计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20160103
53(8):1673-1683,2016

用户在线购买预测:一种基于用户操作序列和选择模型的方法

曾宪宇　刘　淇　赵洪科　徐　童　王怡君　陈恩红
(中国科学技术大学计算机学院　合肥　230027)
(zengxy@mail.ustc.edu.cn)

Online Consumptions Prediction via Modeling User Behaviors and Choices

Zeng Xianyu, Liu Qi, Zhao Hongke, Xu Tong, Wang Yijun, and Chen Enhong
(*School of Computer Science, University of Science and Technology of China, Hefei* 230027)

Abstract　The rise of electronic e-commerce sites and the formation of the user's online shopping habits, have brought a huge amount of online consumer behavioral data. Mining users' preferences from these behavioral logs (e.g. clicking data) and then predicting their final consumption choices are of great importance for improving the conversion rate of e-commerce. Along this line, this paper proposes a way of combining users' behavioral data and choice model to predict which item each user will finally consume. Specifically, we first estimate the optimum substitute in each consumption session by a utility function of users' behavioral sequences, and then we build a latent factor based choice model (LF-CM) for the consumed items and the substitutes. In this way, the preference of users can be computed and the future consumptions can be predicted. One step further, to make full use of users' information of choosing and improve the precision of consumption prediction, we also propose a learning-to-rank model (latent factor and sequence based choice model, LFS-CM), which considers all the items in one session. By integrating latent factors and utility function of users' behavioral sequences, LFS-CM can improve the prediction precision. Finally, we use the real-world dataset of Tmall and evaluate the performance of our methods on a distributed environment. The experimental results show that both LF-CM and LFS-CM perform well in predicting online consumption behaviors.

Key words　online consumption prediction; choice model; behavioral sequence; sequence utility; distributed platform

摘　要　电商网站的兴起与用户在线购物习惯的形成,带来了海量的在线消费行为数据.如何从这些行为数据(如点击数据)中建模用户对相似产品的比较和选择过程,进而准确预测用户的兴趣偏好和购买行为,对于提高产品的购买转化率具有重要意义.针对这一问题,提出了基于用户行为序列数据和选择模型的在线购买预测解决方案.具体而言,1)使用行为序列效用函数估计用户在购买周期(session)中的

收稿日期:2016-03-04;**修回日期**:2016-06-05
基金项目:国家杰出青年科学基金项目(61325010);国家自然科学基金项目(61403358);科技惠民计划项目(2013GS340302);青年创新促进会会员专项基金项目(2014299);多媒体计算与通信教育部-微软重点实验室基金项目
This work was supported by the National Science Fund for Distinguished Young Scholars (61325010), the National Natural Science Foundation of China (61403358), the Plan of Science and Technology to Benefit the People (2013GS340302), the Special Fund for the Member of Youth Innovation Promotion Association of CAS (2014299), and the the Fund for Ministry of Education (MOE)-Microsoft Key Laboratory of USTC.
通信作者:刘淇(qiliuql@ustc.edu.cn)

最佳替代商品,然后对购买商品和最佳替代商品建立基于潜在因子的选择模型(latent factor based choice model, LF-CM),从而得到用户的购买偏好,实现对用户购买行为的预测.更进一步,为了充分地利用用户在每个购买周期的所有选择和比较信息,提高预测精度;2)提出了一种可以作用于购买周期内所有商品的排序学习模型(latent factor and sequence based choice model, LFS-CM),它通过融合潜在因子和行为序列的效用函数,提高了购买预测的精度;3)使用大规模真实数据集在分布式环境下进行了实验,并与参照算法进行了对比,证实了所提出的2个方法在用户在线购买预测上的有效性.

关键词 在线购买预测;选择模型;行为序列;序列效用;分布式平台

中图法分类号 TP391

随着电子商务的迅猛发展,电商消费者规模与在线交易量呈现激增态势.根据第36次中国互联网络发展状况统计报告的数据,截止2015年6月,我国在线购物用户与网络支付用户规模均突破3.5亿.国内最大电商平台阿里巴巴集团的财报亦显示,其2014年全年成交量达2.3万亿元人民币,同比增长47%.然而,电子商务在为人们的生活带来便捷的同时,由于其所承载商品的类别丰富且营销方式多样,这些海量信息也为用户挑选所需商品增添了诸多困难[1-2].因此,借助技术手段准确分析用户购买行为,为用户决策或商家营销提供支撑,已成为当前面向商务智能的数据挖掘领域重要的研究问题,对于改善用户体验、提高电商收益具有重要意义.

在此背景下,协同过滤等定向推荐技术[3]应运而生,并广泛运用于各电商网站的营销系统中.这些传统的推荐技术主要分析用户的购买行为,并以此为依据推荐相似或者相关的商品供用户选择[4].然而,它们仅着眼于孤立的交易(购买)行为,却忽略了这些购买与用户其他类型的行为(如在购买时为了比较相似产品而进行的点击行为)之间的关联.因此,定向推荐技术往往能够分析出用户会购买哪一类型的产品,却不能精准预测用户在一个购买周期(session)内最终选择哪个商品进行购买.这里,购物周期包含用户购买产品过程中所发生的一系列针对多种相似的、竞争产品的点击、比较行为和最终所发生的购买行为所组成的用户行为序列.可以看出,因为在线购物网站商品的多样性,不同用户在不同购买周期所面临的选择都千差万别,而且在同一购买周期内的产品经常是同类型的、相似度很高的产品,造成了传统的推荐技术难以进行用户在线购买的预测.与此同时,在线购买预测对于提高电商网站经济效益(如预测和提高产品的购买转化率、帮助用户进行购买目标(选择集)的筛选)有着重要的意义,所以,如何利用购买周期所包含的产品比较和选择信息序列来描述用户的当前购买意图,准确预测用户将要发生的购买行为是本文所关注的核心问题.

针对在线购买预测的问题场景和购物周期的数据特性,本文提出了基于用户行为序列数据和选择模型[5]的在线购买预测方案.具体而言,本文分别设计了基于潜在因子的选择模型(latent factor based choice model, LF-CM)和基于潜在因子和行为序列效用的选择模型(latent factor and sequence based choice model, LFS-CM).在LF-CM模型中,通过引入机会成本的概念,分析用户行为序列,并使用行为序列效用函数估计用户在购买周期中的最佳替代商品(即在同一周期中,除去最终的购买选择外最可能购买的商品),从而对购买商品和最佳替代品建立选择模型,进而利用用户历史购买行为中同一周期内的商品比较选择信息,对用户当前的购买偏好进行排序学习.在LF-CM模型基础之上,为更好地利用购物周期所包含的所有产品选择和比较信息,还提出了一个可以融合潜在因子和行为序列的效用函数,作用于购买周期内所有商品的排序学习模型LFS-CM.相比于LF-CM模型,LFS-CM能够更充分地分析用户的选择行为,以提升预测精度.

最后,为验证本文所提算法,在国内最大的电商网站Tmall的真实数据集上,借助Tmall提供的大规模、分布式数据处理服务(open data processing service, ODPS)平台进行了分布式实验,并与该领域的多个参照算法进行了对比.实验结果证实了本文所提出的LF-CM和LFS-CM方法在用户在线购买预测上的有效性.

1 相关工作

与本文相关的工作主要分为2个方面:1)离散选择模型;2)在线购买预测.

1.1 离散选择模型

离散选择模型是经济学中的重要概念,又称为定性选择模型(qualitative choice model)[6].离散选择模型表示了从2个或者更多的离散候选项目(产品)中做出选择的过程.在可选项离散的情况下,研究的是选择"哪一个"的问题[7].经典的离散选择模型有逻辑回归(logistic regression, LR)[8]、Probit回归(probit regression)[9]等.离散选择模型将事件发生与否概率解释为特征变量的函数,可以由模型计算得到选择各个物品的概率而得到最终的结果.

由于其在决策建模上的有效性,离散选择模型被广泛应用于社会学、生物统计学、数量心理学、市场营销等多个领域[10].例如,市场调研人员使用选择模型研究用户的需求、预测商品的市场响应[6];交通管理人员使用选择模型来规划交通系统[11-12];社会学研究者使用选择模型来预测职业和训练项目[13].此外,离散选择模型还可用于描述一些新兴应用场景.例如,文献[14]提出了一种基于用户近期偏好的选择模型用于解决兴趣点(point of interest, POI)推荐;文献[15]提出了一种使用绝对距离和相对距离作为特征输入的选择模型来模拟用户在地图上进行餐馆选择过程的方法.

1.2 在线购买预测

在线购买预测对于电商网站提高经济效益有着重要的意义,在工业界已经被广泛重视,Tmall与Recsys都曾举办过关于在线购买预测的比赛.大部分队伍使用了特征工程和模型融合的方法[16-18].这些方法首先提取出用户序列中与购买有关的特征,然后借助不同的模型(例如逻辑回归[8]、梯度提升决策树(GBDT)等)进行拟合训练及模型融合[19].例如,在Tmall举办的移动购买预测比赛中文献[17]首先使用训练数据得到多个GBDT模型,然后用其输出作为LR模型的输入,得到最终的预测结果.同时,相关问题在学术界也正被广泛的研究,例如,文献[20-22]使用社交信息,研究了在用户的选择与消费和社交因素之间的联系.

与此同时,推荐技术在某种程度上也可以解决购买预测问题[23].传统的推荐技术包括基于内容推荐[24]、协同过滤(Model Based CF[25],Memory Based CF[26])和混合推荐策略[27].然而,如在引言中所述,由于这些技术多着眼于孤立的用户交易(购买)行为,却忽略了这些购买与用户其他类型的行为(如在购买时为了比较相似的、竞争的同类型产品而进行的点击行为)之间的关联,因此虽然能够分析出用户在之后一段时间内可能会购买哪一类型的产品,却不能精准预测用户在一个购买周期内最终选择哪个商品.而本文引入购物周期和机会成本等概念,通过使用用户的操作行为序列建模用户在相似产品之间的比较信息,更加清晰地反应了用户的真实偏好,能够更有效地解决用户在线购买预测的问题.

2 预备知识

本节将首先介绍在线购买预测针对的场景和使用的数据形式,然后定义本文中涉及到的基本概念,最后对在线购买预测问题进行形式化描述.

2.1 问题场景和数据描述

以Tmall为例,用户在进行在线购物时,会看到许多待选商品,在对这些商品进行比较和选择时,会产生一系列的行为序列数据,如点击(click)、收藏(collect)、加入购物车(cart)、和购买(buy)等,并被系统记录到用户操作日志中.例如,Tmall提供的日志数据样例如表1所示:

Table 1 Example of Original Records

表1 原始数据示例

User	Item	Brand	Category	Action	Timestamp
U_1	a	M	C_1	Click	2013-05-01T20:30:02
U_1	b	Q	C_1	Click	2013-05-01T20:33:11
U_1	a	M	C_1	Buy	2013-05-01T20:40:42
U_2	c	M	C_2	Click	2013-05-02T12:13:55

表1数据共包含6个字段,分别是用户(user)、交互商品(item)、商品品牌(brand)、商品类别(category)、用户操作类型(action)和时间戳(timestamp).每个用户的行为日志记录了此用户在Tmall平台上完整的行为数据.

为了更加方便地描述用户单次消费的选择情形,按照时间戳信息对用户行为序列进行划分,从而获得用户在线购物中的购买周期(session).本文采用以下的启发式方法进行购买周期划分:以每次的购买行为作为分割点,向此分割点之前的操作记录进行搜索,如果该记录与下一个操作之间的时间间隔小于设定的阈值(如若干小时),就将其归到该次购买商品所对应的周期中.需要注意的是,在该分割方法下,一个购买周期中只有一个被购买的产品.通过周期划分,可以得到每个用户在不同购买周期的行为序列,如表2所示:

Table 2 Example of Session Records

表 2 分段后的数据示例

Session	User	Item Sequence	Item Bought
S_1	U_1	a,b,b,b,a,c,c,b	b
S_2	U_2	d,d,c,c,e,f,f,c	c
S_3	U_2	a,a,c,a,a,a,a	a
S_4	U_3	e,a,a,e,f,c,a,c	c

例如,表 2 的第 1 行表示用户 U_1 在购买周期 S_1 中的产品操作序列为 a,b,b,b,a,c,c,b,而他最终购买的是产品 b.需要注意的是,为了简单起见,表 2 忽略了用户不同的操作类型 click,cart 和 collect 的差异,而将其视为统一的浏览(点击)操作,即本文主要关注于挖掘统一产品操作中的用户序列模式.

2.2 相关概念

本文主要讨论用户在线购买商品的行为及其预测方法.在本文中,所有的用户构成的用户集合 U 表示为 $U=\{u_1,u_2,u_3,\cdots\}$,所有的商品构成的商品集合 I 表示为 $I=\{i_1,i_2,i_3,\cdots\}$.在此基础之上,给出购买周期以及其他相关概念的形式化定义:

定义 1. 购买周期(session).一个购买周期 s 表示一个用户在一定时间范围内经过对比和选择并最终产生购买行为的过程,可以表示为一个三元组 $s=(u,sq,i_b)$.其中 $u\in U$,表示该购买周期的用户;$sq=\{i_{b_1},i_{b_2},i_{b_3},\cdots\}$ 表示用户在该购买周期内的商品操作序列,即表 2 中的 Item Sequence 字段的商品记录,$i_b\in sq$ 表示用户在此次购买周期中经过比较和选择之后最终购买的商品.所有购买周期的集合记为 $S=\{s_1,s_2,\cdots\}$.

定义 2. 商品效用.为了量化用户对商品的购买意愿,在任一个购买周期 s 中,针对某一用户 u,对每件商品 i 定义一个效用值 $w_{s,u,i}=w(s,u,i)$,其表示该用户在当前购买周期中对于该商品价值的衡量.效用值越高,说明用户对商品价值的衡量越高,购买它的可能性也越大.

定义 3. 机会成本.在购买周期 s 中,当用户购买了商品 i_b 时,就意味着会放弃其他的商品.机会成本 $w_{s,u,i_{oc}}$ 即指该购买周期中放弃的所有商品所对应的最大效用值(是个估计值),而具有该最大效用值的产品 i_{oc} 被称为购买周期中的最佳替代商品,即:

$$w_{s,u,i_{oc}}=\max_{i\in I(s)\setminus i_b} w_{s,u,i}.$$

最后,本文所研究的问题可总结为从用户所有的历史在线购买周期的序列模式中挖掘用户对商品的潜在兴趣和心理偏好,并根据当前购买周期中已出现的商品操作序列准确地预测用户可能购买的商品.形式化描述为:给定历史训练数据 S_T,从中学习到预测模型 M;对于任意待预测的购买周期 $s=(u,sq,i_b)$,在 u 和 sq 已知的情况下利用预测模型 M 预测出用户 u 在此次购买周期中最有可能购买的商品 i_b.

3 模型介绍

针对上述的研究问题,本文提出了 2 种排序选择模型:基于潜在因子的选择模型(LF-CM),及基于潜在因子和行为序列效用的选择模型(LFS-CM).具体而言,两者都借鉴了潜在因子模型中的潜在因子(latent factor)概念,即通过对用户操作序列的分析,将用户和商品的属性分别映射到低维空间中(潜在因子).

基于潜在因子的选择模型(LF-CM)首先基于行为序列的特征设计出序列效用函数,接下来根据该效用函数计算购买周期内所有商品的预估效用值,从而在当前购买周期内确定最佳替代商品,然后在当前购买商品与最佳替代商品之间建立选择模型.因此,在单个购买周期内,LF-CM 模型只需要在当前购买商品和最佳替代商品之间进行对比学习(潜在因子学习),从而可以有效地减少学习训练时间.

基于潜在因子和行为序列效用的选择模型(LFS-CM),将商品的潜在因子和在当前购买周期内的行为序列效用结合在一起表示商品的效用,并在当前购买商品和本周期所有候选的商品之间建立选择模型.相比于 LF-CM 模型,由于 LFS-CM 模型使用了购买周期内所有商品,因此能够更加充分地利用用户的比较和选择信息,提升模型的准确度.

3.1 基于潜在因子的选择模型(LF-CM)

LF-CM 模型是建立在当前购买商品和最佳替代商品之间的选择模型.它利用用户行为序列设计效用函数来估计购买周期内所购买商品的机会成本,从而锁定候选商品中的最佳替代品.

3.1.1 商品效用函数

在实际的在线购买过程中,用户的选择取决于他对购买周期内商品的价值评估,即商品效用 $w_{s,u,i}$.本文借鉴潜在因子的思想,将商品对于用户的效用 $w_{s,u,i}$ 表示为代表用户的潜在因子向量 $\boldsymbol{p}_u$ 与代表商品的潜在因子向量 $\boldsymbol{q}_i$ 的内积,即:

$$w_{s,u,i}=\boldsymbol{p}_u^{\mathrm{T}}\boldsymbol{q}_i. \tag{1}$$

用户的偏好与商品特征越契合,即商品越符合用户的选择习惯,则其内积 w 越大;相反,商品越不符合用户的选择偏好,其效用值 w 就越小. Luce 提出的选择公理(Luce's choice axiom)说明了在商品集合中进行选择的方式:选择某一件商品的概率依赖于其在整个商品集合中的相对效用,相对效用越大,选取这件商品的概率就越大[8].

假定用户都是理性的,即他们会考虑到在进行决策时的机会成本和收益,最终选择效用不小于机会成本 $w_{s,u,i_{oc}}$ 的商品购买,即在任意购买周期 s 中,用户所购买商品 i_b 的效用应当不小于最佳替代产品的效用(机会成本):

$$w_{s,u,i_b}\geqslant\max_{i\in I(s)\setminus i_b} w_{s,u,i}=w_{s,u,i_{oc}}.$$

3.1.2 确定最佳替代品

用户最终购买的产品 i_b 是可观察到的,而最佳替代产品则是未知的(本周期内的所有候选商品都有可能),为了估计最佳替代产品必须首先得到商品效用(式(1)). 然而,如果直接使用式(1)进行计算,则需要判断每个周期内所有候选商品的效用值. 事实上,商品效用(用户的选择偏好)可以简单通过用户在每个购买周期内的行为序列来预估(不需要进行式(1)的因子相乘). 在本文中,提出序列效用函数 $f(s,i)$ 的概念,可以使用 $f(s,i)$ 来对商品的效用进行预估. 然后,LF-CM 模型以该效用函数计算得到商品效用为依据,确定预估的机会成本以及相应的最佳替代品 i_{oc}(除去购买商品 i_b 外数值最大的序列效用函数对应的商品). 使用这种预估商品效用的方法,在每个购买周期中只需要在购买商品和最佳替代商品之间利用式(1)建立选择模型.

本文设计的序列效用函数考虑 2 个影响商品选择的因素:

1) 商品在购买周期中出现的频率(frequency). 直观上,一件商品在购买周期中出现的频率越高,说明用户对其越感兴趣,就有越大的可能性去选择购买这件商品. 例如,购买周期的商品点击序列为{a,a,b,b,a,a,a,a,a},那么用户购买商品 a 的可能性应该高于购买商品 b 的可能性.

2) 点击与购买的时间间隔(recency). 在一个购买周期中,用户最近点击的商品更有可能被选择. 假定一个购买周期的点击序列为{a,a,a,b,c,b,a,b,b,b,b},那么商品 b 的购买概率应该高于商品 a. 因为在经过对比 a,b,c 之后,用户将重点放在了商品 b 上,所以用户对商品 b 的满意度可能更高.

综合 frequency 和 recency 因素,给出以下 $f(s,i)$ 的具体定义形式. 将购买周期 s 中用户的点击序列 sq 按照时间排序,其长度为 N,sq 从起始至结束各个位置分别编号为 1,2,…,N,假定商品 i 出现的位置组成集合 $P(s,i)$,则在该购买周期中,商品 i 的序列效用函数可表示为

$$f(s,i)=\sum_{k\in P(s,i)}\mathrm{e}^{-\frac{N-k}{N}}.$$

例如,假定一个购买周期中的商品点击序列为{a,b,b,c,c,a,c},序列长度为 $N=7$,商品出现的位置为 4,5,7,那么商品 c 的行为序列效用由 $f(s,i)$ 计算得到:$\mathrm{e}^{-\frac{7-4}{7}}+\mathrm{e}^{-\frac{7-5}{7}}+\mathrm{e}^{-\frac{7-7}{7}}=2.40$.

通过如上方式,可以计算得到购买周期内所有商品的序列效用 $f(s,i)$,从而预估出最佳替代商品 i_{oc}. 值得一提的是,对于 $f(s,i)$ 的选择也可以是其他的形式,但都需要能够表现出序列的特征.

3.1.3 基于潜在因子的选择模型(LF-CM)

在使用 $f(s,i)$ 确定了每个购买周期中当前购买商品的最佳替代商品之后,便可以在当前购买商品与最佳替代商品二者之间建立比较选择模型. 如下所示,其优化目标是在所有的购买周期中购买商品和最佳替代品的效用差的和最小:

$$\arg\min_{w} opt=\sum_{s\in S}(w_{s,u,i_{oc}}-w_{s,u,i_b})+\lambda\|\theta\|^2, \tag{2}$$

其中,θ 为 w 包含的参数.

将式(1)带入到式(2)中可得 LF-CM 最终所要优化求解的目标:

$$\arg\min_{w} opt=\sum_{s\in S}(\boldsymbol{p}_u^{\mathrm{T}}\boldsymbol{q}_{i_{oc}}-\boldsymbol{p}_u^{\mathrm{T}}\boldsymbol{q}_{i_b})+\frac{\lambda_u}{2}\sum_{u\in U}\|\boldsymbol{p}_u\|^2+\frac{\lambda_i}{2}\sum_{i\in I}\|\boldsymbol{q}_i\|^2. \tag{3}$$

本文采用梯度下降法对式(3)进行求解. 具体地,首先分别在固定 $\boldsymbol{q}_i$ 和 $\boldsymbol{p}_u$ 的情况下对式(3)进行求导:

$$\frac{\partial opt}{\partial \boldsymbol{p}_u}=\sum_{s\in S}(\boldsymbol{q}_{i_{oc}}-\boldsymbol{q}_{i_b})+\lambda_u\boldsymbol{p}_u, \tag{4}$$

$$\frac{\partial opt}{\partial \boldsymbol{q}_i}=\sum_{s\in S}I_i\boldsymbol{p}_u+\lambda_i\boldsymbol{q}_i, \tag{5}$$

其中,I_i 为指示函数:$I_i=\begin{cases}-1, & i=i_b,\\ 1, & i=i_{oc},\\ 0, & \text{other}.\end{cases}$

对 $\boldsymbol{p}_u$ 和 $\boldsymbol{q}_i$ 交替地进行更新优化:

$$\boldsymbol{p}_u\leftarrow\boldsymbol{p}_u-\eta\Big(\sum_{s\in S}\boldsymbol{q}_{i_{oc}}-\boldsymbol{q}_{i_b}+\lambda_u\boldsymbol{p}_u\Big), \tag{6}$$

$$\boldsymbol{q}_i\leftarrow\boldsymbol{q}_i-\eta\Big(\sum_{s\in S}I_i\boldsymbol{p}_u+\lambda_i\boldsymbol{q}_i\Big). \tag{7}$$

其中,η 为步长.

模型训练收敛后即得到用户和商品在隐特征空间的特征表示($\boldsymbol{p}_u$,$\boldsymbol{q}_i$).需要注意的是,LF-CM 模型虽然在每个购买周期中没有考虑非最佳替代品,但是一个购买周期中的非最佳替代品可能是另一个购买周期中的购买商品或者最佳替代品,因此,使用 LF-CM 仍可以得到所有商品的潜在信息.

在得到用户和商品在隐特征空间的特征表示($\boldsymbol{p}_u$,$\boldsymbol{q}_i$)之后,对于任一用户 u 和其待预测的购买周期 s,s 中效用 $w_{s,u,i}=\boldsymbol{p}_u^{\mathrm{T}}\boldsymbol{q}_i$ 最高的商品即该用户在此购买周期中最有可能购买的商品.

3.2 基于潜在因子和行为序列的选择模型(LFS-CM)

LF-CM 模型使用行为序列效用函数估计用户在购买周期中的最佳替代商品,并仅在当前购买商品和最佳替代品之间建立的选择模型.它可能存在 3 个方面的缺点:

1) 损失用户的部分比较选择信息.在线消费的过程中,用户不仅仅在最佳替代品和当前购买商品之间进行比较选择,而是所有购买周期中的商品都需要与当前购买商品进行比较;

2) 采用行为序列效用函数估计机会成本,确定最佳替代品有一定的风险,即行为序列效用函数 $f(s,i)$并不一定能找到真正的最佳替代品;

3) 无法解决冷启动[28-29]问题.如果一个购买周期中的用户和出现的商品在历史记录中没有足够的训练数据,那么无法得到准确稳定的用户和商品的特征表示.

为了克服 LF-CM 模型的上述缺点,本节提出基于潜在因子和行为序列效用的选择模型(LFS-CM).LFS-CM 模型采用效用函数:

$$w_{s,u,i}=\alpha\boldsymbol{p}_u^{\mathrm{T}}\boldsymbol{q}_i+(1-\alpha)f(s,i). \tag{8}$$

式(8)在式(1)的基础上添加了行为序列效用函数 $f(s,i)$项,并采用参数 α 调节潜在因子效用和行为序列效用的权重.

LFS-CM 模型是在用户购买周期内所有候选商品和当前购买商品之间建立的选择模型:

$$\begin{aligned}\arg\min_{w} opt=&\sum_{s\in S}\sum_{i\in I(s)\setminus i_b}(\alpha\boldsymbol{p}_u^{\mathrm{T}}\boldsymbol{q}_i+(1-\alpha)f(s,i)-\\&\alpha\boldsymbol{p}_u^{\mathrm{T}}\boldsymbol{q}_{i_b}-(1-\alpha)f(s,i_b))-\\&\frac{\lambda_u}{2}\sum_{u\in U}\|\boldsymbol{p}_u\|^2+\frac{\lambda_i}{2}\sum_{i\in I}\|\boldsymbol{q}_i\|^2.\end{aligned} \tag{9}$$

类似地,LFS-CM 模型也采用梯度下降法求解,$\boldsymbol{p}_u$ 和 $\boldsymbol{q}_i$ 交替地进行优化:

$$\boldsymbol{p}_u\leftarrow\boldsymbol{p}_u-\eta\Big(\sum_{s\in S}\boldsymbol{q}_{i_{oc}}-\boldsymbol{q}_{i_b}+\lambda_u\boldsymbol{p}_u\Big), \tag{10}$$

$$\boldsymbol{q}_i\leftarrow\boldsymbol{q}_i-\eta\Big(\sum_{s\in S}I_i\boldsymbol{p}_u+\lambda_i\boldsymbol{q}_i\Big), \tag{11}$$

其中,指示函数 $I_i=\begin{cases}-1, & i=i_b,\\ 1, & i\neq i_b.\end{cases}$

同理,模型训练收敛后得到用户和商品的特征表示($\boldsymbol{p}_u$,$\boldsymbol{q}_i$),对于给定用户待预测周期中的商品,效用值 $w_{s,u,i}=\alpha\boldsymbol{p}_u^{\mathrm{T}}\boldsymbol{q}_i+(1-\alpha)f(s,i)$最高的商品即该用户在此购买周期中最有可能购买的商品.

由于 LFS-CM 模型的效用函数包含潜在因子效用和行为序列效用 2 项,所以 LFS-CM 模型在一定程度上可以解决用户和商品的冷启动问题.因而,相比于 LF-CM 模型,LFS-CM 模型能够更加充分、全面地利用用户的选择信息,进而提高预测精度.

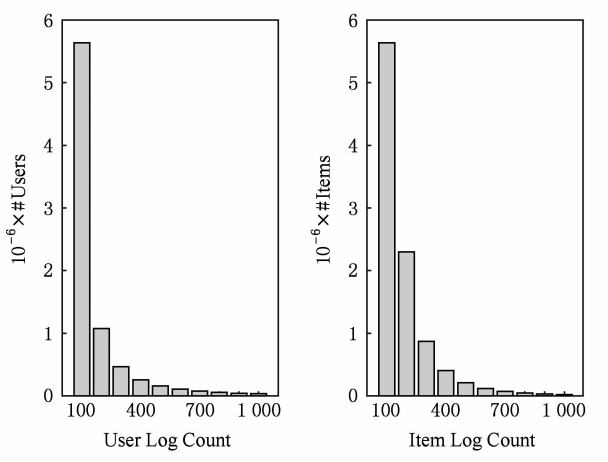
4 实 验

本文实验全部在 Tmall 提供的开放数据处理服务(ODPS)平台上完成.ODPS 平台由阿里云自主研发,提供了针对 TB/PB 级数据分布式处理能力,应用于数据分析、挖掘、商业智能等领域.该平台提供了 SQL 查询接口、MapReduce 编程接口等服务.本文在该平台下利用上述工具以分布式的方式设计和实现了实验所用的算法,能够对大体量的数据进行处理,充分挖掘其中所包含的有价值信息,具有很好的扩展性.

4.1 数据分析与预处理

实验数据采用了 Tmall 网站 2013 年 4 月至 9 月全部的用户行为记录,数据格式如表 1 所示.该数据集一共包含 1 333 729 303 条记录,涉及 9 774 184 位用户、8 133 507 件商品.其中,绝大多数行为记录为点击行为,购买行为仅占不到 1%.

图 1 展示了用户与商品涉及记录数的统计



(a) User Log Count Distribution (b) Item Log Count Distribution

Fig. 1 User and item count distribution.

图 1 用户与商品记录数分布

分布.其中横轴为记录数,纵轴为其数目,绝大部分的操作都在1 000次以内,其中每个用户的平均操作次数约为136,每个商品的平均操作次数约为164.从购买行为所占比例和商品、用户的平均操作次数的分布可以看出,该数据十分稀疏,且呈现显著的长尾特征,因此需要充分地利用用户行为序列的信息以对购买进行预测.

4.1.1 数据预处理

首先对数据进行预处理:1)去除低频商品.为在一定程度上缓解冷启动问题,在实验中只选择保留出现次数在500以上的商品,约占全部商品的10%;2)划分购买周期.按照第3节中提到的方法进行划分,并选择2次操作的间隔阈值为12 h.即若2次操作的间隔时间大于12 h,则它们应当位于不同的购买周期中.同时,由于长度较短的购买周期购买行为存在较强的随机性,为了使模型的训练有效和准确,本文仅考虑长度(行为记录数)大于5的购买周期.表3展示了划分购买周期后的一些统计结果.

Table 3 Data after Pre-Processing

表3 预处理后的数据

Statistics	Value
#Sessions	270 947
#Users	186 907
#Items	68 912
Average Session length	11.93
Average #Items in one Session	4.28

通过表3我们发现,在每个购买周期中,用户通常会比较超过4个商品并在其中进行选择购买,而每个商品的平均点击次数接近3次.由此可见,购买周期的序列中包含着丰富的比较和选择信息.

对于frequency,在每个购买周期中按照不同商品出现频次从大到小进行排序,然后得到购买的商品所占排位的百分比.

对于recency,在每个购买周期中按照不同商品出现的时间从后向前进行排序,然后得到购买的商品所占排位的百分比.

图2展示了所有购买周期中frequency和recency的分布.可以看出,无论是对于frequency还是recency,用户所购买的商品很大一部分都在前30%中,即用户倾向于购买在购买周期中出现频率更高和操作时间更近的商品.这为本文所采用的行为序列效用函数$f(s,i)$考虑recency和frequency因素提供了统计依据.

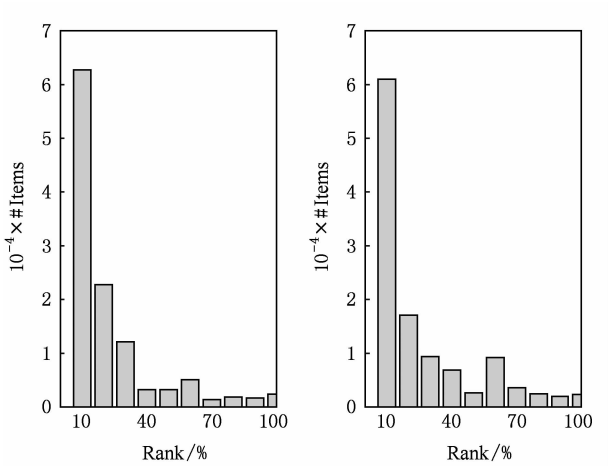


Fig. 2 Recency/frequency rank of purchased items distribution.

图2 购买商品所占排位的分布

4.2 实验设置

本文在ODPS平台上按照分布式的方式[30-31]实现了所提出的LF-CM和LFS-CM两种选择模型以及其他几种对比方法.实验中,按照4:1的比例将用户的购买周期集合划分为训练集和测试集,在训练集的购买周期中对模型参数进行学习和调整,然后在测试集中按照学习得到的参数对新购买周期的商品进行购买预测.

在LF-CM模型和LFS-CM模型中,本文通过比较设置正则化项参数$\lambda_u=\lambda_i=0.01$,用户和商品的特征维度设置为10,在计算过程中动态的调整步长来进行迭代训练,待收敛后得到用户和商品的特征向量$(\boldsymbol{p}_u,\boldsymbol{q}_i)$.

4.2.1 对比方法

为了验证所提出模型的效果,本文选取了7种方法作为对比实验.

1) 随机选择(Rand).对测试集中每个购买周期内出现的商品,随机地选取一个作为用户购买预测的结果.

2) 最流行商品(Pop).在训练集中统计每个商品被用户操作的次数,作为商品流行度的指标.然后对测试集的购买周期,推荐该购买周期中最流行的商品.

3) 基于用户的K近邻(KNN)[32].使用传统推荐系统中协同过滤的方法.具体地,在训练集中,按照用户购买的商品计算出用户之间的相似度.然后在测试集中用相似度计算每个商品可能会被购买的权重,最后按照权重的高低衡量商品购买概率.

4) 使用用户购买周期中所用商品进行排序学习(PQ)[33]. 在训练集中,使用购买周期中的所有商品进行学习,这实际上就是 LFS-CM 模型在 $\alpha=1$ 时的特殊情形.

5) 使用用户在购买周期中表现出的序列效用(FIS). 根据用户购买周期的序列计算得到的 $f(s,i)$作为商品效用,这实际上是 LFS-CM 模型在 $\alpha=0$ 的特殊情形.

6) 逻辑回归(LR)[8,34]. 为了与其他方法进行有效地对比,在 LR 模型中也使用了 frequency 和 recency 两类特征. 第 1 类 frequency 特征包括商品在购买周期中出现的次数以及频率(相对次数);第 2 类 recency 特征包括商品在购买周期中最后出现的位次和相对位次(位次除以购买周期中的长度). 将购买预测转化为二分类问题,购买周期中 LR 预测概率最大的商品即为购买商品.

7) 梯度提升决策树(GBDT)[35]. 与 LR 相同,使用 frequency 和 recency 特征. 购买周期中预测概率最大的商品即为购买商品.

4.2.2 衡量指标

实验中,按照每种方法计算出不同商品的效用值,分别选取效用值最高的 1 个(Top1)、2 个(Top2)、3 个(Top3)商品作为预测购买的结果.

本文使用召回率(*Recall*)和精度(*Precision*)作为衡量预测方法的预测准确性的指标,其计算公式分别为

$$Precision=\frac{TP}{TP+FP},$$

$$Recall=\frac{TP}{TP+FN},$$

其中,TP 为预测购买的商品数量;FP 为预测购买而实际未被购买的商品数量.

4.3 实验结果

本节从多个方面对实验结果进行展示和分析. 首先,通过购买预测结果,再次验证 LF-CM 模型中使用行为序列效用函数 估计最佳替代品的有效性. 然后,测试 LF-CM 模型的效用函数中权重参数 α 的影响,并得到 α 的最优取值. 最后展示所提出的模型以及对比方法在预测准确性上的对比结果.

4.3.1 $f(s,i)$估计最佳替代品的有效性

因为最佳替代商品不是可观测的,所以为了说明使用 $f(s,i)$来估计最佳替代品的有效性,本文的验证方法如下:1)分别使用 $f(s,i)$以及随机选取的方法选择购买周期中的其他商品作为最佳替代商品;2)在最佳替代商品和用户购买商品之间建立选择模型(使用的 $f(s,i)$方法即为 LF-CM);3)在测试集上通过购买预测的效果验证所选择的最佳替代品的准确性. 表 4 给出了 2 种方法效果对比:

Table 4 Effect Comparison of two Methods of Selecting the Optimum Substitutes

表 4 2 种选择最佳替代品方法的效果对比

Prediction	Use $f(s,i)$		Random	
	Precision	*Recall*	*Precision*	*Recall*
Top1	**0.712**	**0.712**	0.645	0.645
Top2	**0.401**	**0.802**	0.366	0.731
Top3	**0.276**	**0.828**	0.262	0.787

从表 4 中可以看出相比于随机选择,使用 $f(s,i)$来估计最佳替代品的购买预测结果在精度和召回率 2 种指标上都有着显著的提升,说明了使用行为序列效用函数 $f(s,i)$估计最佳替代品的有效性.

4.3.2 LFS-CM 参数 α 的影响

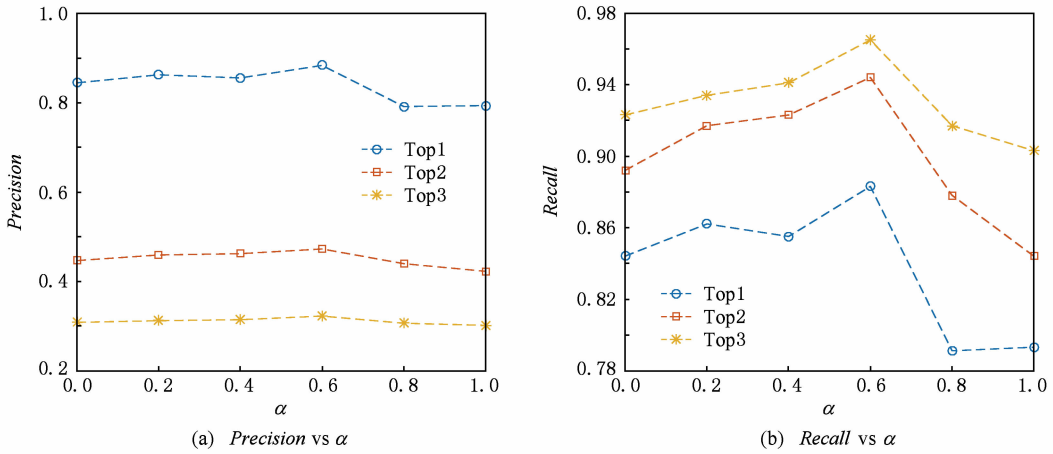
为了测试 LFS-CM 模型对参数 α 的敏感性并得到最佳的 α 值,分别设置 $\alpha=0,0.2,0.4,0.6,0.8,1.0$ 进行了 Top k 预测的实验.

如图 3 所示,在不同 α 的条件下,模型预测效果会有所差别,当 $\alpha=0.6$ 时精度和召回率都有最好的效果,因此在后文与其他方法的对比实验中,LFS-CM 模型的 α 参数设置为 0.6.

4.3.3 预测准确性结果对比

图 4 显示了 LF-CM 模型、LFS-CM 模型以及对比方法(Rand,Pop,KNN,PQ,FIS,LR,GBDT)在精度和召回率 2 种指标上的对比结果.

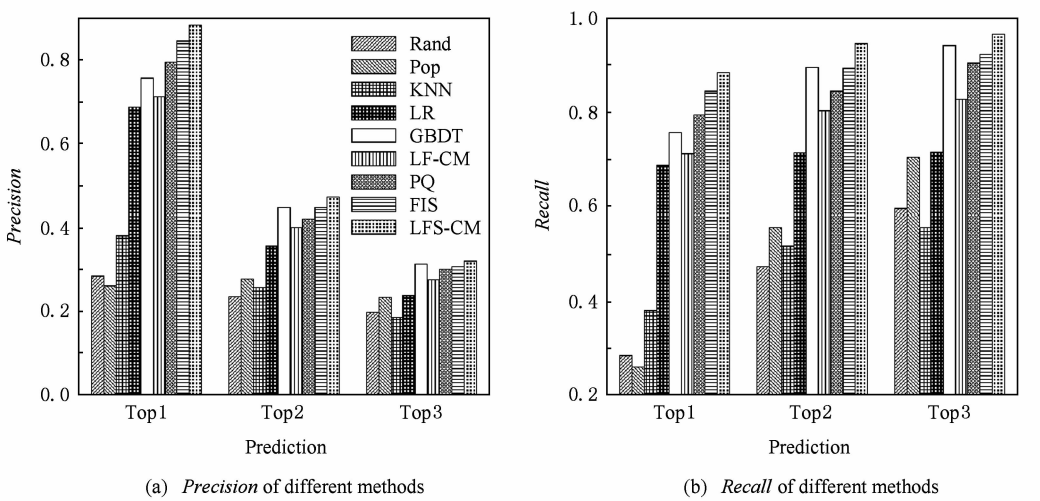
可以看出,Rand 和 Pop 两种方法的表现较差,因为它们并不能分辨出每个用户的个性化偏好;由于数据的稀疏性,以及同一购买周期内的产品的高度相似性,使得传统的 KNN 方法也不能很好地进行预测,在 Top2 和 Top3 上的预测效果甚至不如 Pop 方法;通过使用用户在购买周期中的偏好信息(frequency 和 recency),LR 和 GBDT 在效果上有了显著的提升,由于 GBDT 相对 LR 模型本身的优势,GBDT 取得了相对更高的预测准确度;同样考虑了用户的偏好和比较选择信息以后,LF-CM 模型的预测精度相比于 Rand,Pop,KNN 三种方法有了大幅提升;PQ 方法相对于 LF-CM 模型,由于使用了购买周期内全部商品的信息建立选择模型,使得预测效果有了更进一步的提升,说明了选择模型本身的合理性;而 FIS 使用了用户在购买周期中的 recency



(a) *Precision* vs α (b) *Recall* vs α

Fig. 3 *Precision* & *Recall* vs α.

图 3 精度和召回率随 α 变化情况



(a) *Precision* of different methods (b) *Recall* of different methods

Fig. 4 *Precision* & *Recall* of different methods.

图 4 不同方法的精度和召回率对比

和 frequency 时序行为特征得到效用函数,也取得了不错的表现;最后,LFS-CM 模型充分利用了用户的选择和比较信息以及序列特征,在所有的方法中表现最好,即它可以最准确地预测用户购买偏好和行为.

4.3.4 时间复杂度分析

本节简单介绍所提算法的分布式实现过程,并分析相应的运行时间开销情况. 在本文的算法设计中,将 LF-CM 和 LFS-CM 算法的实现都分为 2 个 Map-Reduce 过程:在第 1 个 Reduce 操作中对用户的潜在因子 $\boldsymbol{p}_u$ 进行更新迭代;在第 2 个 Reduce 中对商品的潜在因子 $\boldsymbol{q}_i$ 进行更新迭代. 将这 2 次操作看做整个算法的一次完整迭代.

记 n 为所有购买周期的数目,k 为周期平均长度. 在一次迭代过程中 LF-CM 只考虑购买周期中的 2 个商品(购买的商品和最佳替代的商品),由式(6)(7)可以看出其需要计算 n 次;而 LFS-CM 模型(包括 PQ 算法,但不包括 $\alpha=0$ 的情况)考虑了购买周期中的所有商品,需要计算 kn 次,然而在实验中由于其他因素的影响(平台内部的任务调度、机器集群的网络状况、磁盘 I/O 占用的时间等),这些方法的运行时间都大致相同. 除了以上一些理论上分析的结果,我们也观察到使用 3 台机器、3 个分布式结点的情况下:在数据集上训练时,LF-CM 以及 LFS-CM 一次完整迭代需要的 2 次 Map-Reduce 作业用时都是在 210 s 左右,10~15 次迭代后基本趋于收敛;KNN,Rand,Pop 等算法(不包括预处理)都是单次 Map-Reduce 作业,运行时间约在 100 s. 而在测试集上,对任一个购买周期所有的方法基本上都可实时地给出预测结果.

5 总 结

本文主要研究电商网站中用户的在线购买行为预测问题.通过引入购物周期和机会成本等概念,试图将用户在相似商品之间的比较和选择行为进行建模,从而解决用户实时偏好和真实购买意图的理解问题.具体来说,本文首先将用户的行为序列分成了不同的购买周期,然后通过考虑商品对用户的效用情况,提出了针对所购买商品和最佳替代商品的、基于潜在因子的选择模型 LF-CM,以及针对所有商品的、基于潜在因子和行为序列效用的选择模型 LFS-CM 两种算法,分别实现对用户购买行为的预测.本文使用了 Tmall 用户的购物行为日志对所提出的模型进行了验证,实验结果证实了所提出的方法在用户在线购买预测上的有效性.

在未来工作中,将重点关注更综合的描述模型,考虑更多的潜在因素.例如,目前本文中没有考虑到用户的不同类型操作行为(点击、购买和收藏)的不同作用,将在未来工作中加以区分;又如,考虑将商品的先验知识,如类别、价格等因素纳入建模,从而进一步提升模型预测的精度.

参 考 文 献

[1] Rassin E, Muris P. Indecisiveness and the interpretation of ambiguous situations [J]. Personality and Individual Differences, 2005, 39(7): 1285-1291

[2] Liu Q. et al. Mining indecisiveness in customer behaviors [C] //Proc of the 15th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2015: 281-290

[3] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms [C] //Proc of the 10th Int Conf on World Wide Web. New York: ACM, 2001: 285-295

[4] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering [J]. IEEE Internet Computing, 2003, 7(1): 76-80

[5] Luce R D. Individual choice behavior: A theoretical analysis [M]. North Chelmsford, Massachusetts: Courier Corporation, 2005

[6] Train K. Qualitative choice analysis: Theory, econometrics, and an application to automobile demand [M]. Cambridge, Massachusetts: MIT Press, 1986

[7] Talluri K, Van Ryzin G. Revenue management under a general discrete choice model of consumer behavior [J]. Management Science, 2004, 50(1): 15-33

[8] Hosmer Jr D W, Lemeshow S. Applied logistic regression [M]. Hoboken, NJ: John Wiley & Sons, 2004

[9] Cappellari L, Jenkins S P. Multivariate probit regression using simulated maximum likelihood [J]. The Stata Journal, 2003, 3(3): 278-294

[10] Harrell F E. Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis [M]. New York: Springer Science & Business Media, 2013

[11] Train K. A validation test of a disaggregate mode choice model [J]. Transportation Research, 1978, 12(3): 167-174

[12] Ramming M S. Network knowledge and route choice [D]. Cambridge, MA: Massachusetts Institute of Technology, 2001

[13] Fuller W C, Manski C F, Wise D A. New evidence on the economic determinants of postsecondary schooling choices [J]. Journal of Human Resources, 1982, 17(4): 477-498

[14] Li Xin, Xu G, Chen E, et al. Learning recency based comparative choice towards point-of-interest recommendation [J]. Expert Systems with Applications, 2015, 42(9): 4274-4283

[15] Kumar R, Mahdian M, Pang B, et al. Driven by food: Modeling geographic choice [C] //Proc of the 8th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2015: 213-222

[16] Chen W, Li Z, Zhang M. Linear and non-linear models for purchase prediction [C] //Proc of the 2015 Int ACM Recommender Systems Challenge. New York: ACM, 2015: No. 9

[17] Li Q, Gu M, Zhou K, et al. Multi-classes feature engineering with sliding window for purchase prediction in mobile commerce [C] // Proc of the 15th Int Conf On Data Mining Workshop. Piscataway, NJ: IEEE, 2015: 1048-1054

[18] Li D, Zhao G, Wang Z, et al. A method of purchase prediction based on user behavior log [C] // Proc of the 15th Int Conf on Data Mining Workshop. Piscataway, NJ: IEEE, 2015: 1031-1039

[19] Rodriguez J J, Kuncheva L I, Alonso C J. Rotation forest: A new classifier ensemble method [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28(10): 1619-1630

[20] Zhang Y, Pennacchiotti M. Predicting purchase behaviors from social media [C] //Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 1521-1532

[21] Tsuboi Y, Jatowt A, Tanaka K. Product purchase prediction based on time series data analysis in social media [C] //Proc of the 2015 IEEE/WIC/ACM Int Conf on Web Intelligence. Piscataway, NJ: IEEE, 2015: 219-224

[22] Wang Y, Li J, Liu Q, et al. Prediction of purchase behaviors across heterogeneous social networks [J]. The Journal of Supercomputing, 2015, 71(9): 3320-3336

[23] Song Q, Cheng J, Yuan T, et al. Personalized recommendation meets your next favorite [C] //Proc of the 24th ACM Int on Conf on Information and Knowledge Management. New York: ACM, 2015: 1775-1778

[24] Balabanović M, Shoham Y. Fab: Content-based, collaborative recommendation [J]. Communications of the ACM, 1997, 40(3): 66-72

[25] Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo [C] //Proc of the 25th Int Conf on Machine Learning. New York: ACM, 2008: 880-887

[26] Yu K, Schwaighofer A, Tresp V, et al. Probabilistic memory-based collaborative filtering [J]. IEEE Trans on Knowledge and Data Engineering, 2004, 16(1): 56-69

[27] Pennock D M, Horvitz E, Lawrence S, et al. Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach [C] //Proc of the 16th Conf on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2000: 473-480

[28] Lam X N, Vu T, et al. Addressing cold-start problem in recommendation systems [C] //Proc of the 2nd Int Conf on Ubiquitous Information Management and Communication. New York: ACM, 2008: 208-211

[29] Yin P, Luo P, Lee W C, et al. Silence is also evidence: Interpreting dwell time for recommendation from psychological perspective [C] //Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 989-997

[30] Ding Xiangwu, Guo Tao, Wang Mei, et al. A clustering algorithm for large-scale categorical data and its parallel implementation [J]. Journal of Computer Research and Development, 2016, 53(5): 1063-1071 (in Chinese)
(丁祥武,郭涛,王梅,等. 一种大规模分类数据聚类算法及其并行实现[J]. 计算机研究与发展, 2016, 53(5): 1063-1071)

[31] Gemulla R, Nijkamp E, Haas P J, et al. Large-scale matrix factorization with distributed stochastic gradient descent [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011: 69-77

[32] Wang J, De Vries A P, Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion [C] //Proc of the 29th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2006: 501-508

[33] Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback [C] // Proc of the 24th Conf on Uncertainty in Artificial Intelligence. Arlington, Virginia: AUAI, 2009: 452-461

[34] Jiang Zhuoxuan, Zhang Yan, Li Xiaoming. Learning behavior analysis and prediction based on MOOC data [J]. Journal of Computer Research and Development, 2015, 52 (3): 614-628 (in Chinese)
(蒋卓轩,张岩,李晓明. 基于数据的学习行为分析与预测[J]. 计算机研究与发展, 2015, 52(3): 614-628)

[35] Friedman J H. Stochastic gradient boosting [J]. Computational Statistics & Data Analysis, 2002, 38(4): 367-378

Zeng Xianyu, born in 1991. MSc candidate. His main research interests include data mining and recommender system.

Liu Qi, born in 1986. PhD, associate professor. His main research interests include data mining and knowledge discovery in database, machine learning method and application.

Zhao Hongke, born in 1988. PhD candidate. His main research interests include data mining, internet-based finance such as crowdfunding and P2P lending.

Xu Tong, born in 1988. PhD candidate, assistant researcher. His main research interests include social network & media analysis, mobile computing, recommender system and other data mining related techniques.

Wang Yijun, born in 1991. MSc candidate. Her main research interests include data mining and recommender system.

Chen Enhong, born in 1968. PhD, professor and PhD supervisor. His main research interests include data mining and machine learning, social network analysis, and recommender systems.