推荐系统中的新颖性问题研究*

徐元萍，陈 翔†
（北京理工大学 管理与经济学院，北京 100081）

摘 要：准确性推荐中存在商品类型单一、流行商品多、缺乏新意的问题，因而新颖性推荐得到重视，但已有研究在设计算法时未考虑项的特征，无法针对不同用户帮其区分和挑选具备较高新颖度的项。为提高推荐系统的性能，对基于随机游走的方法进行改进，提出融合新颖性特征的推荐算法。从兴趣扩展和预测角度分析项的特征，给出完善的新颖度定义，并结合用户需求构建新的转移概率，产生个性化的推荐列表，提高了列表内容的新意。实验结果表明，提出的算法较现有算法对准确率影响较小，同时在新颖性指标上有明显提升，并得出通过融合新颖性特征能够在兼顾准确性的情况下有效改善推荐内容的结论。

关键词：推荐算法；准确性；新颖性；随机游走
中图分类号：TP181 **文献标志码**：A **文章编号**：1001-3695(2020)08-014-2310-05
doi:10.19734/j.issn.1001-3695.2019.03.0046

Research on novelty problems in recommendation systems

Xu Yuanping, Chen Xiang†
(*School of Management & Economics*, *Beijing Institute of Technology*, *Beijing* 100081, *China*)

Abstract: Focusing on problems of the accuracy recommendation system that the recommended commodity type is relatively single, and commodities are mostly popular goods and lack of freshness, the novelty recommendation is gradually gaining attention. However, current researches don't combine item features when designing algorithms, which make them unable to distinguish and select items with higher novelty for different users. In order to improve the performance of the recommendation system, this paper improved the method based on random walk and designed a new recommendation algorithm by fusing novelty features. This algorithm further analyzed features of items and gave the formal definition of the novelty from perspectives of user interest expansion and prediction. This paper analyzed user demands, constructed new transition probability, generated personalized recommendation lists and improved the novelty of the lists. The experimental results show that the proposed algorithm has less influence on the accuracy than existing methods and has significant improvement on novelty indexes. It concludes that by fusing novel features, this system can improve the recommendation contents effectively while taking into account the accuracy.

Key words: recommendation algorithm; accuracy; novelty; random walk

推荐系统能够根据用户的历史行为向用户推荐感兴趣的物品，在缓解信息过载、减少用户搜索时间、提高个性化体验方面发挥着重要作用。目前，常用的推荐算法主要包括协同过滤、基于物品特征的推荐和基于网络结构的随机游走推荐[1]。其中，随机游走方法具备较强的扩展性，能够有效缓解数据稀疏性问题，显著提高推荐效果。然而现有算法大多在推荐时仅注重与用户历史消费记录或行为的相似性，造成推荐结果的冗余和严重的同质化问题，使用户难以接触到新鲜、多样的内容，降低了用户满意度[2]。针对准确性推荐的不足，Herlocker 等人[3]提出新颖性推荐的概念，即向目标用户推荐其有潜在兴趣但不知道的商品。推荐系统的质量和实用性在很大程度上取决于结果的新颖程度。新颖的推荐列表能够扩大用户兴趣范围，促进用户产生惊喜感，提升用户体验。现有的新颖性推荐算法主要基于用户和项的交互信息，将推荐问题转换为网络图，通过补充特定的属性信息或利用参数结合复杂动力学方法一定程度上提高了推荐列表的新颖性；但现有算法没有严格区分同一项目对不同用户的新颖度，未引入明确的新颖性项目特征进行综合建模导致无法帮助用户匹配具备较高新颖度的项，难以满足人们对个性化推荐的需要，新颖性推荐问题亟需解决。

用户兴趣的不同导致产生的历史选择不同，对同一项目的了解程度和感受往往存在差异。因此，随着用户个性化需求与日俱增，推荐系统需要区分不同用户对同一项目的了解程度，得到差异化的项目新颖度。为了结合用户自身兴趣发现新颖项目，开阔用户视野，需要进一步扩大用户兴趣边界，融入和发现新的项目类型以提高用户感兴趣的可能性。在提供个性化服务时，考虑用户兴趣的变化能够增加推荐结果的竞争力[4]。因此，有必要更深入地挖掘用户新的兴趣用于提高推荐结果的新意。本文从兴趣扩展和预测角度提出融合新颖性特征的随机游走算法，结合用户历史选择进一步补充项目特征和度量，在兼顾准确性的基础上针对不同用户产生个性化的新颖结果。

1 相关工作

目前，基于网络结构的随机游走算法广泛应用在推荐系统中，提高了推荐精度。文献[5]应用随机游走方法在用户、项目、标签构成的三元交互图上探索用户和项目之间的关系，得到的实验结果表明提出的方法有效提高了准确度；文献[6]通过整合潜在主题模型和重启动随机游走方法发现用户的隐含偏好和项目潜在的特征信息，有效缓解了数据稀疏性问题。但以上方法仅注重与用户历史偏好接近的推荐，往往限制了用户的选择，缩小用户视野，造成用户对推荐结果表示不满。

针对准确性推荐存在的问题，近年来新颖性推荐逐渐得到关注。为明确相关定义，一些研究尝试运用已有的冷启动方法

收稿日期：2019-03-14；**修回日期**：2019-05-17 **基金项目**：国家自然科学基金资助项目(71572013,71872013)
作者简介：徐元萍(1994-)，女，河南周口人，硕士，主要研究方向为推荐系统、数据挖掘；陈翔(1976-)，男(通信作者)，江西赣州人，教授，主要研究方向为计算机软件及计算机应用(chenxiang@bit.edu.cn).

解释新颖性。文献[7]在解决新项目推荐问题的同时发现由于这部分项目还未被广大用户熟知,流行度很低,对用户来说往往具备新意,并根据推荐列表中新项目所占的比例衡量推荐结果新颖度;文献[8]基于音乐播放数据,将音乐推荐列表中用户所知道艺术家所占的百分比定义为新颖度。然而,以上研究仅评价了推荐结果整体的新颖情况,并没有针对项的定义。部分研究从不同角度进一步给出具体定义。文献[9]考虑评分时间,引入"革新者"和"潜在跟随者"的概念,认为革新者评分的项对目标用户来说具有新颖性;文献[10]认为新颖性推荐应当被严格定义为向用户推荐不知道的、不具备重复性的项目,同时能够为用户带来意想不到的体验;文献[11]认为新颖性项目对系统来说是新加入的、没有评分的项目,对用户来说是用户不知道、不了解的项目或者是被用户遗忘的项目。然而以上研究无法根据项目特征针对性地衡量新颖度。由于项目的流行情况方便计算,常认为较低流行度的商品具有更高的新颖性[12,13]。利用全局值不能反映用户的兴趣,无法表示同一项目对不同用户的新颖度。随着新颖性概念的提出,相关研究将基于网络的随机游走算法应用于新颖性推荐场景中。文献[13]基于用户—项目二分图,利用参数将随机游走与热传导理论结合,提高了准确度和新颖性;文献[14]发现在用户—资源二分图中加入用户关系信息和在资源—标签二分图中加入属性信息,通过参数调节游走结果有效地提高了推荐的新颖性和多样性。以上算法本身依赖节点的度关系,在节点选择时忽略了用户对新颖性节点的偏好,无法引导资源分配到新颖度较高的节点。为寻找具备新颖特征的项目,文献[15]利用项的拓扑信息来帮助用户发现连接用户历史选择和新类型的新颖节点,打开用户视野,但该算法适用于项目的发现,没有考虑推荐的准确度;文献[16]结合协同过滤和重启动随机游走方法,通过调整用户之间的相似性将类别间距离较远的新颖项目推荐给目标用户;文献[17]通过采用 Auralist 推荐框架,构建以项为节点、项相似度为边的关系图,通过将特定的项加入目标用户子图后计算得到的聚集因子值评价项的新颖度,该值越大,项目具备的新颖度则越高。以上方法无法区分和排序同类下物品的新颖度,导致无法推荐相同类别中的商品。

为凸显用户差异性,针对不同用户为其选择新颖性较高的项,本文进一步分析用户兴趣,补充项目特征和相应度量,改进基于网络结构的随机游走算法,提出一种个性化的新颖性推荐系统,在确保准确度的情况下融合新颖性项目特征进行综合建模,改善新颖性推荐问题。

2　问题定义

在现实生活中,由于用户兴趣存在差异,同一项目体现的新颖程度应有所区别。针对目前项目新颖度难以区分的情况,在利用流行度定义项目新颖度的基础上进一步结合用户兴趣,补充新颖性项目度量。

1)流行度　由于用户自身难以发现流行度较低的项目(即评分较少的或者新的项目),此类项目对用户来说较为陌生,具备一定的新意,以往的研究常根据流行情况度量新颖性。

特征 1　流行度较低的项目,即评分较少或者评分为 0 的项目。在项目集合 I 中,项目 $i(i\in I)$ 的流行度为 popularity(i),则 i 具备的新颖性特征 N_x 可表示为

$$N_x=\text{popularity}(i),\text{popularity}(i)\leqslant\delta \tag{1}$$

其中:δ 表示流行度的阈值。

度量 1　基于流行度的新颖性。给定项目集合,根据流行度计算项目 i 具备的新颖度 novelty$_x$,表示为

$$\text{novelty}_x(i)=\text{unpopularity}(i)=1-\text{popularity}(i)=1-\frac{R_i-\min(\boldsymbol{R})}{\max(\boldsymbol{R})-\min(\boldsymbol{R})} \tag{2}$$

其中:R_i 表示项目 i 的评论数量;$\boldsymbol{R}$ 表示所有项目的评论数量矩阵;unpopularity(i)表示反流行度。项目流行度越低,新颖度越高。

2)多重兴趣特征　针对流行度作为全局特征值的不足,需要结合用户自身兴趣补充个性化特征。由于传统推荐方法多根据用户的历史选择匹配高相似度的产品,局限了推荐列表中的产品类型,降低了小众商品变为流行商品的可能性,同时用户易对推荐结果感到枯燥、乏味[2],为此,本文进一步扩展用户兴趣,打开用户视野、提高内容的吸引力。

文献[18]发现以两个项目具备的特征为基础,结合其他项目特征进行特征混合,得到的推荐结果具有惊喜性、新颖性;但其仅以两个项目为输入,没有全面考虑用户选择单个或者多个项目的情况。基于该发现,本文进一步考虑融合用户选择过的所有项目特征来扩展用户兴趣。在设计推荐系统时常将社会网络结构引入用户兴趣模型中以提高推荐效果[19]。受到结构洞理论的启发,将结构洞的概念引入新颖性推荐系统研究,用于发现项目关系网络中连接多个兴趣主题的中介项目。此类项目由于具备所连接兴趣主题的多重混合兴趣特征,能够给用户带来意想不到的体验,具备一定的新颖性。定义方式如下:

特征 2　具备多重兴趣特征的中介项目。基于项目关系构建网络 $G=(V,E)$ 和用户—项目网络 $G_U=(U,I_U)$(其中:V 是项目节点 V_i 集合;E 是边的集合;U 表示用户集合;I_U 表示用户选择过的项目集合),发现连接用户原有兴趣社区 S_U 和其他类型社区 $S_{\overline{U}}$的中介节点 V_B,根据 V_B 在网络图中的中介位置表示新颖性特征 N_y。

$$N_y=|V_B=V_i,V_i\in S_U\&V_i\in S_{\overline{U}}| \tag{3}$$

如图 1 所示,节点 1、2、3 表示用户选择过的项目,S_U 表示用户的兴趣类型。通过兴趣扩展找到中介节点 V_B,V_B 对于用户来说具备新颖性。结构洞的中介中心性表现为节点在社会网络中所处位置的中介程度,节点中介性越强,越需要发挥该节点的作用扩展用户兴趣,找到扩展后用户可能感兴趣的新类型。则节点的中介中心性在网络图中的位置表现为项目新颖程度。中介中心性越强,进行兴趣扩展的重要性越高,新颖性越高。

度量 2　基于多重兴趣特征的新颖性。给定项目关系网络,在用户原始兴趣基础上进行扩展,考虑连接用户原始兴趣和其他兴趣主题的项目节点 V_i 所处位置的中介中心性 betweenness(i),将超过中心性阈值 λ 的节点设定为中介项目,则中介项目 V_i 具备的新颖度 novelty$_y$ 为

$$\text{novelty}_y(i)=\begin{cases}\text{betweenness}(i),\text{betweenness}(i)\geqslant\lambda & V_i\in c\\ 0 & V_i\notin c\end{cases} \tag{4}$$

其中:c 表示用户原有兴趣的邻接兴趣类别;λ 表示中心性阈值。基于电流思想的随机游走算法[20],节点 i 中心性的度量等于电流从起始节点 s 到终节点 t 经过节点 i 的净次数。

$$\text{betweenness}(i)=\frac{\sum_{s<t}I_i^{(st)}}{\frac{1}{2}n(n-1)} \tag{5}$$

其中:n 表示网络中节点数;$I_i^{(st)}$ 表示通过节点 i 的净流量。从兴趣扩展的角度补充新颖性项目特征,打破单一使用流行度定义新颖性项目的局限性,同时通过兴趣扩展挖掘用户潜在的新兴趣。

3)新的兴趣主题　多数研究建立在用户兴趣固定不变的基础上,而实际生活中用户的兴趣常受外界因素的影响,从而产生新的需求。因此考虑用户兴趣变化是提高个性化推荐质量的一个重要方面[7]。现有的兴趣漂移研究主要通过监测用户兴趣的变化,根据变化后的兴趣向用户推荐可能感兴趣的商品,很少有研究从用户潜在变化的兴趣角度定义新颖性。考虑到用户对新兴趣中包含的项目并不熟悉或者需要花费时间了解,又存在潜在需求,对用户来说具备新鲜感。对此,本文进一

步预测用户兴趣可能变化的方向,发现用户可能感兴趣的新类型,并推荐其中的商品。用户新兴趣的出现一般与历史选择有密切的联系,利用具备特征 2 的中介项目发现与用户历史兴趣密切相关的新兴趣,其中包含的商品具备新颖性,定义如下:

特征 3 新兴趣主题中包含的项目。给定项目关系网络 $G=(V,E)$ 和用户—项目网络 $G_U=(U,I_U)$。通过中介节点 $V_B(V_B\in S_{\bar{U}})$ 的连接关系发现用户未选择的社区 S,S 中的节点具备新颖性 N_z。

$$N_z=|V_i,V_i\in S,S\in(S_{\bar{U}}|V_B\in S_{\bar{U}})| \tag{6}$$

如图 1 所示,通过与用户原有兴趣相连的中介节点 V_B 预测用户下阶段兴趣可能改变为 S_1 所在的兴趣类别,S_1 类别中包含的项目对用户来说具备新颖性。

对于用户来说,新兴趣与用户历史兴趣相差越大,兴趣主题间距离较远,用户对新兴趣了解越少,新兴趣包含的项目更有新意。因此与用户历史兴趣下的项目平均距离越远,相似度越低,项目新颖性越高。

度量 3 基于新兴趣主题的新颖性。给定项目相似性关系网络,定义在用户新兴趣主题中项目的新颖度为与用户选择过项目的平均距离,即节点 V_i 的新颖度为

$$\text{novelty}_z(i)=\begin{cases}\sum\limits_{j\in I_U}\dfrac{d(i,j)}{\text{num}(I_U)}=1-\sum\limits_{j\in I_U}\dfrac{\text{sim}(i,j)}{\text{num}(I_U)} & V_i\in S\\ 0 & V_i\notin S\end{cases} \tag{7}$$

其中:$\text{novelty}_z(i)$ 表示节点 i 在用户新的兴趣主题中的新颖性;$\text{num}(I_U)$ 表示用户选择的商品数量;$\text{sim}(i,j)$ 表示基于内容的相似性度量函数;$d(i,j)$ 表示距离函数。综合以上分析,将基于流行度、多重兴趣特征、新兴趣主题中的新颖性特征进行融合,进一步提出新颖性推荐系统的形式化定义。

定义 1 新颖性推荐系统。由较低流行度 N_x、多重兴趣特征 N_y、包含在新兴趣主题领域中 N_z 三种特征在量化函数 $F(x,y,z)$ 作用下产生的推荐系统 NOVELTY 表示为

$$\text{NOVELTY}=|N_x,N_y,N_z,F(N_x,N_y,N_z)| \tag{8}$$

其中:$F(N_x,N_y,N_z)$ 表示 N_x、N_y、N_z 三种特征间的相互作用关系;N_x、N_y、N_z 通过式(2)(4)(7)量化表示。新颖性推荐系统能够向用户推荐具备潜在兴趣但用户不知道且难以依靠自身发现的商品,此类商品需要花费时间寻找,具备新意。

3 融合新颖性特征的随机游走算法

一般的随机游走算法仅考虑了节点间的相似关系或节点度,限制了推荐列表的内容类型;而已有基于网络结构的新颖性算法多利用参数结合不同方法或加入辅助信息在保证准确度的基础上降低推荐结果的流行度,算法本身没有综合考虑项目的新颖性特征,无法在游走过程中选择用户偏好的新颖度较高的节点。本文进一步融合新颖性项目度量用于计算转移概率,提出融合新颖性特征的随机游走算法(random walk by fusing novelty features,RWFNF),以满足用户对新颖性的需求。

3.1 相似度计算

通过相似度计算为用户匹配与其历史偏好有高相似度的内容,从而能够保证推荐的准确性。提取项目的描述信息,将其表示为向量形式,即描述信息 $\boldsymbol{d}=(w_1,w_2,w_3,\cdots,w_n)$。可利用 LDA 方法划分项目所属的主题,采用 TF-IDF 方法提取特征词,计算特征权重。项目之间的相似度采用余弦相似度函数度量,得到项目之间的相似关系。相似度计算公式为

$$\text{sim}(d_i,d_j)=\cos(d_i,d_j)=\frac{d_i\times d_j}{|d_i|\times|d_j|}=\frac{\sum_t w_{ik}\times w_{jk}}{\sqrt{\sum_t w_{ik}^2}\sqrt{\sum_t w_{jk}^2}} \tag{9}$$

其中:d_i、d_j 分别为项目 i 和 j 的描述信息;w_{ik} 表示词 $t_k(k\in[1,n])$ 在项目 i 描述信息中的权重;w_{jk} 表示词 t_k 在项目 j 描述信息中的权重。与用户评分过的项目越相似,推荐准确度越高。

3.2 构造转移概率

通过分析用户历史行为,针对不同用户确定不同的新颖性需求权重。当用户倾向于新颖性推荐时,具备较高新颖度的节点应该以较大的概率被选择;当用户倾向于准确性推荐时,下一节点与当前节点的相似度越高,则被选择的可能性越大。权衡相似度,融合项目新颖性特征因素,构建转移概率。

定义 2 概率权重。在随机游走过程中,假设当前所处位置为节点 i 时,节点 j 在下一步游走时被选择的概率权重为

$$x(i,j)=\omega_1(\tilde{H},\tilde{Q})E^{\partial(\text{novelty}_x(j)-\text{novelty}_x(i))}\times E^{\beta(\text{novelty}_y(j)-\text{novelty}_y(i))}\times E^{\gamma(\text{novelty}_z(j)-\text{novelty}_z(i))}+\omega_2(\tilde{H},\tilde{Q})E^{\text{sim}(i,j)} \tag{10}$$

其中:调节因子 $\omega_1(\tilde{H},\tilde{Q})$ 表示新颖性推荐的权重函数;$\omega_2(\tilde{H},\tilde{Q})$ 表示准确性推荐的权重函数;α、β、γ 表示三种新颖性特征的权重,且 $\alpha+\beta+\gamma=1$。对于新颖性权重函数 $\omega_1(\tilde{H},\tilde{Q})$,结合信息熵[21]和用户购买情况建立函数关系。将用户选择过的商品所属类别总数 ε 定义为用户的度,信息熵为

$$H(\varepsilon)=\sum_{i=1}^{n(\varepsilon)}P_i\log_2\frac{1}{p_i} \tag{11}$$

其中:$H(\varepsilon)$ 表示度为 ε 的用户所选商品的信息熵值;随机变量 P 有 n 种可能的结果,概率分别为 $\{p_1,p_2,p_3,\cdots,p_n\}$;$P_i$ 表示度为 ε 的用户选择过的所有商品中商品的度为 i 的概率;$n(\varepsilon)$ 表示用户度为 ε 的情况下 i 的最大取值。信息熵越大,表示用户选择行为越不确定,偏好的商品类型越多样,新颖性需求越大[11]。另外,用户的新颖性需求除用信息熵间接表示以外,用户购买数量也是影响新颖性需求的关键因素。

用户对不同类别产品购买的数量越多,表明用户对多样性产品的要求越高,新颖性推荐更易满足用户需要,对准确性需求较小;相反,不同类别产品购买的数量相对较少,则系统更应该综合用户选择的产品特征为用户准确推荐符合兴趣的产品。此外,对同类产品购买的数量相对越多,准确性推荐给用户选择的空间就更小,结果越易让用户感到无趣,用户对新颖性的需求较高;相反,同类产品用户购买的数量越少,系统则倾向于准确性推荐,帮助用户获取同类型下喜欢的产品。根据以上信息,用户 u 的调节因子 $\omega_1(\tilde{H},\tilde{Q})$ 的计算公式如下:

$$\omega_1(\tilde{H}_u,\tilde{Q}_u)=\frac{\tilde{H}_u\times\tilde{Q}_u-\min(\tilde{H}\times\tilde{Q})}{\max(\tilde{H}\times\tilde{Q})-\min(\tilde{H}\times\tilde{Q})} \tag{12}$$

其中:$\tilde{H}$、$\tilde{Q}$ 分别表示信息熵、用户购买情况标准化后的结果。则准确性权重函数可表示为

$$\omega_2(\tilde{H},\tilde{Q})=1-\omega_1(\tilde{H},\tilde{Q}) \tag{13}$$

定义 3 转移概率。将概率权重归一化处理后的结果作为每个项目在下一步被选择的概率,如式(14)所示。

$$P(i,j)=(1-a)\times x(i,j)/\textstyle\sum_{j\in I}x(i,j)+a\times f \tag{14}$$

其中:a 为跳转发生概率;跳转分布 f 满足均匀分布。

3.3 基于随机游走方法的新颖性推荐

推荐系统中用户和项目之间的交互为评分,用矩阵 $\boldsymbol{R}$ 表示,其中 $R_{u,i}$ 反映某一用户对某一项目的评分值。根据用户—项目历史评分记录构建标准化评分矩阵,反映用户历史偏好。用户对项目评分的初始分布表示为 $R_{u,i}^*=\Pr(X_{u,0}=i)$。令 $P(i,j)=\Pr(X_{u,k}=j|X_{u,k-1}=i)$,将用户历史评分 $R_{u,i}^*$ 作为初始状态向量,代入构建的转移概率 $P(i,j)$ 进行随机游走,产生预测评分。在第 k 步,用户 u 在项目 j 上的概率为

$$\Pr(X_{u,k}=j)=(1-a)\sum_{i=1}^{N}\Pr(X_{u,k-1}=i)P(i,j) \tag{15}$$

其中:$\Pr(X_{u,k-1}=i)$ 表示第 $k-1$ 步,随机变量 X 取值为 i 时,用户 u 在项目 i 上的概率;N 表示项目总数。迭代式(15)直到收敛。

定义 4 用户 u 在项目 j 上的预测值。定义用户 u 在项目 j 上经过 k 次迭代的总概率值与随机游走过程中所有步长的概

率之和成正比,即

$$\Pr(X_{u,k}=j)=\frac{\sum_{k=1}^{\infty}\Pr(X_{u,k}=j)}{\sum_{k=1}^{\infty}\sum_{i=1}^{N}\Pr(X_{u,k}=i)}=m\sum_{k=1}^{\infty}\Pr(X_{u,k}=j) \tag{16}$$

其中:m 为比例系数。

定义 5　基于新颖性的评分预测。在考虑项目新颖性特征情况下,用户对项目 j 的评分预测值为 $\mathrm{rank}(u,j)$,则

$$\mathrm{rank}(u,j)=\sum_{k=1}^{\infty}\Pr(X_{u,k}=j) \tag{17}$$

所有的用户—项目对构成的评分矩阵可以表示为

$$\widetilde{\boldsymbol{R}}=\widehat{\boldsymbol{\Pr}} \tag{18}$$

对评分结果排序,选取 top k 项目推荐给目标用户。

4　实验

4.1　实验数据分析

本文利用 Python 3.6 爬取 8 070 条亚马逊图书信息,361 条用户记录,每本书包含书名、作者、商品类别、商品描述和评论数量信息,每条用户记录包含用户购买所有的图书以及评分信息。对每个用户评论信息选取 80% 作为训练集,20% 作为测试集。设定 a 为 0.2,新颖性特征权重值相等的情况进行实验。通过获取的图书类型信息划分为 16 种图书类型,结合 TF-IDF 方法计算图书文本之间的相似性,设定相似性阈值为 0.1,将相似性大于阈值的节点间建立连边,得到图书相似性关联图,如图 2 所示。在图书相似性关联图上,需找到具备多重兴趣特征的中介节点。利用式(5),设定中心性阈值 λ 为 0.27,在图中确定 111 个具有多重兴趣特征的中介项目,如图 3 所示。以中介节点 1 为例,假设用户选择过的图书为图中节点 2 ~ 12,中介节点 1 连接用户历史兴趣和其他兴趣类型中的节点,即图 4 中黑色实心点。根据连接节点所在图书类型确定用户下阶段读书兴趣可能改变方向为传记、科技、时尚、旅游与地图、烹饪美食与酒、教材教辅与参考书、艺术、文学、经济管理等。

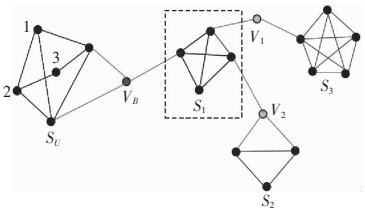

图 1　项目关系网络
Fig. 1　Item relationship network

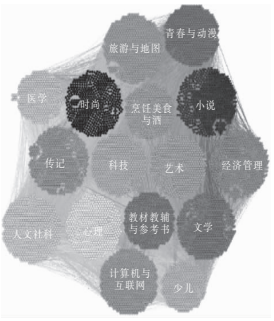

图 2　图书相似性关联图
Fig. 2　Similarity correlation graph of books

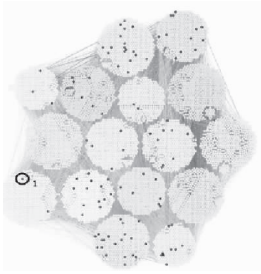

图 3　中介项目
Fig. 3　Intermediary items

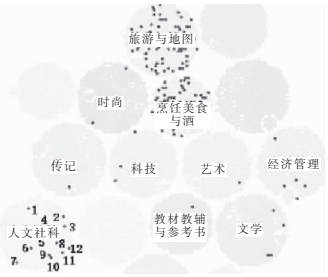

图 4　节点 1 的中介作用
Fig. 4　Intermediation of node 1

4.2　推荐系统评价

4.2.1　新颖性评价指标

广为接受的新颖性评价方式为基于流行度和基于距离的新颖性[2]。

a)基于平均流行度的新颖性分析。

$$\mathrm{novelty}_p=\frac{1}{UL}\sum_{u=1}^{U}\sum_{i\in K_u}\mathrm{popularity}(i) \tag{19}$$

其中:K_u 表示用户 u 的推荐列表;L 表示推荐列表长度;U 表示用户数量。

b)基于平均距离的新颖性分析。基于距离的新颖性评价直接考虑推荐的物品 $i\in K$ 与已评价物品 $j\in I_u$ 的平均距离。

$$\mathrm{novelty}_d=\frac{1}{U}\sum_{u=1}^{U}\frac{\sum_{i\in K}\sum_{j\in I_u}d(i,j)}{L\times|I_u|} \tag{20}$$

其中:d 表示距离度量函数,计算公式为 $d(i,j)=1-\mathrm{sim}(i,j)$,$\mathrm{sim}(i,j)$ 表示项目间的相似性,可利用协同过滤方法计算得到;$|I_u|$ 为用户已评价项目数。

c)基于冷启动的新颖性分析。冷启动方法在新颖性推荐场景中应用时,常将目标用户推荐列表中包含的冷启动项目个数 num 与列表中包含的项目总个数的比值作为新颖度[7],具体公式如下:

$$\mathrm{novelty}_{\mathrm{cold}}=\mathrm{num}/L \tag{21}$$

4.2.2　准确性评价指标

a)平均绝对误差指标(mean absolute error,MAE),度量预测评分与用户实际评分的平均绝对误差。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|v_{i,u}-r_{i,u}| \tag{22}$$

其中:n 是用户评分项目的个数;$v_{i,u}$ 表示预测用户 u 对项目 i 的评分;$r_{i,u}$ 表示用户 u 对项目 i 的实际评分。

b)均方根误差,即平均平方误差(root mean squared error,RMSE)。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{(i,u)}|v_{i,u}-r_{i,u}|^2} \tag{23}$$

4.3　结果分析

本文采用基于项目评分预测的协同过滤算法(item based collaborative filtering,IBCF)[22]、基于内容和项目的混合推荐算法(content-item collaborative filtering,CICF)[23]和随机游走推荐算法(random walk,RW)[24]作为实验对比。在实验中将冷启动项目选为评分数量小于等于 1 的项目,选取推荐列表长度 $L=20$ 的情况下,评价参数对比结果如表 1 所示。

表 1　RWFNF 算法与 IBCF、CICF、RW 算法对比
Tab. 1　Comparison of different algorithms

算法	MAE	RMSE	$\mathrm{novelty}_p$	$\mathrm{novelty}_d$	$\mathrm{novelty}_{\mathrm{cold}}$
IBCF	2.689 7	3.053 6	0.013 1	0.249 9	0.178 9
CICF	3.365 2	3.560 4	0.046 8	0.270 2	0.097 4
RW	2.645 0	3.024 7	0.015 5	0.283 3	0.144 9
RWFNF	**2.645 2**	**3.024 5**	**0.009 6**	**0.361 3**	**0.250 7**

从实验对比结果可以看出,融合新颖性特征的随机游走算法相比 IBCF 算法在预测准确度 MAE 和 RMSE 指标分别下降 1.65% 和 0.96% 的情况下,基于冷启动和距离的新颖性评价指标分别提高了 40.09% 和 44.57%,基于流行度的指标降低了 26.70%,说明提出算法能够有效缓解数据稀疏性问题,较大幅度地提高推荐结果的新颖性。同理,与 CICF、RW 算法相比,RWFNF 算法产生的推荐结果在准确性指标 MAE 和 RMSE 差别较小的情况下,在各项新颖性指标上均具备明显优势。

基于冷启动角度的新颖性分析如图 5 所示,随着推荐项目数量 L 的增加,RWFNF 算法在该指标上有较大幅度的提高,相比 IBCF、CICF 和 RW 算法能够有效覆盖冷启动项目。从图 6 中可以看出,随着 L 值的不断增大,四种算法的平均流行度变化相对稳定,同时 RWFNF 算法的平均流行度始终处于较低水平。以上结果表明 RWFNF 算法能够有效降低推荐列表的流行度,改善热门商品反复推荐导致推荐效果不显著的问题。随

着 L 值的提高，尽管 IBCF、CICF 和 RW 算法的平均距离逐渐提高，但 RWFNF 算法的平均距离始终保持在较高水平，如图 7 所示。可见 RWFNF 算法能够有效降低推荐列表中商品与用户历史选择的相似程度，用户了解推荐物品的可能性较低，推荐效果更能给用户带来意想不到的体验。

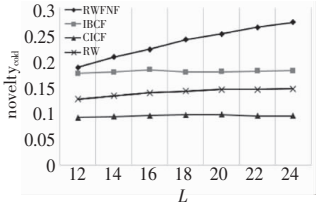

图 5 不同 L 值下 $\text{novelty}_{\text{cold}}$ 比较

Fig. 5 $\text{novelty}_{\text{cold}}$ comparison under different L values

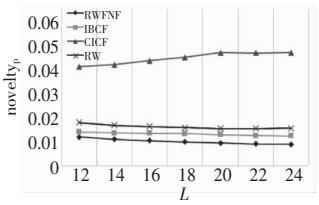

图 6 不同 L 值下 novelty_{p} 比较

Fig. 6 novelty_{p} comparison under different L values

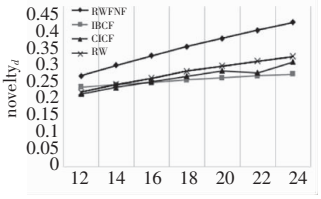

图 7 不同 L 值下 novelty_{d} 比较

Fig. 7 novelty_{d} comparison under different L values

综上分析，融合新颖性特征的随机游走算法在对平均绝对误差 MAE 和均方根误差 RMSE 影响较小的范围内，能够较大幅度地提高推荐结果的新颖性。因此，该方法可用于弥补准确性推荐的同质化问题，有利于扩展用户视野，改善用户的推荐列表，提升推荐效果。

5 结束语

本文利用亚马逊网站真实的用户和书籍数据对传统的基于网络结构的随机游走算法进行改进，提出一种融合项目流行度、多重兴趣特征、包含在新兴趣领域中三种特征的随机游走算法。通过从多重兴趣特征和新兴趣角度补充项的定义，强调用户的差异化，并将其用于构建新的转移概率，引导随机游走过程，满足用户对项的新颖性偏好。通过实验对比证明本文算法较大程度地提高了推荐列表的新颖度，同时保证了推荐的准确性。

参考文献：

[1] 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)：634-647. (Liu Mengjuan，Wang Wei，Li Yangxi，*et al*. AttentionRank$^+$：a graph-based recommendation combining attention relationship and multi-behaviors[J]. Chinese Journal of Computers，2017，40(3)：634-647).

[2] Han J，Yamana H. A survey on recommendation methods beyond accuracy[J]. IEICE Trans on Information & Systems，2017，E100D(12)：2931-2944.

[3] Herlocker J L，Konstan J A，Terveen L G，*et al*. Evaluating collaborative filtering recommender systems[J]. ACM Trans on Information Systems，2004，22(1)：5-53.

[4] 吕学强，王腾，李雪伟，等. 基于内容和兴趣漂移模型的电影推荐算法研究[J]. 计算机应用研究，2018，35(3)：717-720，802. (Lyu Xueqiang，Wang Teng，Li Xuewei，*et al*. Research on movie recommendation algorithm based on content and interest drift model[J]. Application Research of Computers，2018，35(3)：717-720，802.)

[5] Zhang Zhu，Zeng D D，Abbasi A，*et al*. A random walk model for item recommendation in social tagging systems[J]. ACM Trans on Management Information Systems，2013，4(2)：article No. 8.

[6] Feng Shanshan，Cao Jian，Wang Jie，*et al*. Recommendations based on comprehensively exploiting the latent factors hidden in items' ratings and content[J]. ACM Trans on Knowledge Discovery from Data，2017，11(3)：35.

[7] 于洪，李俊华. 一种解决新项目冷启动问题的推荐算法[J]. 软件学报，2015，26(6)：1395-1408. (Yu Hong，Li Junhua. Algorithm to solve the cold-start problem in new item recommendations[J]. Journal of Software，2015，26(6)：1395-1408.)

[8] Chou S Y，Yang Y H，Jang J S R，*et al*. Addressing cold start for next-song recommendation[C]//Proc of the 10th ACM Conference on Recommender Systems. New York：ACM Press，2016：115-118.

[9] Chen Lingjiao，Gao Jian. A trust-based recommendation method using network diffusion processes[J]. Physica A：Statistic Mechanic and Its Application，2018，506(9)：679-691.

[10] Adamopoulos P，Tuzhilin A. On unexpectedness in recommender systems：or how to better expect the unexpected[J]. ACM Trans on Intelligent Systems and Technology，2014，5(4)：article No. 54.

[11] Kapoor K，Kumar V，Terveen L，*et al*. "I like to explore sometimes"：adapting to dynamic user novelty preferences[C]//Proc of the 9th ACM Conference on Recommender Systems. New York：ACM Press，2015：19-26.

[12] Ma Wenping，Feng Xiang，Wang Shanfeng，*et al*. Personalized recommendation based on heat bidirectional transfer[J]. Physica A：Statistical Mechanics & Its Applications，2016，444(2)：713-721.

[13] 胡吉明，林鑫. 基于热传导能量扩散的社会化小众推荐融合算法设计[J]. 情报理论与实践，2016，39(4)：119-123. (Hu Jiming，Lin Xin. Design of fusion algorithm for socialized minority recommendation based on energy diffusion theory of heat spreading[J]. Information Studies：Theory & Application，2016，39(4)：119-123.)

[14] Wu Hao，Cui Xiaohui，He Jun，*et al*. On improving aggregate recommendation diversity and novelty in folksonomy-based social systems[J]. Personal & Ubiquitous Computing，2014，18(8)：1855-1869.

[15] Onuma K，Tong H，Faloutsos C. TANGENT：a novel，esurprise me' recommendation algorithm[C]//Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM Press，2009：657-666.

[16] Nakatsuji M，Fujiwara Y，Tanaka A，*et al*. Classical music for rock fans? Novel recommendations for expanding user interests[C]//Proc of the 19th ACM International Conference on Information and Knowledge Management. New York：ACM Press，2010：949-958.

[17] Zhang Yuancao，Séaghdha D Ó，Quercia D，*et al*. Auralist：introducing serendipity into music recommendation[C]//Proc of the 5th ACM International Conference on Web Search and Data Mining. New York：ACM Press，2012：13-22.

[18] Oku K，Hattori F. Fusion-based recommender system for serendipity-oriented recommendations[J]. Journal of Japan Society for Fuzzy Theory & Intelligent Informatics，2013，25(1)：524-539.

[19] 夏立新，郑路，翟姗姗，等. 基于结构洞理论的虚拟社区边缘用户信息资源推荐模型构建研究[J]. 情报理论与实践，2017，40(2)：1-6. (Xia Lixin，Zheng Lu，Zhai Shanshan，*et al*. Research on construction of information resource recommendation model of periphery user in virtual community based on structural holes[J]. Information Studies：Theory & Application，2017，42(2)：1-6.)

[20] Newman M E J. A measure of betweenness centrality based on random walks[J]. Social Networks，2003，27(1)：39-54.

[21] 王茜，喻继军. 基于消费性格的商品多样性推荐研究[J]. 信息系统学报，2017(1)：23-37. (Wang Qian，Yu Jijun. Study on recommendation of commodity diversity based on consumer's character[J]. China Journal of Information Systems，2017(1)：23-37.)

[22] 曾艳，麦永浩. 基于内容预测和项目评分的协同过滤推荐[J]. 计算机应用，2004，24(1)：111-113. (Zeng Yan，Mai Yonghao. Collaborative filtering recommendation based on content and item rating prediction[J]. Journal of Computer Applications，2004，24(1)：111-113.)

[23] 于波，陈庚午，王爱玲，等. 一种结合项目属性的混合推荐算法[J]. 计算机系统应用，2017，26(1)：147-151. (Yu Bo，Chen Gengwu，Wang Ailing，*et al*. Hybrid recommendation algorithm combined with the project properties[J]. Computer Systems & Applications，2017，26(1)：147-151.)

[24] Yildirim H，Krishnamoorthy M S. A random walk method for alleviating the sparsity problem in collaborative filtering[C]//Proc of ACM Conference on Recommender Systems. New York：ACM Press，2008：131-138.