文章编号：2095-2163(2022)03-0200-04 中图分类号：TP391 文献标志码：A

基于隐式数据和 Apriori 的协同过滤推荐算法

王君威，余 粟
（上海工程技术大学 机械与汽车工程学院，上海 201620）

摘 要：针对传统协同过滤推荐算法对目标客户进行个性化推荐时，因用户评价数据和物品属性等显式数据稀疏，造成推荐商品的准确率和质量相对较差的问题，本文基于隐式数据和 Apriori 算法对协同过滤推荐算法做出改进。首先，算法基于隐式数据中用户对商品的行为和用户对商品的评价，建立用户对商品的评分偏好模型，用以构建原始评分数据；其次，利用 Apriori 算法找出用户行为数据集中商品的强关联规则，利用输出的关联规则对原始评分数据进行降维，并进行相似度计算，确定用户之间的相似性，根据计算结果来确定目标用户的近邻集合；最后，算法通过度量后的最近邻居来计算目标用户对特定商品的预测评分。从数据集中分别采取 70 000 条数据和 30 000 条数据进行算法测试，测试结果表明改进后的推荐算法与基于用户的协同过滤算法相比准确率和召回率分别提高了 1.56% 和 0.23%；和基于项目的推荐算法相比准确率和召回率分别提高了 4.39% 和 0.92%，证明基于隐式数据和 Apriori 算法改进的协同过滤算法，在缓解数据稀疏的同时，能提高推荐的准确率。

关键词：Apriori 算法；关联规则；协同过滤算法

Collaborative filtering recommendation algorithm based on implicit data and Apriori algorithm

WANG Junwei, YU Su
(School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 In order to improve the traditional collaborative filtering recommendation algorithm in the personalized recommendation of target customers, the accuracy and quality of recommended products are relatively poor due to sparse explicit data such as user evaluation data and item attributes. This paper improves the collaborative filtering recommendation algorithm based on implicit data and Apriori algorithm. First, the algorithm builds the user's scoring preference model for the product based on the user's behavior on the product and the user's evaluation of the product in the implicit data to construct the original score data. Second, the algorithm uses Apriori to find the strong association rules of the products in the user behavior data set. The output association rules are used to reduce the dimensionality of the original score data, and the similarity calculation is performed to determine the similarity between users, and the neighbor set of the target user is determined according to the calculation result. Finally, the algorithm uses the measured nearest neighbors to calculate the target user's prediction score for a specific product. The experiment takes 70,000 pieces of data and 30,000 pieces of data from the data set for algorithm testing. The test results show that the improved recommendation algorithm has increased accuracy and recall rate by 1.56% and 0.23% compared with the user-based collaborative filtering algorithm, and Compared with the item-based recommendation algorithm, the accuracy and recall rate are increased by 4.39% and 0.92%, respectively. Experiments prove that the collaborative filtering algorithm based on implicit data and Apriori algorithm improves the accuracy of recommendation while alleviating data sparseness.

【Key words】 Apriori algorithm; association rules; collaborative filtering algorithm

0 引 言

随着互联网的兴起以及淘宝、京东等电子商务网站的发展，商品数据以爆发式的速度积累，信息量也随之增长。当顾客面对如此庞大体量的商品数据信息时，往往会出现“信息迷失”，个性化推荐也就应运而生[1]。个性化的推荐系统能够根据用户和商品的特有属性和用户对商品的行为信息，向目标客户提供能够符合其兴趣的商品和信息，帮助目标客户完成购物。在数据科学中，能在个性化的推荐系统中使用的主流技术有关联规则，聚类算法，协同过滤算法等等，协同过滤算法是目前在推荐算法领域应用最多的一类算法。

传统的协同过滤算法不需要用户提出明确的需求，而是基于显式数据即用户对商品的直接评价，给目标客户进行商品的个性化推荐。传统的协同过滤算法主要包含近邻模型和隐向量模型两种实现方式[2]，近邻模型主要是寻找用户或项目间的关系，

作者简介：王君威(1997-)，男，硕士研究生，主要研究方向：大数据、推荐算法；余 粟(1962-)，女，博士，教授，主要研究方向：大数据、机电控制、计算机视觉等。

收稿日期：2021-08-22

借此构建近邻集合，由近邻集合的评分产生预测，近邻模型的算法实现可以分为基于用户和基于项目的协同过滤两种。基于用户的协同过滤算法是通过用户对商品的评分来估测用户的兴趣大小，通过对所有用户兴趣的估测作为相似度的衡量标准，根据计算出的相似度，来获取目标用户的最近邻居，再依靠最近邻居的兴趣估测，产生目标用户对其他商品的预测评分，最后按照预测评分从商品集合中产生商品推荐。基于项目的协同过滤算法和基于用户的协同过滤算法相似，都是根据相似度计算产生目标客户商品推荐。基于项目的过滤算法的不同之处是根据喜欢商品的用户量来计算商品的相似度[3]。

针对用户的评价数据规模大、数据稀疏、推荐精度和准确度低，以及冷启动等问题，专家们在经典协同过滤算法的基础上进行了改进和修正。李红梅等[4]采用多探寻机制改进的 P 稳态分布的局部敏感哈希，对用户评分数据进行降维与索引，来实现近似近邻搜索，快速获取目标用户的近邻用户集合，在一定程度上提高了准确率；任永功等[5]提出了混合填充算法来缓解数据稀疏，通过相似物品的评分来填补稀疏数据，基于填充后的数据来确定目标的近邻用户，虽然提高了用户推荐的精度，但在数据高度稀疏的情况下，类似商品的评分填充较为单一。本文在传统协同过滤和关联规则等推荐算法的基础上，对隐式数据进行调整，通过改进后用户评分公式来计算用户兴趣，通过 Apriori 算法获取的强关联规则来降低原始评分数据索引，构建目标用户的近邻集合，可以有效提高推荐算法的准确度。

1　相关工作

1.1　协同过滤推荐算法

通过个性化的协同过滤算法对目标用户提供合适的商品推荐，帮助目标用户完成购物过程，协同过滤推荐算法的实现步骤分为 3 步：获取目标用户信息，寻找目标的最近邻居，获取前 N 个($top-N$) 的商品推荐；获取用户信息并不是指要获取用户的自有属性，而是要获取用户对商品的评价，这种评价一般以数字的形式呈现用户对商品的兴趣大小，以此计算用户之间的相似性；通过相似性计算寻找目标客户的最近邻居，就是寻找与目标客户在兴趣评分上最相近的用户；最后产生 $top-N$ 的推荐是根据最近邻居对商品的评分信息，产生目标用户的预测评分，从中选出前 N 个评分高的预测商品，完成针对目标用户的个性化推荐。

假设系统中的用户集合 $U=\{U_1,U_2,U_3,\cdots,U_m\}$，商品集合 $I=\{I_1,I_2,I_3,\cdots,I_n\}$，用户评分矩阵 R 表示用户集合中的每个用户对商品集合中感兴趣的商品的兴趣大小，见表 1。n 表示商品的数目，m 表示用户的人数，$R_{m,n}$ 表示第 m 个用户对第 n 个商品的兴趣值。

表 1　用户商品偏好模型

Tab. 1　User product preference model

	I_1	I_2	I_3	…	I_n
U_1	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	…	$R_{1,n}$
U_2	$R_{2,1}$	$R_{2,2}$	$R_{2,3}$	…	$R_{2,n}$
⋮	…	…	…	…	…
U_m	$R_{m,1}$	$R_{m,2}$	$R_{m,3}$	…	$R_{m,n}$

用来衡量用户是否相似的计算公式很多，常用余弦相似度、修正余弦相似度以及皮尔逊相关系数(Pearson 相关系数)等。余弦相似度常用来计算向量间的余弦夹角，通过角度的大小来衡量两个向量的相似性；修正余弦相似度在余弦相似度的基础上调整了不同用户对商品的评分尺度的差异；Pearson 相关系数用来表示两个变量之间的线性相关的程度，且该系数在度量精度上比修正余弦相似度更有优势[6]。所以本文使用 Pearson 相关系数来测量用户间的相似性，计算公式(1)：

$$Sim(u,v)=\frac{\sum_{I\in i_{uv}}(R_{u,i}-\overline{R_u})(R_{v,i}-\overline{R_v})}{\sqrt{\sum_{I\in i_{uv}}(R_{u,i}-\overline{R_u})^2\sum_{I\in i_{uv}}(R_{v,i}-\overline{R_v})^2}} \tag{1}$$

其中，$Sim(u,v)$ 表示用户 u 和用户 v 的相似度；I_{uv} 表示用户 u 和用户 v 共同的评分项目集合；$R_{u,i}$ 表示用户 u 对商品 i 的评分；$\overline{R_u}$ 表示用户 u 对所有商品的评分均值；$R_{v,i}$ 表示用户 v 对商品 i 的评分；$\overline{R_v}$ 表示用户 v 对所有商品的评分均值。

为了评估用户对特定商品的兴趣，在找到目标用户的最近邻居后，通过获取最近邻居对该特定商品的兴趣评分，来预测目标用户对商品的兴趣程度，商品兴趣预测公式(2)：

$$P_{u,i}=\overline{R_u}+\frac{\sum_{v\in Q}sim(u,v)(R_{v,i}-\overline{R_v})}{\sum_{v\in Q}sim(u,v)} \tag{2}$$

其中，$P_{u,i}$ 表示目标用户 u 对商品的兴趣预测；$\overline{R_u}$ 则表示用户 u 的兴趣评分均值；Q 表示目标用户的近邻用户集合；$R_{v,i}$ 表示用户 v 对物品 i 的兴趣评分；$\overline{R_v}$ 则表示用户 v 的兴趣评分均值；$Sim(u,v)$ 则

表示用户 u 和用户 v 的相似度。

1.2 Apriori 算法

Apriori 算法作为关联规则的经典算法,经常会用来寻找物与物之间的隐含关系,输出物品的频繁项集或强关联规则。频繁项集一般指经常一起出现的物品集合,常用支持度来评估,是支持度大于最小支持度的项集;强关联规则一般指物品间有着比较强的关系,常用置信度来评估,是置信度大于最小置信度的项集[7]。

支持度用来表示物品出现在数据集中的次数或概率,式(3)。

$$support(A \Rightarrow B) = P(A \cup B) \tag{3}$$

置信度则表示某物品已经出现的条件下,另一个物品出现的概率,式(4)。

$$confidence(A \Rightarrow B) = support(A \cup B)/support(A) \tag{4}$$

Apriori 算法的实现即寻找频繁项集,在频繁项集中找出强关联规则。寻找频繁项集就是遍历数据集,计算遍历到的每一项数据的频数,以最小支持度为判断标准,从数据集中找出频繁 1 项集;在频繁 1 项集的基础上继续以最小支持度为衡量标准,产生频繁 2 项集;不断反复扫描数据集,直到 $N+1$ 项集不在满足最小支持度,取第 N 项集为结果。从频繁项集中找到强关联规则。遍历第一步获取的频繁项集,根据支持度来计算每个频繁项集的置信度,在所有满足最小置信度标准的频繁项集中挖掘出强关联规则,进而构建规则库。

2 改进的协同过滤推荐算法

2.1 算法的改进

在电子商务网站中,用户和商品以及用户对商品的行为,每日都会产生大量的数据。用户会对商品产生浏览、点击、关注、加入购物车、下单和删除购物车等 6 种行为,除此之外,用户还会在商品到货后对商品的质量给出评论。

假设 1 用户对商品的不同行为可以衡量用户对商品的兴趣大小;

假设 2 其他用户对商品的评论会对目标用户对商品的兴趣产生影响;

根据用户的行为不同来确定用户对特定商品的兴趣度,将用户对商品的这 6 种行为分别定为{1,2,3,4,5,0}。根据假设,通过用户对商品不同的行为来评估用户对商品偏好程度,加入其他用户对商品评论造成的影响,由此来构建用户评分偏好模型,式(5):

$$Is = Ua + (1 - b) * Uc \tag{5}$$

其中,Is 表示用户对商品的评分;Ua 表示用户对商品的兴趣度;Uc 是一个常数,用来表示商品受欢迎程度;b 表示目前其他用户对该商品产生差评的概率。

改进算法通过使用 Apriori 算法获取数据集中商品的强关联规则,对相似度计算进行改进,将用户间在强关联规则中的共同商品作为集合,改进后的公式为式(6):

$$S(u,v) = \frac{\sum_{j \in A_{uv}} (R_{u,j} - \overline{R_u})(R_{v,j} - \overline{R_v})}{\sqrt{\sum_{j \in A_{uv}} (R_{u,j} - \overline{R_u})^2 \sum_{j \in A_{uv}} (R_{v,j} - \overline{R_v})^2}} \tag{6}$$

其中,A_{uv} 表示用户 u 和用户 v 在强关联规则中的共同商品集合。

2.2 算法实现

使用改进算法针对用户进行商品个性化推荐,主要步骤:根据改进的用户评分偏好模型对商品进行兴趣预测,使用 Apriori 算法获取商品集合中的强关联规则,根据兴趣预测和强关联规则,使用改进后的相似度计算方法计算用户间的相似度,获取目标用户的最近邻居,从而得到目标用户的 $top-N$ 的商品推荐。改进的协同过滤算法见表 2。

表 2 改进的协同过滤算法

Tab. 2 Improved collaborative filtering algorithm

算法:改进的协同过滤算法
输入: 用户和商品的数据集
输出: 用户 u 对商品 i 的预测评分
(1)FOR(each, $u \in$ 用户集合) FOR(each, $i \in$ 商品集合) DO 计算 兴趣评分 By formula(5) #遍历数据集,计算用户对商品的兴趣评分
(2)FOR(each, $l \in$ 数据集合) DO Apriori 计算 置信度,支持度 By formula(3),(4) #通过 Apriori 算法,获取强关联规则
(3)Replace 用户的共同评分集合 with 获取的强关联规则 FOR(each, $u \in$ 用户集合, $v \in$ 用户集合) DO 计算 用户间的相似度 By formula(6)#使用强关联规则改进相似度计算 $H = \{G_1, G_2, G_3, G_4 \cdots G_k\}$ #获取最近邻居 K
(4)计算 $P_{u,i}$ By formular(2)#预测用户对商品的兴趣评分
(5)获取 $top-N$ 的商品推荐

3 实验设计与结果分析

3.1 实验数据集和算法评价指标

3.1.1 实验数据集

本文的实验数据集来自京东网站 2016 年 2 月的

用户、商品和相应的行为数据脱敏后形成的数据集，该数据集一共包含了 105 321 用户，共计 11 485 424 条用户行为数据和 558 552 条用户对商品的评论信息。对该数据集进行清洗和预处理后，从中随机抽取了十万条用户行为数据，近五百位用户，将数据集按 7∶3 随机分为训练集和测试集。

3.1.2　算法评价指标

推荐算法的衡量标准有很多，本次实验将准确率(*Accuracy*) 和召回率(*Recall*) 作为算法的衡量标准式(7) 和式(8)。准确率用来衡量用户对算法推荐商品的满意程度，召回率是算法推荐满意商品与数据集中所有目标用户满意商品的比值。

$$P(Acc) = \frac{A}{A + B} \tag{7}$$

$$P(Rec) = \frac{A}{A + C} \tag{8}$$

其中，A 表示向用户推荐的商品正是用户满意的数量；B 表示向用户推荐的商品不是用户满意的数量；C 表示没有向用户推荐的商品，但该商品是用户满意商品的数量。

3.2　结果分析

为了验证基于隐式数据和 Apriori 改进的协同过滤算法的性能，本次算法验证使用基于用户和项目的协同过滤算法，在采用的数据集上进行不同近邻数目下算法准确率和召回率的对比。

在实验中，算法分别采取最近邻用户集合数为：50，45，40，35，30，25，20，15，10，实验十次获取准确率和召回率的平均值，3 种算法的准确率和召回率的影响如图 1 和图 2 所示；U_CF 表示基于用户的协同过滤算法；I_CF 表示基于项目的协同过滤算法；USER 表示改进后的算法，可以明显看到 3 种算法的准确率和召回率随着最近邻数目的扩大，都有着明显的变化。USER 算法在开始时召回率和准确率随着最近邻居数的增大呈现上升的趋势，随着近邻数目的变化，召回率和准确率表现出阶段性下降和回升的趋势；在最近邻居数较小时，USER 算法的召回率较小，但随着最近邻居数量的逐渐增多，召回率开始回升。

改进算法与传统算法的结果对比见表 3，可以看出改进后的算法与基于用户的协同过滤算法相比，准确率和召回率分别提升 1.56% 和 0.23%；与基于项目的协同过滤算法相比，准确率和召回率分别提升 4.39% 和 0.92%。

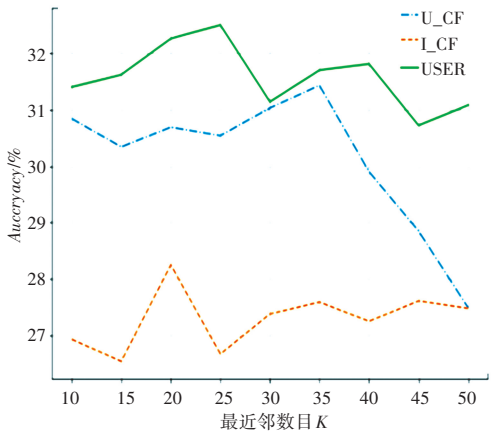

图 1　不同近邻数目对准确率的影响

Fig. 1　Impact on accuracy

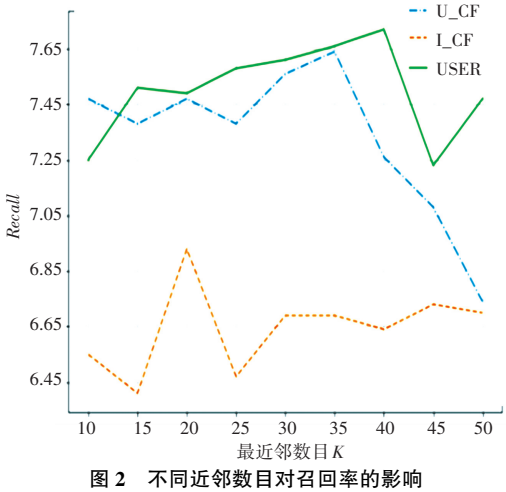

图 2　不同近邻数目对召回率的影响

Fig. 2　Impact on recall rate

表 3　改进算法与传统算法的结果对比

Tab. 3　Comparison of the results of the improved algorithm and the traditional algorithm

	Accuracy/%	Recall/%
U_CF	30.12	7.28
I_CF	27.29	6.59
User	31.68	7.51

4　结束语

本文基于用户隐式数据和 Apriori 算法对协同过滤推荐算法做出了改进。针对用户推荐商品过程中，通过关注用户对商品的隐式行为信息和其他用户的评价来评估用户对商品的兴趣程度。对数据进行清洗和预处理以后，通过使用 Apriori 算法来获取商品中的强关联规则，从而改进了用户间的相似度计算方式，对推荐结果进行优化。本文使用用户对商品的行为属性来判断用户的兴趣，今后将尝试加入用户的显式行为数据来判断用户对商品的兴趣，进一步改进推荐算法。

(下转第 207 页)