第31卷第1期 管理科学 ISSN 1672-0334 Vol. 31 No. 1 3-14
2018年1月 *Journal of Management Science* January, 2018

基于随机游走的电子商务退货风险预测研究

刘冠男[1],张 亮[1],马宝君[2]
1 北京航空航天大学 经济管理学院,北京 100191
2 北京邮电大学 经济管理学院,北京 100876

摘要:随着电子商务的迅猛发展,电子商务环境下的退货问题日益重要。产品销售中的高退货比例会为企业带来额外的物流和返修等成本,影响企业的正常运营。因此,有效防范退货风险、识别潜在的退货倾向对于提高电子商务企业的管理决策水平具有重要意义。在大数据背景下,电子商务企业积累了大量销售、退货和客户等多源异构数据,可基于此挖掘用户在电子商务平台中购买和退货的行为模式,进而预测退货风险。

针对电子商务环境下的退货风险建模,引入二部图结构组织历史退货记录,并将问题形式化为二部图上的节点排序问题。根据退货的用户和被退货产品的结构特点,在退货二部图中定义随机游走规则,以表征退货风险在不同产品与用户之间的传递,即将客户的退货风险表征为其退货的产品,而产品的风险表征为相关的客户。此外,考虑到退货记录的稀疏性问题,进一步引入影响退货的用户和产品等各类特征,作为随机游走的先验信息,从而提出一种融合特征的退货风险预测方法。

针对提出的预测算法,采用淘宝网一户商家的真实退货数据进行实验,实验结果表明,提出的退货风险预测方法与同类方法(如SVD和NMF等)相比具有更高的预测精度,同时相关特征的引入有效提升了模型的预测能力,特别是产品价格和质保证书对于预测精度具有显著的提升。

该方法对于电子商务企业有效防范退货风险、加强客户管理等方面具有较强的实际意义。一方面,可以帮助企业识别具有高退货风险的客户,并加强特定客户的关系管理;另一方面,可以改进对于高退货风险产品的规划,如改进质量、加强包装等。

关键词:电子商务;退货;二部图;随机游走
中图分类号:F713.36 **文献标识码**:A **doi**:10.3969/j.issn.1672-0334.2018.01.001
文章编号:1672-0334(2018)01-0003-12

引言

近年来,网上购物因其具有方便快捷的属性,已经成为人们的一种生活方式,极大地推动了电子商务的发展。然而,电子商务的虚拟特性使顾客无法获得商品的现场体验,只能依赖于商品描述、照片等媒介,所以顾客对商品的认知容易产生偏差,导致实际产品与需求不一致。在这种情况下,顾客便可能产生退货行为。退货率过高会给制造商和零售商带

收稿日期:2017-09-20 **修返日期**:2017-12-10
基金项目:国家自然科学基金(71701007,71772017,71402007);北京市社会科学基金(17GLB009)
作者简介:刘冠男,管理学博士,北京航空航天大学经济管理学院讲师,研究方向为数据挖掘与商务智能、社会网络等,代表性学术成果为"Fused latent models for assessing product return propensity in online commerce",发表在2016年第91卷《Decision Support Systems》,E-mail:liugn@buaa.edu.cn
张亮,北京航空航天大学经济管理学院硕士研究生,研究方向为数据挖掘与商务智能、电子商务等,E-mail:bhjg_zl@163.com
马宝君,管理学博士,北京邮电大学经济管理学院副教授,研究方向为数据挖掘与商务智能、移动用户行为大数据分析、政策信息学等,代表性学术成果为"Content & structure coverage: extracting a diverse information subset",发表在2017年第4期《INFORMS Journal on Computing》,E-mail:mabaojun@bupt.edu.cn

来巨大的损失,例如零售商必须根据退回产品的情况进行库存调整,带来巨大的运输和返修等成本,制造商可能要调整其生产计划等。有数据表明,在美国,每年因为产品退货产生的成本损失达到上千亿美元[1]。据零售咨询公司统计,电商企业产品的平均退货率高达三分之一。因此,防范电子商务环境中的退货风险是电子商务企业需要高度重视的问题。

实际上,随着大数据分析在商务管理中的深入应用,大多数电子商务企业已经开始重视大数据对于管理的重要意义,并具备了较为成熟的客户关系管理系统、库存管理和销售管理系统,也因此积累了大量的销售、客户和退货记录等数据。但是对于电子商务环境下大规模退货行为模式的分析和研究仍然较缺乏,也难以为管理者提供有效的退货风险预警。

有鉴于此,本研究针对个体用户和产品在电子商务环境下的退货风险进行分析和建模。电子商务的退货环境中包含了用户和产品两种基本的实体类型,因而可以将退货记录构造为二部图,而二部图的结构及节点的排序可以通过定义实体间互相表示的随机游走来发现。基于此,本研究设计了关于用户和产品的随机游走过程,进而将用户和产品的退货风险进行迭代直至收敛。同时,考虑到影响退货的用户以及产品本身的各类因素,提出一种融合特征的退货风险预测方法,并采用真实数据进行实验,验证方法的有效性。

1 相关研究评述

1.1 退货的影响因素

目前针对退货的相关研究主要是从营销和运作管理的角度出发,分析影响退货的各类因素,并且探究不同退货政策对于运营管理的影响。在退货的影响因素研究方面,LI et al.[2]设计了不同的模型检验在线购物中退货政策、商品价格、商品质量对于消费者购买意愿和退货意愿的影响,发现这些要素的影响是相互作用和耦合的;WALSH et al.[3]运用风险理论,通过实验检验退款保证、产品评论和免费退货标签3种工具对用户退货行为的影响,发现退款保证的使用增加了产品的退货率,而产品评论与之相反,降低了产品的退货率,提供免费退货标签对退货行为没有产生显著影响。这些研究说明产品价格、产品质量等产品本身的属性在退货行为的预测中占据着重要的地位。孙永波等[4]通过实证分析研究用户的购买行为与退货行为之间的关联,发现有过退货经历的用户其后续的购买行为是可以被零售商善意"操控"的。这启发研究者可以从用户特质的角度去探讨对退货行为的预测。特别地,DE et al.[5]通过实证方法研究电商平台中信息技术的使用对退货的影响,包括图片、网站排版、文字描述等;FU et al.[6]认为退货的发生是由两种不一致导致的,顾客期望的商品属性与实际的商品属性之间不一致,实际的商品属性与顾客收到的商品属性之间不一致,在此基础上利用带有隐变量的概率矩阵分解预测了交易的退货概率。

在退货政策方面,PASTERNACK[7]研究定价策略和退货政策,提出一种对于短期寿命商品的层次定价模型;张霖霖等[8]将用户的退货行为引入到在线零售企业的单周期和多周期定价订货策略研究中,发现退货率与在线零售企业定价正相关,而与订货量和收益负相关。这些研究都只聚焦于产品价格对于退货的影响,没有很好地探讨其他属性对结果的影响。李勇建等[9]研究在产品需求和消费者产品估价均不确定的情况下,报童零售商的预售策略和无缺陷退货问题,发现最优的退货策略是部分退款退货策略,且最优退货价格为产品的残余价值。但却在模型中忽略了产品需求与产品本身特征和消费者类型之间的联系,类似的缺陷也存在于孙军等[10]的研究中。赵晓敏等[11]着重从产品生命周期的视角探讨不同的退货政策对企业供应链系统运作绩效的影响;MUKHOPADHYAY et al.[12]发现提供友好的退货政策能够增加收入,但同时也会由于高昂的退货和设计费用增加成本,并基于此提出一种优化退货政策的最大化模型;ANDERSON et al.[13]提出一个用来识别最优退货政策的结构化模型,使零售商可以在销售需求和退货成本之间进行取舍。与本研究不同的是,这些关于退货政策的研究都是从较为宏观的角度出发,在电子商务的环境下不容易进行个性化的应用和推广。更进一步地,卢美丽等[14]将退货视为一种促进销售的服务策略,讨论不同商品的服务敏感系数、销量退货率和退货量对于价格敏感系数和最优利润的影响;单汨源等[15]聚焦于退运险这一细分领域,通过构建数学模型分析不提供退运险服务、赠送退货运费险和消费者购买退货运费险3种退货策略下零售商的盈利能力,证明了赠送退货运费险这种策略的有效性。这些研究启发我们在对退货的预测研究中,零售商的服务水平和品牌效应等因素也应当融入到建模过程中。

以上研究一般仅从统计意义上分析影响退货的各类因素,无法针对特定用户对特定商品的退货倾向性进行分析。有鉴于此,本研究从更为微观和个性化的角度出发,挖掘用户在退货过程中的行为模式,进而预测用户对特定商品的退货风险,指导电子商务企业的运营管理实践。

1.2 二部图

现实世界中的许多行为活动都可以转换为二部图结构,如用户购买产品和用户评价等。因而,关于二部图的结构分析和模式发现等研究一直是热点问题。MOONESINGHE et al.[16]基于实体之间的相似性构造二部图,为每个实体分配异常得分,并假设与其他实体之间的关系较少的实体更有可能是异常点;BEUTEL et al.[17]对社交网络中的异常"点赞"行为进行研究,他们将用户与社交网络的页面根据"点赞"关系构造为二部图,并将疑似的非法"点赞"行为定

义为一种基于时间的子图结构,从而将问题转化为在二部图中的结构搜索问题。这类异常检测的研究一定程度上证明了二部图的结构可以很有效地对退货这类数据进行建模。ZHU et al.[18]通过构建用户和产生内容的二部图,利用随机游走的方法研究社交网络中用户影响力的识别和度量;FOUSS et al.[19]将用户和产品构建成为二部图,并定义了在图结构上的马尔科夫链的随机游走过程,他们通过定义一些马尔科夫链上的基本度量,如第一次经过的时间、成本和平均的游走时间等,以度量不同节点之间的相似性,提供了一种利用随机游走方法对二部图中节点进行排序的基本思路。HE et al.[20]提出一套贝叶斯框架,可以基于图的链接结构和节点信息来研究二部图上的节点排序问题,他们通过引入查询向量来平滑二部图,在优化正则化函数的同时动态地更新各节点的得分,进而实现排序的目的。查询向量的引入能够很好地平滑异常点的影响,大幅提高算法的鲁棒性,具有很强的借鉴意义。蔡小雨等[21]提出一种采用群体信息的二部图链接预测方法,通过对二部图进行投影,抽取二部图中节点对的局部结构属性,并运用群体检测技术抽取节点对的群体属性,融合二者作为相似度的度量标准,有效地提高了二部图链接预测的准确率。在推荐领域,关雲菲[22]通过构建用户项目二部图,引入用户的点击、收藏、加入购物车和购买4种行为数据优化评分系统,实现了对传统的基于二部图的推荐算法的改进;黄熠姿等[23]根据用户的评论数以及与该用户对项目评分相同的评论数量定义该用户的专家信任度,根据传统的评分信息定义用户的偏好程度,提出融合专家信息的二部图推荐算法,实验结果表明该算法表现出了优良的性能。但这些工作的研究重点主要是对推荐算法本身的改进,没有聚焦于用户在电子商务环境中的退货行为模式的建模。

以上研究均说明,基于二部图研究具有较好的泛化能力,可以适应多种场景下针对不同实体之间交互关系的建模。因此,本研究以二部图结构组织用户的产品退货记录,进而对个体用户在电子商务中的退货行为进行预测分析。

1.3　基于随机游走的推荐算法

自从随机游走被提出,就一直受到研究者的青睐,现已被广泛应用于图像分割[24]、图挖掘[25-26]和文本挖掘[27]等领域。近年来研究者通过构建用户网络和产品网络,利用随机游走等模型,定义不同节点之间的相似性,从而设计推荐算法,以解决稀疏性和冷启动等传统推荐中常见的问题。PUCCI et al.[28]提出一种基于随机游走的评分算法ItemRank,可以根据潜在目标用户的偏好对产品进行得分排序,进而实现推荐的目的。但是该方法并没有考虑到与目标用户相似的其他用户的偏好,对偏好的建模不够完备。针对冷启动问题,SHANG et al.[29]提出一种基于马尔科夫随机游走的混合协同过滤模型,发现与传统的协同过滤模型相比,该算法能够更好地适应冷启动的情况;施海鹰[30]利用关联规则挖掘的特性,挖掘用户属性与项目之间的关联,为新用户构造初始的评分向量,弥补了传统推荐算法的不足。这类基于协同过滤的模型难以处理极端稀疏的数据,且对异常点十分敏感,不适合用来建模退货这类数据集。张光前等[31]尝试从消费心理学的角度解决冷启动问题,提出基于消费者购物记录分析其消费性格、基于消费者消费性格进行新商品推荐的方法,通过消费心理这一纽带建立起消费者与新商品之间的联系。但该方法在应用时需要收集较多的额外信息,在电子商务环境下难以有效实施。JAMALI et al.[32]认为,基于信任网络的推荐比传统的基于用户评分的推荐包含更多的信息,有利于解决冷启动和稀疏性问题,他们提出TrustWalker算法,即基于信任网络的随机游走,并在游走的过程中返回预测的用户产品评分;张萌等[33]在此基础上提出一种基于用户偏好的PtTrustWalker算法,该算法在TrustWalker的基础上通过细化信任度量,引入权威度等信息加强了信任网络,使推荐变得更有针对性和可解释性,并且一定程度上增强了模型的稳定性。这类方法一般仅使用二部图本身的信息,缺乏利用丰富的先验信息提高算法性能的机制。MO et al.[34]将随机游走方法引入到基于事件的社交网络的推荐中,通过构建异构图来表示社交网络中不同类型的实体之间的交互作用,并提出一种重启动的反向随机游走方法,以获得每个用户的评分列表。类似的,曹云忠等[35]将社交网络中用户间的交互行为引入信任的计算,通过基于信任的随机游走模型实现了微博粉丝的精准推荐。与之类似,在退货二部图中,用户间通过产品而产生的交互行为也需要被引入到偏好的计算中。张怡文等[36]采用共同项目和用户打分项目数量的共同性质体现用户兴趣度,提出一种基于用户兴趣度的二部图随机游走方法;李镇东等[37]在传统的二部图推荐算法的基础上,提出一种以单调饱和函数为权重,利用目标用户和其他项目共同评分个数相对用户总数均值的正切值作为相似性度量的推荐算法。这类研究大多只从用户角度出发,没有将产品一侧的相似度融入到模型之中。杨华等[38]将推荐网络的拓扑结构从二部图延伸到更一般的网络,根据商品、品牌、店铺及其关联关系构建混合图,通过重启动的随机游走算法确定节点间的转移概率,实现商品推荐,证明了随机游走方法在图排序问题上良好的泛化能力。

上述研究仅针对用户的购买记录进行建模,并未考虑用户特征和产品本身的特征。而对于退货问题来说,需要同时考虑与购买和退货相关的行为,融合影响退货的用户特征和产品特征,从而提升模型的预测精度。

2　基于二部图的退货风险模型

退货是用户的一项综合决策过程,与产品的购买过程类似,在一定程度上反映了用户对于产品的偏好特征和个性化的退货行为模式,同时也涉及到

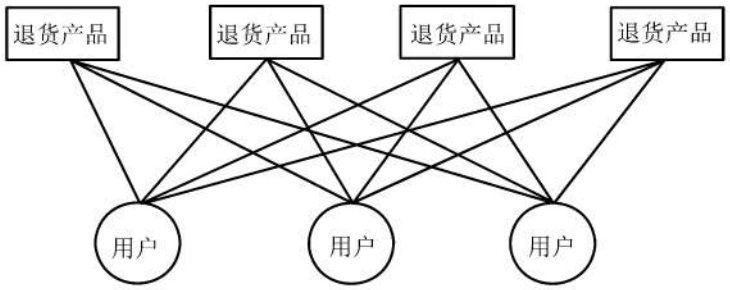


图 1 退货二部图结构示例

Figure 1 Example for Product Return Bipartite Network

用户和产品等不同实体。不同的用户对于不同类型商品评价的侧重点不同,对应的退货行为也存在特定的模式,因此需要针对用户购买和退货的行为数据进行深度挖掘,进而对用户在购买各类产品时发生退货的风险进行预测。对于具体的目标用户来说,退货风险即为针对不同产品的退货倾向。

2.1 退货二部图与随机游走

如前所述,二部图能够有效地表征不同类型实体间的交互活动。实际上,电子商务中的退货场景中所包含的用户和产品符合二部图刻画不同实体类型间交互行为的结构。令由“用户 - 产品”的退货记录构成的退货二部图为$G,G=(U\cup I,E)$,U为电子商务平台中的用户集合,I为平台上的产品集合,E为该二部图的边集。二部图中的边由历史退货记录集合T生成,形如$(u_j,i_k,w_{jk})\in E,u_j$为用户,$u_j\in U,1\leqslant j\leqslant|U|$;$i_k$为产品,$i_k\in I,1\leqslant k\leqslant|I|$;$w_{jk}$为$u_j$用户对$i_k$产品的退货次数。对二部图中的每一个用户节点和每一个产品节点而言,度是图上的重要属性,因此可以引入两个由权重矩阵$\boldsymbol{W}$生成的对角矩阵$\boldsymbol{D}_U$和$\boldsymbol{D}_I$。

基于如上定义的退货二部图,可以根据二部图的结构特征对图中的节点按照一定的规则进行排序。因此,对于退货风险的预测问题可以转换为基于二部图的结构发现问题。具体而言,对于特定用户的退货风险的预测问题可以定义为:给定目标用户节点u_j,根据该节点在二部图中与不同产品的连接以及与其他用户节点的相似性,得到该用户对于不同产品的潜在风险退货列表。

随机游走提供了一种根据二部图中节点间的相关性进行排序的方法,其基本思想是根据特定的概率游走规则,在不同类型的节点间进行转移,直至收敛,能够在一定程度上减小稀疏性的影响。因此,在对用户和产品的退货风险进行建模时,本研究构建二部图,并通过随机游走模型实现对用户和产品的循环表示。具体而言,对应于本研究所关注的退货二部图,可以将用户到产品的一条退货记录边作为一条随机游走的路径,而在退货网络中的随机游走则可以视作是退货风险在用户与用户之间、产品与产品之间的传递。其中相似的用户具有相似的退货行为,而相似的产品也会被相似的用户退货。图 1 为一个退货二部图结构的示意图,直接反映用户与产品退货关系的结构特点。

于是,令u_j用户为待预测的目标用户,由退货二部图可以得到其对应的产品集合为$I(u_j)$,$I(u_j)=\{i_k\}$,$(u_j,i_k)\in T$。显然,$I(u_j)$中的产品与目标用户具有较强的相关性。因此,基于随机游走的基本思想,退过$I(u_j)$中产品的u_p用户则与目标用户具有较强的相似性。与此同时,u_p用户所退的产品集合$I(u_p)$也与目标用户产生了相关性,循环迭代,则可以生成与目标用户最相似的用户节点集以及最相关的产品节点集。上述过程可形式化地描述为以下两个迭代规则,即

$$r_{u_j}=\sum_{k=1}^{|I|}w_{jk}r_{i_k} \tag{1}$$

$$r_{i_k}=\sum_{j=1}^{|U|}w_{jk}r_{u_j} \tag{2}$$

其中,r_{u_j}为u_j用户的退货风险,可以用其对应的退货产品和退货次数表示;r_{i_k}为i_k产品的退货风险,可以用退过该产品的用户和退货次数表示。但是,根据 ZHOU et al.[39-40]的研究,上述形式的迭代规则不容易平稳地收敛,很容易受到异常点和参数设置的影响,所以需要进行形式上的正则化处理。因此,本研究使用对于图的对称正则方法进行平滑处理,正则化后的迭代规则为

$$r_{u_j}=\sum_{k=1}^{|I|}\frac{w_{jk}}{\sqrt{d_j}\sqrt{d_k}}r_{i_k} \tag{3}$$

$$r_{i_k}=\sum_{j=1}^{|U|}\frac{w_{jk}}{\sqrt{d_j}\sqrt{d_k}}r_{u_j} \tag{4}$$

其中,d_j为二部图中u_j用户的度,d_k为二部图中i_k产品的度。

本研究涉及的变量及其含义见表 1。

2.2 融合退货特征的二部图排序模型

2.2.1 影响退货的特征分析

本研究针对用户和产品的各类特征进行观测。在淘宝网中,平台根据用户的购买记录对用户的信用水平进行评分。图 2 给出不同信用评分用户的退货率分布,其中高退货率的用户主要集中在低信用评分区段,当信用评分超过 2 000 时,退货率基本稳定在 0 附近,总体呈现出负相关的趋势。由此可见,用户的信用评分与退货有很强的相关性。不同信用评分区段的用户具有不同的退货特征,信用评分较

表 1　变量及其含义

Table 1　Variable and Definition

符号	描述
u_j	索引为 j 的用户
i_k	索引为 k 的产品
w_{jk}	u_j 用户对 i_k 产品的退货次数
U	用户集合,$U=\{u_j \mid 1 \leqslant j \leqslant \mid U \mid\}$
I	产品集合,$I=\{i_k \mid 1 \leqslant k \leqslant \mid I \mid\}$
$\boldsymbol{F}_{u_j}$	u_j 用户的特征属性向量
$\boldsymbol{F}_{i_k}$	i_k 产品的特征属性向量
T	历史退货记录集合,$T=\{(u_j,i_k)\}$
$\boldsymbol{u}$	迭代收敛后用户退货风险向量
$\boldsymbol{i}$	迭代收敛后产品退货风险向量
$R(u_j)$	对 u_j 用户退货风险预测得到的产品集合
G	用户–产品退货二部图,$G=(U\cup I,E)$
E	二部图的边集
$\boldsymbol{W}$	二部图的权重矩阵,$\boldsymbol{W}=\{w_{jk}\}$
$\boldsymbol{D}_U$	用户节点的度矩阵
$\boldsymbol{D}_I$	产品节点的度矩阵

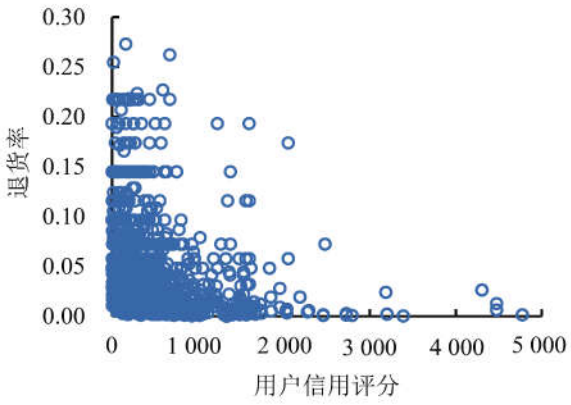


图 2　不同信用评分用户的退货率分布

Figure 2　Product Return Rate Distribution for Users with Different Credit Scores

低的用户退货倾向更明显。

图 3 给出不同价格的产品呈现出的不同的退货特征。由图 3 可知,随着产品价格的升高,产品的退货率也逐渐升高,呈现出正相关的特征。一般来说,对于价格较为便宜的产品,用户的期望相对较低,退货风险较小;而对于价格较高的产品,用户要求较高,发生退货的风险也更高。因此,产品价格可以作为预测退货风险的一大特征。

图 4 给出产品运费的支付方与退货频次分布之间的关系。由图 4 可知,当运费支付方为用户时退货

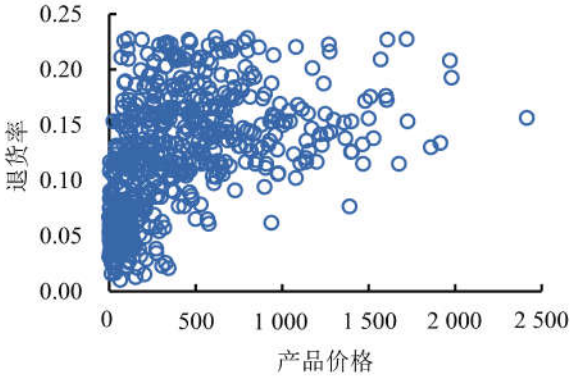


图 3　不同价格产品的退货率分布

Figure 3　Product Return Rate Distribution with Different Price

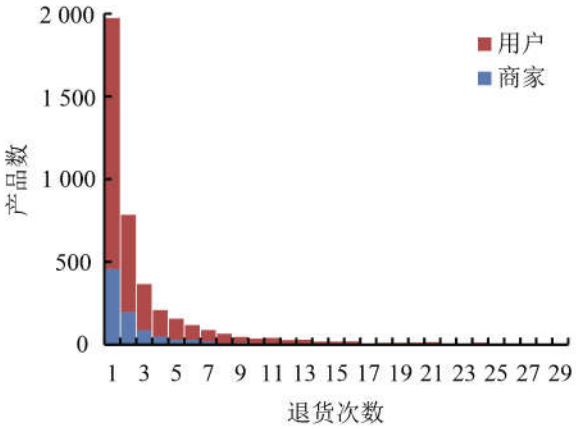


图 4　不同产品运费支付方的退货频次分布

Figure 4　Product Return Frequency Distribution When Shipping Rate Paid by Different Parties

风险更高。因此,产品包邮与否也可以作为测量退货风险的特征。

此外,在电子商务环境中,用户只能通过产品的简介和描述来判定产品的质量,其中是否拥有质保证书是一项重要的指标,图 5 给出是否拥有质保证书的产品被退货的频次分布。由图 5 可知,无质保证书的产品被退货的风险高于有质保证书的产品。可能

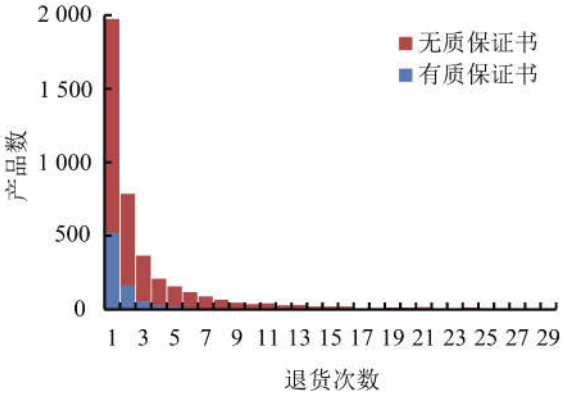


图 5　产品是否拥有质保证书的退货频次分布

Figure 5　Product Return Frequency Distribution over Whether Product Has Warranty

无质保证书的产品总体上质量较差,也可能因为用户对无质保证书的产品持负面态度。因此,有无质保证书也可以作为影响退货的重要特征融入到退货风险的预测模型中。

2.2.2 退货特征相似性度量

随机游走测量用户与产品之间的相关性,表示退货风险在二部图中传递。因此,为了将上述相关特征融入到随机游走过程中,需要度量用户与产品在不同特征间的相似性,并将相似性作为随机游走的先验信息,指导游走过程。

(1)用户静态相似性的度量

根据图2可知,不同信用评分的用户具有不同的退货行为特征,可以很好地用来量化用户的静态相似性。对于任意的目标用户u_j,任取用户集合U中的一个元素记为u_l,设计如下的相似性函数计算该用户与目标用户之间的相似度,即

$$S_U(u_j,u_l)=\frac{|S_{u_j}-S_{u_l}|}{\max(S_{u_j}-\min\limits_{1\leqslant x\leqslant|U|}S_{u_x},\max\limits_{1\leqslant x\leqslant|U|}S_{u_x}-S_{u_j})} \tag{5}$$

其中,$S_U(u_j,u_l)$为基于用户的相似性度量函数,S_{u_j}为u_j用户的信用评分,S_{u_l}为u_l用户的信用评分,S_{u_x}为除u_j和u_l用户外其他任一用户的信用评分。当u_l用户是目标用户时,$S_U(u_j,u_l)$的取值为0;当u_l用户不是目标用户,但与目标用户信用评分差距最大时,$S_U(u_j,u_l)$的取值为1。且$S_U(u_j,u_l)$在0~1之间具有良好的线性变化性质。

(2)产品相似性的度量

根据之前的观测,产品的相关特征属性主要包括价格、产品包邮与否和是否有质保证书3项,价格是连续性变量,其他两项是[0,1]变量。为了消除量纲的影响,先对价格属性进行归一化处理,归一化函数为

$$f(i_k,i_p)=\frac{|P_{i_k}-P_{i_p}|}{\max(P_{i_k}-\min\limits_{1\leqslant y\leqslant|I|}P_{i_y},\max\limits_{1\leqslant y\leqslant|I|}P_{i_y}-P_{i_k})} \tag{6}$$

其中,i_k为目标产品,i_p为产品集合I中的任意一个元素,P_{i_k}为i_k产品的价格,P_{i_p}为i_p产品的价格,P_{i_y}为除i_k和i_p产品外其他任一产品的价格。

令i_k产品经过归一化后的特征属性向量为$\boldsymbol{F}_{i_k}$,i_p产品经过归一化后的特征属性向量为$\boldsymbol{F}_{i_p}$,采用调整的相关系数作为产品之间相似性的度量函数,记为$S_I(i_k,i_p)$,即

$$S_I(i_k,i_p)=0.5\cdot\frac{cov(\boldsymbol{F}_{i_k},\boldsymbol{F}_{i_p})}{\sqrt{Var(\boldsymbol{F}_{i_k})}\sqrt{Var(\boldsymbol{F}_{i_p})}}+0.5 \tag{7}$$

(3)退货特征的随机游走

在测量退货特征相似性的基础上,可将其作为算法的先验信息融入到随机游走中。具体而言,通过$S_U(u_j,u_l)$函数计算所有用户与目标用户u_j的相似性,可以生成用户的先验信息$\boldsymbol{u}^0$,从而将用户特征融合到用户端退货风险的测量中,即

$$\boldsymbol{u}^0=\{S_U(u_j,u_l)\}\quad 1\leqslant l\leqslant|U| \tag{8}$$

产品在退货特征上的相似性也可以作为产品端游走过程的先验信息,以此改进(2)式中对于产品退货风险的测量。同时,由于退货风险预测的目标是寻找目标用户最可能退货的产品列表,所以产品的先验信息还应包含产品与目标用户之间的相关性,这里采用退货次数占比作为相关性的度量,记为$r(u_j,i_k)$,即

$$r(u_j,i_k)=\frac{w_{jk}}{(D_U)_{jj}} \tag{9}$$

其中,$(D_U)_{jj}$为u_j用户的总退货次数。但是,用户的退货记录矩阵是较为稀疏的矩阵,即目标用户对很多产品的退货次数可能为0,难以进行有效的区分。因此,本研究在产品特征相似性的基础上,引入基于产品特征相似性的平均退货次数占比,记为$C(u_j,i_k)$,即

$$C(u_j,i_k)=\begin{cases}\dfrac{\sum\limits_{i_p\in I(u_j)}S_I(i_k,i_p)r(u_j,i_k)}{|I(u_j)|}, & i_k\notin I(u_j)\\ r(u_j,i_k), & i_k\in I(u_j)\end{cases}$$

$$I(u_j)=\{i_p\},(u_j,i_p)\in T \tag{10}$$

根据(10)式可以测量u_j目标用户与所有产品之间的相关性,进而生成产品的先验信息$\boldsymbol{i}^0$,从而将产品特征融合到产品端退货风险的测量中,即

$$\boldsymbol{i}^0=\{C(u_j,i_k)\},\ 1\leqslant k\leqslant|I| \tag{11}$$

进一步地,引入超参数α和β对原有的随机游走过程和退货特征的相似性进行线性组合,得到融合的迭代规则。

$$r_{u_j}=\beta\sum_{k=1}^{|I|}\frac{w_{jk}}{\sqrt{d_j}\sqrt{d_k}}r_{i_k}+(1-\beta)u_j^0,\ 0\leqslant\beta\leqslant1 \tag{12}$$

$$r_{i_k}=\alpha\sum_{j=1}^{|U|}\frac{w_{jk}}{\sqrt{d_j}\sqrt{d_k}}r_{u_j}+(1-\alpha)i_k^0,\ 0\leqslant\alpha\leqslant1 \tag{13}$$

其中,α和β为超参数,α表示产品先验信息的重要性,β表示用户先验信息的重要性。上述规则可以使用向量形式更为简洁地表达为

$$\boldsymbol{u}=\beta(\boldsymbol{D}_U^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}_I^{-\frac{1}{2}})\boldsymbol{i}+(1-\beta)\boldsymbol{u}^0,\ 0\leqslant\beta\leqslant1 \tag{14}$$

$$\boldsymbol{i}=\alpha(\boldsymbol{D}_U^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}_I^{-\frac{1}{2}})^T\boldsymbol{u}+(1-\alpha)\boldsymbol{i}^0,\ 0\leqslant\alpha\leqslant1 \tag{15}$$

其中,$\boldsymbol{u}$为按与目标用户相似性排序的用户向量,$\boldsymbol{i}$为按退货风险排序的产品向量。

上述迭代规则是基于二部图的退货风险预测模型的核心,根据迭代规则可以设计如算法1(ReRank)所示的退货风险预测方法。具体而言,输入目标用户、权重矩阵、超参数α和β,经过多次的迭代直至收敛,最终输出$\boldsymbol{u}$和$\boldsymbol{i}$,其中排名前N的产品集合$R(u_j)$作为预测的退货风险列表。

算法1 基于二部图的退货风险预测模型(ReRank)

Input: Target User u_j, Weight matrix $\boldsymbol{W}$, hyperparameters α,β;

Output: User vector $\boldsymbol{u}$, Product vector $\boldsymbol{i}$;

1: Compute User prior information $\boldsymbol{u}^0$ according to equation(8)

2: Compute Product prior information $\boldsymbol{i}^0$ according to equation(11)

3: Symmetrically normalize $\boldsymbol{W}$: $\boldsymbol{D}_U^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}_I^{-\frac{1}{2}}$;

4: Randomly initialize $\boldsymbol{i}$ and $\boldsymbol{u}$;

5:while Stopping criteria is not met do:

6: $\boldsymbol{u} = \beta(\boldsymbol{D}_U^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}_I^{-\frac{1}{2}})\boldsymbol{i} + (1-\beta)\boldsymbol{u}^0$;

7: $\boldsymbol{i} = \alpha(\boldsymbol{D}_U^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}_I^{-\frac{1}{2}})^T\boldsymbol{u} + (1-\alpha)\boldsymbol{i}^0$;

8:end

9:return $\boldsymbol{i}$ and $\boldsymbol{u}$

3 实验

3.1 实验设计

本研究从淘宝网的在线商家中获取交易数据,淘宝网是阿里巴巴旗下的电子商务B2C购物网站,是目前中国最大的电子商务平台之一。该在线商家主要经营护肤产品,包括面霜、面膜、香水等。该数据集包含用户记录、产品记录和2013年全年的退货记录。为了更好地发现用户退货的潜在行为模式,本研究对发生频繁退货的用户进行采样,保留退货次数超过2的用户及其退货记录。并抽取用户的信用评分作为用户特征,以产品价格、运费支付方和证书状态作为产品特征。抽样后形成的新数据集的统计数据见表2。

表2 数据集描述

Table 2 Description for Dataset

描述	数值
用户数	6 433
产品数	16 184
退货记录数	18 823
用户平均退货次数	2.93
产品平均退货次数	1.16

将退货记录划分为5份,取其中的4份划入训练集,其余的划入测试集。对于无法等分的部分,向上取整划入训练集中。在此基础上进行实验。

3.2 实验比较方法和评价指标

3.2.1 实验比较方法

为了验证本研究提出的算法ReRank的实际预测效果,选取一些常用的推荐方法作为基准比较方法。

(1)基于产品的协同过滤(ItemCF)

基于产品的协同过滤的基本思想是向用户推荐与他们之前偏好的产品相似的产品。该算法认为,A产品与B产品具有很强的相似性是因为偏好A产品的用户也更倾向于偏好B产品。记A产品的退货向量为$\boldsymbol{V}_A$,B产品的退货向量为$\boldsymbol{V}_B$,采用余弦夹角计算二者之间的相似度可以得到产品的相似度矩阵。对于目标用户,利用产品相似度对用户偏好程度进行加权平均,经排序后可输出推荐列表$R(u_j)$。

$$S_I(\boldsymbol{V}_A,\boldsymbol{V}_B) = \frac{\sum_{j=1}^{|V_A|}(V_{A_j}V_{B_j})}{\sqrt{\sum_{j=1}^{|V_A|}(V_{A_j})^2}\sqrt{\sum_{j=1}^{|V_B|}(V_{B_j})^2}} \tag{16}$$

其中,V_{A_j}为退货向量$\boldsymbol{V}_A$的第j个分量的值,V_{B_j}为退货向量$\boldsymbol{V}_B$的第j个分量的值。

(2)基于用户的协同过滤(UserCF)

基于用户的协同过滤的基本思想是向用户推荐与其相似的用户所偏好的产品。该算法认为,C用户与D用户很相似是因为二者偏好同样的产品。记C用户的退货向量为$\boldsymbol{V}_C$,D用户的退货向量为$\boldsymbol{V}_D$,采用余弦夹角计算二者之间的相似度可以得到用户之间的相似度矩阵。对于目标用户,利用用户相似度对产品偏好程度进行加权平均,经排序后可输出推荐列表$R(u_j)$。

$$S_U(\boldsymbol{V}_C,\boldsymbol{V}_D) = \frac{\sum_{j=1}^{|V_C|}(V_{C_j}V_{D_j})}{\sqrt{\sum_{j=1}^{|V_C|}(V_{C_j})^2}\sqrt{\sum_{j=1}^{|V_D|}(V_{D_j})^2}} \tag{17}$$

其中,V_{C_j}为退货向量$\boldsymbol{V}_C$的第j个分量的值,V_{D_j}为退货向量$\boldsymbol{V}_D$的第j个分量的值。

(3)奇异值分解(SVD)

奇异值分解是一种矩阵分解的方法,它可以将推荐问题映射到一个隐含空间进行求解。对于本研究关注的退货问题,给定退货矩阵$\boldsymbol{W}$,w_{jk}为矩阵中任意元素。SVD假设用户和产品都可以被映射到一个低维度的隐含空间,而退货矩阵可以分解为用户对各个隐含因子的偏好程度$\boldsymbol{L}$以及产品包含各个隐含因子的程度$\boldsymbol{M}$。典型的奇异值分解公式为

$$\boldsymbol{W} = \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{M}^T \tag{18}$$

其中,$\boldsymbol{\Sigma}$为分解后的中间矩阵。

(4)非负矩阵分解(NMF)

与SVD方法类似,NMF也是将消费者对于产品的评分矩阵分解为消费者与产品的隐含矩阵。NMF要求输入矩阵元素非负,目标是最小化消费者对于产品的评分矩阵与多个隐含矩阵乘积之间的距离。

3.2.2 评价指标

(1)准确率(*Precision*)

准确率是反映预测精度的单值指标,表示预测的退货风险列表中实际发生退货的产品数在预测列表中所占的比例。因此对于u_j用户,退货风险预测得到的产品集合为$R(u_j)$,$R(u_j)$中实际发生退货的产品集合为$hits(u_j)$,对应的准确率为

$$Precision_{u_j} = \frac{|hits(u_j)|}{R(u_j)} \tag{19}$$

(2)召回率(*Recall*)

召回率是指预测的退货风险列表中实际发生退货的产品数在用户实际发生退货的产品数中所占的比例。对于u_j用户,其实际发生退货的产品集合记为$I(u_j)$,$R(u_j)$中实际发生退货的产品集合为$hits(u_j)$。

$$Recall_{u_j} = \frac{|hits(u_j)|}{I(u_j)} \tag{20}$$

(3)*nDcg*

该指标用来测量算法能否将实际发生的退货产品置于预测风险列表的顶端,该指标值越大,说明得到的预测精度越高。对于u_j用户,退货风险预测得到的产品集合为$R(u_j)$,长度为N。计算*Dcg*的公式为

$$Dcg = q_1 + \sum_{k=2}^{N} \frac{q_k}{\log_2(k)} \quad k = 1,2,\cdots,N \tag{21}$$

其中,当排序列表中的第k件产品在交易记录中被实际购买时,$q_k = 1$;反之,$q_k = 0$。为了得到*nDcg*,需要对*Dcg*进行标准化,即

$$nDcg = \frac{Dcg}{Idcg} \tag{22}$$

其中,*Idcg*为在最理想的排序情形时*Dcg*的取值,即最大化的取值。当有多个目标用户时,计算不同用户*nDcg*的均值即可。

3.3 退货风险预测实验结果

3.3.1 算法收敛性分析

基于随机游走算法的特点,在实验中首先利用用户和产品的退货风险向量平均值的变化率对算法的收敛进行分析。取$\alpha = 0.5$,$\beta = 0.8$,根据(14)式和(15)式计算迭代后得到的退货风险向量$\boldsymbol{u}$和$\boldsymbol{i}$,同时计算与上次迭代得到的向量的平均值的变化率。收敛性分析见图6,随着迭代次数的增加,用户和产品退货向量的变化率都在同时减小,当迭代次数大于10时,$\boldsymbol{u}$和$\boldsymbol{i}$平均值的变化率同时趋近于0,算法趋于收敛。

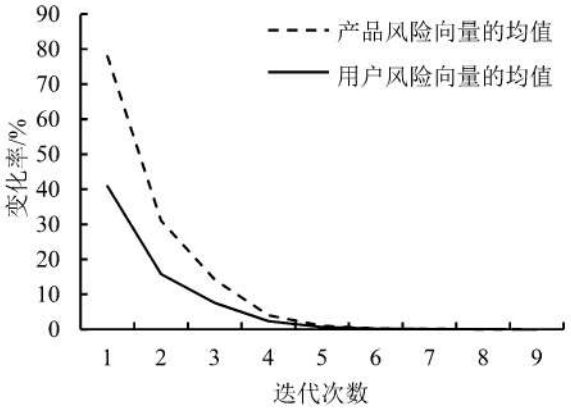


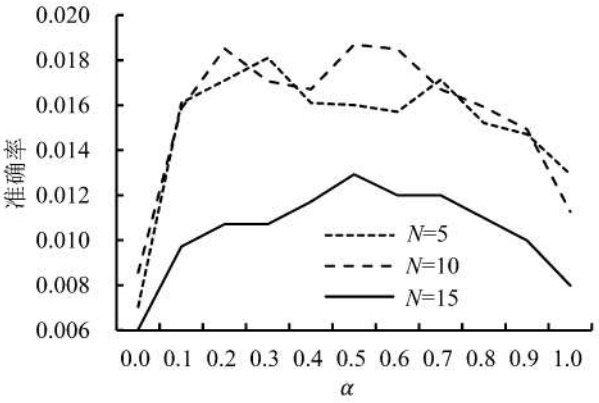
图6 算法的收敛性分析结果

Figure 6 Convergence Analysis Results for the Algorithm

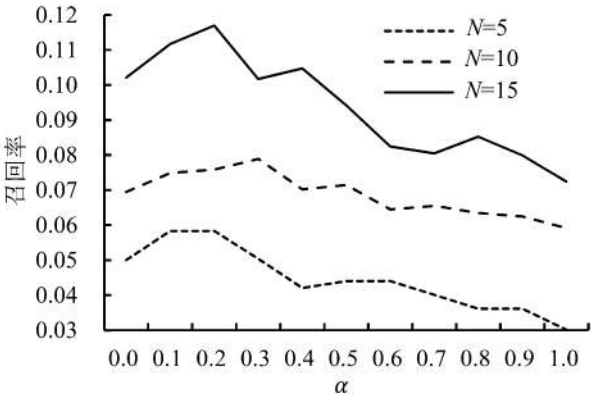
3.3.2 参数敏感性分析

本研究提出的ReRank算法中包含α和β两个超参数,分别用来衡量产品先验信息和用户先验信息的重要性,可以根据实际的使用情况自由设置。不同的参数设置可以导致不同的推荐结果,因此在本实验中着重分析模型对超参数的敏感性。

取$\beta = 0.8$并保持不变,分析α对模型性能的影响,见图7。由图7可知,分别在列表长度为5、10和15的情形下进行参数分析,随着α值的增大,模型的召回率呈现不断下降的趋势,准确率先升后降。当$\alpha = 1$,即无任何产品先验信息时,与包含一定的先验信息时相比,模型的准确率和召回率都有明显的下降,可见先验信息对于模型性能的重要影响。



(a)准确率



(b)召回率

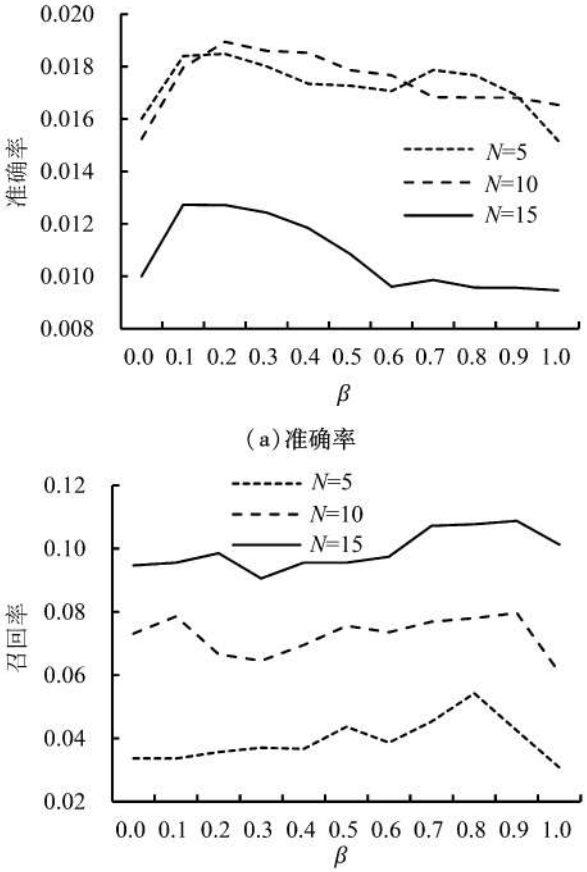
图7 α的敏感性分析

Figure 7 Sensitivity Analysis Results for the α

取$\alpha = 0.5$,分析β对模型性能的影响,见图8。由图8可知,在退货预测列表长度分别为5、10和15时,随着β值的不断增加,模型的召回率总体呈上升的趋势,准确率总体呈下降的趋势。同样的,当$\beta = 1$,即无任何用户先验信息时,与包含一定的先验信息时相比,模型的准确率和召回率也都有明显的下降。另外,准确率和召回率曲线的变化幅度都很小,说明在该数据集上ReRank算法对β不敏感。

3.3.3 算法性能分析

进一步地,设定最优参数($\alpha = 0.5$, $\beta = 0.8$),对所有用户的退货风险进行预测,即根据用户对于产品的退货风险预测用户的退货列表。将预测结果与UserCF、ItemCF、SVD和NMF等算法进行对比,分析结果见图9。整体上看,本研究提出的算法在所有指标上均表现得最好,当列表长度为15时,与NMF相比,ReRank的准确率提高了16%,召回率提高了17%,*nDcg*提高了11%。另外,基于产品的协同过滤表现



(a)准确率

(b)召回率

图 8　β 的敏感性分析

Figure 8　Sensitivity Analysis Results for the β

出较差的性能,可能是因为在该数据集中产品的退货记录较为分散,所以基于产品的相似度计算区分度不高。

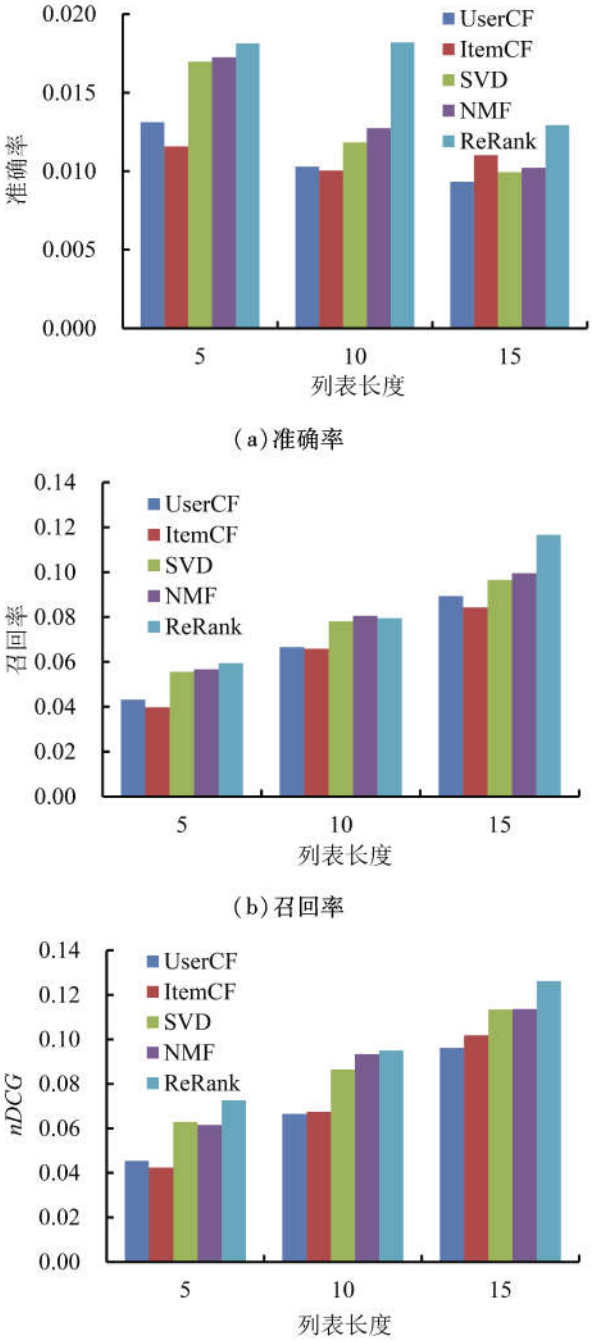
3.3.4　退货特征的预测能力分析

为了进一步分析融合到随机游走过程的各个退货特征对于退货风险的预测能力,分别在初始的随机游走模型中加入各个特征,得到各自的预测精度,见表3。在模型中加入所有特征后,各项预测指标均达到最高,而不加入任何退货特征的模型整体表现最差。单独加入用户的信用评分或产品价格均从较大程度上提升了算法的精度,并且偏重不同的精度指标,信用评分有效提升了准确率,产品价格提升了召回率。运费支付方式和质保证书也从一定程度上改进了算法的预测精度,但精度的提升幅度有限。分析结果再次表明,融合了退货特征的随机游走模型能对退货风险进行更细致的建模。

表3　不同退货特征的预测能力

Table 3　Predictive Power for Different Product Return Feature

退货特征	准确率	召回率	*nDCG*
所有特征	0.018	0.078	0.095
信用评分	0.013	0.062	0.074
产品价格	0.012	0.076	0.062
运费支付方	0.009	0.054	0.058
证书状态	0.008	0.054	0.062
无特征	0.009	0.051	0.055



(a)准确率

(b)召回率

(c) *nDCG*

图 9　不同算法的性能比较结果

Figure 9　Results for Comparing Performance for Different Algorithms

实际上,本研究提出的ReRank算法对于不同类

型的退货特征有较好的可扩展性,各类特征均可以根据相似性的测量融入到随机游走的先验信息中。

4 结论

4.1 研究结果

本研究聚焦于电子商务环境下的退货问题,针对电子商务企业的交易、用户和退货数据,提出一种分析和预测用户对于特定产品退货风险的方法。①退货行为中包含的用户和产品两种实体类型,通过引入二部图结构来组织历史退货记录,将问题形式化为二部图上的节点排序问题。②设计退货风险的随机游走过程,实现用户与产品退货风险的互相表示。基于实际退货数据的观测,发掘影响退货的各类特征属性,并将其转化为先验信息引入模型,有效引导退货风险在用户与产品间的游走过程。③通过在真实数据集上的实验表明,本研究提出的模型比其他方法具有更高的性能,并且相关退货特征的引入可以提升退货风险的预测精度。本研究主要适用于退货率较高且退货难度较低的电子商务环境。

4.2 理论和实践意义

本研究的意义包含两个方面。①提供了一种识别潜在高退货风险的用户和高风险产品的方法,对于电子商务企业的运营管理决策具有较强的实践意义。相关企业可以利用本研究提出的分析和预测方法对相关的用户购买各类产品时进行风险判断,有针对性地加强客户关系管理。同时可以加强对高退货风险产品的管理和规划,如采用加强包装、改善产品质量等方式,以规避退货风险。②本研究针对电子商务退货数据,创新性地将二部图随机游走模型应用到退货风险管理中,为电子商务领域相关研究提供一种新的视角,具有一定的理论意义。

4.3 研究的局限和不足

①受数据本身的限制,本研究采用的退货特征相对有限,因此仅针对部分用户和产品的相关特征进行融合。但实际上仍存在大量影响退货的因素,如产品的选择过程、产品退货的难易程度等,可以更有效地识别退货风险。虽然本算法对各类特征具有较强的可扩展性,但仍无法全面验证和分析退货特征对于风险的预测能力。②本研究仅针对截面时间上的退货数据进行分析,但实际上用户的退货行为和产品的被退货模式可能随时间发生变化,因此未来研究需对模型进行动态性的扩展。③后续研究可以结合一些行为学研究范式,补充个体用户对于电子商务环境下退货的主观认知,从而更好地揭示退货的管理意义。

参考文献:

[1] ANDERSON E T, HANSEN K, SIMESTER D. The option value of returns: theory and empirical evidence. ***Marketing Science***, 2009, 28(3): 405-423.

[2] LI Y, XU L, LI D. Examining relationships between the return policy, product quality, and pricing strategy in online direct selling. ***International Journal of Production Economics***, 2013, 144(2): 451-460.

[3] WALSH G, MÖHRING M. Effectiveness of product return-prevention instruments: empirical evidence. ***Electronic Markets***, 2017, 27(4): 341-350.

[4] 孙永波,李霞. 网购退货后续购买行为的实证研究. **企业经济**, 2017, 36(2): 149-155.
SUN Yongbo, LI Xia. An empirical study on the follow-up purchases of online shopping returns. ***Enterprise Economy***, 2017, 36(2): 149-155. (in Chinese)

[5] DE P, HU Y J, RAHMAN M S. Product-oriented web technologies and product returns: an exploratory study. ***Information System Research***, 2013, 24(4): 998-1010.

[6] FU Y, LIU G, PAPADIMITRIOU S, et al. Fused latent models for assessing product return propensity in online commerce. ***Decision Support Systems***, 2016, 91: 77-88.

[7] PASTERNACK B A. Optimal pricing and return policies for perishable commodities. ***Marketing Science***, 2008, 27(1): 133-140.

[8] 张霖霖,姚忠. 考虑顾客退货时在线企业的定价与订货策略. **管理科学学报**, 2013, 16(6): 10-21.
ZHANG Linlin, YAO Zhong. Pricing and order decisions with customer returns in online retailing. ***Journal of Management Sciences in China***, 2013, 16(6): 10-21. (in Chinese)

[9] 李勇建,许磊,杨晓丽. 产品预售、退货策略和消费者无缺陷退货行为. **南开管理评论**, 2012, 15(5): 105-113.
LI Yongjian, XU Lei, YANG Xiaoli. Advance selling, return policy and false failure return for a newsvendor retailer. ***Nankai Business Review***, 2012, 15(5): 105-113. (in Chinese)

[10] 孙军,徐路恒,刘宇. 退货问题下的在线零售商最优采购量研究. **管理科学**, 2014, 27(6): 114-120.
SUN Jun, XU Luheng, LIU Yu. Optimal purchase quantity of on-line retailers under returns issue. ***Journal of Management Science***, 2014, 27(6): 114-120. (in Chinese)

[11] 赵晓敏,高方方,林英晖. 基于顾客退货的闭环供应链运作绩效研究. **管理科学**, 2015, 28(1): 66-82.
ZHAO Xiaomin, GAO Fangfang, LIN Yinghui. Research on operational performance of a closed-loop supply chain with customer returns. ***Journal of Management Science***, 2015, 28(1): 66-82. (in Chinese)

[12] MUKHOPADHYAY S K, SETOPUTRO R. Optimal return policy and modular design for build-to-order products. ***Journal of Operations Management***, 2005, 23(5): 496-506.

[13] ANDERSON E T, HANSEN K, SIMESTER D. The option value of returns: theory and empirical evidence. ***Marketing Science***, 2009, 28(3): 405-423.

[14] 卢美丽,叶作亮,王芳. 考虑退货的在线零售价格和服务水平决策. **系统工程**, 2017, 35(1): 102-109.
LU Meili, YE Zuoliang, WANG Fang. Online retail prices and service level decision considering returns. ***Systems Engineering***, 2017, 35(1): 102-109. (in Chinese)

[15] 单汨源,江黄山,刘小红. 在线零售商盈利能力及其退货策略研究. **华东经济管理**, 2016, 30(11): 123-128.
SHAN Miyuan, JIANG Huangshan, LIU Xiaohong. Research on profitability and return policy of online retailers. ***East Chi-***

na Economic Management, 2016, 30(11): 123-128. (in Chinese)

[16] MOONESINGHE H D K, TAN P N. OutRank: a graph-based outlier detection framework using random walk. ***International Journal on Artificial Intelligence Tools***, 2008, 17(1): 19-36.

[17] BEUTEL A, XU W H, CURUSWAMI V, et al. CopyCatch: stopping group attacks by spotting lockstep behavior in social networks // ***Proceedings of the 22nd International Conference on World Wide Web***. Brazil, 2013: 119-130.

[18] ZHU Z, SU J, KONG L. Measuring influence in online social network based on the user-content bipartite graph. ***Computers in Human Behavior***, 2015, 52: 184-189.

[19] FOUSS F, PIROTTE A, RENDERS J M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. ***IEEE Transactions on Knowledge and Data Engineering***, 2007, 19(3): 355-369.

[20] HE X, GAO M, KAN M Y, et al. BiRank: towards ranking on bipartite graphs. ***IEEE Transactions on Knowledge and Data Engineering***, 2017, 29(1): 57-71.

[21] 蔡小雨,陈可佳,安琛. 采用群体信息的二部图链接预测方法. **计算机工程**, 2016, 42(10): 187-191.
CAI Xiaoyu, CHEN Kejia, AN Chen. Bipartite graph link prediction method using community information. ***Computer Engineering***, 2016, 42(10): 187-191. (in Chinese)

[22] 关雲菲. 改进的基于二部图网络结构的推荐算法. **信息技术**, 2015(9): 196-199.
GUAN Yunfei. Improved recommendation algorithm based on bipartite networks. ***Information Technology***, 2015(9): 196-199. (in Chinese)

[23] 黄熠姿,杨金鑫,孙维. 基于改进二部图与专家信任的混合推荐算法. **价值工程**, 2017, 36(19): 160-164.
HUANG Yizi, YANG Jinxin, SUN Wei. Research of hybrid recommendation algorithm based on improved bipartite network and expert trust. ***Value Engineering***, 2017, 36(19): 160-164. (in Chinese)

[24] 田东平. 融合PLSA和随机游走模型的自动图像标注. **小型微型计算机系统**, 2017, 38(8): 1899-1905.
TIAN Dongping. Integrating PLSA and random walk model for automatic image annotation. ***Journal of Chinese Computer Systems***, 2017, 38(8): 1899-1905. (in Chinese)

[25] LIU X, CHEUNG G, WU X, et al. Random walk graph laplacian-based smoothness prior for soft decoding of JPEG images. ***IEEE Transactions on Image Processing***, 2017, 26(2): 509-524.

[26] SHEN R, CHENG I, SHI J, et al. Generalized random walks for fusion of multi-exposure images. ***IEEE Transactions on Image Processing***, 2011, 20(12): 3634-3646.

[27] 李鹏,王斌,石志伟,等. Tag-TextRank:一种基于Tag的网页关键词抽取方法. **计算机研究与发展**, 2012, 49(11): 2344-2351.
LI Peng, WANG Bin, SHI Zhiwei, et al. Tag-TextRank: a webpage keyword extraction method based on Tags. ***Journal of Computer Research & Development***, 2012, 49(11): 2344-2351. (in Chinese)

[28] PUCCI A, GORI M, MAGGINI M. A random-walk based scoring algorithm applied to recommender engines // ***Advances in Web Mining and Web Usage Analysis***, 2007, 4811: 127-146.

[29] SHANG S, KULKARNI S R, CUFF P W, et al. A random-walk based model incorporating social information for recommendations // **2012** ***IEEE International Workshop on Machine Learning for Signal Processing***. Santander, Spain, 2012: 1-6.

[30] 施海鹰. 基于关联规则挖掘的分类随机游走算法. **计算机技术与发展**, 2017, 27(9): 7-11.
SHI Haiying. Random-walk classification algorithm with association rules mining. ***Computer Technology and Development***, 2017, 27(9): 7-11. (in Chinese)

[31] 张光前,白雪. 基于消费性格的新商品推荐方法. **管理科学**, 2015, 28(2): 60-68.
ZHANG Guangqian, BAI Xue. Method of new commodities recommendation based on consuming personalities. ***Journal of Management Science***, 2015, 28(2): 60-68. (in Chinese)

[32] JAMALI M, ESTER M. TrustWalker: a random walk model for combining trust-based and item-based recommendation // ***Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining***. Paris, France, 2009: 397-406.

[33] 张萌,南志红. 基于用户偏好的信任网络随机游走推荐模型. **计算机应用**, 2016, 36(12): 3363-3368.
ZHANG Meng, NAN Zhihong. Trust network random walk model based on user preferences. ***Journal of Computer Applications***, 2016, 36(12): 3363-3368. (in Chinese)

[34] MO Y, LI B, WANG B, et al. Event recommendation in social networks based on reverse random walk and participant scale control. ***Future Generation Computer Systems***, 2017, 79(1): 383-395.

[35] 曹云忠,邵培基,李良强. 基于信任随机游走模型的微博粉丝推荐. **系统管理学报**, 2017, 26(1): 117-123, 132.
CAO Yunzhong, SHAO Peiji, LI Liangqiang. Microblogging fans recommendation based on trust random walk model. ***Journal of Systems & Management***, 2017, 26(1): 117-123, 132. (in Chinese)

[36] 张怡文,王冉,程家兴. 基于用户兴趣度的改进二部图随机游走推荐方法. **计算机应用与软件**, 2015, 32(6): 76-79.
ZHANG Yiwen, WANG Ran, CHENG Jiaxing. Improved recommendation algorithm of bipartite graph random walk based on user interest degree. ***Computer Applications and Software***, 2015, 32(6): 76-79. (in Chinese)

[37] 李镇东,罗琦,施力力. 基于增加相似度系数的加权二部图推荐算法. **计算机科学**, 2016, 43(7): 259-264.
LI Zhendong, LUO Qi, SHI Lili. Weighted bipartite network recommendation algorithm based on increasing similarity coefficient. ***Computer Science***, 2016, 43(7): 259-264. (in Chinese)

[38] 杨华,周琪云,汤青,等. 混合图随机游走算法的商品推荐. **小型微型计算机系统**, 2016, 37(11): 2433-2436.
YANG Hua, ZHOU Qiyun, TANG Qing, et al. Hybrid graph random walk algorithm for commodity recommendation. ***Journal of Chinese Computer Systems***, 2016, 37(11): 2433-2436. (in Chinese)

[39] ZHOU D, HUANG J, SCHÖLKOPF B. Learning from labeled

and unlabeled data on a directed graph // ***Proceedings of the 22nd International Conference on Machine Learning***. Bonn, Germany, 2005:1036-1043.

[40] ZHOU D, SCHÖLKOPF B. Regularization on discrete spaces // KROPATSCH W, SABLATNIG R, HANBURY A. ***Pattern Recognition: 27th Annual Meeting of the German Association for Pattern Recognition***. Vienna, AustriaBerlin Heidelberg, 2005:361-368.

Risk Prediction for Product Return in Electronic Commerce Based on Random Walk

LIU Guannan[1], ZHANG Liang[1], MA Baojun[2]

1 School of Economics and Management, Beihang University, Beijing 100191, China

2 School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: Recently product return has become a focal issue in e-commerce platforms with the thorough development of e-business. The high ratio of product returns can cause additional costs in logistics, maintenance, etc., which can impact the normal operation of enterprises. Therefore, it is extremely necessary to prevent the product return risk and identify the return propensity for e-business to improve the decision-making for e-commerce operation. In the era of big data, e-commerce enterprises have accumulated large amount of heterogeneous data including sales, returns and customers, which can be utilized to mine consumers' purchase and return pattern, and further predict the return risks.

With respect to modeling the return risks in e-commerce, a bipartite network is introduced to organize the product return records, with the consumers and items representing the two types of nodes, and the edges representing the return event. Then, the prediction problem can be formulated as a ranking problem in the bipartite network. According to the structural characters of the returning consumers and returned products, random walk, which is a typical ranking method, is defined to represent the passing of risk information between consumers and products. The return risks of consumers can be represented as the products they have ever returned, while the return risk of products can be represented by the corresponding consumers. In addition, considering the sparsity issues in product return records, the innate features of the consumers and products are further incorporated by computing the similarity between the products and the product, and then the similarity is fed into the random walk as prior information. Thus, a prediction approach with the features fused is developed to improve the accuracy of prediction.

The model is validated on the real-world e-commerce product return data, which is obtained from an online merchant in Taobao. The experiments demonstrate the effectiveness of the proposed prediction method ReRank in comparison with other baseline methods including SVD, NMF, etc. Moreover, the experiments also show that the related features of both the consumers and products can improve the predictive power, among which the product warranty, product price can contribute significantly to the predictive accuracy.

The proposed approach is applicable for e-commerce enterprises. On one hand, the enterprises can utilize the approach to identify the return risks and enhance customer relationship management toward particular customers. On the other hand, they can improve the planning and management for products with high return risk and take measures such as improving the quality, strengthening the wrapping.

Keywords: electronic commerce; product return; bipartite network; random walk

Received Date: September 20th, 2017 **Accepted Date**: December 10th, 2017

Funded Project: Supported by the National Natural Science Foundation of China(71701007, 71772017, 71402007) and the Beijing Social Science Foundation(17GLB009)

Biography: LIU Guannan, doctor in management, is a lecturer in the School of Economics and Management at Beihang University. His research interests include data mining and business intelligence, and social network. His representative paper titled "Fused latent models for assessing product return propensity in online commerce" was published in the ***Decision Support Systems*** (Volume 91, 2016). E-mail: liugn@buaa.edu.cn

ZHANG Liang is a master degree candidate in the School of Economics and Management at Beihang University. His research interests include data mining and business intelligence, and e-commerce. E-mail: bhjg_zl@163.com

MA Baojun, doctor in management, is an associate professor in the School of Economics and Management at Beijing University of Posts and Telecommunications. His research interests cover data mining and business intelligence, big data analytics on mobile user behaviors, and policy informatics. His representative paper titled "Content & structure coverage: extracting a diverse information subset" was published in the ***INFORMS Journal on Computing*** (Issue 4, 2017). E-mail: mabaojun@bupt.edu.cn □