,包含丰富商品的属性以及文本内容,能够给所提算法提供充足的数据,从而实现有效的商品特征融合表示。但是在推荐系统的应用中尚且存在冷启动的问题,即当有新的物品加入时,如何对其准确地表示,仍然是后续亟待解决的另一重要问题。

参 考 文 献

[1] WANG J J,MA Y Q,CHEN S T,et al. Fragmentation knowledge processing and networked artificial intelligence[J]. Scientia Sinica Informations,2017,47(2):171-192.

[2] HUA B L,LI G J. Discussion on Theory and Application of Multi-Source Information Fusion in Big Data Environment[J]. Library and Information Service,2015,59(16):5-10.

[3] ZHU Z T J. A Multi-source Heterogeneous Vector Space Data Integration Scheme Based on GeoJSON[C]// 26th International Conference on Geoinformatics. IEEE,2018:1-4.

[4] TEZGIDER M,YLDZ B,AYDN G. Improving Word Representation by Tuning Word2Vec Parameters with Deep Learning Model[C]// International Conference on Artificial Intelligence

and Data Processing(IDPA). IEEE, 2018: 1-7.
[5] WANG X, LIAO Y, ZHU J, et al. A Low-Dimensional Representation Learning Method for Text Classification and Clustering[C]// IEEE Fifth International Conference on Data Science in Cyberspace (DSC). IEEE, 2020: 214-217.
[6] CHU Y, FENG C, GUO C. Social-Guided Representation Learning for Images via Deep Heterogeneous Hypergraph Embedding[C]// IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018: 1-6.
[7] ZHONG P, GONG Z, LI S, et al. Learning to Diversify Deep Belief Networks for Hyperspectral Image Classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(6): 3516-3530.
[8] HUA Y, GUO J, ZHAO H. Deep Belief Networks and Deep Learning[C]// International Conference on Intelligent Computing and Internet of Things (ICIT). IEEE, 2015: 1-4.
[9] KENTER T, DE RIJKE M. Short Text Similarity with Word Embeddings[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015: 1411-1420.
[10] YE J M, LUO D X, CHEN S. Short-text Sentiment Enhanced Achievement Prediction Method for Online Learners[J]. Acta Automatica Sinica, 2020, 46(9): 1927-1940.
[11] ZHANG Q, GAO Z M, LIU J Y. Research of Weibo Short Text Classification Based on Word2vec[J]. Netinfo Security, 2017(1): 57-62.
[12] ZHANG P, HE Z S. Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification[J]. Journal of Information Science. 2015, 41(4): 531-549.
[13] LAI S W, XU L H, LIU K, et al. Recurrent Convolutional Neural Networks for Text Classification[C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015: 2267-2273.
[14] CHEN Q, YAO L, YANG J. Short text classification based on LDA topic model[C]// International Conference on Audio, Language and Image Processing (ICALIP). IEEE, 2016: 749-753.
[15] ZADEH A, CHEN M, PORIA S, et al. Tensor Fusion Network for Multimodal Sentiment Analysis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1103-1114.
[16] WANG Y Y. Relationship Between Linear Convolution and Circular Convolution of Discrete Sequence[J]. Sichuan University of Arts and Science Journal, 2015, 25(5): 32-35.
[17] WANG J H, LIU X Q, LI R X. Summary of Understanding and Calculation of Discrete Linear Convolution[J]. Science & Technology Vision, 2016(27): 300-304.
[18] YANG Y, WANG J, YANG Y. Improving SVM classifier with prior knowledge in microcalcification detection1[C]// The International Conference on Image Processing (ICIP). IEEE, 2012: 2837-2840.
[19] JOELSSON S R, BENEDIKTSSON J A, SVEINSSON J R. Feature Selection for Morphological Feature Extraction using Random Forests[C]// Norwegian Signal Processing Symposium. IEEE, 2006: 10-13.
[20] NAPA K K, VIGNESWARI M, KRISHNA M V, et al. An Optimized Random Forest Classifier for Diabetes Mellitus[M]// Emerging Technologies in Data Mining and Information Security. Berlin: Springer, 2018: 765-773.
[21] PATIL S, KULKARNI U. Accuracy Prediction for Distributed Decision Tree using Machine Learning Approach[C]// Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2019: 1365-1371.
[22] RADHIKA P R, NAIR R A S, VEENA G. A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms [C]// IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). IEEE, 2019: 1-4.
[23] SINGH G, KUMAR B, GAUR L, et al. Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification [C]// International Conference on Automation, Computational and Technology Management (ICACTM). IEEE, 2019: 593-596.
[24] ZHANG D, WANG J, ZHAO X, et al. A Bayesian Hierarchical Model for Comparing Average F1 Scores[C]// International Conference on Data Mining. IEEE, 2015: 589-598.

JI Nan-xun, born in 1994, postgraduate. His main research interests include natural language processing and machine learning.

SUN Xiao-yan, born in 1978, Ph.D professor. Her main research interests include interactive evolutionary computation, big data and intelligence optimization.