文章编号：1003-0077(2017)05-0185-09

基于迭代回归树模型的跨平台长尾商品购买行为预测

白 婷[1,2]，文继荣[1,2]，赵 鑫[1,2]，杨伯华[1,2]

(1. 中国人民大学 信息学院，北京 100872；
2. 大数据管理与分析方法研究北京市重点实验室，北京 100872)

摘 要：长尾商品是指单种商品销量较低，但是由于种类繁多，形成的累计销售总量较大，能够增加企业盈利空间的商品。在电子商务网站中，用户信息量较少且购买长尾商品数量较少、数据稀疏，因此对用户购买长尾商品的行为预测具有一定的挑战性。该文提出预测用户购买长尾商品的比例，研究单一用户购买长尾商品的整体偏好程度。利用社交媒体网站上海量的文本信息和丰富的用户个人信息，提取用户的个人属性、文本语义、关注关系、活跃时间等多个种类的特征；采用改进的迭代回归树模型 MART(Multiple Additive Regression Tree)，对用户购买长尾商品的行为进行预测分析；分别选取京东商城和新浪微博作为电子商务网站和社交媒体网站，使用真实数据构建回归预测实验，得到了一些有意义的发现。该文从社交媒体网站抽取用户特征，对于预测用户购买长尾商品的行为给出一个新颖的思路，可以更好地理解用户个性化需求，挖掘长尾市场潜在的经济价值，改进电子商务网站的服务。

关键词：长尾商品；电子商务；社交媒体；购买行为预测

中图分类号：TP391　　**文献标识码**：A

Connecting Social Media to E-Commerce: Predicting Long-tail Purchase Behaviors using Multiple Additive Regression Tree

BAI Ting[1,2], WEN Jirong[1,2], ZHAO Xin[1,2], YANG Bohua[1,2]

(1. School of Information, Renmin University of China, Beijing 100872, China;
2. Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing 100872, China)

Abstract: Long-tail products, with low demands, occupy a significant share of total revenue in total. It is challenging to analyze the long-tail purchase behaviors due to the data sparsity resulted from few purchase behaviors. This paper proposes to leverage online social media information for predicting the long-tail purchase behaviors. In specific, we collect the user profiles form the social media information, including the status text, following links and temporal activity distributions, and predict their purchases by a weighted Multiple Additive Regression Trees (MART). Experimented on the data from JingDong and SinaWeibo, the effectiveness of the proposed method are revealed, together with several interesting findings.

Key words: long-tail products; e-commerce shopping; social media; purchase prediction

1 引言

长尾商品是指单种商品销量较低，但由于种类繁多，形成的累计销售总量较大，能够增加企业盈利空间的商品[1]。随着信息科技的发展，人们能够较容易地在电子商务网站中找到实体市场中因为冷门而几乎没有消费者的长尾产品。如图 1① 所示，人们比较关注曲线主体的那些热门商品，而将处于曲线

① 维基百科 http://zh.wikipedia.org/wiki/%E9%95%BF%E5%B0%BE

收稿日期：2015-04-15　**定稿日期**：2016-03-08

基金项目：国家自然科学基金青年科学基金(61502502)；国家重点基础研究发展计划(2014CB340403)；北京市自然科学基金(4162032)；中国人民大学 2016 年度拔尖创新人才培育资助计划

尾部的商品忽略，但被忽略的较长的尾部商品累计产生的总体效益甚至可以与主体热销商品抗衡。首先，长尾商品的销量可观，例如，有学者研究过亚马逊网站的书本销售量和销售排名的关系，发现亚马逊40%的图书销量来自于本地书店里不卖的图书[2]；Deniz Oktar① 也指出，商家盈利的增加在于对长尾市场的开发，他认为热门商品因为很多商家竞价出售而导致商品的利润降低，而长尾商品若能找到对其偏好的消费者，商家获利的空间将会很大；Anderson也提出通过让商品种类全面，并帮助用户找到它，可以推动长尾市场的繁荣[3]。

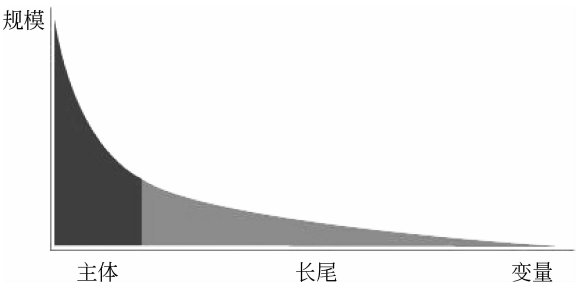

图1 长尾理论中商品销量图

对用户购买长尾商品的行为进行预测，就是探究哪些用户更倾向于购买长尾商品，分析用户购买长尾商品时的偏好、购买习惯等特点。长尾商品由于购买量少导致数据稀疏，传统的基于内容推荐和协同过滤、关联规则、聚类等方法适用性较差，所以对用户购买长尾商品行为的预测具有一定的挑战性。本文提出一种基于社交媒体信息对用户购买长尾商品行为做预测的方法，探究如何利用社交媒体上海量的文本信息和丰富的用户信息，对用户购买长尾商品的行为做预测，以更好地理解用户的个性化需求，从而挖掘长尾商品的潜在经济价值。

本文主要有三点贡献：①针对长尾商品的购买行为，形式化地给出了研究问题的定义，提出利用社交媒体上海量的文本信息和丰富的用户信息，对用户购买长尾商品的比例做预测；②针对数据样本分布的偏置性问题，改进MART模型，显著地提高了模型的预测效果；③在真实的数据集（新浪微博、京东商城）上构建大量的实验，与LR（linear regression）模型，SVR（support vector regression）模型，CART（classification and regression Tree）模型，神经网络多层感知机模型MLP（multilayer perceptron）对比，验证了预测的效果，并详细分析用户特征对其购买长尾商品比例的影响。

2 相关工作

目前，对用户购买行为的研究大多基于用户的购买记录，为用户推荐可能购买的商品，通常采用基于内容推荐、协同过滤推荐、关联规则、聚类等方法。基于内容的推荐[4]是根据用户过去喜欢的物品内容，为用户推荐相似的物品，长尾商品由于购买量少，基于内容推荐的算法并不适用；协同过滤算法是利用用户喜好之间的相似性进行推荐[5]，不依赖于商品的实际内容，但需要用户对商品的喜好信息，在长尾商品的购买中，用户喜好差别很大，所以也不适用。长尾商品由于购买量少，数据稀疏，关联规则、聚类等方法也都适用性较差，这使得对长尾商品的研究具有一定的挑战性。目前针对长尾商品推荐的研究较少，且都是基于用户购买记录本身，如文献[6]中提出了一种基于用户购买记录的图模型长尾商品推荐算法，文献[7]中是基于长尾商品在所有商品购买图中的位置进行分析。

基于购物网站上信息，对长尾商品的购买行为分析存在以下不足：第一，电子商务网站用户注册信息一般比较简单，如京东商城，用户只需填写用户名和密码，进行邮箱或手机号的验证，就可以进行购物；第二，每个用户长尾商品的购买记录少，数据稀疏。购物网站上简单的用户信息，较少的长尾商品购买记录，是研究长尾商品购买行为的挑战所在，而在社交媒体上，虽然无法得知用户的购买记录，但有丰富的用户信息，如年龄、性别、职业及海量文本信息，将这些丰富的信息用于对用户购买长尾商品的预测，是长尾商品购买行为研究的一个新思路。文献[8]初步验证了社交媒体网站中用户人口统计学特征、喜好，与用户在电子商务网站中购买商品类别有一定的联系，本文针对长尾商品，进一步挖掘社交媒体网站中用户的信息，对用户购买行为进行预测，并构建实验，给出验证。

3 问题描述及定义

在传统经济里，“二八定律”认为公司80%的利润来自20%的畅销产品，其余20%的利润则来自于80%的普通产品[1]，因货架空间的局限和成本问题，

① D. Oktar. Recommendation Systems: Increasing Profit by-Long Tail. http://en.webrazzi.com/2009/09/18/

那些由于销量低而周转速度不足以抵消货架单位面积成本的长尾商品，将会被排斥在市场之外。随着电子商务网站的兴起，用户只需简单搜索，即可看到大量可选的商品，这使得种类丰富的长尾商品有较大机会面向庞大的目标消费群体。本文研究用户对长尾商品的整体偏好程度，利用用户特征预测其购买长尾商品的比例，定义如下。

长尾商品　根据“二八定律”，将长尾商品定义为销量排名大于 20% 的商品，定义如下：给定商品集 P，商品总数为 N，对商品按照销量进行倒序排序 $P=\{p_1,p_2,\cdots,p_N\}$，使得 $\forall p_i\in P, S_i\geqslant S_{i+1}$，其中 S_i 是商品 p_i 的销量。长尾商品集 P_{LT}(Long Tail Product)可以定义为式(1)。

$$P_{\mathrm{LT}}=\{p_i \mid i\geqslant N\times 20\%, p_i\in P\} \tag{1}$$

长尾商品购买比例　给定用户集 U，对于 $\forall u\in U$，用户 u 购买的商品集为 P_u，则该用户购买的长尾商品比例 y_u 可以定义为式(2)。

$$y_u=\frac{|P_{\mathrm{LT}}\cap P_u|}{|P_u|} \tag{2}$$

问题定义　用户长尾商品的购买行为预测的问题可以定义为：将用户特征刻画为长度为 n 的特征向量 $x_u=\{x_1,x_2,\cdots,x_n,\}$，学习映射函数 $F: R^n\rightarrow R$，将用户 u 的 n 维特征向量映射到一维空间，即用户购买长尾商品的比例。由用户特征预测该用户的长尾商品购买比例，预测值 $\hat{y}_u$ 由式(3)给出。

$$\hat{y}_u=F(\boldsymbol{x}_u) \tag{3}$$

$\hat{y}_u$ 表示预测得到的用户购买长尾商品的比例，比例越高，表示该用户越愿意购买长尾商品。

4　模型描述

社交媒体中含有丰富的用户信息，如年龄、性别、喜好、文本信息等，这些信息很难从电子购物网站得到，因此，本文从社交媒体中提取用户 u 的特征向量 $\boldsymbol{x}_u=\{x_1,x_2,\cdots,x_n\}$，并从电子商务网站中得到用户 μ 实际购买长尾商品的比例 y_u，构造训练数据集 $\{\boldsymbol{x}_u,y_u\}_{u\in U}$，则问题转化为输入为用户特征向量 $\boldsymbol{x}_u$，期望输出为用户实际购买长尾商品比例 y_u 的预测问题。机器学习中有很多模型可以解决此类问题，如线性回归、支持向量机、决策树等[9]，迭代回归树模型 MART(multiple additive regression tree)是由多个回归树加权合并成的回归树模型，在很多任务中都有不错的效果，例如在解决互联网搜索排序(Web search ranking)[10]、推荐和预测系统[11]中，都有较好的预测效果和较低的错误率。在本文中，用户特征种类较多，特征的不同组合会导致不同的预测结果，与线性回归、支持向量机等方法相比较，MART 模型是由多个简单的决策树组合而成的模型，能够充分利用用户特征信息，有效学习特征表示[12-14]，故本文中采用 MART 模型，并通过引入样本权重的方法对 MART 模型进行改进，使得改进后的模型预测效果有了显著的提升。

4.1　MART 简介

MART(multiple additive regression tree)又叫做 GBDT(gradient boosting decision tree)，是采用梯度迭代算法实现的回归树。

MART 模型的输入为 n 维特征向量 $\boldsymbol{x}$，由映射函数 $F: R^n\rightarrow R$ 将其映射到预测值。在第 m 次迭代中，有：

$$F_m(\boldsymbol{x})=F_{m-1}(\boldsymbol{x})+\rho_m h_m(\boldsymbol{x};\boldsymbol{a}) \tag{4}$$

其中 $h_m(\boldsymbol{x};\boldsymbol{a})$ 表示以 $\boldsymbol{a}$ 为参数的 $\boldsymbol{x}$ 的函数，$\rho_m\in R$ 表示第 m 个函数的权重。

令 $\{(\boldsymbol{x}_i,y_i)\}_1^N$ 表示包含 $|U|$ 个实例的训练数据集，其中 $\boldsymbol{x}_i$ 表示输入的特征向量，y_i 表示期望输出值。梯度迭代算法的学习过程中，每一次迭代包含两个主要步骤，计算参数 $\boldsymbol{a}_m$ 和 ρ_m 如式(5)～(6)所示。

$$\boldsymbol{a}_m=\arg\min_{\boldsymbol{a},\beta}\sum_{i=1}^{|U|}[-g_m(x_i)-\beta h_m(\boldsymbol{x}_i;\boldsymbol{a})]^2 \tag{5}$$

$$\rho_m=\arg\min_{\rho}\sum_{i=1}^{|U|}L(y_i,F_{m-1}(\boldsymbol{x}_i)+\rho h_m(\boldsymbol{x}_i;\boldsymbol{a}_m)) \tag{6}$$

其中 $g_m(\boldsymbol{x})$ 表示函数 $F_{m-1}(\boldsymbol{x})$ 梯度下降方向，计算公式如式(7)所示。

$$g_m(\boldsymbol{x}_i)=\left[\frac{\delta L(y_i,F(\boldsymbol{x}_i))}{\delta F(\boldsymbol{x}_i)}\right]_{F(\boldsymbol{x})=F_{m-1}(\boldsymbol{x})} \tag{7}$$

4.2　MART 的改进

MART 模型中假设所有实例(用户)同等重要，本数据集中用户购买长尾商品的比例非常不均匀，购买长尾商品比例较低的用户占绝大多数，为了更好地学习用户的特征与其购买长尾商品的关系，本文对 MART 模型进行改进，对购买长尾商品比例大的用户着重学习，即根据用户购买长尾商品的比例对用户进行加权。

定义如下损失函数：

$$\min\sum_{i=1}^{|U|}L(y_i,\hat{y}_i)=\min\sum_{i=1}^{|U|}w_i(y_i-\hat{y}_i)^2 \tag{8}$$

其中，$\hat{y}_i$ 表示模型对于第 i 个实例的预测值，w_i 表示第 i 个实例的权重，在模型的训练中，根据用户购买长尾商品的比例，确定该用户对于模型训练的重要性，取值范围区间为[0,1]，用户购买长尾商品的比例越大，其对应的实例的权重也就越大，其对于模型的训练越重要。根据文献[15]中定义权重的思想，w_i 定义为：

$$w_i = \frac{\log(1+y_i) - \log(1+\min(y_i))}{\log(1+\max(y_i)) - \log(1+\min(y_i))} \quad (9)$$

模型的权重 w_i 由训练数据确定，在训练完成得到 MART 模型参数 $\boldsymbol{a}_m$ 和 ρ_m 后，实际预测过程按照式(4)计算，并不需要得待预测样本的权重。

下一节介绍如何从社交媒体中提取用户的特征向量 x。

5 特征选择

本文利用社交媒体中海量的文本信息和丰富的用户信息，如年龄、性别、职业及大量文本信息，对用户购买行为进行预测，构建用户社交媒体中特征向量，分析用户特征对购买行为的影响。

5.1 购买行为分析

商品的价格、目标用户的类别(如男士用品、女士用品)、适用的年龄段(如幼儿产品、老年产品)、功能类别(如日用品、专业领域用品)等因素都会影响到商品的销量。因此，用户的年龄、性别、婚姻状况、教育背景、职业等个人属性，关注的话题、兴趣爱好等特征都是影响其购买行为的因素[16]。

5.2 特征向量构建

如何在社交媒体中提取与购买长尾商品有关的特征，是特征提取面临的一个挑战。在众多的社交媒体中，本文选择涵盖娱乐、体育、生活等多方面，具有庞大用户群体的新浪微博作为提取用户特征的数据来源，通过对用户的社交习惯和购买行为的分析，在新浪微博中提取用户的四大类 12 种特征，用户微博特征见表 1。

用户微博特征详细说明如下：

(1) 个人属性特征

年龄：1～11，12～17，18～30，31～45，46～59，60+；

性别：男，女；

婚姻状况：单身、订婚、暗恋、结婚、追求、丧偶、分居、离婚、热恋、暧昧；

表 1 用户微博特征表

特征类型	特征	维度
个人属性特征	年龄	6
	性别	2
	婚姻状况	10
	教育背景	6
	职业	8
	兴趣爱好	5
文本特征	话题分布	50
关系特征	群组	50
	权威性	1
	互动率	2
微博活跃时间特征	每天分布	24
	每周分布	7

教育背景：自然科学、工程、社会科学、医学、艺术、其他；

职业：互联网、设计、服务业、生产业、医药业、科学工作、管理者、其他；

兴趣爱好：由微博用户标签得到，包括旅游、摄影、音乐和电影、电脑游戏、其他。

(2) 文本特征

话题分布：采用主题模型(topic model)可以从用户所发的博文中，获取该用户的主题分布。采用隐含狄利克雷分布(LDA)，将每个用户的博文聚合成一个文档。提取用户的原创、转发、评论的文本信息，得到每个微博用户的主题分布[17]。

(3) 关系特征

群组：在微博中，有相似关注关系的用户具有相似兴趣爱好的可能性较大，可以根据用户的关注关系，将用户分为群组。与文献[17]中思路相似，采用 LDA 模型，将被关注的用户当做单词，关注者当做文档，发现被关注者的潜在群组，得到每个微博用户的关注偏好分布。

权威性：用户权威性即用户在微博关系图中的 PageRank 值，可以定义为：微博中的用户关注关系用图 $G_U(V,E)$来表示，图中的每一个顶点 $v \in V$ 代表微博中的每一个用户，图中的边 E 则为 $V \times V$ 的子集，代表两个顶点之间的关系。对于顶点 v_i：

$$\mathrm{PageRank}(v_i) = \mu \sum_{v_j \in M(v_i)} \frac{\mathrm{PageRank}(v_j)}{L(v_j)} + \frac{1-\mu}{|V|} \quad (10)$$

其中 $M(v_i)$是指向 v_i 的所有节点，$L(v_j)$是 v_j 链出的节点数量，$|V|$是节点总数。

互动率：由用户@他人的次数和用户参与话题讨论 Hashtag 的使用率构成。

(4) 微博活跃时间特征

每天分布：每天用户活跃在微博上的时间分布；

每周分布：每周用户活跃在微博上的时间分布。

6　实验设置及结果分析

6.1　数据准备

本文分别选取京东商城和新浪微博作为电子商务网站和社交媒体网站，利用京东商城的用户购买记录和新浪微博用户信息构建实验。

电子商务网站数据　从国内最大的 B2C 电子商务网站京东商城爬取商品的评论信息，获得 1 200 万用户对 17.5 万商品的 1.389 亿条商品评论。在京东上只有购买商品的用户才可以对该商品做出评论，每一条评论均可以得到一个用户的 ID(基于用户隐私考虑，ID 均为加密处理)，根据用户的 ID 对商品进行分组，得到每个用户所购买的商品的列表。

社交网络数据　从国内最大的社交媒体网站新浪微博获取用户的社交信息，提取从 2013 年 1 月 1 日到 2013 年 6 月 30 日的微博数据，去除非正常微博用户，例如，极度不活跃用户；粉丝数低于五个、微博数低于五条的用户；活跃度很高但互动率很少的用户；如一天内同一条微博发布五次或自转发五次以上、所发微博中半数以上的微博他人转发以及评论人数少于五人的用户。最后从 500 万正常活跃的微博用户中提取 17 亿条博文信息。

京东-微博用户关联　用户在京东购买商品时，有时会采用第三方账号登录，如果采用新浪微博账号登录，就可获得一个用户的新浪微博 ID 和京东 ID；此外，微博用户有时会将在购物网站上购买的商品链接分享到微博上，根据其分享的链接，我们就可以将该用户的京东 ID 和新浪微博 ID 相关联，本文从 17 亿条博文信息中抽取京东商品分享信息，从 500 万微博用户中找出 23 917 个同时具有京东购物记录和新浪微博信息的用户，去除微博和购买记录中异常的噪声数据，如微博中博文数量极少或极多的数据，购买记录中少于 10 条的用户，最终我们得到有长尾商品购买记录的 15 853 个关联用户。

长尾商品集的构建　考虑到不同种类的商品销售量会有很大的差距，比如电子产品和日常生活用品，若将商品整体按销量排序提取长尾商品，可能会导致销量少的某一种类如电子产品，都会被划分到长尾商品中。所以本实验先将商品按照京东购物网站的 16 个大类目进行分类，在每个类别中按照公式(1)去除销量最高的前 20%的热门商品，以及销售量极少的噪声数据，重新构建得到长尾商品集。找出购买这些长尾商品的用户 ID 中属于关联用户的 ID。长尾商品的销量区间分布如图 2 所示，用户购买长尾商品比例人数统计结果如图 3 所示，最终得到统计结果如表 2 所示。

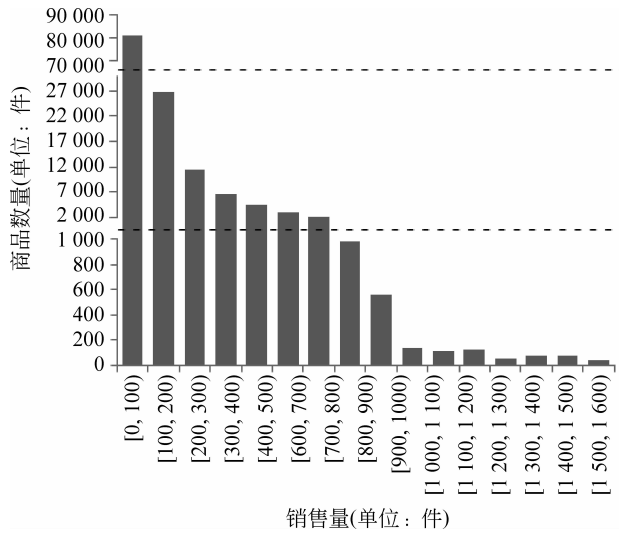

图 2　长尾商品销量区间分布

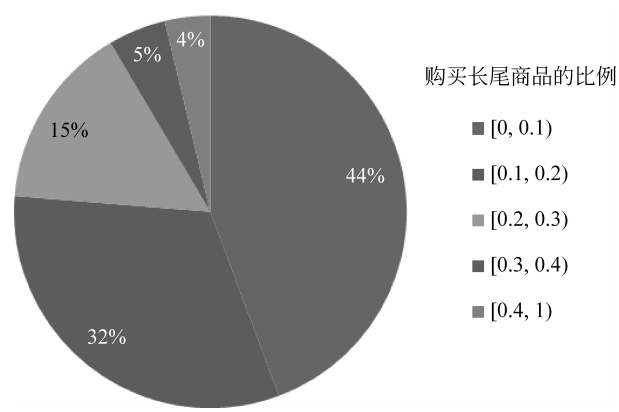

图 3　购买长尾商品用户所占百分比

表 2　关联用户实验数据集统计结果

关联用户数/人	长尾商品总数/件	商品平均购买量/件	长尾商品平均购买量/件	平均博文数量/条
15 853	138 015	52	8.2	41

6.2 评价标准

采用改进后的MART模型，通过十折交叉验证的方法进行测试，由关联用户 u 的新浪微博中提取特征 $\boldsymbol{x}_u$，计算其购买长尾商品比例的预测值为 $\hat{y}_u$，并与真实值 y_u 做比较。采用推荐系统中常用的模型评价指标[18]，即平均绝对误差(MAE)[19]、均方根误差(RMSE)[20]、确定系数(R-square)[21]，三种评价指标，计算如式(11)～(13)所示。

$$\mathrm{MAE}=\frac{\sum|\hat{y}_u-y_u|}{|U|} \tag{11}$$

$$\mathrm{RMSE}=\sqrt{\frac{\sum(\hat{y}_u-y_u)^2}{|U|}} \tag{12}$$

$$\text{R-square}=1-\frac{\sum(\hat{y}_u-y_u)^2}{\sum(\bar{y}-y_u)^2} \tag{13}$$

其中平均绝对误差(MAE)、均方根误差(RMSE)反映的是预测值与真实值的拟合程度，数值越小，表示预测效果越好，确定系数(R-square)反映的是预测值与真实数据的平均值的比较，正常取值范围区间为[0,1]，越接近1，表示模型的预测效果越好。

6.3 实验结果与分析

用改进后的MART模型对用户特征向量 $\boldsymbol{x}_u=\{x_1,x_2,\cdots,x_n\}$ 进行训练和测试，与解决回归问题的LR(linear regression)模型[22]、SVR(support vector regression)模型[23]、CART(classification and regression tree)模型[24]、神经网络多层感知机MLP(multilayer perceptron)模型[25]对比，结果如表3所示。

表3 三种模型评价指标对比

评价标准 \ 模型	LR	SVR	MLP	CART	MART	加权MART
MAE(↓)	0.091 3	0.089 4	0.113 9	0.091 5	0.091 3	0.019 7
RMSE(↓)	0.123 5	0.124 6	0.158 5	0.123 8	0.123 6	0.064 9
R-square(↑)	0.98%	−0.71%	−62.94%	0.64%	0.81%	31.03%

(↑表示值越大，预测效果越好；↓表示值越小，预测效果越好)

实验结果表明：

① MAE和RMSE两个评价指标考虑的是预测值和真实数据的拟合程度，可以看出：对比MART与LR、SVR、MLP、CART模型，五种模型预测效果相似，改进后的加权MART模型，预测效果有了明显的提升。

② R-square反映的是预测值与真实数据平均值的比较，数据的分布影响预测效果的好坏。R-square正常取值范围区间为[0,1]，越接近1，表示模型的预测效果越好。当预测值与真实值相等时，R-square值为1，在本实验中：

• 用户长尾商品的购买比例分布见图3，购买长尾商品比例小于0.3的用户占总用户数的91%，比例在0.3到0.4之间的用户为5%，比例大于0.4的用户为4%。可以看出，购买长尾商品比例比较低的用户占绝大多数，购买长尾商品比例较高的用户由于数量少，数据的分布非常不均匀，导致模型预测效果较差。当预测值偏差较大时，就可能出现负值的情况，表3中SVR模型、MLP模型因数据分布的极度不均匀，R-square的值为负值。

• 采用改进后的加权MART模型，即按照用户购买长尾商品比例加权后，购买长尾商品比例较高的用户权重得到提高，模型预测效果有了显著的提升。

6.4 特征分析

本节主要分析从微博中提取的用户特征对预测其购买长尾商品比例的贡献，选取贡献值最大的四种用户特征，进行详细的统计分析。

6.4.1 特征贡献

在决策树模型中，可以计算属性的贡献值，如论文[26]中介绍的方法，在MART的所有的回归树上，计算每个特征对节点分类的贡献之和，作为该特征的贡献值，如图4所示。

由图4可以看出，用户微博中抽取的特征对其购买长尾商品的影响，话题分布影响最大，用户年龄、群组、性别次之，而用户的兴趣爱好、婚姻状况、职业、教育背景等特征对其购买长尾商品的贡献值非常小。特征的贡献值可能与特征的维度有关，话题和群组维度均为50维，而其他特征维度相对较

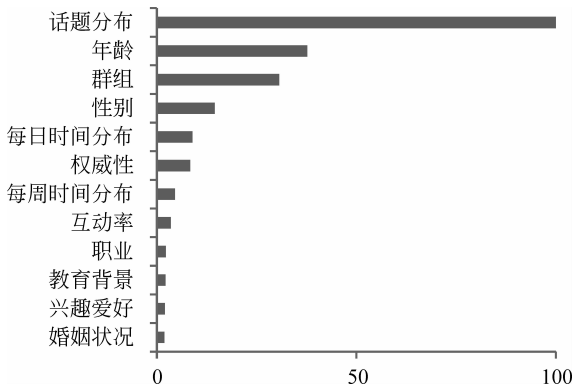

图 4　用户微博特征贡献值

小；也可能与实验所用的数据集有关，在关联用户的新浪微博爬取的数据集中，每类特征的完整度分别为：性别(100%)、兴趣爱好(65.7%)、年龄(36.7%)、教育背景(26.3%)、职业(12.9%)、婚姻状况(4.6%)、微博的文本特征(99.1%)。

由图 4 可以看出，对用户购买长尾商品影响最大的四个特征是话题分布、年龄、群组、性别，下面分别对这几种特征进行分析。

6.4.2　话题、群组的特征分析

定义话题、群组对用户购买长尾商品的影响度如式(14)所示。

$$f(i)=\sum_{u=1}^{|U|}\frac{p_{u,i}}{p_i}y_u \tag{14}$$

其中 $p_{u,i}$ 是用户 u 在话题(群组)i 上的概率分布，p_i 是所有用户在话题(群组)i 上的概率分布之和，y_u 是用户 u 购买长尾商品的比例。对于每一个话题(群组)i，计算所有用户购买长尾商品的影响值之和作为该话题(群组)对长尾商品购买的影响度，选取对长尾商品购买影响度最大的五个话题和群组，每个话题和群组中选取 10 个词，如表 4、表 5 所示。

表 4　对长尾商品购买影响度最大的五个话题

排名	话题	内容
1	话题 27	优惠、不错、非常、发现、套餐、凭券、商品、手机、优惠券、原价
2	话题 5	中国、新闻、头条、日本、记者、凤凰、人民、媒体、视频、报告
3	话题 4	活动、参加、机会、奖品、推荐、大奖、抽奖、参与、快来、快来
4	话题 24	设计、摄影、创意、微刊、作品、制作、赶紧、DIY、手工、玻璃
5	话题 49	签到、上海、点评、图片、广场、酒店、餐厅、美食、咖啡、终于

表 5　对长尾商品购买影响度最大的五个群组

排名	群组	内容	类别
1	群组 23	赵薇、姚晨、小 S 、谢娜、王力宏、何炅、大 S 、文章、林心如、马伊琍	娱乐圈明星
2	群组 35	360 安全卫士、微博 Android 客户端、360 手机卫士、360 云盘、微相册、微博赛事、360 安全浏览器、微博会员、小米手机、京东	软件类
3	群组 19	NBA、明道、微相册、微博客户端、黄健翔、鞍钢、郭明义、潘石屹、李东生、蔡文胜、鲍春来	科技体育明星
4	群组 21	冷笑话精选、精彩语录、搞笑排行榜、生活小智慧、我们爱讲冷笑话、全球热门排行榜、微博经典语录、创意工坊、时尚经典语录、奇闻趣事	语录类
5	群组 26	凤凰卫视、南方都市报、南方周末、人民日报、头条新闻、中国新闻周刊、央视新闻、新周刊、财经网、新浪财经	新闻类

由表 4、表 5 分析话题、群组对用户购买长尾商品的影响，结论如下：

① 排名第一的话题 27 中提及优惠、套餐、凭券、优惠券，可以推测，喜欢购买长尾商品的用户更倾向于关注优惠信息；话题 24 中，提及设计、创意、DIY、手工，推测长尾商品具有新奇、独特的特点；话题 4 中提及奖品、抽奖、大奖、机会等词，推测愿意购买长尾商品的用户也更愿意去参与一些博彩类的话题；

② 对群组分析发现，购买长尾商品比例较高的用户更倾向于去关注娱乐、体育、科技领域的明星，也比较愿意去关注一些语录和新闻类的用户。

6.4.3　年龄、性别的特征分析

不同年龄、性别的用户对长尾商品的购买比例统计结果如图 5 所示。

由图 5 可以看出：

① 46～59 岁年龄段的用户更喜欢购买长尾商品，31～45 岁的用户次之，18～30 岁的用户购买长尾商品的比例最少；

② 女性用户比男性用户更倾向于购买长尾商品。

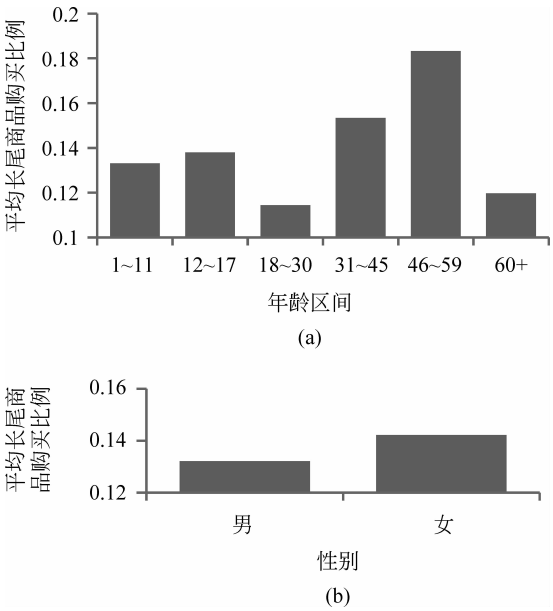

图5 年龄、性别对其购买长尾商品的影响

7 总结与展望

本文充分利用社交媒体网站上海量的文本信息和丰富的用户信息,抽取用户特征,对预测用户购买长尾商品的行为给出了一个新颖的解决思路,并分析用户特征,如年龄、性别、关注的话题和喜好等因素对其购买长尾商品的影响,可以更好地理解用户个性化需求,可据此改进电子商务网站的服务,探究长尾商品的个性化推荐,挖掘长尾市场潜在的经济价值。

然而,本文仍然存在一些需要改进的地方,例如,在单一社交媒体上抽取用户的特征还不够全面,通过对多个媒体网络的用户信息整合,我们可以获得更丰富的用户属性,用来提高预测精度。

近年来,随着深度学习的广泛应用,在跨平台的用户购买行为预测中也取得不错的效果[27],未来我们也将探索利用深度学习模型来对长尾商品进行建模。

针对长尾商品这一较为新颖的研究领域,今后我们还会对以下问题进行研究:

① 探究用户社交媒体上的用户特征对其购买长尾商品的类别的影响;

② 如何有效利用用户社交媒体信息和购买历史记录提高长尾商品的推荐准确度;

③ 如何在社交媒体网站上进行长尾商品的个性化推广。

在后续的研究中,我们将对用户特征处理和长尾商品购买行为进行更深入的分析,继续探究如何有效利用社交媒体信息,对用户购买长尾商品的行为做出更为精准的预测。

参考文献

[1] 克里斯·安德森. 长尾理论[M]. 北京:中信出版社, 2006.12.

[2] Brynjolfsson E, Hu Y, Smith M D. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers[J]. Working Papers, 2003, 49(11):1580-1596.

[3] Jansen B J, Chris Anderson. The Long Tail: Why the Future of Business is Selling Less or More.[J]. Information Processing & Management, 2007, 43(4): 1147-1148.

[4] Ricci F, Rokach L, Shapira B. Introduction to Recommender Systems Handbook[M]. Springer US, 2011: 1-4.

[5] Linden G, Smith B, York J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering[J]. IEEE Internet Computing, 2003, 7(1):76-80.

[6] Yin, Hongzhi, Cui, Bin, Li, Jing, et al. Challenging the Long Tail Recommendation[J]. Proceedings of the Vldb Endowment, 2012, 5(9):896-907.

[7] Oestreichersinger G, Sundararajan A. Recommendation Networks and the Long Tail of Electronic Commerce[J]. Social Science Electronic Publishing, 2009, 36(1):65-84.

[8] Zhang Y, Pennacchiotti M. Predicting purchase behaviors from social media[C]//Proceedings of the 22nd International Conference on World Wide Web. 2013: 1521-1532.

[9] 陈凯, 朱钰. 机器学习及其相关算法综述[J]. 统计与信息论坛, 2007, 22(5):105-112.

[10] S Ankit, S Bhanderi. Survey on Feature Engineering of Author-Paper Pair Matching in Bibliography Data [J]. International Journal of Computer Applications in Engineering Sciences, 2014,6(2):035-039.

[11] Zhang H, Spoelstra J, Spoelstra J, et al. Committee based Prediction System for Recommendation[C]// Proceedings of the 17th International Conference on Kdd Cup, 2011:215-229.

[12] Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine[J]. The Annals of Statistics, 2001, 29(5):1189-1232.

[13] Chen T, Li H, Yang Q, et al. General Functional Matrix Factorization Using Gradient Boosting[C]// Proceedings of the 31st International Conference on

Machine Learning. 2014:436-444.

[14] Zhou K, Yang S H, Zha H. Functional Matrix Factorizations for Cold-start Recommendation[C]//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011:315-324.

[15] Yan R, Huang C, Tang J, et al. To Better Stand on the Shoulder of Giants[C]//Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries. ACM, 2012:51-60.

[16] Zhao X W, Guo Y, He Y, et al. We know what you want to buy: a demographic-based system for product recommendation on microblogs[C]//Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, 2014: 1935-1944.

[17] Lin J, Sugiyama K, Kan M Y, et al. Addressing cold-start in app recommendation: latent user models constructed from twitter followers[C]//Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. ACM, 2013:283-292.

[18] 朱郁筱，吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报，2012，41(2):163-175.

[19] Shardanand U. Social information filtering: algorithms for automating "word of mouth"[C]//Proceedings of the 13th Sigchi Conference on Human Factors in Computing Systems. ACM Press/Addison-Wesley Publishing Co. 1995:210-217.

[20] Balabanovic, Marko, Shoham, Yoav. Fab: content-based, collaborative recommendation[J]. Communications of the Acm, 1997, 40(3):66-72.

[21] STEEL, R. G. D, TORRIE, J. H. Principles and procedures of statistics.[M]. McGraw-Hill, 1960.

[22] Ellis D M, Draper N P, Smith H S. Applied Regression Analysis[J]. Biometrics, 1998, 17(1):83.

[23] Jing Geng, Min-Liang Huang, Ming-Wei Li, et al. Hybridization of seasonal chaotic cloud simulated annealing algorithm in a SVR-based load forecasting model[J]. Neurocomputing, 2015, 151:1362-1373.

[24] L. Breiman. Classification and regression trees[C]//Proceedings of the Chapman & Hall/ CRC, 1984.

[25] Mirjalili S, Mirjalili S M, Lewis A. Let a biogeography-based optimizer train your Multi-Layer Perceptron[J]. Information Sciences, 2014, 269(8):188-209.

[26] Annabi H, Mcgann S T. Social Media as the Missing Link: Connecting Communities of Practice to Business Strategy[J]. Journal of Organizational Computing & Electronic Commerce, 2013, 23(1-2):56-83.

[27] Ting Bai, Hongjian Dou, Wayne Xin Zhao, Dingyi Yang, Ji-Rong Wen. An Experimental Study of Text Representation Methods for Cross-Site Purchase Preference Prediction Using the Social Text Data.. Journal of Computer Science and Technology[J]. 2017,32(4): 828-842.

白婷(1992—)，博士研究生，主要研究领域为数据挖掘、商品推荐。
E-mail: baiting@ruc.edu.cn

赵鑫(1985—)，通信作者，博士，副教授，主要研究领域为社交媒体数据挖掘、自然语言处理。
E-mail: batmanfly@ruc.edu.cn

文继荣(1972—)，博士，博士生导师，教授，主要研究领域为信息检索、数据库。
E-mail: jirong.wen@gmail.com