SMART:一种面向电商平台快速消费品的图推荐算法

卿 勇[1] 刘梦娟[2] 银 盈[2] 李杨曦[2]
(达州职业技术学院 达州 635001)[1] (电子科技大学信息与软件工程学院 成都 610054)[2]

摘 要 提出一种针对电商平台快速消费品的图推荐算法SMART。该算法在传统二部图推荐算法的基础上增加商品种类节点及其与用户和商品两类节点的关联边,且利用每个用户对商品种类和单个商品的兴趣倾向设置无向边的权重,通过节点间转移概率的差异实现有倾向性的随机游走;经过多次迭代,用户节点到所有节点的游走概率会收敛到稳定值,收敛后的游走概率能够在一定程度上体现用户对商品的购买概率;最后考虑每个用户对商品所属商家的兴趣偏好,对用户节点到各商品节点的游走概率进行调整,并根据调整后的游走概率计算每个用户的TOP-N推荐列表。在京东生鲜类商品的评论数据集上对所提出的推荐算法进行性能评价,实验结果表明该算法的确能够提供高质量的推荐,与基本二部图推荐算法相比,准确率提高了1.32%,召回率提高了1.48%。

关键词 图推荐,随机游走,购买兴趣,重复推荐
中图法分类号 TP399 **文献标识码** A

SMART:A Graph-based Recommendation Algorithm for Fast Moving Consumer Goods in E-commerce Platform

QING Yong[1] LIU Meng-juan[2] YIN Ying[2] LI Yang-xi[2]
(Dazhou Vocational and Technical College,Dazhou 635001,China)[1]
(School of Information and Software Engineering,University of Electronic Science and Technology of China,Chengdu 610054,China)[2]

Abstract This paper proposed a new graph-based recommendation algorithm for fast consuming goods in e-commerce platform,called SMART. Different from the traditional recommendation based on user-item bigraph,there are four types of nodes including users,items,categories,and their corresponding edges. In the new user-item-category graph,the weight of each undirected edge is set according to each user's interests in the items and their categories,so we can run biased random walks on the weighted graph. After many iterations,the probabilities from each user node walking to all other nodes would converge to stable values. It is believed that the convergence probabilities can reflect the probabilities with which users purchase the goods. At last,we explored the user's preference for shops to adjust the convergence probabilities and compute the TOP-N recommendation list. We evaluated the performance of the proposed algorithm based on the dataset of comment records in JD fresh goods,and the results show that our algorithm can provide high quality recommendation. Compared with the basic bigraph-based recommendation,the accuracy and the recall of our algorithm have increased by 1.32% and 1.48% respectively.

Keywords Graph-based recommendation,Random walk,Purchase interests,Repeated recommendation

1 引言

随着网络购物的广泛普及,越来越多的用户在电商平台上购买各种商品,从家电、家具、电子设备等大件商品到服装、清洁护理、生鲜食品等日常生活用品,网络购物已经成为人们现代生活的一种重要方式。然而在大型电商平台上,面对数以亿计的商品,如何帮助用户快速发现其真正感兴趣的商品,已经成为电商平台运营商亟待解决的关键问题。目前克服该问题的主要途径有两种:一种是主动方式,由用户给出关键词,通过搜索和筛选功能提供与关键词相匹配的商品信息给用户;另一种是被动方式,由平台根据用户的浏览历史、购买行为分析用户当前的购买兴趣,从而为用户推荐购买概率高的商品,这就是个性化推荐系统。本文主要针对个性化推荐系统展开研究。

目前国内外主流的电商平台,例如亚马逊、淘宝、京东等都已经提供个性化推荐服务。目前商业化的推荐系统以基于用户行为的协同过滤算法(Collaborative Filtering,CF)为主[1],特别是基于物品的协同过滤算法(Item-based CF,Item-CF)[2]已经得到了亚马逊、京东等大型电商平台的广泛部署。其基本思想是:首先根据用户对物品的行为(例如浏览、购买、评价等),计算物品之间的相似度;然后找到与目标物品最相似的若干物品,从而给用户推荐与其之前喜欢物品相似的物

本文受国家自然科学基金(61202445,61272527),达州市科技惠民计划项目,中央高校基本科研业务费项目(ZYGX2016J096)资助。
卿 勇(1960-),男,硕士,讲师,主要研究方向为数据挖掘、物联网;刘梦娟(1979-),女,博士,副教授,主要研究方向为数据挖掘、推荐系统,E-mail:mjliu@uestc.edu.cn(通信作者);银 盈(1994-),女,硕士生,CCF会员,主要研究方向为数据挖掘;李杨曦(1987-),男,硕士,主要研究方向为数据挖掘。

品。ItemCF 的优点是简单,且它在用户行为比较密集的数据集中能获得较好的推荐性能。然而当用户行为稀疏或者存在冷启动用户或物品时,它的推荐性能会大幅下降,扩展性也较差,为此研究者提出了基于机器学习模型的协同过滤算法[3-6]。

基于模型的方案主要采用机器学习算法建立用户的偏好模型,利用历史用户行为学习得到模型参数,进而利用该模型进行预测和推荐。基于模型的方案主要包括:基于聚类的方法[3]、基于矩阵分解的方法[4]、基于贝叶斯信念网络的方法[5]以及基于马尔科夫决策过程的方法[6]等。其中基于马尔科夫决策过程的推荐算法又称为图推荐算法,其基本思想是根据用户对物品的行为建立“用户-物品”二部图,在二部图上进行随机游走,从而得到稳定状态下用户与每个物品的游走概率,进而根据游走概率(被认为是用户购买该物品的概率)为用户进行推荐。图推荐算法比较灵活,在构建二部图时,可以根据场景融入时间、地点、社交关系等信息,因此对于用户行为比较稀疏的数据集通常能够获得较好的推荐性能。

本文的目标是提出一种针对电商平台快速消费商品的用户购买习惯和兴趣偏好的图推荐算法,以提高该类商品的推荐效果。快速消费品是指使用寿命短、消费速度快的商品,主要包括个人护理用品、食品饮料、保健品、烟酒等。不同于普通的衣物、家电、书籍等商品的推荐,用户通常会多次重复购买该类商品,且对于商品的品牌、所属商家、价格、产地等属性,用户通常会有比较明确的购买兴趣和购买习惯。

首先对数据集中用户的评论行为进行分析,总结每个用户对购买商品、商品种类、商家等的兴趣偏好,同时分析用户对每类商品的购买周期和商品产地等信息;然后尝试将每个用户的购买习惯和兴趣偏好融入图推荐算法中,并在此基础上提出一个基于“用户-商品-种类”三部图的推荐算法 SMART。该算法利用图推荐算法易于融合附加信息的优点,不仅包含了用户、商品节点,还增加了商品种类节点及其关联边,并通过边的权重体现用户对商品、用户对种类的兴趣偏好;然后利用节点间转移概率的差异性在三部图上进行带倾向性的随机游走;最后得到稳态下每个用户与所有商品的游走概率,同时利用每个用户对商家的兴趣偏好,对游走概率进行调整,将调整后的游走概率按照降序排列,从而计算每个用户的 TOP-N 推荐列表。不同于普通商品推荐,SMART 更倾向于向用户推荐其曾经多次购买的种类的商品或者多次购买的商家的商品,甚至用户已经多次购买的商品,因此更符合用户在电商平台购物的习惯和兴趣。本文在京东生鲜类商品的数据集上对所提推荐算法进行性能评价,离线实验结果表明 SMART 确实能够提供高质量的推荐。

本文第 2 节介绍与本文相关的研究工作;第 3 节以京东生鲜类商品的评论数据集为例,分析用户的购买习惯及兴趣;第 4 节介绍本文提出的基于用户购买兴趣的三部图推荐算法的基本框架;第 5 节在京东生鲜类商品的数据集上对所提算法进行性能评价;最后总结全文并提出下一步工作计划。

2　相关工作

由于推荐系统在产业界获得了极大的成功,近年来学术界也出现了大量关于推荐算法的研究,其中应用最为广泛的是基于用户行为的协同过滤算法 CF。CF 的早期研究主要集中于基于邻域的方案,该类方案又可进一步细分为基于用户的协同过滤算法(User-based CF,UserCF)[1]和基于物品的协同过滤算法 ItemCF[2]。邻域方案的优点是易于实现,不用考虑用户和物品的具体内容,缺点在于数据稀疏时推荐准确性大幅降低,且不易于与其他信息融合。为此,研究者在后期提出了大量的基于机器学习模型的方案,特别是基于矩阵分解模型的推荐算法[4]在近几年得到了广泛关注。基于模型的推荐方案通常具有较高的推荐性能,且能更好解决稀疏性和扩展性问题,但是建模过程相对复杂,可能丢失重要信息。

本文采用的图推荐算法是一种利用马尔科夫决策过程的方法,它不仅能有效地计算图中节点的相似性,还能发掘节点之间传递的关联性。最早的图推荐算法是文献[6]提出的基于随机游走的二部图推荐算法 (Topic-Sensitive Page-Rank,TSPR)。该算法采用一阶 Markov-chain 计算游走概率,具体方法如下:分别从每个用户节点出发,在“用户-物品”二部图上进行随机游走;每到达一个节点时,需要判断是以概率 β 继续游走,还是以概率 $1-\beta$ 返回出发点重新游走;重复游走过程,直到二部图上的用户节点到所有节点的游走概率收敛到稳定值;最终根据用户对物品的游走概率生成每个用户的物品推荐列表。TSPR 算法实现简单,具有较高的推荐准确率,但是算法的时间复杂度较高,因此通常采用矩阵理论的方法来求解。

文献[7]在 TSPR 的基础上,综合考虑用户的长期偏好和短期偏好,提出一种可改善推荐准确率的图推荐算法。文献[8]提出一种基于图模型的多维度推荐算法,在二部图中不仅包含物品和用户节点,还包含时间和地点信息,因此在推荐时可以根据不同场景需求,结合不同的节点信息进行推荐。文献[9]利用用户的点击与查询行为,建立“用户-查询-视频”三部图,通过在三部图中进行概率游走,得到用户节点到各视频节点的访问概率,作为视频推荐的基础。由于用户与视频之间的游走路径变得更多,因此数据稀疏性问题得到了一定的缓解。文献[10]提出一种将多用户行为与用户间关注关系相结合的图推荐算法,利用差异化的用户行为表示用户对视频兴趣的差异化,利用关注关系反映用户的兴趣倾向,从而能在一定程度上解决稀疏和冷启动问题。遗憾的是目前已有的绝大多数推荐算法都是针对不重复推荐的物品的,即不考虑用户对物品的多次评分或多次购买行为,这对于电影、书籍、家电等物品的推荐场景是适用的,但对于快速消费品的推荐往往不太适合。

与已有工作相比,本文提出的图推荐算法具有以下创新点:1)通过对用户的评论数据进行分析,提出一个针对购买习惯和兴趣的用户标签标注方法,包括用户的购买兴趣及其对应的权重;2)提出一个“用户-商品-种类”三部图构图法,根据用户对商品及商品种类的购买兴趣设置“用户-商品”边和“用户-种类”边的权重,以实现带倾向性的随机游走,得到每个用户与所有商品在稳态时的游走概率;3)提出一个考虑用户对购买商家偏好的游走概率调整策略。

3　购买兴趣分析

本节将以京东生鲜类食品的评论数据为例,对其中的禽蛋、水果、猪牛羊肉、蔬菜四大类商品的信息及评论行为进行统计,分析结果如图 1 所示。由于只能爬取到京东平台上用

户对商品的评论信息，为了避免用户行为太稀疏，本文对数据集进行了预处理，只保留了评论数大于或等于 14 条的用户及其购买商品的评论信息。

图 1(a)给出了 1568 个用户购买的商品种类情况，可以发现 47.8%的用户都只购买了一个种类的商品，4 个种类都购买的用户只有 3 个，这说明绝大多数用户对在电商平台上购买的商品种类具有一定的倾向性，例如有的用户可能只在平台上购买肉类或者蔬菜类商品，有的用户只购买水果类商品；图 1(b)给出了是用户购买商品所属的商家数目的统计，可以发现 32.72%的用户只购买一个商家的商品，59.5%的用户在平台购买的商家数目小于或等于 2，只有 37 个用户购买的商家超过了 10 家；图 1(c)给出了用户购买商品的重复次数的统计，发现只有 117 个用户没有重复购买商品的行为，71.75%的用户重复购买商品的次数达到 4 次及以上。

图 2 给出了一个具体用户的购买行为的分析，图 2(a)给出了该用户 57 次购买行为在 5 个商家的分布，以及每个商家所购种类的分布，如图 2 所示，该用户只购买蔬菜和肉类两个种类，且全部蔬菜以及 65.71%的肉类都在名为“鲜农乐”的商家购买，这是一家位于北京的著名生鲜电商；图 2(b)给出了该用户购买肉类商品的时间序列，观察发现每隔一段时间用户就会再次购买肉类，但是并没有精确固定的周期；图 2(c)给出了该用户购买商品的重复次数，观察发现蔬菜中只有土豆购买了两次，其余反复购买的都是肉类，特别是“鲜农乐”的肉类商品，每种商品重复购买的时间都至少在 20 天以上。

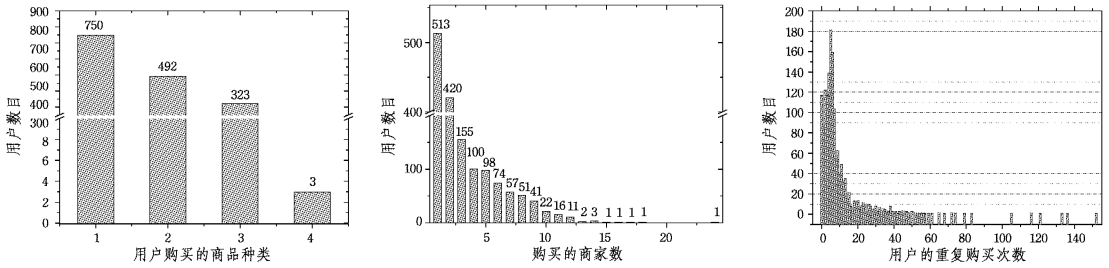

图 1 用户购买商品的行为统计

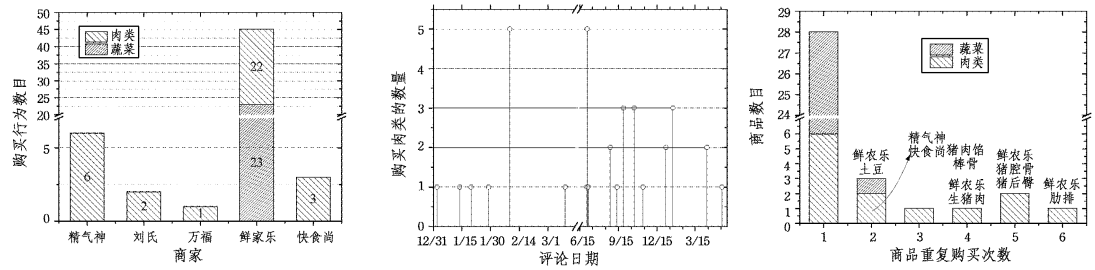

图 2 一个具体用户购买商品的行为统计

根据上述统计结果，可以得出如下分析：1)用户根据习惯通常只在电商平台上购买某些种类的商品，因此用户对于感兴趣的商品种类带有一定的倾向性；2)每个用户通常会有比较认可的品牌商家，会更倾向于购买这些品牌商家的商品；3)由于是快速消费品，用户通常会重复购买曾经购买且评分较高的商品。此外，本文还分析了用户购买商品的产地信息，发现由于冷链物流的广泛应用，除新鲜蔬菜类的商品外，其余商品几乎无产地方面的限制；而对于新鲜蔬菜，无论是商家还是用户都会有所限制。

4 SMART 算法框架

本节将详细介绍本文提出的基于购买习惯和兴趣偏好的三部图推荐算法 SMART。其基本思路如下：首先根据已有的用户购买行为，标注用户的购买兴趣及对应权重；其次构建“用户-商品-种类”三部图，根据兴趣标签及权重进行带倾向性的随机游走，得到稳定状态下用户节点与图中所有商品节点的游走概率；然后根据用户对购买商家的倾向性对收敛后的游走概率进行调节；最后按照调节后的游走概率对商品降序排列，计算 TOP-N 推荐列表。

4.1 标注用户的购买兴趣

针对每个用户的购买行为，统计其已购买商品的种类、商家及商品产地，得到该用户的购买兴趣标签及权重。假设用户 u 购买了 N 件商品，分属于 S 个种类和 M 个商家，其中每个种类购买的商品数记为 $\{num(S_1),\cdots,num(S_S)\}$，每个商家购买的商品数记为 $\{num(M_1),\cdots,num(M_M)\}$，因此可利用式(1)和式(2)分别计算该用户的种类标签和商家标签的权重。

$$w(u,S_i)=num(S_i)/\sum_{j=1}^{S}num(S_j) \tag{1}$$

$$w(u,M_i)=num(M_i)/\sum_{j=1}^{M}num(M_j) \tag{2}$$

进一步针对每个商品 I_i，利用式(3)计算用户对该商品的兴趣权重，对商品的兴趣权重不仅考虑了用户单次购买的评分，还考虑了用户对该商品重复购买的次数。假设用户对商品的评价分为 K 个等级，每个等级 l 对应的权重为 φ_l，n_l 表示用户对商品评价的等级为 l 的次数，$\{Non-K\}$ 为评价次数大于 0 的等级集合，$|Non-K|$ 为评价次数大于 0 的等级数目。此处兴趣标签的权重越大，说明该用户对该兴趣标签感兴趣的程度越高。

$$w(u,I_i)=\frac{1}{|Non-K|}\sum_{l\in\{Non-K\}} l\times\varphi_l^{n_l} \tag{3}$$

通过一个简单示例对兴趣标签标注方法进行说明。假设用户购买的商品列表为$\{I_1,I_2,I_3,I_4,I_1,I_3,I_1,I_3,I_5,I_6\}$，其中$\{I_1,I_2,I_3\}$属于种类$S_1$，购买于商家$M_1$，产地为$A_1$，$\{I_4,I_5,I_6\}$属于种类$S_2$，购买于商家$M_2$，产地为$A_2$，用户对商品的评价分为 5 个等级，每个等级对应的购买次数累积权重记为$\{1.01,1.00,0.95,0.90,0.85\}$，所有购买商品的评价均为 5 分。

根据式(1)，可计算用户对种类的兴趣权重：

$$w(u,S_1)=num(S_1)/\sum_{j=1}^{2}num(S_j)=\frac{7}{10}=0.7$$

$$w(u,S_2)=num(S_2)/\sum_{j=1}^{2}num(S_j)=\frac{3}{10}=0.3$$

根据式(2)，可计算用户对商家的兴趣权重：

$$w(u,M_1)=num(M_1)/\sum_{j=1}^{2}num(M_j)=\frac{7}{10}=0.7$$

$$w(u,M_2)=num(M_2)/\sum_{j=1}^{2}num(M_j)=\frac{3}{10}=0.3$$

根据式(3)，可计算用户对商品的兴趣权重：

$$w(u,I_1)=\frac{1}{|Non-K|}\sum_{l\in\{Non-K\}} l\times\varphi_l^{n_l}=5\times1.01^3=5.15$$

本例中由于所有购买商品的评价均为 5 分，因此$\{Non-K\}=\{5\}$，$l=5$，$\varphi_5=1.01$，由于用户对物品I_1评价了 3 次，因此$n_5=3$。综上，可得该用户的购买兴趣：种类兴趣标签及权重为$\{(S_1,0.7),(S_2,0.3)\}$，商家兴趣标签及权重为$\{(M_1,0.7),(M_2,0.3)\}$，商品兴趣标签及权重为$\{(I_1,5.15),(I_2,5.05),(I_3,5.15),(I_4,5.05),(I_5,5.05),(I_6,5.05)\}$。

4.2　构建“用户-商品-种类”三部图

图推荐算法的基础是构建用户与商品的关系图，本文中构建的三部图中主要包括三类节点和三类关联边，分别是用户节点V_U、商品节点V_I、商品种类节点V_S以及“用户-商品”边E_{UI}、“用户-种类”边E_{US}、“商品-种类”边E_{IS}。V_U包含数据集中所有的用户节点，V_I包含数据集中所有的商品节点，V_S包含数据集中所有的种类节点，$V=V_U\cup V_I\cup V_S$。E_{UI}是根据训练集中用户对商品的评价行为建立的，如果用户u评价了商品i，则在用户节点v_u和商品节点v_i之间建立一条无向边$e(v_u,v_i)$。E_{US}也是根据训练集中用户购买该种类商品的评价行为建立的，如果用户u评价的商品i属于种类s，则在用户节点v_u和种类节点v_s之间建立一条无向边$e(v_u,v_s)$。E_{IS}则直接根据该商品属于的种类信息建立边，无论该商品是否在训练集中。各边的权重计算公式如式(4)—式(6)所示。

$$w(e(v_u,v_i))=w(u,I_i) \tag{4}$$

$$w(e(v_u,v_s))=\frac{1}{|I(u,s)|}\sum_{i\in I(u,s)}w(e(v_u,v_i)) \tag{5}$$

$$w(e(v_i,v_s))=\frac{1}{|U(i,s)|}\sum_{u\in U(i,s)}w(e(v_u,v_i)) \tag{6}$$

其中，$w(u,I_i)$表示用户u对商品i的兴趣权重，如式(3)所示；$I(u,s)$表示训练集中用户u评价过的且种类为s的商品集合；$U(i,s)$表示训练集中评价过商品i的用户集合。图 3 是根据 4.1 节的示例建立的三部图。

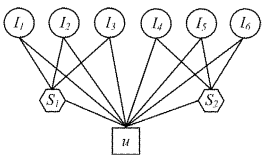

图 3　“用户-商品-种类”三部图示例

图 3 中各边的权重计算如下：

$$w(e(u,I_1))=w(e(u,I_3))=5.15$$

$$w(e(u,I_2))=w(e(u,I_4))=w(e(u,I_5))=w(e(u,I_6))=5.05$$

$$w(e(u,S_1))=(5.15+5.05+5.15)/3=5.1$$

$$w(e(u,S_2))=(5.05+5.05+5.05)/3=5.05$$

$$w(e(I_1,S_1))=w(e(I_3,S_1))=5.15$$

$$w(e(I_2,S_1))=w(e(I_4,S_2))=w(e(I_5,S_2))=w(e(I_6,S_2))=5.05$$

4.3　随机游走过程

本文中随机游走过程基本沿用了 TSPR 算法的随机游走规则，即从每个用户节点v_u出发，在三部图上随机游走。游走到任何一个节点时，首先按照概率β决定是继续游走，还是以概率$1-\beta$停止本轮游走并从节点v_u开始重新游走。如果决定继续游走，则从当前节点相连的节点中按照边的权重比例，随机选择一个节点作为下次游走经过的节点。这样经过多次随机游走后，v_u与图中所有节点的游走概率会收敛到一个稳定值。假设图中有$|V|$个节点，$r^0(u)$表示用户节点v_u与图中所有节点的游走概率的初始分布，本文中定义初始时用户节点v_u到v_i的游走概率为 1，到其余节点的游走概率为 0，如式(7)所示，m_i^0表示初始时从用户节点v_u到节点v_i的游走概率。上述随机游走过程可以用式(8)进行计算，m_i^n表示第n步后用户节点v_u到节点v_i的游走概率，$in(v_i)$表示与节点v_i直接相连的节点集合，$out(v_j)$表示与节点v_j直接相连的节点集合。

$$r^0(u)=[m_i^0]_{1\times|V|} \tag{7}$$

其中，$m_i^0=\begin{cases}1, & i=u\\ 0, & i\neq u\end{cases}$。

$r^n(u)=[m_i^n]_{1\times|V|}$，其中：

$$m_i^n=\begin{cases}(1-\beta)+\beta\sum\limits_{v_j\in in(v_i)}\dfrac{m_j^{n-1}*w(e(v_j,v_i))}{\sum\limits_{v_k\in out(v_j)}w(e(v_j,v_k))}, & i=u\\ \beta\sum\limits_{v_j\in in(v_i)}\dfrac{m_j^{n-1}*w(e(v_j,v_i))}{\sum\limits_{v_k\in out(v_j)}w(e(v_j,v_k))}, & i\neq u\end{cases} \tag{8}$$

图推荐算法的问题在于时间复杂度很高，因此本文利用矩阵运算进行求解。假设节点间的概率转移矩阵$M=[q_{ij}]_{|V|\times|V|}$定义如下：

$$q_{ij}=\begin{cases}\dfrac{w(e(v_i,v_j))}{\sum\limits_{v_k\in out(v_i)}w(e(v_i,v_k))}, & \text{if } |out(v_i)|\neq0\\ 0, & \text{otherwise}\end{cases}$$

式(8)所述的随机游走过程可转化为：

$$r^n(u)=\beta\times M^{\mathrm{T}}\times r^n(u)+(1-\beta)\times r^0(u)$$

$$\Rightarrow r^n(u)=(1-\beta)(I-\beta\cdot M^{\mathrm{T}})^{-1}r^0(u)$$

因此，只需要求解$(I-\beta\cdot M^{\mathrm{T}})^{-1}$即可求解出稳态时的随机游走概率。

4.4　游走概率调节策略

游走过程结束后，得到稳态时每个用户节点与所有商品节点的游走概率。由于在随机游走的三部图中已经将用户对商品和种类的偏好融入“用户-商品”边和“用户-种类”边的权重中，因此本节将提出一个考虑用户对商家兴趣倾向的游走概率调整策略。一种简单的调节方法是根据该商品所属的商

家与目标用户对商家的兴趣权重进行调节，即将用户与每个商品的游走概率乘以该用户对商品所属商家的兴趣权重，具体如式(9)所示，其中 m_i^n 表示稳态时用户节点 v_u 到商品节点 $v_i(v_i \in V_I)$ 的游走概率，$\eta(u,i)$ 表示商品 i 所属商家 M_i 对于用户 u 的调节权重。尝试采用 3 种简单的单调递增函数(线性递增函数、Sigmoid 函数、二次函数)来计算基于商家兴趣倾向的调节权重，如图 4 所示。其物理含义是如果用户从未购买商家的商品，则该商家的权重为 0.5，用户在商家购买的商品越多，其兴趣权重越大，调节权重越大。需要说明的是，最后需要将用户到所有商品调节后的游走概率进行归一化，使其和为 1。

$$p(u,i)=\eta(u,i) * m_i^n \tag{9}$$

其中，3 种单调递增函数的定义如下：

$$\eta(u,i)=\begin{cases}1/(1+e^{-[10*w(u,M_i)]})\\0.5 \cdot w(u,M_i)+0.5\\0.5 \cdot w(u,M_i)^2+0.5\end{cases}$$

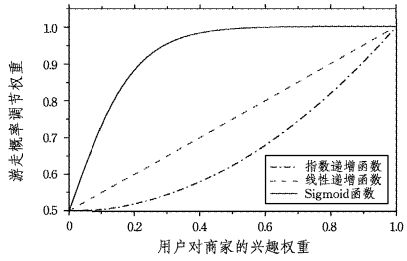

图 4 计算商家调节权重的 3 种单调递增函数

4.5 计算推荐列表

SMART 支持对商品重复推荐，因此计算推荐列表时不会排除已经购买的商品，而是针对所有商品按照调整后的游走概率进行降序排列，得到 TOP-N 推荐列表。考虑到在重复推荐已购买商品时应根据用户在重复购买该类商品时的最小时间间隔来加以调整，我们将在后续工作中进一步研究。

5 性能评价

采用京东的生鲜类商品的评论数据集对所提推荐算法 SMART 进行性能评价。评论数据集中包括 1568 个用户、250 个商家、2426 个商品、38449 条评论记录，所有商品属于 4 个种类，分别是禽蛋、水果、猪牛羊肉、蔬菜，且每个用户的评论记录至少为 14 条。训练集和测试集的划分方法为将每个用户的评论按照 5∶5 的比例随机划分。实验执行 TOP-N 推荐，性能评价指标采用准确率和召回率，其定义如式(10)和式(11)所示，其中 $R(u)$ 表示给用户 u 推荐的商品集合，$T(u)$ 表示用户 u 在测试集中实际评论的商品的集合。由于允许重复推荐，因此测试集中用户 u 可能会多次评论同一个商品，实验将测试集中重复评价的商品合并为一个商品。推荐列表中的命中个数越多，推荐准确率和召回率越高。为了进行对比，同时执行了几种经典推荐算法，对比算法的说明如表 1 所列，其中 WRMF 和 BPR 采用的是 LibRec 工具包[11]完成的实验。

$$\text{准确率}=\frac{\sum\limits_{u\in V_U}|R(u)\cap T(u)|}{\sum\limits_{u\in V_U}|R(u)|} \tag{10}$$

$$\text{召回率}=\frac{\sum\limits_{u\in V_U}|R(u)\cap T(u)|}{\sum\limits_{u\in V_U}|T(u)|} \tag{11}$$

表 1 几种对比算法的说明

算法名称	算法说明
ItemCF	基于物品的协同过滤算法(k=10)
WRMF	加权矩阵分解算法[12]
BPR	基于贝叶斯后验概率学习模型的推荐算法[13]
TSPR-0	基于二部图的推荐算法(最低评分作为边权重)
TSPR-1	基于二部图的推荐算法(本文提出的边权重计算方法)
SMART-0	基于三部图的推荐算法(最低评分作为边权重)
SMART-1	基于三部图的推荐算法(本文提出的边权重计算方法)

实验 1 主要考察增加商品种类节点以及增加无向边权重对重复购买商品的倾向性对算法性能的影响。实验分别执行算法 TSPR-0，TSPR-1，SMART-0，SMART-1，进行 TOP-10，20，30 推荐，其中随机游走时的继续游走概率 β=0.5，算法性能如表 2 所列。实验结果显示，只考虑“用户-物品”二部图的基本随机游走算法的性能是最差的，在 TOP-10 时，TSPR-0的准确率和召回率分别为 30.50%和 34.19%；通过增加“种类”节点及“用户-种类”和“商品-种类”边可以在一定程度上提高推荐性能，SMART-0 的准确率和召回率分别提升到 31.14%和 34.91%；SMART-1 通过增加用户重复购买商品的边的权重，可以略微增加推荐算法的准确率和召回率。

表 2 增加种类节点以及调整边的权重对算法性能的影响/%

评价指标	TSPR-0	TSPR-1	SMART-0	SMART-1
准确率@10	30.50	30.72	31.14	31.31
召回率@10	34.19	34.44	34.91	35.09
准确率@20	21.55	21.67	21.71	22.03
召回率@20	48.31	48.59	48.68	49.39
准确率@30	16.52	16.62	16.54	16.86
召回率@30	55.54	55.90	55.61	56.70

实验 2 考察在基于商家偏好的游走概率调整策略中，调节权重的 3 种计算方法对算法性能的影响，实验结果如表 3 所列。SMART-1 是只依据稳态游走概率进行推荐，SMART-2 是基于 Sigmoid 函数计算调整权重的推荐，SMART-3 是基于线性函数计算调整权重的推荐，SMART-4 是基于二次函数计算调整权重的推荐。实验结果显示，基于商家偏好进行游走概率调整策略确实有助于提高推荐性能，其中基于 Sigmoid 函数计算调整权重的 SMART-2 算法能够获得最优的推荐性能，SMART-3 算法次之，SMART-4 略低于 SMART-3。

表 3 3 种调整权重计算方法对推荐性能的影响/%

评价指标	TSPR-1	TSPR-2	SMART-3	SMART-4
准确率@10	31.31	31.82	31.75	31.73
召回率@10	35.09	35.67	35.59	35.56
准确率@20	22.03	22.23	22.14	22.12
召回率@20	49.39	49.82	49.62	49.58
准确率@30	16.86	17.01	16.98	16.96
召回率@30	56.70	57.20	57.10	57.02

实验 3 将 SMART 算法与其他经典推荐算法进行对比，结果如表 4 所列。其中 ItemCF，WRMF，BPR 3 种算法的推荐性能都明显低于本文提出的 SMART 算法和 TSPR 算法，这是因为后两种算法都支持重复推荐，即算法产生的推荐列表中可以包含已经在训练集中评价过的商品，而前 3 种算法都是不支持重复推荐的，因此在产生列表时会排除训练集中已经评价的商品。为了更加完整地给出评价，表 4 列出了不支持重复推荐的 SMART 和 TSPR 算法的性能指标，结果显示在不考虑重复推荐的情况下，本文提出的 SMART 算法

仍然能够取得最优的推荐性能。

表 4　几种经典推荐算法的性能对比/%

算法	准确率@10	召回率@10	准确率@20	召回率@20	准确率@30	召回率@30
ItemCF	15.29	17.14	9.71	21.76	7.32	24.61
BPR	15.64	17.53	10.79	24.18	8.40	28.27
WRMF	15.50	17.38	10.74	24.08	8.42	28.32
TSPR-0(不重复)	16.19	18.14	10.93	24.50	8.59	28.89
SMART-2(不重复)	16.68	18.69	11.24	25.20	8.82	29.64
TSPR-0(可重复)	30.50	34.19	21.55	48.31	16.52	55.54
SMART-2(可重复)	31.82	35.67	22.23	49.82	17.01	57.20

综上分析,本文提出的 SMART 算法确实能够捕捉到电商平台用户对快速消费商品的购买兴趣和购买习惯,从而在一定程度上提升了推荐性能,特别是针对用户重复购买商品的场景,准确率和召回率都达到了较高的水平。

结束语　本文针对电商平台上用户对快速消费品的购买习惯和购买兴趣,提出一个基于"用户-商品-种类"三部图的推荐算法,通过将用户对商品和商品种类的偏好融入节点间的转移概率中,实现了带有兴趣倾向性的随机游走,使得稳态时的游走概率能够在一定程度上体现用户的购买习惯和兴趣,从而提升算法的推荐性能;此外,文中还针对用户对商家的兴趣偏好,提出一个游走概率调整策略,使商品推荐更倾向于该用户感兴趣的商家。本文利用真实的京东生鲜类商品的评论数据集对所提的推荐算法进行了性能评价,结果显示 SMART 算法确实能够捕捉到电商平台用户对快速消费商品的购买兴趣和购买习惯,特别是针对用户重复购买商品的场景,准确率和召回率都达到了较高的水平。但是电商平台上的商品推荐还应该考虑用户对于重复购买同类商品的时间间隔,这些工作都将在后续研究中完成。另外,本文提出的图推荐算法不仅适用于快速消费品的推荐,同样适用于餐馆推荐或者菜品推荐,因此后续研究的另一个工作是将算法推广到餐馆、菜品推荐等其他场景中。

参 考 文 献

[1] BELLOGIN A, PARAPAR J. Using graph partitioning techniques for neighbor selection in user-based collaborative filtering [C] // Proceedings of the Sixth ACM Conference on Recommender Systems. Dublin, Ireland, 2012: 213-216.

[2] LINDEN G, SMITH B, YORK J. Amazon. com Recommendations: Item-to-Item Collaborative Filtering [J]. IEEE Internet-Computing, 2003, 7(1): 76-80.

[3] MORADI P, AHMADIAN S, AKHLAGHIAN F. An effective trust-based recommendation method using a novel graph clustering algorithm[J]. Physica A: Statistical Mechanics and its Applications, 2015, 436: 462-481.

[4] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009(8): 30-37.

[5] SU X, KHOSGOFTAAR T M. Collaborative filtering for multi-class data using belief nets algorithms [C]// 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI). 2006.

[6] HAVELIWALA T H. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 784-796.

[7] XIANG L, YUAN Q, ZHAO S, et al. Temporal recommendation on graphs via long- and short-term preference fusion [C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 723-732.

[8] LEE S, PARK S, KAHNG M, et al. PathRank: Ranking nodes on a heterogeneous graph for flexible hybrid recommender systems [J]. Expert Systems with Applications, 2013, 40(2): 684-697.

[9] CHEN B S, WANG J D, HUANG Q H, et al. Personalized Video Recommendation Through Tripartite Graph Propagation[C]// Proceedings of the 20th ACM International Conference on Multimedia. Nara, Japan, 2012: 1133-1136.

[10] 刘梦娟,王巍,李杨曦,等. AttentionRank$^+$:一种基于关注关系与多用户行为的图推荐算法[J]. 计算机学报, 2017, 40(3): 634-648.

[11] LibRec 工具箱[OL]. http://www.librec.net.

[12] DIAZ-AVILES E, DRUMOND L, SCHMIDT-THIEME L, et al. Real-time top-n recommendation in social streams[C]// ACM Conference on Recommender Systems. ACM, 2012: 59-66.

[13] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking from implicit feedback [C]// Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. Chicago, USA, 2009: 452-461.

(上接第 463 页)

[11] 刘燕兵,谭建龙,郭莉. 可动态增删关键词的串匹配算法[J]. 计算机工程与应用, 2005, 41(35): 138-141.

[12] 张莉婧,李业丽,曾庆涛,等. 基于改进 TextRank 的关键词抽取算法[J]. 北京印刷学院学报, 2016, 24(4): 51-55.

[13] ZHU L, WANG G J, ZOU X C. A Study of Chinese Document Representation and Classification with Word2vec[C]// 9th International Symposium on Computational Intelligence and Design(ISCID). 2016: 298-302.

[14] LIU C H, LIU Q, LEE C H. Valence-arousal ratings prediction of Chinese words using similarity measures based on Word2Vec [C]// International Conference on Asian Language Processing (IALP). 2016: 317-319.

[15] 游博. 词语语义相关度计算研究[D]. 武汉:华中师范大学, 2013.

[16] 李璐,张国印,等. 基于 SVM 的主题爬虫技术研究[J]. 计算机科学, 2015, 42(2): 118-122.

[17] 蒋华荣,郁雪. 应用遗传算法优化子空间的 SVM 分类算法分类算法[J]. 计算机科学, 2013, 40(11): 255-260, 275.

[18] 王良芳. 文本挖掘关键词提取算法的研究[D]. 杭州:浙江工业大学, 2013.