文章编号:1007－6735(2021)03－0313－06　DOI: 10.13255/j.cnki.jusst.20200908002

一种基于用户关注行为的标签预测方法研究

邓春燕[1]，　郭　强[1]，　林青轩[1]，　王雅静[1]，　刘建国[2,3]
（1. 上海理工大学 管理学院，上海 200093；2. 上海财经大学 会计与财务研究院，上海 200433；
3. 新浪微热点大数据研究院 上海 201204）

摘要：为从互联网用户的关注行为中抽离出更有效的用户标签，通过挖掘用户行为的关注特性，对标签进行预测，完善用户画像系统。首先，构建用户和关注行为对象的邻接矩阵，对其进行奇异值分解，得到行为特征矩阵；然后，利用逻辑斯蒂回归模型训练特征矩阵，预测用户行业属性标签；最后，针对微博上 673 144 条用户行为数据进行实验研究。结果表明，利用用户关注行为特征矩阵预测行业标签，15 类行业标签中预测准确性的最大值可达到 0.657。研究结果缓解了用户关注行为的稀疏性，并且提升了行业标签的预测效果，改善了用户关注行为不能很好反映用户重要标签的缺陷，可为互联网用户画像贴标系统提供借鉴。

关键词：用户画像；关注行为；特征分解；标签预测
中图分类号：TP 391　　**文献标志码：**A

Label prediction based on user’s attention behaviors

DENG Chunyan[1], GUO Qiang[1], LIN Qingxuan[1], WANG Yajing[1], LIU Jianguo[2,3]
(1. *Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*; 2. *Institute of Accounting and Finance, Shanghai University of Finance and Economics, Shanghai 200433, China*;
3. *Institute of Sina WRD Big Data, Shanghai 201204, China*)

Abstract: In order to extract more effective user tags from the attention behaviors of Internet users, attention characteristics of user behaviors were explored, and tags were predicted to improve the user portrait system. By constructing the adjacency matrix of the user and the attention behavior object, singular value decomposition was then performed to obtain the behavior feature matrix, and the logistic regression model was finally used to train the feature matrix and to predict the user's industry label. Experiments with 673 144 user behavior data on Weibo were carried out. The results show that the maximum prediction accuracy of 15 industry labels can reach 0.657 by using feature matrix of the user attention behavior to predict industry labels. The innovation is to alleviate the sparseness of user attention behaviors and improve the prediction effect of industry labels. The defects of users’ important labels can not be well reflected by improving user attention behaviors. The research provides a reference for the portrait labeling system of Internet users.

Keywords: *user portrait*; *attention behavior*; *feature decomposition*; *label prediction*

收稿日期: 2020−09−08
基金项目: 国家自然科学基金资助项目（71771152，617773248）；国家社会科学基金资助项目（18ZDA088，20ZDA060）
第一作者: 邓春燕（1997−），女，硕士研究生．研究方向：数据挖掘．E-mail：dcy02130213@163.com
通信作者: 刘建国（1979−），男，教授．研究方向：媒体大数据建模与分析、知识管理．E-mail：liujg004@ustc.edu.cn

对用户行为进行建模分析，有助于精准刻画互联网用户标签，全方位掌握用户画像以实施科学准确的营销方案[1-2]。Quintana等[3]认为用户画像是一种用户形象集合，依赖用户历史行为数据的形象集合，可以用来分析用户的喜好、需求以及潜在的行为。因此，用户画像对于识别隐藏在互联网后的用户有了很大的帮助，而同时用户在互联网上产生的繁多行为数据，则给构建有效的用户画像、准确预测更优质的标签带来了困难[4]。已有Adomavicius等[5]构建了1∶1的Pro系统，利用相似性规则对用户行为进行分组，建立个人档案。Mueller等[6]针对Twitter用户的昵称文本和词语结构，统计分析用户行为特征。然而互联网平台用户兴趣不断变化，用户行为的特征、内容在时间线上皆发生改变[7]。Iglesias等[8]进一步融合用户动态行为信息，构建演化主体行为分类模型。Wu等[9]注意到用户历史记录的稀疏性，将用户社交关系和用户爱好结合寻找相似用户。这些研究都从行为相似的角度将互联网的用户进行划分，在相似的区域内给用户贴上标签，构建用户画像。但用户的历史行为除了相似的属性，用户在产生行为的同时还具有不同关注程度的偏向，通过研究用户行为的不同偏向程度，可以预测更准确的用户标签。

费鹏等[10]通过系统融合用户网络消费行为特征，把用户行为总结为特定特征，识别价格敏感型客户。王洋等[11]为了提升用户标签精度，利用Hadoop大数据平台发现用户历史行为的区别并分析规律。胡媛等[12]主要加入用户活跃度的因素，在用户画像中建立关联模型进行分析。黄文彬等[13]在这些研究基础上，着重刻画体现目标用户行为规律和移动情况的动态画像，考虑行为特征和动态变化情况。Cho等[14]通过将用户行为的上下文信息内容转换为数值，改进了推荐系统的准确性，也表明挖掘用户行为中的信息并以特征数值来表示，可以提高用户标签预测的准确性。因此，通过将行为统计为特征的维度，以特征来给用户贴标、构建用户画像也是值得研究的方向。

在将用户行为统计为特征的过程中，用户行为的偏向性、互联网社交平台中的内容多样性都给特征统计带来了困难[15]。如何最好代表用户的内心想法，挖掘出用户在不同内容上的偏向度，从而将行为转化为有效的数值，缓解用户行为的稀疏性，对于预测用户标签非常重要。本文尝试从微博用户的点赞行为记录，构建用户和关注内容对象的0-1矩阵，利用奇异值分解方法对0-1矩阵进行计算重构，得到用户行为特征矩阵，侧重用户关注行为的偏向性。然后利用逻辑斯蒂回归模型对用户行为特征矩阵进行训练，预测用户对应行业标签。从缓解用户关注行为中的稀疏性出发，利用提取的行为特征值，预测有效的用户标签，并且提高预测标签的准确值，从而为互联网用户贴标提供一定参考依据。

1 模型与方法

1.1 构建用户行为矩阵

考虑到用户在互联网平台上对内容有赞同、喜爱或者想要关注的想法，才会发生点赞行为，因此，行为具有自发性、认可性，本文将用户与点赞内容的交互关系定义为关注行为。关注行为中一个用户可对多个内容对象表示点赞，一个内容对象也可受到多个用户的点赞。假设集合$\boldsymbol{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{w}_3,\cdots,\boldsymbol{w}_i\}$表示不重复的关注行为对象，$i$表示行为对象数量，集合$\boldsymbol{U}=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\boldsymbol{u}_3,\cdots,\boldsymbol{u}_j\}$表示不重复用户，$j$表示用户数量。以行为对象集合和用户集合为用户关注矩阵的行坐标与列坐标，构建用户和关注文本内容对象的邻接0-1矩阵，若第j个用户对第i条对象有关注行为，则矩阵中$(j,i)=1$，否则为0。据此，建立用户关注行为的矩阵$\boldsymbol{C}_{j\times i}$，如式(1)所示。

$$\boldsymbol{C}_{j\times i}=\begin{array}{c} \\ \boldsymbol{u}_1\\ \boldsymbol{u}_2\\ \boldsymbol{u}_3\\ \vdots\\ \boldsymbol{u}_j\end{array}\begin{array}{c}\begin{array}{ccccc}\boldsymbol{w}_1 & \boldsymbol{w}_2 & \boldsymbol{w}_3 & \cdots & \boldsymbol{w}_i\end{array}\\ \begin{pmatrix}1&0&0&\cdots&1\\0&1&1&\cdots&0\\1&0&0&\cdots&0\\\vdots&\vdots&\vdots&&\vdots\\0&1&0&\cdots&1\end{pmatrix}\end{array} \tag{1}$$

1.2 矩阵特征分解

通过上文将用户关注行为数字化，构建0-1矩阵，关注行为已经被转化为初始数值矩阵。但由于用户的关注行为有偏重的对象，初始数值矩阵的表示方法具有不太明显的特征。因此将0-1矩阵进行分解，目的是将矩阵中的0和1转化为不同大小的特征值，通过特征值序列反映用户关注行为以及不同程度的倾向。在分解过程中，本文选取机器学习中常用的SVD(singular value decomposition)奇异值分解方法，降低0-1矩阵维度，然后保留最大特征值，去掉无效或者可利用程度较低的关注信息数值。由此，本文对用户关注行为矩阵$\boldsymbol{C}_{j\times i}$通过式(2)进行SVD分解处理。

$$\boldsymbol{C}_{j\times i}=\boldsymbol{U}_{j\times j}\boldsymbol{\Lambda}_{j\times i}\boldsymbol{V}^{\mathrm{T}}{}_{i\times i} \tag{2}$$

式中：$\boldsymbol{U}$表示左奇异矩阵；$\boldsymbol{\Lambda}$是奇异值对角矩阵；$\boldsymbol{V}$表示右奇异矩阵。

SVD 分解依靠$\boldsymbol{\Lambda}$矩阵来收缩原矩阵范围，在收缩过程中，需要对降维的数据判断有效与无效的矩阵范围。本文通过信息量阈值σ来判断降维矩阵内数值的有效范围，信息量阈值σ为不大于对角矩阵中前S个奇异值与总奇异值的比值[16]。设o_S为$\boldsymbol{\Lambda}$中前S个奇异特征值的总和，o为$\boldsymbol{\Lambda}$中所有奇异特征值总和，阈值公式如式(3)所示。

$$\sigma \leqslant \frac{o_S}{o} \tag{3}$$

为了取得最佳的实验结果，遍历S的全部取值组合，以A_{AUC}(area under curve)值的平均值为因变量。当A_{AUC}均值达到最大时，σ的取值即作为奇异特征值取数的最优取值组合。然后根据式(3)算出最优的S值，从$\boldsymbol{\Lambda}$中取出前S个奇异值，对应左右奇异矩阵中前S个奇异向量，相乘得到降维的最大关注行为特征矩阵$\boldsymbol{C}_{j\times S}$，其中特征值用$\boldsymbol{Z}$表示，$\boldsymbol{Z}$表示降维后的用户关注对象，如式(4)所示。

$$\boldsymbol{C}_{j\times S}=\boldsymbol{U}_{j\times S}\boldsymbol{\Lambda}_{S\times S}=\begin{array}{c} \\ \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \\ \vdots \\ \boldsymbol{u}_j \end{array}\begin{array}{c} \begin{array}{ccccc} \boldsymbol{Z}_1 & \boldsymbol{Z}_2 & \boldsymbol{Z}_3 & \cdots & \boldsymbol{Z}_S \end{array} \\ \begin{pmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1S} \\ z_{21} & z_{22} & z_{23} & \cdots & z_{2S} \\ z_{31} & z_{32} & z_{33} & \cdots & z_{3S} \\ \vdots & \vdots & \vdots & & \vdots \\ z_{j1} & z_{j2} & z_{j3} & \cdots & z_{jS} \end{pmatrix} \end{array} \tag{4}$$

1.3　基于特征分解的用户标签预测

通过构建用户关注矩阵和矩阵分解的过程，用户的关注行为在原始数据上维度变小，表示的特征信息得到放大。为了将用户行为矩阵的特征值转化为有效的行为信息，预测出具体的用户标签，本文选择基于逻辑斯蒂回归模型、随机森林模型、支持向量机的机器学习方法[17]，以特征值矩阵为自变量对大范围用户进行特征训练和标签预测。

1.3.1　逻辑斯蒂回归模型

逻辑斯蒂回归模型是广义线性回归模型中的一种，对因变量假设运用范围较大[18]，该模型通过把自变量样本的特征数值，折射成概率归类来预测样本结果，被广泛运用在二分类预测中，通常样本数量与变量特征数量越多模型结果越稳健[19]。逻辑斯蒂回归模型通过将对数几率函数代入基础的线性回归方程，从而将线性输出压缩在 0 和 1 之间，如式(5)所示。

$$y_j=\frac{\mathrm{e}^{\theta_{j1}z_{j1}+\theta_{j2}z_{j2}+\cdots+\theta_{jS}z_{jS}}}{1+\mathrm{e}^{\theta_{j1}z_{j1}+\theta_{j2}z_{j2}+\cdots+\theta_{jS}z_{jS}}} \tag{5}$$

式中，向量$\boldsymbol{\theta}_j=(\theta_{j1},\theta_{j2},\cdots,\theta_{jS})^{\mathrm{T}}$为解释变量$\boldsymbol{u}_j=(z_{j1},z_{j2},\cdots,z_{jS})^{\mathrm{T}}$的回归系数。

通过将用户行为特征值矩阵输入逻辑斯蒂回归模型，计算每个样本用户的概率值，以阈值 0.5 判断样本用户的分类标签属性。大于等于 0.5 为正类，即属于某一类标签；小于 0.5 为负类，即不属于某一类标签，模型流程图如图 1 所示。

1.3.2　随机森林模型

随机森林是 Breiman 利用自助重采样技术，随机生成决策树，再汇总决策树得出分类结果的集成分类方法，随机性的引入使该模型在面对数据不平衡时具有稳健性[20]。而用户关注行为中关注对象的特征存在不平衡性，因此，运用随机森林模型对用户关注特征进行预测，具有一定优势。在随机森林模型训练中，对$\boldsymbol{C}_{j\times S}$中的数据样本进行同等数量的行采样，然后对$\boldsymbol{C}_{j\times S}$中的用户样本标签进行特征列采样，构成子样本数据集并建立子决策树。重复上述过程形成随机森林，模型的预测结果则由子决策树的决策结果统计得出，少数服从多数，最终得到模型来预测用户的标签分类。

1.3.3　支持向量机模型

支持向量机是基于有监督学习对数据进行二分类的判别模型[21]，通过把特征向量映射到一个高维空间，并在空间内找到一个间隔最大的分离超平面将正负类分开，从而得到判别结果。支持向量机借助统计学学习理论而获得的强学习能力和泛化能力，被经常运用在数据深度挖掘方面[22]。在画像标签模型训练过程中，通过支持向量机求解最优判别分类函数，如式(6)所示，同时满足约束目标函数，如式(7)所示，输入同为式(4)和已知标签，从而得到预测结果。

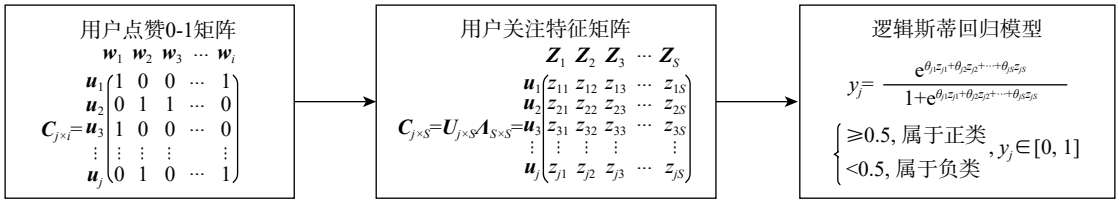

图 1　预测标签模型流程图

Fig.1　Predictive label model flowchart

$$y_j[(\boldsymbol{a}\cdot\boldsymbol{u}_j)+\boldsymbol{b}]\geqslant 1-\delta_j \tag{6}$$

$$\min\left(\left[\frac{1}{2}\|\boldsymbol{a}\|^2+B\sum_{j=1}^{S}\delta_j\right]\right) \tag{7}$$

式中：$\boldsymbol{a}$为划分数据样本超平面的法向量；$\boldsymbol{b}$为位移向量；δ_j为调节数据线性不可分的松弛变量；B为惩罚系数，控制模型训练过程中的误差范围，确保超平面的分类间隔最大。

2 用户标签预测评价标准

为了判断本文基于用户关注行为标签预测方法的准确度，选用A_{AUC}来判定准确性。A_{AUC}常被用来分析二分类结果的准确性，是指同时给定一个正类样本和负类样本，计算正类样本预测为正的可能性比负类样本预测为正的可能性大的概率，取值结果在0.5～1之间，越靠近于1说明二分类结果越好[23]。因此，本文以A_{AUC}来判断用户标签预测结果的准确性，计算公式如式(8)所示。

$$A_{\mathrm{AUC}}=\sum_{t\in(0,j)}\frac{1}{2}\left[\left(\left(\frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FN}}}\right)_t+\left(\frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FN}}}\right)_{t-1}\right)\cdot\left(\left(\frac{F_{\mathrm{FP}}}{F_{\mathrm{FP}}+T_{\mathrm{TN}}}\right)_t-\left(\frac{F_{\mathrm{FP}}}{F_{\mathrm{FP}}+T_{\mathrm{TN}}}\right)_{t-1}\right)\right] \tag{8}$$

式中：T_{TP}为某一标签用户中，经预测为该标签的用户数量；F_{FN}则为某一标签用户中，经预测为非该标签的用户数量；F_{FP}表示真实不在某一标签的用户中，经预测为该标签的用户数量；T_{TN}表示预测为非该标签的用户数量。

3 实证数据分析

3.1 数据简介

本文使用微博上用户点赞的行为数据来验证用户标签预测方法的可行性。数据集为15类行业用户的点赞行为记录，包含了多名用户对多个微博的点赞。一个用户可以对多条微博点赞，一条微博也可以受到多个用户的点赞。然后删除赞数小于20以及微博文本内容过短的点赞对象，最后统计用户数量为6242名，共673144条点赞记录。另外数据已进行脱敏，分别按照行业类别数字进行编号，如表1所示。

表1 实验数据统计量

Tab.1 Experimental data statistics

行业类别	用户数	点赞行为数
类别1	398	42171
类别2	462	49084
类别3	448	51008
类别4	349	39632
类别5	340	35487
类别6	376	29095
类别7	671	72971
类别8	497	55992
类别9	172	15324
类别10	200	21142
类别11	322	33474
类别12	488	59157
类别13	733	82084
类别14	425	47404
类别15	271	29095

3.2 特征分解结果

在矩阵特征分解中，为了保留用户关注行为的最大特征值，结合式(3)得到分解结果如表2所示。以所有行业计算得出的A_{AUC}均值为指标，可以发现A_{AUC}均值最大值为0.55，此时信息量阈值σ值为0.0046。由此根据式(3)计算出S值最佳结果为70，然后保留S为70时的最大特征信息，用最大特征信息来进行用户标签预测训练。

然后，以最佳S值为70，得到用户行为特征矩阵，如式(9)所示。用户关注矩阵里面的值从0和1变成有正有负的特征值，从而将用户关注行为矩阵降维成有区别的特征矩阵，代表用户在不同行为对象上的关注程度。

$$\boldsymbol{C}_{6\,242\times 70}=\begin{pmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\\ \boldsymbol{u}_3\\ \vdots\\ \boldsymbol{u}_{6\,242}\end{pmatrix}=\begin{pmatrix}-0.172\,3 & 0.058\,3 & -0.014\,2 & \cdots & 0.008\,5\\ -0.250\,6 & 0.095\,6 & -0.044\,7 & \cdots & 0.006\,3\\ 0.056\,7 & -0.019\,6 & 0.008\,2 & \cdots & 0.000\,1\\ \vdots & \vdots & \vdots & & \vdots\\ 0.014\,7 & -0.027\,0 & 0.002\,1 & \cdots & 0.000\,0\end{pmatrix} \tag{9}$$

3.3 标签预测结果

得到用户行为矩阵后，以矩阵和已知用户行业标签为主键训练模型，再通过划分训练集和测试集，确保模型测试结果的准确性[24]。本文采用交叉验证的方式切分数据，数据集平均切分为M份，每份数据集在训练循环中有一次被作为测试

表 2　信息量阈值σ实验结果

Tab.2　Information threshold σ experiment results

σ	S	A_{AUC}均值
0.0046	70	0.549
0.0559	100	0.544
0.0840	200	0.540
0.1120	300	0.543
0.1380	400	0.543

集，利用无重复抽样的优势对模型进行评估，计算训练循环的准确性平均值。另外和其他标签预测方法的准确性进行对比，对比方法包括基于贝叶斯网络的 K2 算法[25] 和 BP 网络模型融合算法[26]。基于贝叶斯网络的 K2 算法是通过用户浏览行为构建先验概率，然后给用户行为后验概率设定阈值判断用户标签的方法，着重于用户行为的因果分析，但忽略了用户行为的多样性和稀疏性。BP 网络模型融合算法使用 TF-IDF 算法和 Doc2Vec 模型处理用户的搜索文本内容，利用处理好的文本特征来训练 BP 神经网络模型，输出预测结果，但没有对输入模型的特征重点进行凸显和筛选。因此，本文选用这两个方法作为预测方法有效性的对比。本文实验中数据集切分 M 值取值分别为 3，5，10，A_{AUC} 均值结果如表 3 所示。

表 3　不同训练集和测试集比例下 A_{AUC} 均值

Tab.3　Mean A_{AUC} at different training and test set ratios

M	SVD+逻辑斯蒂回归模型	SVD+随机森林	SVD+支持向量机	基于贝叶斯网络的K2算法	BP网络模型融合算法
3	0.552	0.541	0.547	0.518	0.526
5	0.548	0.540	0.549	0.521	0.518
10	0.549	0.510	0.548	0.519	0.521

由表 3 可发现，针对用户标签预测，同种方法的 A_{AUC} 均值在 3 类训练集与测试集比例下，都在 M=3 时达到最大值。再以 M=3 为最佳标准，对比不同方法的标签预测准确性，可以发现 SVD 和逻辑斯蒂回归模型预测准确性最大，A_{AUC} 均值为 0.552，比随机森林和支持向量机的高 0.011，0.005；和其他标签预测方法相比，SVD 和逻辑斯蒂回归模型方法比基于贝叶斯网络的 K2 算法预测和 BP 网络模型融合方法的准确性高 0.034，0.026。因此，综合表 3 的全部实验结果，M=3 时的 SVD 和逻辑斯蒂回归模型方法的预测准确性最大，效果最好。进一步将 A_{AUC} 均值展开观察 15 类行业的预测准确性，各行业 A_{AUC} 值如图 2 所示。

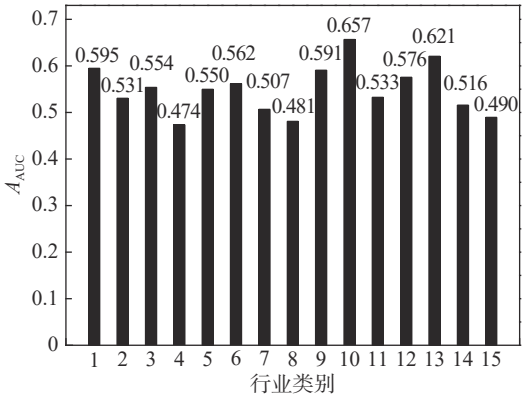

图 2　各行业预测准确值

Fig.2　Prediction accuracy for various industries

从图 2 中可以发现，在矩阵特征分解选择维度 S=70 和模型训练预测比例 M=3 时，用户行业标签预测准确值 A_{AUC} 前 3 最高分别为 0.657，0.621，0.595，分别是类别 10、类别 13 和类别 1。当预测准确率为 0.657 时，说明从所有用户的关注行为记录中，预测出原本属于类别 10 的用户有 65.7% 的概率，类别 13 和类别 1 的用户标签预测则有 62.1% 和 59.5% 的概率。其他行业类别的标签预测准确性结果显示在 50%～59% 之间，类别 4 和类别 15 的预测准确率分别为 47.4% 和 49%，在所有用户行业标签预测中准确率不是很高。对不同行业类别预测准确值的差异进行分析发现，表征性越明显的行业类别用户，预测准确度越高。比如行业类别 10，该类别用户点赞关注的内容大多为司法相关、政策相关内容，且内容集中度很高，行业类别 10 的预测准确率也最高，为 65.7%；行业类别 13 的点赞关注内容则多为三农类、时事政治类新闻，内容偏向性明显，但混杂其他主题的微博也较明显，内容偏向比起行业类别 10 分类更广；行业类别 1 的点赞内容集中在传媒类上，但是混杂的主题微博内容则多于行业类别 10 和类别 13；剩余其他行业类别的点赞关注内容更为分散，导致用户关注内容的集中度不同，形成特征值矩阵和模型训练预测上有不同的效果，集中性越强，用户行业标签的预测准确度越高。

4　结　论

根据互联网用户自发性和偏向性的关注行为，提出了以用户关注行为矩阵为基础，然后进行特征分解和逻辑斯蒂回归模型训练，最后进行用户标签预测的方法。通过将用户关注行为的点赞记录，处理为关注行为 0-1 矩阵，然后进行矩阵

特征分解，筛选用户点赞的特征数值，转化为用户行为特征矩阵，再利用逻辑斯蒂回归模型预测用户行业标签。经过微博上6 242名用户的673 144条点赞行为数据实验表明，用户行业标签预测准确性最大值可达0.657，在所有行业标签预测上准确性均值也高于K2算法和BP网络模型的方法。本文创新点在于缓解了用户关注行为的稀疏性，从中获取用户关注行为的特征值，凸显用户行为来预测用户的行业标签。如今互联网用户的行为数据繁多，本文可为用户贴标、用户管理提供一定借鉴。同时对于用户关注行为还可以进一步分析，比如关注行为聚类、主题偏向对比、模型调节参数优化等角度，从而提高用户标签预测效果，构建更准确的用户画像标签系统，这些工作将在以后的研究中再分析。

参考文献：

[1] GU H Q, WANG J, WANG Z W, et al. Modeling of user portrait through social media[C]//Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME). San Diego, CA, USA: IEEE, 2018: 1–6.

[2] HAN Z H, CHEN X S, ZENG X M, et al. Detecting proxy user based on communication behavior portrait[J]. The Computer Journal, 2019, 62(12): 1777–1792.

[3] QUINTANA R M, HALEY S R, LEVICK A, et al. The persona party: using personas to design for learning at scale[C]//Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. Denver, Colorado, USA: ACM, 2017: 933–941.

[4] 刘海鸥, 孙晶晶, 苏妍嫄, 等. 国内外用户画像研究综述 [J]. 情报理论与实践, 2018, 41(11): 155–160.

[5] ADOMAVICIUS G, TUZHILIN A. Using data mining methods to build customer profiles[J]. Computer, 2001, 34(2): 74–82.

[6] MUELLER J, STUMME G. Gender inference using statistical name characteristics in twitter[C]//Proceedings of the 3rd Multidisciplinary International Social Networks Conference on Social Informatics 2016. Union, NJ, USA: ACM, 2016.

[7] LÜ L Y, MEDO M, YEUNG C H, et al. Recommender systems[J]. Physics Reports, 2012, 519(1): 1–49.

[8] IGLESIAS J A, ANGELOV P, LEDEZMA A, et al. Creating evolving user behavior profiles automatically[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(5): 854–867.

[9] WU L, GE Y, LIU Q, et al. Modeling users' preferences and social links in Social Networking Services: a joint-evolving perspective[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix: AAAI Press, 2016: 279–286.

[10] 费鹏, 林鸿飞, 杨亮, 等. 一种用于构建用户画像的多视角融合框架 [J]. 计算机科学, 2018, 45(1): 179–182, 204.

[11] 王洋, 丁志刚, 郑树泉, 等. 一种用户画像系统的设计与实现 [J]. 计算机应用与软件, 2018, 35(3): 8–14.

[12] 胡媛, 毛宁. 基于用户画像的数字图书馆知识社区用户模型构建 [J]. 图书馆理论与实践, 2017(4): 82–85, 97.

[13] 黄文彬, 徐山川, 吴家辉, 等. 移动用户画像构建研究 [J]. 现代情报, 2016, 36(10): 54–61.

[14] CHO S, LEE M, JANG C, et al. Multidimensional filtering approach based on contextual information[C]//2006 International Conference on Hybrid Information Technology. Cheju, Korea (South): IEEE, 2006: 497–504.

[15] SARWAR B M, KARYPIS G, KONSTAN J A, et al. Application of dimensionality reduction in recommender system—a case study[C]//Proceedings of WebKDD 2000 Web Mining for E-Commerce Workshop. New York: ACM, 2000.

[16] 郭强, 岳强, 李仁德, 等. 基于四阶奇异值分解的推荐算法研究 [J]. 电子科技大学学报, 2019, 48(4): 586–594.

[17] 黄志超, 乔振华. 基于机器学习模型的手写数字识别 [J]. 电脑知识与技术, 2019, 15(33): 215–217.

[18] 邢秋菊, 赵纯勇, 高克昌, 等. 基于GIS的滑坡危险性逻辑回归评价研究 [J]. 地理与地理信息科学, 2004, 20(3): 49–51.

[19] VAN SMEDEN M, MOONS K G M, DE GROOT J A H, et al. Sample size for binary logistic prediction models: beyond events per variable criteria[J]. Statistical Methods in Medical Research, 2019, 28(8): 2455–2474.

[20] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5–32.

[21] BREIMAN L. Statistical modeling: the two cultures (with comments and a rejoinder by the author)[J]. Statistical Science, 2001, 16(3): 199–231.

[22] HSU C W, LIN C J. A comparison of methods for multiclass support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415–425.

[23] KOSINSKI M, STILLWELL D, GRAEPEL T. Private traits and attributes are predictable from digital records of human behavior[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(15): 5802–5805.

[24] 项亮. 推荐系统实践 [M]. 北京: 人民邮电出版社, 2012.

[25] 赵会群, 李子木, 郭峰, 等. 基于K2算法的精准营销研究 [J]. 计算机应用与软件, 2019, 36(7): 43–49.

[26] 郭梁, 王佳斌, 马迎杰, 等. 基于模型融合的搜索引擎用户画像技术 [J]. 科技与创新, 2020(7): 17–19.

（编辑：丁红艺）