



基于多源大数据的 个性化推荐系统效果研究

姚凯¹, 涂平², 陈宇新³, 苏萌⁴

1 中央财经大学 商学院, 北京 100081

2 北京大学 光华管理学院, 北京 100871

3 上海纽约大学 商学部, 上海 200122

4 北京百分点信息科技有限公司, 北京 100101

摘要: 个性化推荐系统已成为各大电商向消费者提供个性化购物体验的重要工具之一, 通过推荐系统, 商家可以提高收入和消费者满意度。但传统推荐系统通常只利用消费者在当前网站的历史信息推荐个性化商品, 无法获得消费者在其他网站的数据来优化推荐效果。大数据时代, 一些第三方公司抓住机遇, 利用不同公司的多源大数据提供更好的个性化推荐服务。然而, 这种新型的推荐系统对消费者购物行为的影响存在极大的未知性。

探究基于多源大数据的个性化推荐系统对消费者购物行为的影响。为了建立推荐系统与消费者购物行为之间的因果关系, 采用实地实验有效地避免传统研究方法存在的内生性问题, 并具有较好的外部有效性。一方面, 基于内部数据和外部数据构造解释性变量, 探究内部数据特征和外部数据特征与推荐效果之间的关系; 另一方面, 通过检验消费者特征与内外部数据的推荐效果间的交互效应, 进一步分析外部数据和内部数据的推荐效果如何随消费者的特征变化, 帮助企业更好地利用多源大数据提升推荐效果。

研究表明, 基于内部数据的推荐系统能够显著提升消费者点击个性化推荐商品的概率, 可以降低消费者决策时间, 激励消费者浏览更多的商品。外部数据的推荐效果不仅与外部公司网站的用户数量相关, 也会受到外部网站与当前网站的关联程度的影响。消费者特征对基于内部数据和外部数据的推荐效果起调节作用, 如果消费者是当前网站的老用户, 利用该消费者在当前网站的内部数据提供个性化推荐的效果更佳。

通过分析基于多源大数据的推荐效果对消费者购物行为的影响, 进一步完善个性化推荐领域的理论框架。研究结果对如何利用多源数据构建更加有效的推荐系统具有重要指导价值, 并为不同网站之间的数据共享机制提供重要的管理建议。

关键词: 多源大数据; 个性化推荐; 实地实验; 电子商务; 互联网营销

中图分类号: F713.36 **文献标识码:** A **doi:** 10.3969/j.issn.1672-0334.2018.05.001

文章编号: 1672-0334(2018)05-0003-13

收稿日期: 2018-01-31 **修返日期:** 2018-06-28

基金项目: 国家自然科学基金(71332006, 71702208)

作者简介: 姚凯, 管理学博士, 中央财经大学商学院讲师, 研究方向为大数据营销、个性化营销和社会网络分析等, 代表性学术成果为“基于商品关联网络的销量预测方法”, 发表在2015年第4期《营销科学学报》, E-mail: yaokai@cufe.edu.cn

涂平, 社会学博士, 北京大学光华管理学院教授, 研究方向为新产品扩散模型、社会网络分析和顾客满意度等, 代表性学术成果为“产品属性对顾客满意影响的实证研究”, 发表在2007年第6期《管理科学》, E-mail: taping@gsm.pku.edu.cn

陈宇新, 管理学博士, 上海纽约大学商学部教授, 研究方向为数据驱动市场营销、网络市场营销和竞争策略等, 代表性学术成果为“Modeling credit card ‘share of wallet’: solving the incomplete information problem”, 发表在2012年第5期《Journal of Marketing Research》, E-mail: yc18@nyu.edu

苏萌, 管理学博士, 北京百分点信息科技有限公司董事长, 研究方向为客户终身价值和新品竞争等, 代表性学术成果为“Coupon trading and its impacts on consumer purchase and firm profits”, 发表在2014年第1期《Journal of Retailing》, E-mail: meng.su@baifendian.com

引言

个性化推荐系统已成为各大电商平台向消费者提供个性化购物体验的重要工具之一,超过77%的电商平台会对消费者的历史信息进行分析,估计消费者购物偏好并提供个性化推荐服务^[1]。电商通过个性化推荐系统,一方面可以提高消费者的转化率,另一方面能够提升消费者满意度和忠诚度^[2-3]。大数据时代,消费者除在当前网站上的消费信息外,可能也在其他网站存在大量购物信息。特别是当新的消费者在当前网站上没有任何历史信息时,他们在其他网站的偏好信息就显得尤为重要^[4]。即便消费者在当前网站没有历史数据,电商也可以利用他们在其他网站的数据估计消费者购物偏好,提高当前网站推荐商品的准确性。

然而,由于电商之间存在相互竞争或需要保护用户隐私,外部电商通常不愿意将消费者的数据分享给当前电商。就本研究所了解的,目前还没有利用消费者在其他电商的网购数据为当前电商网站提供个性化推荐的实证研究。本研究主要探讨一种新型的推荐系统对消费者购物行为的影响效果,该系统能够使用消费者在当前电商网站的内部数据(简称内部数据)和其他电商网站的外部数据(简称外部数据),为他们在当前网站购物过程中提供个性化商品推荐服务。本研究通过实地实验探索该推荐系统对消费者购物行为的影响,可以有效地避免传统研究方法存在的内生性问题以及缺乏外部有效性的缺点^[5]。该研究基于消费者在当前网站和外部网站的多源大数据进行建模分析,不但能够探究内部数据和外部数据中影响推荐效果的主要因素,而且完善了推荐系统如何影响消费者购买过程的理论体系,相关结论对电商的个性化营销和公司间制定数据共享策略具有重要的指导意义。

1 相关研究评述和研究假设

本研究探讨的个性化商品推荐属于个性化营销的研究范畴,已有大量关于个性化营销的研究主要关注公司与消费者之间个性化互动的效果。ANSARI et al.^[6]提出一种优化方法对邮件的布局和内容进行个性化变化,以提高消费者点击邮件中链接的概率;SIMESTER et al.^[7]研究发现,发送邮件的频率对点击率的影响随消费者特征的不同而发生变化;SAHNI et al.^[8]发现,促销邮件确实能够提高消费者的总消费金额,但大部分的收入不是来自促销券,而是由于促销邮件为公司产品起到了一种广告效果,进而提高了消费者的消费金额。与个性化邮件类似,在线精准广告也主要是通过分析消费者数据,为他们推送个性化的广告,以吸引他们打开广告链接并跳转到对应的网站去购买产品或者服务^[9-10]。广告主可以将消费者的行为数据或者商品信息整合到个性化广告系统中,提升广告效果^[11]。商家不仅可以通过在一个网站上打广告提高营业收入,还能通过广告联盟的方式获得新的消费者或保留现有的消费者^[12]。

本研究与个性化邮件或个性化广告的最大区别在于,这两种个性化营销方式都是在消费者离开商家的网站后为消费者展示邮件或广告。个性化邮件在消费者离开商家网站后,通过邮件中的链接将消费者重新跳转到商家网站。在线精准广告是消费者离开电商网站后,当他们访问其他网站(如新闻网站、社交媒体网站)时为他们显示之前在电商网站浏览的商品链接,以此跳转回原电商网站^[13]。而个性化推荐系统只为当前正在访问电商网站的消费者提供决策支持服务,并且推荐的内容都是电商网站上自己销售的商品^[14]。

早期关于个性化推荐的研究主要关注对个性化推荐算法的改进,张莉等^[15]对协同过滤算法进行改进,取得比传统的基于用户的协同过滤算法更优的推荐效果。为了深入了解推荐系统对消费者购物行为的影响,SENECAL et al.^[16]通过实验室实验发现,推荐系统的有效性受到产品类型和网站间相对独立性的影响。然而,这种实验室实验的方法得到的研究结果存在多方面的局限性,外部有效性较低,很难在实际应用环境中得到一致的结果。此外,部分研究在分析推荐系统的作用时,屏蔽控制组中消费者浏览网页中的推荐栏,造成研究结果出现偏差,因为推荐效果可能是由于是否在推荐栏区域显示商品导致,而不是个性化推荐系统产生的影响^[17]。PASANT^[18]使用关联数据(多源数据)提高消费者的购物体验。但这种关联数据需要采用统一的协议进行存储,很少有商业公司采用这种协议^[19],外部有效性较低。为了避免前人研究中存在的不足,本研究通过实地实验探究个性化推荐系统对消费者购物行为的影响,提高研究结果的外部有效性。同时,利用消费者在当前网站的内部数据和其他网站的外部数据为消费者提供个性化推荐服务,结合消费者行为理论深入探究该类推荐系统对消费者购物行为的影响。

1.1 内部数据的影响

整个网购过程中,消费者会在多个阶段响应电商网站上的个性化刺激,包括注意过程、认知过程、决策结果^[20]。前人研究主要通过实验室实验探究推荐系统对消费者购物行为的影响,但在实际应用过程中外部有效性较低^[17,21]。本研究基于消费者行为学相关理论,通过实地实验对推荐系统的效果进行实证分析。自我是心理学的一个基本概念,并在该学科已被广泛研究^[22]。自我的一个重要性质是自我参照效果,主要指人们对信息进行编码的结果取决于自我会多大程度在信息中得到暗示^[23]。它有利于人们将记忆编码得更加精细,并且更容易检索出来。在本研究的个性化推荐中,自我参照指消费者以前在该网站的购物经验,这类信息在消费者记忆中能够长期被获取。BARGH^[24]通过研究发现,人的注意力会自发地关注与自我相关的信息。由于自我参照具有容易检索的性质,使消费者对自我参照相关的回忆速度更快^[25]。因此,如果网页个性化内容与自

我相关,将降低消费者的搜索成本,消费者花费在决策上的时间会较少。相对于普通的商品,个性化推荐系统推荐的商品主要基于消费者的历史购物数据,与消费者的自我相关程度更高。为了研究推荐系统对消费者购物行为的影响,已有研究中将屏蔽推荐栏的实验组设定为控制组^[17],这样无法判断是否是由于推荐栏这一网页元素导致不同组中消费者购物行为的差异。因此,本研究在控制组采用随机推荐的方式,在网页同样位置显示同样数目的商品,避免结果出现偏差。基于自我参照理论,本研究提出假设。

H_{1a} 与随机推荐商品的方式相比,利用消费者的内部数据推荐个性化商品可以达到更高的点击率。

H_{1b} 与随机推荐商品的方式相比,如果消费者看到的商品是基于内部数据推荐产生的,他们花费在决策上的时间更短,点击推荐商品的速度更快。

除自我参照会影响消费者购物过程外,商品间的差异性也会影响消费者的决策过程^[26]。如果消费者连续浏览的商品之间的吸引力差异越小,他们终止当前的搜索行为的可能性就会越低。本研究中,推荐系统利用消费者历史数据推荐与消费者偏好相似的商品,这些商品之间的吸引力比较接近^[27]。因此,消费者在存在个性化推荐商品的情况下,吸引力相似的商品会造成消费者搜索更多的商品。此外,已有研究表明,商品间差异性较小的情况下,消费者难以得到他们比较喜欢的商品,并造成选择延迟^[28]。因此,本研究提出假设。

H₂ 与随机推荐商品的方式相比,消费者在推荐系统协助下会浏览更多基于内部数据产生的个性化网站内容。

1.2 外部数据的影响

除了自身内部数据源外,如果公司能从其他公司获取消费者的偏好信息,那么该公司将会获得更全面的消费者偏好,并提供更好的精准营销服务^[29]。大量研究表明,不同公司间的营销活动存在相互影响。以此为基础,本研究主要探究利用外部数据在公司网站上提供个性化推荐服务如何影响消费者在当前公司的购物行为。CHEN et al.^[30]的研究表明,如果公司与竞争对手之间共享消费者数据,将会同时提高两个公司获得消费者的能力,形成双赢。消费者触及能力较弱的小公司也能在分享数据的过程中获利。JENTZSCH et al.^[31]认为,外部数据的有效性取决于数据类型和消费者的偏好。本研究中,基于外部数据的推荐系统的效果与外部公司的自身特征以及当前公司与外部公司之间的关系密切相关。本研究利用公司拥有的消费者数量衡量公司的市场影响力,公司的市场影响力越大,消费者在网购过程中感知的风险越低,并且消费者对该公司的忠诚度也越高^[32]。消费者对公司的忠诚度越高,购买该公司商品的概率也越高,即该公司的内部数据更能够反映消费者偏好^[33]。如果消费者来自市场影响力大的外部公司,由于他们对于外部公司的忠诚度较高,

所以消费者点击当前网站商品的概率较低。但消费者在外部公司的历史数据能够更好地代表消费者的购物偏好,如果当前网站利用这类消费者在外部公司的历史数据来为他们提供个性化推荐服务,他们点击个性化推荐商品的概率更高。因此,本研究提出假设。

H₃ 根据外部公司市场影响力特征,如果消费者数据来自市场影响力较大的外部公司,那么消费者点击当前网站上商品的概率较低。但如果利用他们在外部公司的数据提供个性化推荐服务,消费者点击个性化推荐商品的概率较高。

外部数据的推荐效果,除与外部公司的特征紧密相关外,外部公司与当前公司之间的关联度也会影响最终推荐效果。WANG et al.^[34]认为,公司之间如果具有较强的网络效应,消费者在不同公司间的购买行为会相互影响。本研究使用两个公司之间的共同用户数量表示公司间的关联度,如果两个公司之间的关联度越高,即两个公司之间存在大量共同用户,那么两个公司为消费者提供的商品或服务相似^[35]。如果消费者来自关联度较高的外部公司,由于当前网站提供的商品或服务与他之前浏览的网站内容相似,那么当前网站能够提供与该消费者之前偏好接近的商品。在这种情况下,消费者点击当前网站上商品的概率较高。但如果使用该消费者在外部网站的历史数据提供个性化推荐服务,该消费者点击由此产生的推荐结果的概率反而较低。由于消费者是离开关联度高的外部网站后来到当前网站,表明该消费者并不满意外部网站的商品。而当前网站与外部网站关联度高,如果再用该消费者不喜欢的偏好信息推荐商品,消费者对推荐结果的点击率会比较低。根据当前公司与外部公司间的关系,本研究提出假设。

H₄ 如果消费者来自与当前网站关联度高的外部公司,他们点击当前网站商品的概率较高。如果利用他们在外部网站的数据提供个性化推荐服务,消费者点击由此产生的个性化推荐商品的概率较低。

2 推荐系统和实验设计

个性化推荐是过去20年中计算机科学领域发展出来的一个热门研究话题。然而,已有推荐系统的研究工作主要集中在开发和评估不同的个性化推荐算法,以此为消费者产生个性化的商品建议^[36-37]。考虑到各种个性化算法的有效性和流行程度,一种简单而被广泛使用的推荐系统分类方法是将推荐算法分为基于内容的推荐方法和基于协同过滤的推荐方法。基于内容(content based, CB)的推荐系统利用产品信息(如商品名、书籍作者等)推荐与消费者的偏好类似的其他商品。基于内容的推荐系统在媒体或音乐这类网站被广泛应用,因为其中的商品具有大量的信息^[38],如Reel.com。与基于内容的推荐方法相对立的是协同过滤方法(collaborative filtering,

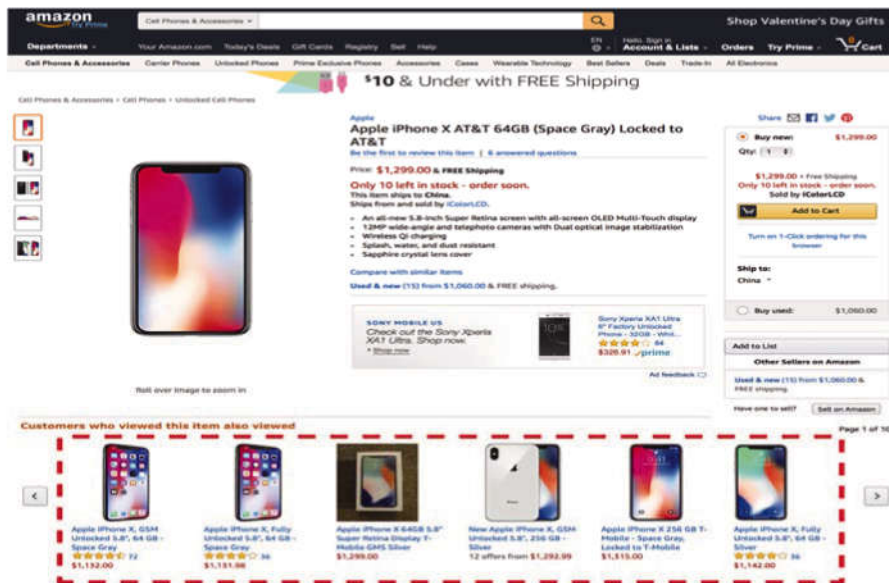


图1 亚马逊网站的推荐栏
Figure 1 Recommendation Bar of Amazon

CF),该方法不依赖于产品的属性信息,而是利用其他消费者的偏好信息识别当前消费者有可能购买的商品^[39-40]。具体地,有两种协同过滤方法,即基于商品的协同过滤和基于用户的协同过滤^[41]。本研究采用基于用户的协同过滤算法提供个性化推荐服务,该方法的核心思想是利用消费者的购物历史信息计算用户间的相似度,然后根据与当前消费者比较相似的其他消费者的偏好信息估计当前消费者可能喜欢的商品,进而为当前消费者生成个性化推荐商品。考虑到推荐系统的推荐过程较为复杂,为了使读者更好地了解推荐系统,本研究对个性化推荐产生过程和实验设计进行简要描述。

2.1 推荐过程

通常情况下,消费者在网站推荐栏内看到的个性化推荐商品主要通过3个阶段得到。第1阶段,推荐请求阶段,当消费者打开网页时,网页中的脚本将为消费者发送个性化商品推荐请求到服务器端的推荐系统。第2阶段,推荐商品产生阶段,推荐系统接收到请求之后,根据消费者的偏好信息产生推荐结果并将结果返回刚才打开的网页。第3阶段,推荐曝光阶段,推荐系统产生的个性化产品展示在网页的推荐栏中。例如,图1给出亚马逊商品信息页,消费者在浏览商品信息的同时,系统在商品下方会推荐消费者可能喜欢的其他商品。本研究中,受屏幕大小的限制,实验中的消费者只看到产品的描述信息,而看不到推荐栏。只有当消费者向下滑动网页时,推荐栏才会出现在消费者视野中。为了记录消费者是否看到推荐栏,网页中的JavaScript脚本会在推荐栏出现在消费者视野时发送一条信息到服务器,记录推荐栏的曝光情况。如果消费者在购物过程中没有看到推荐栏,那么推荐系统对消费者的购买行为不产生影响。最后是消费者响应阶段,当消费者看

到推荐栏中的商品后,他们可以选择点击喜欢的商品或浏览网页中的其他商品。如果消费者点击了推荐栏中的商品,他们将跳转到商品介绍页,并浏览该商品的详细信息。消费者可以购买他们喜欢的商品或离开该网站,消费者的每个行为信息都会被记录到服务器,每条记录中包含了商品特征和时间戳等信息。

2.2 实验设计

为了得到推荐系统影响消费者购物行为的因果关系,本研究采用实地实验探究推荐系统对消费者购物行为的影响,避免实验室实验缺乏外部有效性等问题。本研究使用Cookie ID标识每个消费者,使用Cookie的好处之一是即便消费者没有用自己的账号登陆网站,推荐系统依然可以通过匿名的方式识别他们的历史信息 and 偏好。此外,Cookie不包含消费者的任何人口统计学信息(如姓名、住址等),有利于保护消费者隐私。本实验部署在一个卖包的电商网站,该网站在实施实验前1个月平均每天有83 000位不同消费者在网站上产生20万次的点击行为。实地实验部署在网站首页,可以有效地控制内生性问题。因为实验中的推荐系统实时为消费者提供个性化推荐商品,一旦消费者在当前购物过程中浏览过任何商品信息,消费者的偏好便会被更新,后面看到的推荐商品也会随之发生改变。所以,本研究将实验控制放在在首页,能够在消费者还没有看过任何商品信息的情况下就使用历史数据为他们提供商品推荐。除了使用消费者在当前网站的内部数据提供个性化推荐服务外,本研究还使用消费者在其他8个卖包网站的数据为消费者在当前网站产生个性化推荐商品。由于外部网站数据量较多,为了保证推荐速度,实验过程中,推荐系统只利用消费者最近一次购物的网站数据提供个性化推荐服务。图2给出实验期

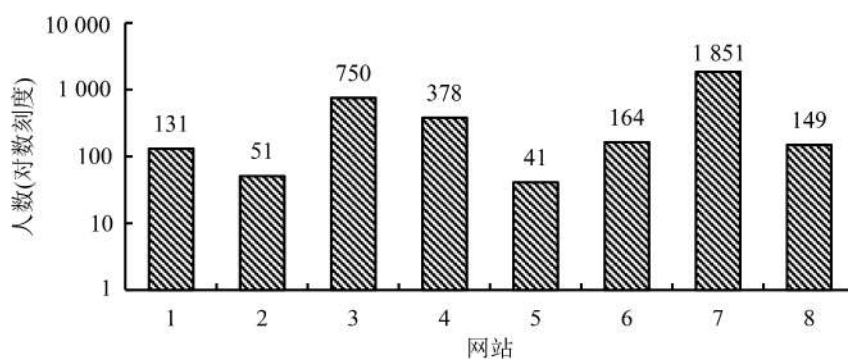


图2 来自外部网站的消费者数量

Figure 2 Number of Consumers from Outside Websites

间在外部8个网站上存在消费历史数据的消费者数量,纵轴用对数刻度表示不同网站的人数,可以看到,来自第7个网站的消费者最多,共1 851人,而来自第5个网站的消费者最少,仅有41人。

当消费者访问网站首页时,他们被随机分配到不同的实验组中,本研究通过设置不同组中推荐系统使用的推荐算法操控实验。实验中,消费者被随机分为3组,每组中的推荐栏大小和位置相同,区别是推荐栏内显示的商品是根据不同的推荐规则产生。图3为实验设计示意图,第1组为随机组,该组中消费者看到的推荐结果使用随机算法产生,即不使用个性化推荐。第2组中展示给消费者的推荐结果由协同过滤算法产生,主要利用消费者在当前网站的内部数据。第3组中,消费者看到的个性化推荐结果通过基于内容的推荐方法产生,主要利用消费者在其他网站的外部数据。由于3个组中的推荐结果同时受数据源和推荐方法的影响,为了便于理解,下面对每个实验组中使用的推荐规则和数据源进行详细描述。

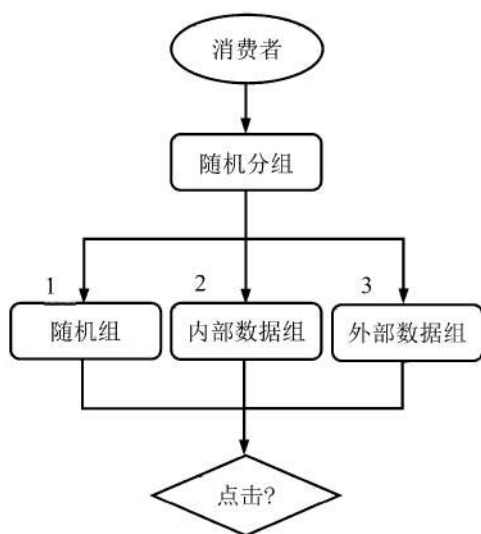


图3 实地实验示意图

Figure 3 Diagram of Field Experiment

(1)随机组。该组为实验中的第1组,也是该研究的控制组,该组中的推荐系统随机选择10个热门商品推荐给消费者,每次出现的商品和顺序都不一样。在正式开展实地实验前,本研究做了1次准实验,结果表明这种随机推荐的方法比按照商品热门程度推荐的方法效果好,而后者正是该网站以前推荐商品的惯用方式(按照商品热门程度推荐,是按商品销量排序,向消费者推荐前10个商品,短期内所推荐的商品和顺序不变)。本研究将该组作为控制组,以此减少该实验对公司造成的损失。与已有研究中屏蔽控制组推荐栏的方式不同^[17],本研究将随机组作为控制组能够排除是否显示推荐栏这个网页元素导致推荐效果的差异,最终得到推荐系统的真实作用。

(2)内部数据组。第2个实验组中,本研究利用消费者在当前网站的内部数据产生个性化推荐结果,采用协同过滤算法。首先根据消费者的历史购物数据计算消费者之间的相似度,然后利用其他消费者的偏好信息估计当前消费者可能喜欢的商品,进而为当前消费者推荐商品。如果消费者是第1次访问当前网站(即在当前网站不存在任何历史数据)或协同过滤算法无法产生足够的推荐商品,推荐系统将利用第1组中使用的随机推荐的方法补充推荐结果,以保证不同实验组中推荐栏内显示的商品数量相等。

(3)外部数据组。第3个实验组中,推荐系统主要基于消费者在其他网站的外部数据产生个性化推荐商品,采用基于内容的个性化推荐算法。由于不同网站间消费者不同,不能对消费者的外部数据使用协同过滤算法。因此,本研究利用消费者在外部网站的数据提供个性化推荐时,使用基于内容的推荐算法为消费者在当前网站提供个性化服务。基于内容的推荐算法利用消费者在外部网站浏览的商品属性信息(如类别、品牌或名称)推荐当前网站上类似的商品。如果当前网站上没有找到相似的商品,借鉴第2组的解决方法,使用随机推荐的方法对推荐结果进行补充,保证各组中推荐栏内显示的商品数量相等。

表1 样本描述性统计结果
Table 1 Descriptive Statistics Results for Samples

实验组	推荐规则	消费者数量	快速关闭	快速切换	最终用户	点击次数	点击率/%	看到推荐栏	条件点击率/%
随机组	随机推荐	31 639	180	5 991	25 468	359	1.410	16 210	2.215
内部数据组	协同过滤	31 495	173	6 117	25 205	599	2.377	16 083	3.724
外部数据组	基于内容	31 107	156	6 057	24 894	422	1.695	15 815	2.668
合计		94 241	509	18 165	75 567	1 380	1.826	48 108	2.869

3 数据描述

3.1 数据描述和清理

本研究中的实地实验由一家提供推荐服务的第三方公司配合完成,该公司为超过1 000家在线零售商提供过推荐系统服务。目前,已有约6亿个Cookie和超过300 TB(1 TB=1000 GB)的数据存储在基于Hadoop分布式文件系统服务器。由于本研究中的数据量很大,为了有效地节省构造变量的计算时间,使用数据仓库管理工具Hive提取和构造分析中需要的相关变量,分布式地使用数百台服务器同时完成计算任务,使需要很多天才能完成的计算任务缩减到几个小时。

从2014年8月29日至2014年9月5日,实验在网站首页持续开展8天,为了减小实验对电商造成的负面影响,实验过程中将该网站总用户的20%随机分配到3个实验组中,其他用户继续使用网站原有的推荐规则。如实验设计中所述,每位消费者登陆到网站后会被随机分配到3个实验组中,以此避免样本选择偏差。如果消费者在试验期间多次访问网站,该消费者会被分配到之前的组中。然后,系统会记录每位消费者在网站的所有行为信息,保存到Hadoop数据仓库中。实验中采集的数据主要由3个数据集构成,第1个是产品信息数据集,包含每个产品的详细信息,如产品ID、类别、品牌、名称和价格等。第2个是点击流数据集,记录消费者的每个在线行为,如浏览和购买行为。第3个是实验控制数据集,包含每个消费者被分配到的实验组的具体信息,以此保证消费者如果在实验期间再次返回网站,仍然会被分配到之前的实验组。在对消费者行为分析和建模过程中,可以使用每个数据集的关键字将3个数据集整合起来构造相应的变量。例如,假设有得到每次会话期间每位消费者的点击行为,可以将实验控制数据集和点击流数据集通过消费者的Cookie ID和会话ID进行整合,这样就可以知道消费者每次会话内的点击行为。如果希望比较不同实验组中每一位消费者浏览或购买了多少产品,可以将这3个数据集合并,以满足分析需求。详细的变量构建过程参考3.2。

尽管有些消费者向服务器发送了推荐请求,但他们的网络购物行为存在异常,如果忽略他们的影响,对推荐系统效果分析将会出现估计偏差。首先,

部分消费者虽然登录了网站,但服务器上只有推荐请求的数据,而没有推荐结果被展示的数据,因为这些消费者在网页完全打开之前就关闭了浏览器。其次,很多消费者是通过搜索引擎登陆到当前网站,可能除了当前网站还有很多同类网站吸引消费者,导致部分消费者打开当前网站后,没有点击任何商品就离开了当前网站。最后,在实验期间,有些消费者可能再次回到当前网站,但他们上一次的网购数据会影响下一次购物的推荐结果。从DIAS et al.^[2]的研究可知,推荐系统不仅提高了电商的收入,同时也刺激消费者再次返回网站。如果忽略该影响,将会错误地估计推荐系统对消费者购物行为的真实影响。

表1给出实验期间每组消费者的样本信息。由表1可知,每组大约有31 000位消费者访问过该网站,过滤掉登录网站就离开和网页还未显示就关闭的用户,每组剩下约25 000位消费者。造成如此多异常用户的原因可能是投放广告的效果不好或网页设计的吸引力不强,所以消费者会快速离开。因此,公司应检查自身广告投放和网站布局的效果,以便能够吸引更多的消费者访问该网站。从3组消费者的点击次数可知,随机组的点击率最低,为1.410%;内部数据组的点击率最高,为2.377%;外部数据组的点击率为1.695%。然而,该结果仍然存在偏差。本研究中,消费者需要滚动网页才能看到推荐栏,而有些消费者从未看到过推荐栏,这样会低估推荐系统的作用。可以发现,仅有63.663% ($\frac{48\ 108}{75\ 567}$)的消费者看到了推荐栏。如果过滤掉没有看到推荐栏的消费者,可以得到3组消费者的点击率分别为2.215%、3.724%和2.668%。该结果表明,与第1组随机推荐方法的点击率2.215%相比,利用内部数据组产生个性化推荐结果的系统确实能够提高消费者的点击率,达到3.724%。此外,在当前网站利用消费者的外部数据为他们提供个性化推荐也能得到比随机组高的点击率,为2.668%。

然而,表1中的描述性统计结果无法清楚地解释使用不同数据源的推荐系统与消费者的购物行为之间的因果关系。因此,本研究基于消费者在当前网站和8个外部网站的数据构建解释变量,对消费者网购行为进行建模,深入剖析推荐系统如何影响消费

者购买行为。

3.2 变量构建和描述

为了更好地了解本研究中的数据结构,图4举例给出消费者在购物过程中产生的点击流数据。为了深入探究推荐系统对消费者购物行为的影响,本研究对点击流数据进行处理,构建下列变量用于建模分析。

用户 ID	一次上网过程
1	H C P P O P H P P P P P P O P P E
2	H P P P H P O E
3	P P P H C P H P E
4	H P P O P O H P H P P E
5	P P P O H C P E
6	H P O P P H P P E

注: H为首页, C为点击推荐栏, P为产品页, O为订购, E为退出。

图4 点击流数据示意图

Figure 4 Diagram of the Click-stream Data

①显示, 0-1 变量, 表示消费者是否能够看到推荐栏, 如果消费者能看到推荐栏, 取值为1, 否则取值为0。

②点击, 0-1 变量, 用来测量推荐系统有效性的主要因变量, 如果消费者点击了推荐栏, 取值为1, 否则取值为0。

③老用户, 0-1 变量, 表示消费者是否为当前网站的老用户, 如果用户是老用户, 取值为1, 否则取值为0。

④商品数量, 测量消费者在网购这段时间内浏览过多少个不同的商品。

⑤点击速度, 表示从消费者看到推荐栏到点击它之间等待的时间, 消费者可能在一次购物过程中多次看到推荐栏, 点击速度计算的是从消费者看到推荐栏开始至消费者最近一次点击推荐栏的时间。因此需要对图4中点击数据流进行拆分, 得到该变量精确的测量值, 这也是本研究用Hadoop集群进行数据处理的重要原因之一。

为了判断哪些因素影响推荐系统的有效性, 本研究利用当前网站的内部数据和外部网站的数据构建以下解释变量。

①外部用户, 0-1 变量, 表示消费者是否同时是外部网站的用户。如果该用户同时是外部网站用户, 取值为1, 否则取值为0。

②最近访问时间, 表示消费者最近一次访问当前网站到现在的时间长度。

③访问频率, 用消费者访问当前网站的天数表示访问频率。

④外部公司影响力, 表示外部公司的市场影响

力, 本研究用外部网站的用户数量测量, 为了避免该变量的估计系数太小, 统计结果为真实人数除以1 000。

⑤外部公司关联度, 表示当前公司与外部公司之间关联程度, 用两个公司的共同用户数量表示, 为了避免该变量估计系数太小, 统计结果为真实人数除以1 000。

⑥第*i*组, 表示消费者被分配到第*i*个组中, 本研究中的*i*取值范围为1、2、3, 分别对应随机组、内部数据组和外部数据组。在建模过程中, 分别用哑变量随机组、哑变量内部数据组和哑变量外部数据组表示消费者属于哪个组, 用户属于该组取值为1, 否则取值为0。

⑦星期几, 分类变量, 用来指示当天是一周内的星期几。

⑧小时, 表示消费者在一天的具体访问小时数。

表2给出相关变量的描述性统计结果, 以消费者每次会话为单位统计消费者购买行为特征。由表2可知, 消费者看到推荐栏的概率均值为0.637, 并且所有消费者点击推荐栏的概率均值为0.018, 如果可以将消费者看到推荐栏的概率提高一点, 相对于原来的情况, 可以在很大程度上提高消费者点击推荐栏的概率。平均只有0.156的用户在当前网站存在历史数据, 即如果可以使用消费者的外部数据提供个性化推荐, 将有很大一部分在当前网站没有历史数据的消费者可以从中受益。从消费者的购物特点可

表2 变量描述性统计结果

Table 2 Descriptive Statistics Results for Variables

	观测数	均值	标准差	最小值	最大值
会话层面的变量					
显示	75 567	0.637	0.481	0	1
点击	75 567	0.018	0.134	0	1
老用户	75 567	0.156	0.363	0	1
商品数量	75 567	3.435	7.690	0	239
点击速度	961	294.335	1 204.314	4	21 542
外部用户	75 567	0.109	0.312	0	1
老用户特征					
最近访问时间	11 824	34.731	42.605	1	186
访问频率	11 824	4.137	8.360	1	172
外部公司特征					
外部公司影响力	8 236	1 385.945	962.640	8.678	2 548.944
外部公司关联度	8 236	6.627	3.928	0.090	11.165

知,消费者平均浏览的商品个数为3.435个,停留795.680秒。其中,只有0.109的用户存在外部数据,如果能够获取更多外部网站的数据,拥有外部数据的人数将更多。老用户平均最近购买时间为34.731天,平均访问4.137次。外部公司的平均用户数量为1385945人,外部网站与当前网站的平均共同用户数量为6627人。

4 结果和分析

4.1 内部数据的推荐效果

推荐系统的主要作用是协助消费者购买,为他们推荐与之偏好相关的商品。BARGH^[24]认为,人的注意力会自发地关注与自我关联的信息。由于推荐系统的协助,消费者看到的推荐商品与他们的偏好紧密相关,这样可以缩短消费者决策时间,他们会在较短的时间内给予反馈^[17]。本研究中,消费者的点击速度通过消费者看到推荐栏至消费者点击个性化推荐商品之间的时间测量。由于在一次购物过程中,消费者可能多次点击推荐栏,并且前一次点击结果可能会影响下一次推荐结果。为了避免内生性问题,本研究仅测量消费者在一个购物会话中第一次看到推荐栏的反馈时间。

图5给出随机组和内部数据组中消费者看到推荐栏中商品后的响应速度,纵轴表示消费者从看到推荐栏到点击推荐栏的时间,单位为秒。随机组(控制组)的平均响应速度为106.024秒,比内部数据组的平均响应速度(84.469秒)更长,并且两组之间的响应速度存在显著差异, $p < 0.010$ 。由于内部数据组中的个性化推荐系统根据消费者历史数据推荐商品,与消费者自我相关程度高,使消费者点击推荐栏的响应时间更短。因此, H_{1b} 得到验证。

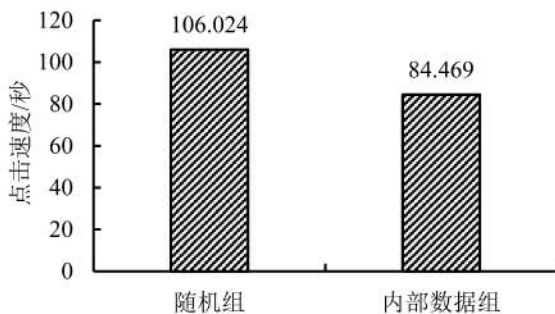


图5 响应速度分析结果

Figure 5 Analysis Results for Response Speed

商品间的差异性也会影响消费者的决策过程,如果连续浏览的商品之间的吸引力差异较小,消费者需要找更多的商品进行对比,终止搜索过程的概率更低。本研究中,与控制组中随机产生的推荐结果相比较,内部数据组中使用个性化推荐算法产生与消费者偏好接近的商品,所以推荐商品之间的差异性较小。因此,消费者看到含有个性化推荐商品的推荐栏时,比较难选出自己最喜欢的商品,所以他

们会比较更多的商品。图6给出消费者浏览商品数量的分析结果,纵轴为消费者浏览商品的数量。由图6可知,在没有点击推荐栏内商品的情况下,内部数据组消费者浏览的商品数量为2.503,随机组消费者浏览的商品数量为2.597,且两组均值差异不显著。然而,当两组消费者点击推荐栏中的商品后,如果他们点击的商品是相似度较高的个性化推荐结果,内部数据组消费者会显著地比随机组消费者浏览更多的商品,浏览的商品数量分别为7.455和6.572, $p < 0.010$ 。因此, H_2 得到验证。

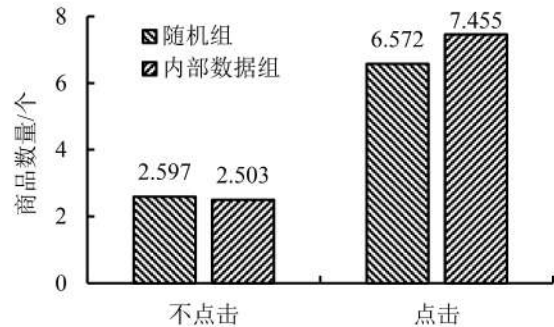


图6 商品数量分析结果

Figure 6 Analysis Results for Number of Products

4.2 外部数据的推荐效果

除了研究基于内部数据的推荐系统如何影响消费者的购物行为,本研究也分析基于消费者外部数据的推荐系统对他们在当前网站上购物行为的影响。表3给出基于外部数据的推荐系统对消费者购物行为的影响,因变量为消费者点击推荐商品。模型1检验使用外部数据对推荐商品点击率的影响,模型2检验外部公司特征对外部数据推荐效果的影响,模型3在模型2的基础上加入外部公司特征与外部数据组的交互项,检验外部公司特征对消费者点击率的影响。由模型1可知,外部数据组的系数为0.079, $p < 0.010$,表明利用外部数据为消费者提供个性化推荐,能够显著提高消费者点击推荐栏的概率。模型2中外外部公司影响力的系数为-0.0003, $p < 0.100$,即如果消费者来自市场影响力大的外部公司,他们点击当前网站商品的概率较低。该结果可能的解释是,若外部公司具有较高的市场影响力,消费者对该企业的忠诚度更高,即他们选择外部公司商品的概率更高,而点击当前网站上商品的概率较低。此外,外部公司关联度的估计系数为0.064, $p < 0.100$,表明如果消费者来自与当前网站的公司关联度较高的外部网站,他们点击当前网站商品的概率较高。该结果的一种解释为,由于外部网站的公司与当前网站的公司关联度高,当前网站提供的商品与之前浏览的外部网站内容相似。既然消费者对外部网站的内容感兴趣,他们也很可能对当前网站的商品感兴趣。因此,如果消费者来自与当前网站关联度高的外部

表3 外部数据推荐效果分析结果
Table 3 Analysis Results of Recommendation Effectiveness for External Data

	因变量:点击		
	模型1	模型2	模型3
截距项	-2.011*** (0.022)	-2.020*** (0.023)	-2.027*** (0.023)
随机组	-	-	-
外部数据组	0.079*** (0.030)	0.080*** (0.030)	0.093*** (0.032)
外部公司影响力		-0.0003* (0.0002)	-0.0005** (0.0002)
外部公司关联度		0.064* (0.034)	0.116*** (0.044)
外部公司影响力 × 外部数据组			0.001 (0.0003)
外部公司关联度 × 外部数据组			-0.120* (0.069)
观测数	32 019	32 019	32 019
自由度	2	4	6
对数似然函数值	-3 668.178	-3 665.858	-3 664.174
AIC	7 340.356	7 339.717	7 340.347
BIC	7 357.104	7 373.213	7 390.592

注:分析中只用到随机组合外部数据组中看到推荐栏的用户数据,因此观测数为32 019。括号中的数据为标准差,*为 $p < 0.100$,**为 $p < 0.050$,***为 $p < 0.010$,下同。

网站,他们在当前网站点击商品的概率较高。

与模型2的结果相比,模型3中相同变量的估计系数符号保持一致。外部公司影响力与外部数据组的交互项系数为0.001,表示如果使用来自具有较高市场影响力的外部公司数据在当前网站为消费者提供个性化推荐服务,有利于提高推荐系统推荐的商品的点击率。虽然该结果与假设一致,但统计指标不显著, H_3 部分得到验证。外部公司关联度与外部数据组的交互项系数为-0.120, $p < 0.100$,表明如果推荐系统使用消费者在关联度高的外部网站数据产生个性化推荐商品,消费者点击推荐的概率反而较低。导致该结果的一种解释为,既然消费者已经离开了与当前网站相似的外部网站,说明外部网站的商品没有满足他们的偏好。倘若当前网站仍然使用消费者不满意的商品信息推荐商品,他们点击推荐栏的概率会较低。结合模型2中外部网站关联度的估计结果可知, H_4 得到验证。

4.3 消费者异质性

为了深入探究个性化推荐系统效果的影响因

素,本研究利用Probit模型对消费者的点击行为进行建模,表4给出分析结果,因变量为消费者点击推荐栏。模型4中仅包含消费者分组信息,检验使用不同数据对推荐效果的影响,可以发现,内部数据组和外部数据组消费者点击推荐栏的概率显著高于没有个性化推荐服务的随机组消费者。相对于随机为消费者推荐商品这一常用的策略,利用消费者内部数据和外部数据都能显著提高消费者点击个性化推荐商品的概率。因此, H_{1a} 得到验证。模型5控制了星期几和小时的固定效应,与模型4相比发现,内部数据组的系数为0.228,外部数据组的系数为0.078,即内部数据组的推荐效果整体优于外部数据组。如表1所给出的,内部数据组与外部数据组的点击率分别为3.724%和2.668%, $p < 0.010$ 。从表2外部用户的统计结果可知,仅有10%左右的用户有外部数据。所以,尽管使用外部数据推荐的效果比随机推荐的效果好,但比使用内部数据推荐的效果差。

模型6加入消费者特征变量(仅有老用户的相关数据),检验推荐系统对不同消费者购物行为的影响差异,老用户的系数为0.372, $p < 0.010$,即老用户比新用户点击个性化推荐商品的概率更高。该结果表明,老用户在网站上存在历史信息,可以用来估计他们的购物偏好,提供更准确的推荐结果。最近访问时间的系数为-0.001, $p < 0.050$,表明如果消费者离开网站很长一段时间,他们点击推荐商品的概率较低。访问频率的系数为-0.002,表明如果消费者访问频次太高,点击推荐栏的概率会更低。一种可能的解释是,这类消费者对网站布局比较熟悉,能够自己找到想要的商品,而不需要推荐栏,所以他们点击推荐商品的概率较低。为了探究什么情况下内部数据与外部数据的推荐效果有所差异,模型7将内部数据组作为控制组进行建模分析,探究消费者异质性如何影响不同数据源的推荐效果。可以发现,外部数据组的估计系数为-0.093, $p < 0.010$,说明外部数据组消费者比内部数据组消费者点击个性化推荐栏的概率低;在模型中加入老用户与外部数据组的交互项,其系数为-0.219, $p < 0.010$,表明如果消费者是当前网站的老用户,使用他们的外部数据提供个性化推荐服务,点击推荐栏的概率较低。因此,公司在个性化实践过程中需要考虑消费者异质性,结合数据源和用户特征提供更精准的个性化推荐服务。

5 结论

本研究通过实地实验,探究基于内部数据和外部数据的个性化推荐系统对消费者购物行为的影响。一方面可以避免传统实验室研究中外部有效性低的问题,另一方面可以得到推荐系统与消费者购物行为之间的因果关系。

5.1 研究结果

(1)根据消费者在网站的内外部数据为消费者产生的个性化推荐商品与消费者自我相关程度高,因此消费者点击由此产生的推荐商品的概率比随机推荐

表4 消费者点击行为分析结果
Table 4 Analysis Results for Consumers' Click Behaviors

	因变量:点击			
	模型4	模型5	模型6	模型7
截距项	-2.011*** (0.022)	-1.921*** (0.111)	-2.007*** (0.112)	-1.938*** (0.141)
随机组	-	-	-	-
内部数据组	0.228*** (0.029)	0.228*** (0.029)	0.223*** (0.029)	
外部数据组	0.079*** (0.030)	0.078*** (0.030)	0.078*** (0.030)	-0.093*** (0.032)
老用户			0.372*** (0.039)	0.535*** (0.053)
最近访问时间			-0.001** (0.001)	-0.002** (0.001)
访问频率			-0.002*** (0.003)	-0.001 (0.004)
老用户×外部数据组				-0.219*** (0.066)
星期几	未控制	控制	控制	控制
小时	未控制	控制	控制	控制
观测数	48 102	48 102	48 102	27 637
自由度	3	33	36	36
对数似然函数值	-6 226.899	-6 210.680	-6 148.739	-3 116.425
AIC	12 459.800	12 487.360	12 369.480	6 304.850
BIC	12 486.140	12 777.140	12 685.600	6 601.019

注:模型4~模型仅使用了3组中看到推荐栏的消费者数据,因有个别数据缺失,所以观测数为48 102;模型7中仅保留了内部数据组和外部数据组的用户,观测数为27 637。

的商品更高。由于消费者的注意力会自发关注与自我相关的内容,花费在决策上面的时间会更短,所以消费者点击个性化推荐的速度会更快。

(2)基于内部数据推荐的个性化商品之间差异较小,消费者需要对比更多的商品才会停止搜索过程。因此,为消费者提供个性化推荐的情况下,消费者会浏览更多的商品。

(3)外部数据对消费者的推荐效果与外部公司的市场影响力密切相关,如果外部公司的消费者数量比较多,这类公司的消费者在当前网站点击商品的概率相对较低。

(4)基于外部数据的推荐系统的效果与外部公司网站与当前公司网站之间的相互关系有关,如果外部公司网站与当前公司网站存在大量共同用户,来自这类外部公司的消费者在当前网站点击的概率较高,但假如利用他们的外部数据产生个性化推荐,消

费者点击推荐商品的概率反而更低。

(5)消费者特征对基于内部数据和外部数据的推荐效果起调节作用。具体而言,不管是利用内部数据还是外部数据,消费者点击推荐商品的概率都会高于采用随机推荐的结果。但如果消费者是一位老用户,利用该消费者在网站的内外部数据提供个性化推荐服务取得的效果优于利用他的外部数据推荐个性化商品的效果。

5.2 理论意义和实践意义

大量的已有研究通过实验室实验探究推荐对消费者行为的影响,以此得到的研究结果外部有效性较低,实际应用中容易出现偏差。本研究在实地实验过程中,将登陆网站之后的消费者随机分配到不同的实验组。即使消费者多次访问网站,依然将该消费者分配在同一实验组内,避免结果出现偏差。已有利用多源数据构造推荐系统的研究主要关注推

荐算法的改进,没有探索推荐系统对消费者购物行为的影响。并且外部数据必须满足一定格式要求,很难在实际应用中推广。本研究通过从内部数据和外部数据中构造变量,解释基于不同数据源的推荐系统如何影响消费者的购物行为,一方面对前人实验室实验得出的结论进行验证,另一方面通过分析外部公司特征如何影响基于外部数据的推荐系统的效果,填补个性化推荐研究在这方面的空白。此外,通过分析消费者特征对不同数据源的推荐效果的调节作用,进一步完善个性化推荐领域消费者行为的理论框架。

目前虽然很多电子商务公司都为消费者提供个性化推荐服务,但尚不清楚个性化推荐如何影响消费者的网购行为。依靠经验或简单的统计数字判断个性化推荐的作用,很可能会高估或低估个性化推荐的真实作用。本研究通过解决分析中存在的估计偏差,可以使公司改进现有对推荐效果的统计指标,掌握推荐系统的真实效果。一方面,在评估推荐效果时,需要过滤异常用户和没有看到推荐栏的用户,否则会低估推荐效果。另一方面,消费者多次访问网站的情况会导致公司高估推荐系统的效果。此外,将消费者特征引入模型分析推荐系统对消费者购物行为的影响,可以帮助公司进一步完善个性化推荐算法。在设计推荐系统时考虑消费者访问习惯,访问网站频繁的用户点击推荐栏的概率反而偏低,这类用户对于网站内容比较熟悉,可以不用依靠推荐系统就能找到自己需要的商品。

当公司希望利用外部数据提升精准营销的效果时,本研究结果可以为公司在选择外部数据和利用数据两方面提供重要的指导。一方面,本研究在分析外部数据如何影响推荐系统效果的过程中加入外部公司特征,可以帮助公司明确外部公司数据的哪些特征对于提升自己个性化推荐服务的效果更好。公司在选择外部数据时,不仅要看外部企业消费者的规模,还要看外部公司与当前公司之间的共同用户数量。另一方面,假如公司之间需要进行数据交易,本研究结果可以对数据交易过程中的定价策略提供指导,定价时要考虑公司自身的数据特点,还要参考公司间的相互关联程度,即一个公司的数据对于不同的公司作用也不同。此外,本研究通过分析消费者对不同数据源推荐效果的调节作用,可以帮助公司进一步改进推荐系统,根据消费者特征选择恰当的数据源,以为消费者提供个性化推荐服务。

5.3 研究局限和展望

虽然本研究有上述多方面的贡献,但也存在一些不足,需要在未来研究中改进。①本研究在一家网上卖包的电商实施实地实验,一定程度上限制了研究结果的外部有效性。未来研究可以部署该实地实验到其他行业的网站,如服装或餐饮行业的网站,并与本研究结果进行对比,以发现不同行业特点对研究结果的影响。②本研究只使用经典的算法以避免由算法的特殊性带来的系统性偏差,在实际的商

业环境中,电商可以结合业务需求,使用更为复杂的推荐算法以获得更高的性能。③为了避免实地实验对电商造成较大的损失,本研究的实地实验只持续了8天。未来可以开展更长时间的实验,研究结果的鲁棒性会更高。④由于本研究使用Cookie ID识别同一消费者在不同网站的网购行为信息,如果用户使用多个终端设备(如手机、电脑)或多个浏览器访问网站,会存在用户的数据缺失,尽管这类用户数量很少,但未来可以结合用户登录ID等方式,将同一消费者在不同设备或渠道的数据进行统一,获取更全面的数据,分析推荐系统对消费者购物行为的影响。

参考文献:

- [1] Internet Retailer. *Top 500 guide*. (2014-01-30) [2018-01-31]. <http://www.internetretailer.com/top500/>.
- [2] DIAS M B, LOCHER D, LI M, et al. The value of personalised recommender systems to e-business: a case study // *Proceedings of the 2008 ACM Conference on Recommender Systems*. Lausanne, Switzerland, 2008:291-294.
- [3] 梁昌勇,冷亚军,王勇胜,等. 电子商务推荐系统中群体用户推荐问题研究. *中国管理科学*, 2013, 21(3):153-158.
LIANG Changyong, LENG Yajun, WANG Yongsheng, et al. Research on group recommendation in e-commerce recommender systems. *Chinese Journal of Management Science*, 2013, 21(3):153-158. (in Chinese)
- [4] 冯芷艳,郭迅华,曾大军,等. 大数据背景下商务管理研究若干前沿课题. *管理科学学报*, 2013, 16(1):1-9.
FENG Zhiyan, GUO Xunhua, ZENG Dajun, et al. On the research frontiers of business management in the context of big data. *Journal of Management Sciences in China*, 2013, 16(1):1-9. (in Chinese)
- [5] 姜树广, 谯倩. 实地实验及其在经济学中的应用. *经济评论*, 2012(5):134-143.
JIANG Shuguang, QIAO Qian. Field experiments and their applications in economics. *Economic Review*, 2012(5):134-143. (in Chinese)
- [6] ANSARI A, MELA C F. E-customization. *Journal of Marketing Research*, 2003, 40(2):131-145.
- [7] SIMESTER D, HU Y, BRYNJOLFSSON E, et al. Dynamics of retail advertising: evidence from a field experiment. *Economic Inquiry*, 2009, 47(3):482-499.
- [8] SAHNI N, ZOU D, CHINTAGUNTA P. *Effects of targeted promotions: evidence from field experiments*. Stanford, California: Stanford University, 2014.
- [9] 鞠宏磊,黄琦翔,王宇婷. 大数据精准广告的产业重构效应研究. *新闻与传播研究*, 2015(8):98-106.
JU Honglei, HUANG Qixiang, WANG Yuting. The reconstruction effect of big data precision advertising. *Journalism & Communication*, 2015(8):98-106. (in Chinese)
- [10] SCHUMANN J H, WANGENHEIM F V, GROENE N. Targeted online advertising: using reciprocity appeals to increase acceptance among users of free web services. *Journal of Marketing*, 2014, 78(1):59-75.
- [11] TUCKER C E. Social networks, personalized advertising, and

- privacy controls. *Journal of Marketing Research*, 2014, 51(5):546-562.
- [12] SUMMERS C A, SMITH R W, RECZEK R W. An audience of one; behaviorally targeted ads as implied social labels. *Journal of Consumer Research*, 2016, 43(1):156-178.
- [13] BRAHIM N B E-B, LAHMANDI-AYED R, LAUSSEL D. Is targeted advertising always beneficial?. *International Journal of Industrial Organization*, 2011, 29(6):678-689.
- [14] LAMBRECHT A, TUCKER C E. When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research*, 2013, 50(5):561-576.
- [15] 张莉, 秦桃, 滕丕强. 一种改进的基于用户聚类的协同过滤算法. *情报科学*, 2014, 32(10):24-27.
ZHANG Li, QIN Tao, TENG Piqiang. An improved collaborative filtering algorithm based on user clustering. *Information Science*, 2014, 32(10):24-27. (in Chinese)
- [16] SENECAI S, NANTEL J. The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, 2004, 80(2):159-169.
- [17] TAM K Y, HO S Y. Understanding the impact of web personalization on user information processing and decision outcomes. *MIS Quarterly*, 2006, 30(4):865-890.
- [18] PASSANT A. dbrec-Music recommendations using DBpedia// PATEL-SCHNEIDER P F. *The Semantic Web-ISWC 2010*. Springer Berlin Heidelberg, 2010:209-224.
- [19] KLYNE G, CARROLL J J. *Resource description framework (RDF): concepts and abstract syntax*. W3C Recommendation, World Wide Web Consortium, 2004.
- [20] NICOLAOU A I, IBRAHIM M, HECK E V. Information quality, trust, and risk perceptions in electronic data exchanges. *Decision Support Systems*, 2013, 54(2):986-996.
- [21] 马庆国, 王凯, 舒良超. 积极情绪对用户信息技术采纳意向影响的实验研究:以电子商务推荐系统为例. *科学学研究*, 2009, 27(10):1557-1563.
MA Qingguo, WANG Kai, SHU Liangchao. Influence of positive emotion on users' adoption intention on information technology: an experimental study with RA. *Studies in Science of Science*, 2009, 27(10):1557-1563. (in Chinese)
- [22] ROGERS T B, KUIPER N A, KIRKER W S. Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, 1977, 35(9):677-688.
- [23] 杨文静, 刘培朵, 崔茜, 等. 自我参照对情绪性记忆定向遗忘的影响. *心理学报*, 2014, 46(2):156-164.
YANG Wenjing, LIU Peiduo, CUI Qian, et al. Self-reference influence the directed forgetting of emotional memory. *Acta Psychologica Sinica*, 2014, 46(2):156-164. (in Chinese)
- [24] BARGH J A. Attention and automaticity in the processing of self-relevant information. *Journal of Personality and Social Psychology*, 1982, 43(3):425-436.
- [25] WELLS A, MATTHEWS G. *Attention and emotion: a clinical perspective*. East Sussex: Psychology Press, 2014:171-196.
- [26] DELLAERT B G C, HÄUBL G. Searching in choice mode: consumer decision processes in product search with recommendations. *Journal of Marketing Research*, 2012, 49(2):277-288.
- [27] SMYTH B, MCCLAVE P. Similarity vs. diversity // *Proceedings of the 4th International Conference on Case-Based Reasoning*. Vancouver, BC, 2001:347-361.
- [28] DHAR R, SIMONSON I. The effect of forced choice on choice. *Journal of Marketing Research*, 2003, 40(2):146-160.
- [29] GEORGE G, HAAS M R, PENTLAND A. Big data and management. *Academy of Management Journal*, 2014, 57(2):321-326.
- [30] CHEN Y, NARASIMHAN C, ZHANG Z J. Individual marketing with imperfect targetability. *Marketing Science*, 2001, 20(1):23-41.
- [31] JENTZSCH N, SAPI G, SULEYMANOVA I. Targeted pricing and customer data sharing among rivals. *International Journal of Industrial Organization*, 2013, 31(2):131-144.
- [32] WU L Y, CHEN P Y, CHEN K Y. Why does loyalty-cooperation behavior vary over buyer-seller relationship?. *Journal of Business Research*, 2015, 68(11):2322-2329.
- [33] HSU M H, CHANG C M, CHU K K, et al. Determinants of repurchase intention in online group-buying: the perspectives of DeLone & McLean IS success model and trust. *Computers in Human Behavior*, 2014, 36:234-245.
- [34] WANG Q, XIE J H. Will consumers be willing to pay more when your competitors adopt your technology? The impacts of the supporting-firm base in markets with network effects. *Journal of Marketing*, 2011, 75(5):1-17.
- [35] HOBERG G, PHILLIPS G. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 2016, 124(5):1423-1465.
- [36] 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)
- [37] 刘先红, 李纲. 科研社交网络的推荐系统对比分析. *图书情报工作*, 2016, 60(9):116-122.
LIU Xianhong, LI Gang. Comparative analysis of recommender systems of research social networking service. *Library and Information Service*, 2016, 60(9):116-122. (in Chinese)
- [38] 安悦, 李兵, 杨瑞泰, 等. 基于内容的热门微话题个性化推荐研究. *情报杂志*, 2014, 33(2):155-160.
AN Yue, LI Bing, YANG Ruitai, et al. Content-based personalized recommendation on popular micro-topic. *Journal of Intelligence*, 2014, 33(2):155-160. (in Chinese)
- [39] 王伟, 王洪伟, 孟园. 协同过滤推荐算法研究:考虑在线评论情感倾向. *系统工程理论与实践*, 2014, 34(12):3238-3249.
WANG Wei, WANG Hongwei, MENG Yuan. The collaborative filtering recommendation based on sentiment analysis of online reviews. *Systems Engineering - Theory & Practice*, 2014, 34(12):3238-3249. (in Chinese)
- [40] BOBADILLA J, HERNANDO A, ORTEGA F, et al. A framework for collaborative filtering recommender systems. *Expert Systems with Applications*, 2011, 38(12):14609-14623.
- [41] LINDEN G, SMITH B, YORK J. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 2003, 7(1):76-80.

Research on the Effectiveness of Personalized Recommender System Based on Multi-source Big Data

YAO Kai¹, TU Ping², CHEN Yuxin³, SU Meng⁴

1 Business School, Central University of Finance and Economics, Beijing 100081, China

2 Guanghua School of Management, Peking University, Beijing 100871, China

3 Business, New York University Shanghai, Shanghai 200122, China

4 Percent Corporation, Beijing 100101, China

Abstract: Personalized recommender system has become one of the important tools for online firms to offer personalized service and unique experience to the consumers. Through the personalized recommender system, merchants can improve their performance and consumer's satisfaction. The traditional recommender systems only use consumers' historical information of the focal firm to suggest personalized products. They cannot access to the extra data from external firms to achieve higher recommendation performance. In the era of big data, some third-party companies seize the opportunity to use the multi-source big data from different companies to provide better personalized recommendation service. However, how does this new type of recommender system influences consumers' shopping behavior is largely unknown.

This study mainly explores the influence of personalized recommender system based on multi-source big data on consumers' online shopping behaviors. In order to establish the causal relationship between recommender system and consumers' shopping behaviors, this study adopts field experiment to avoid endogenous problem and achieve high external validity. Meanwhile, according to constructing explanatory variables from internal and external big data, this study illustrates how the characteristics of multi-source big data influence the recommendation effectiveness. Finally, this research examines the interaction between consumers' characteristics and performance of recommender system based on internal and external data to explore how the effectiveness of recommender system varies according to consumers' characteristics.

The results demonstrate that the recommender system based on internal data can significantly improve consumers' click through rate on the personalized products. Meanwhile, the recommender system can help consumers decrease the time cost on decision making and stimulate them to browse more products. The performance of recommender system based on external data is related to the numbers of consumers of external websites and the relationships between the focal website and external websites. In addition, the effectiveness of the recommender system based on internal and external data is moderated by the characteristics of consumers. If the consumers are old users of the focal website, recommendations based on their internal data, rather than external data, can achieve better performance.

This study supplements the theoretical framework of personalized recommender system by analyzing the influence of recommender system based on multi-source big data on consumers' online shopping behaviors. The results have great guide value for how to construct better recommender system based on multi-source big data and provide important managerial suggestions for the strategy of sharing data between different websites.

Keywords: multi-source big data; personalized recommendation; field experiment; e-commerce; internet marketing

Received Date: January 31st, 2018 **Accepted Date:** June 28th, 2018

Funded Project: Supported by the National Natural Science Foundation of China(71332006,71702208)

Biography: YAO Kai, doctor in management, is a lecturer in the Business School at Central University of Finance and Economics. His research interests cover big data marketing, personalized marketing, social network analysis. His representative paper titled "Sales forecasting based on products association network" was published in the *Journal of Marketing Science*(Issue 4, 2015). E-mail: yaokai@cufe.edu.cn

TU Ping, doctor in sociology, is a professor in the Guanghua School of Management at Peking University. His research interests cover new product diffusion model, social network analysis and customer satisfaction. His representative paper titled "The impact of product attributes on customer satisfaction: an empirical study" was published in the *Journal of Management Science*(Issue 6, 2007). E-mail: tuping@gsm.pku.edu.cn

CHEN Yuxin, doctor in management, is a professor in Business at New York University Shanghai. His research interests cover data driven marketing, online marketing and competition strategy. His representative paper titled "Modeling credit card 'share of wallet': solving the incomplete information problem" was published in the *Journal of Marketing Research*(Issue 5, 2012). E-mail: yc18@nyu.edu

SU Meng, doctor in management, founder of Percent Corporation. His research interests include customer lifetime value and new product competition. His representative paper titled "Coupon trading and its impacts on consumer purchase and firm profits" was published in the *Journal of Retailing*(Issue 1, 2014). E-mail: meng.su@baifendian.com □