文章编号:1006-2467(2018)07-0770-07 DOI: 10.16183/j.cnki.jsjtu.2018.07.003

基于偏好度特征构造的个性化推荐算法

黄金超[1], 张佳伟[2], 陈 宁[2], 陈毅鸿[2], 江 文[2], 李生红[1,3]
(1. 上海交通大学 网络空间安全学院, 上海 200240; 2. 携程旅游网络技术有限公司 平台商务部, 上海 200335; 3. 上海市信息安全综合管理技术研究重点实验室, 上海 220240)

摘 要: 随着在线旅游业酒店数量的日益增多,用户点评信息稀疏问题愈加严重,这不仅导致推荐准确度大幅下降,而且使传统推荐算法的计算负荷随之增加,难以满足实时性要求. 基于此,从挖掘用户历史信息与待推荐物品之间潜在相关性的角度出发,对基于内容的推荐算法进行改进,提出了一种基于偏好度特征构造的个性化推荐算法. 该算法通过计算偏好分来构造偏好度特征,并借助机器学习领域的分类算法得以实现. 将该算法应用于线上旅游业的个性化子房型推荐,通过对真实数据集的实验与分析,验证了所提出个性化推荐算法的简便与有效性,且较传统推荐算法更具实时性和通用性.

关键词: 基于内容的推荐; 潜在相关性; 偏好度构造; 子房型推荐
中图分类号: TP 391 **文献标志码**: A

Preference Degree Based Personalized Recommendation Algorithm

HUANG Jinchao[1], *ZHANG Jiawei*[2], *CHEN Ning*[2], *CHEN Yihong*[2]
JIANG Wen[2], *LI Shenghong*[1,3]
(1. School of Cyber Security, Shanghai Jiao Tong University, Shanghai 200240, China; 2. Platform Business Department, Ctrip Travel Network Technology Co., Ltd., Shanghai 200335, China; 3. Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai 220240, China)

Abstract: Faced with increasing number of hotels in online tourism, the problem of sparse data is becoming more and more serious. On one hand, it leads to a significant decrease in recommendation accuracy; on the other hand, the computational load of traditional recommendation algorithm is increased, which is difficult to meet the real-time requirement. So, this paper firstly proposed a preference degree based personalized recommendation algorithm which mined the potential correlation between user historical data and recommend items. The novel algorithm utilized users' historical data to calculate preference degree and then construct new features, and its realization is based on classification algorithm. Besides, the new method is applied to make personalized recommendations in online tourism. Results from real data sets showed that the proposed preference degree based personalized recommendation algorithm is effective and universal.

Key words: content-based recommendation; potential correlation; construction of preference degree; room recommendation

收稿日期:2017-05-02
基金项目:国家重点研发计划资助项目(2016YFB0801003)
作者简介:黄金超(1992-),女,河北省保定市人,博士生,主要从事推荐系统研究. E-mail:hjc2015@sjtu.edu.cn.
通信作者:李生红,男,教授,博士生导师,E-mail:shli@sjtu.edu.cn.

互联网迅速发展带来的信息过载问题变得日益严重，使得用户难以在众多信息中挑选出自己的真正所需. 在此背景下，基于用户兴趣爱好的个性化推荐技术应运而生，个性化推荐系统也逐渐成为具有挑战性的研究热点之一[1-2]. 随着国内旅游业的兴起，旅游企业也将目光转向迅猛发展的电子商务平台[3-4]. 在线旅游的出现将线下的商务机会与互联网结合，依托于互联网来满足旅游消费者信息咨询、产品预定及服务评价等需求. 但是，随着平台用户及酒店数量的日益增加，传统推荐算法的计算负荷越来越重，难以满足实时性的要求[5-6]. 另外，该领域中大部分用户预订酒店的数量较少（即大多数属于低频用户），且对酒店的点评信息也较少，导致数据稀疏性问题更为严峻[7-8]，推荐质量迅速下降.

在上述应用场景中，协同过滤算法在推荐精度和计算开销两方面相比基于内容的推荐算法均有着一定不足，而且考虑到传统推荐算法大多没有综合考虑物品、用户、用户对物品的喜好和上下文等特征，本文对传统基于内容的推荐算法进行改进，并提出一套简单、有效的推荐算法. 首先，将原始数据进行结构化处理；然后从挖掘用户历史信息与物品对应维度间相关性的角度出发计算偏好分，构造偏好度特征，对维度进行有效扩充. 最后，利用机器学习领域的分类方法对待推荐物品进行排序，满足实时性要求. 此外，手机端的酒店预定业务是在线旅游各类服务中一个重要部分，而当前该业务酒店页面的内容展示没有任何侧重. 为此，对手机端酒店预定业务中酒店信息的展示部分也进行了改进，将用户最可能预定的子房型置顶，进行个性化子房型推荐，并将新算法应用于该场景. 实验结果显示，新的推荐算法能在有效处理数据稀疏性问题的同时满足实时性要求，使得推荐结果的命中率有了 5％以上的提升.

1　相关工作

1.1　基于内容的推荐

基于内容的推荐是通过学习得到用户的喜好以及物品的特性，侧重于计算待推荐物品与用户的特征匹配度. 最初提出的基于关键词的向量空间算法[9]在新闻推荐领域[10]的应用较为广泛，其利用 TF-IDF 方法构造特征向量. 随后语义分析及利用外源知识来进行语义分析方法被提出，并在图书推荐领域有着广泛应用[11-12]. 针对图书特征信息稀疏问题，文献[13]中提出一种结合社会化标签的基于内容的推荐算法，利用图书的社会化标签来补充其特征项. 为了提高传统推荐算法的准确性，文献[14-15]中将 k-Means 方法与传统的基于内容的推荐算法相结合，通过对物品的聚类处理来提升推荐的准确性.

基于内容的推荐算法大多是一些简单的模型，需要对每个用户计算相似产品进行推荐. 传统算法并没有综合考虑物品、用户、用户对物品的偏好和上下文等特征，因此导致其总体的准确率不高，存在着较大的提升空间.

1.2　其他推荐算法

其他的推荐算法中主要对协同过滤算法进行介绍. 数据的稀疏性和系统的延伸性是协同过滤推荐算法一直以来面临的两大难题，对此，研究人员陆续提出一系列的改进方法. 例如：文献[16-17]中提出利用基于领域最近邻和不确定近邻方法将未评分物品进行评分预测；文献[18]中通过对隐式用户反馈数据流进行收集与处理，提出一种基于隐式用户反馈的推荐系统；文献[19]中则把聚类方法应用于协同过滤推荐系统中，设计了一种基于在线学习的个性化推荐算法；文献[20]中将大数据集和推荐计算分解到多台计算机上并行处理，进而实现对系统延伸性的提升. 近些年，人们逐渐开始将神经网络与传统的协同过滤算法进行结合，如：文献[21]中提出一个 2 层的无向图模型来处理大数据量的推荐问题；文献[22]中从利用隐式反馈的角度出发，借助深层神经网络及多层感知器来模拟协同过滤推荐系统中用户与物品间的交互模型.

以上算法虽然对传统算法固有难题进行了一定程度的改进，但在针对在线旅游的实际应用中，其仍难以同时满足对稀疏数据的处理和实时性的要求. 特别是其中涉及到神经网络的方法均建立于用户与物品间充足的交互信息，在点评信息稀疏的在线旅游领域并不适用. 从一定程度上来讲，在本文特定应用场景中，协同过滤算法相比于基于内容的推荐算法有着明显的不足.

2　基于偏好度特征构造的推荐算法

传统基于内容的推荐系统没有综合考虑物品、用户、用户对物品的偏好和上下文等特征，因此导致其总体的推荐准确率不高. 本文将从挖掘用户历史信息与待推荐物品之间潜在相关性的角度出发对基于内容的推荐算法进行改进，并提出一种基于偏好度特征构造的个性化推荐算法.

假设推荐系统中，有一系列用户 $\mathrm{User}=\{\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_M\}$ 和待推荐物品 $\boldsymbol{I}=[\boldsymbol{I}_1\ \ \boldsymbol{I}_2\ \ \cdots\ \ \boldsymbol{I}_N]^{\mathrm{T}}$. 对于用户 $\boldsymbol{U}_m(m=1,2,\cdots,M)$，其历史信息表示为

$\boldsymbol{H}^m=[\mathrm{hist}_{ij}^m]_{S\times K}$，其中一行代表该用户一条历史信息，$K$ 则为每条信息包含的维度数量，可通过学习 $\boldsymbol{H}^m$，得到用户自身特征向量

$$\boldsymbol{U}_m=[u_{m1}\quad u_{m2}\quad\cdots\quad u_{mz}]$$

共 z 个属性. 对于物品 $\boldsymbol{I}_n$，其自身特征用向量

$$\boldsymbol{I}_n=[i_{n1}\quad i_{n2}\quad\cdots\quad i_{nK}]$$

表示. 用户每条历史信息

$$\boldsymbol{h}_{i*}^m=[\mathrm{hist}_{i1}^m\quad\mathrm{hist}_{i2}^m\quad\cdots\quad\mathrm{hist}_{iK}^m]$$

与物品向量 $\boldsymbol{I}_n$ 有着相同的特征数量 K，且一一对应，即 hist_{ij}^m 与 i_{nj} 表示同一个特征.

每个用户对应的 $\boldsymbol{H}^m$ 不同，因而 $\boldsymbol{U}_m$ 也不同，传统基于内容的个性化推荐，就是利用 $[\boldsymbol{U}_m:\boldsymbol{I}_n]$ 来对待推荐物品进行打分并排序，如图 1 中矩阵虚线左边内容所示(图中表示某一用户，故用户维度每行均相同). 传统推荐算法将用户与物品视为 2 个独立的个体，实际上，用户历史信息 $\boldsymbol{H}^m$ 与物品特征 $\boldsymbol{I}_n$ 的对应维度之间存在着极大的相关性. 因此，本文提出一种基于偏好度特征构造的推荐算法，通过计算 $\boldsymbol{H}^m$ 与 $\boldsymbol{I}_n$ 的偏好分向量 $\boldsymbol{S}_n^m$，挖掘用户与待推荐物品之间潜在的相关特性，并加入到原有的特征矩阵中(见图 1). 最后，将 $[\boldsymbol{U}_m:\boldsymbol{I}_n:\boldsymbol{S}_n^m]$ 作为特征矩阵，借助机器学习领域的分类器实现对物品的打分并排序，排名越靠前表示用户越喜欢改物品. 具体过程如图 2 所示.

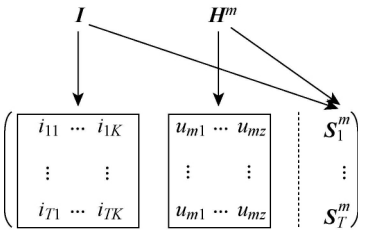

图 1 基于内容推荐的特征矩阵

Fig. 1 The feature matrix of content-based recommendation system

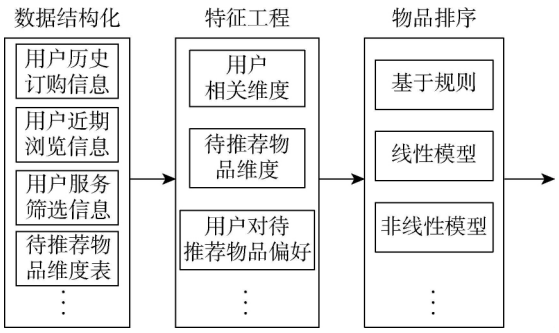

图 2 基于偏好度特征构造推荐算法的整体框架

Fig. 2 The frame of proposed content-based recommendation system

图中，包含数据结构化、特征工程和物品排序等部分.

2.1 数据结构化

内容分析能够从无结构的信息中提取特征并结构化表示，是基于内容推荐的一个重要部分. 本文将原本的酒店、房型及各类相关信息转换成上述 $\boldsymbol{H}$ 和 $\boldsymbol{I}$ 的形式. 用户历史信息矩阵为

$$\boldsymbol{H}^m=\begin{bmatrix}\boldsymbol{c}_1\\ \vdots\\ \boldsymbol{c}_v\end{bmatrix}=\begin{bmatrix}\mathrm{hist}_{11}^m & \cdots & \mathrm{hist}_{1K}^m\\ \vdots & & \vdots\\ \mathrm{hist}_{S1}^m & \cdots & \mathrm{hist}_{SK}^m\end{bmatrix}$$

由用户历史订购信息、用户近期所浏览信息以及用户最近的筛选服务信息等构成. 其中，每类信息 $\boldsymbol{c}_v=[\mathrm{hist}_{ij}^m]_{*\times K}$，表示信息条数不固定(由实际情况而定)，每条信息有 K 个特征(具体包含酒店/子房型的静态信息及线上展示的实时信息等).

类似地，对于某一待推荐物品 $\boldsymbol{I}_n$，每个物品均可用一个 K 维的向量表示，即

$$\boldsymbol{I}=\begin{bmatrix}\boldsymbol{I}_1\\ \vdots\\ \boldsymbol{I}_T\end{bmatrix}=\begin{bmatrix}i_{11} & \cdots & i_{1K}\\ \vdots & & \vdots\\ i_{T1} & \cdots & i_{TK}\end{bmatrix}$$

针对子房型问题，所有待推荐子房型信息均可从用户下单前最后一次浏览的记录信息及用户的历史订单中得到. 此外，将用户最终预定的子房型作为正样本(label=1)，其余子房型为负样本(label=0).

2.2 特征工程

基于内容的推荐是通过一系列的学习得到用户的喜好，即利用每个用户所对应的 $\boldsymbol{H}^m$，提取出用户的兴趣特征向量 $\boldsymbol{U}_m$，再根据用户和物品的特征向量 $[\boldsymbol{U}_m:\boldsymbol{I}_t]$ 来预测用户 m 对物品 t 的喜好程度. 整个过程中，用户的历史信息 $\boldsymbol{H}^m$ 和物品 $\boldsymbol{I}_t$ 被视为 2 个独立的个体，分别提取各自特征，没有考虑两者间潜在的相关性. 实际上，$\boldsymbol{H}^m$ 与 $\boldsymbol{I}_t$ 对应维度之间的相似程度会表达更多有效信息. 因此，本文从挖掘用户历史信息 $\boldsymbol{H}^m$ 与物品 $\boldsymbol{I}_t$ 对应维度之间相关性的角度出发，构造了一系列偏好度特征加入到原有特征中.

用户每条历史信息 $\boldsymbol{h}_{i*}^m$ 与 $\boldsymbol{I}_t$ 对应一个偏好分向量 $\boldsymbol{s}_{it}$(K 维)，那么 $\boldsymbol{H}^m$ 与 $\boldsymbol{I}_t$ 之间则对应一个偏好分矩阵 $\boldsymbol{S}_t^m$；本文利用最简单的均值融合方法处理 $\boldsymbol{S}_t^m$，得到用户 m 对物品 t 的偏好分向量 $\boldsymbol{S}_t^m$. 具体偏好度特征构造如算法 1 所示.

算法 1 用户对待推荐物品偏好度特征构造算法.

输入：用户 m 历史信息矩阵 $\boldsymbol{H}^m$，大小为 $S\times K$；待推荐物品矩阵 $\boldsymbol{I}$，大小为 $T\times K$；E 为 K 维特征中含“序”关系的集合，所有用户历史信息矩阵各维度

的均值

$$\mathbf{A}=[\mathrm{avg}_1 \quad \mathrm{avg}_2 \quad \cdots \quad \mathrm{avg}_K]$$

$$\mathrm{avg}_k=\sum_{m=1}^{M}\sum_{s=1}^{S}\mathrm{hist}_{sk}^{m}$$

输出：用户 m 与待推荐物品 $\boldsymbol{I}$ 的偏好分矩阵（$\boldsymbol{S}_t^m$），大小为 $T\times K$.

(1) For $t<T$

(2)　For $s<S$

(3)　第 s 条历史信息 $\boldsymbol{h}_{s*}^m$ 与物品向量 $\boldsymbol{I}_t$ 对应维度 k 的偏好分计算

$$\mathrm{sim}_{sk}=\begin{cases}I(\mathrm{hist}_{sk}^m,i_{tk}), & k\notin E\\ \exp\left(-\dfrac{|\ \mathrm{hist}_{sk}^m-i_{tk}\ |}{\mathrm{avg}_k}\right), & k\in E\end{cases} \tag{1}$$

(4) 得到偏好度向量 $[\mathrm{sim}_{s1} \quad \mathrm{sim}_{s2} \quad \cdots \quad \mathrm{sim}_{sK}]$

(5) 计算

$$\mathrm{sim\_final}_s=k_1\ \mathrm{sim}_{s1}+k_2\ \mathrm{sim}_{s2}+\cdots+k_K\ \mathrm{sim}_{sK} \tag{2}$$

得到 $\boldsymbol{h}_{s*}$ 与 $\boldsymbol{I}_t$ 的偏好度向量

$$\boldsymbol{s}_s=[\mathrm{sim}_{s1} \quad \mathrm{sim}_{s2} \quad \cdots \quad \mathrm{sim}_{sK} \quad \mathrm{sim\_fianl}_s]$$

$$k_1+k_2+\cdots+k_K=1$$

(6)　End For s

(7)　得到偏好分矩阵

$$\boldsymbol{S}_t^m=[\boldsymbol{s}_{1t} \quad \boldsymbol{s}_{2t} \quad \cdots \quad \boldsymbol{s}_{St}]^{\mathrm{T}}$$

用均值方法将该矩阵中同维度进行融合，

$$\mathrm{sim}_{tk}^m=\frac{1}{S}\sum_{s=1}^{S}\mathrm{sim}_{sk} \tag{3}$$

(8)　得到 $\boldsymbol{H}^m$ 对物品 $\boldsymbol{I}_t$ 的偏好度向量

$$\boldsymbol{S}_t^m=[\mathrm{sim}_{t1}^m \quad \mathrm{sim}_{t2}^m \quad \cdots \quad \mathrm{sim}_{tK}^m \quad \mathrm{sim\_final}_t^m]$$

(9) End For t

(10) 得到 $\boldsymbol{H}^m$ 对物品矩阵 $\boldsymbol{I}$ 的偏好分矩阵

$$\boldsymbol{S}_{mt}=[\boldsymbol{S}_1^m \quad \boldsymbol{S}_2^m \quad \cdots \quad \boldsymbol{S}_T^m]^{\mathrm{T}}$$

即输出结果.

本文以用户历史 Y 年数据计算用户对母基础房型的偏好度为例，具体介绍偏好度特征的构造过程. 假设待推荐母基础房型矩阵 $\boldsymbol{I}=\begin{bmatrix}\boldsymbol{I}_1\\ \boldsymbol{I}_2\end{bmatrix}$，历史数据对应 $\boldsymbol{H}^m=\begin{bmatrix}\boldsymbol{h}_{1*}^m\\ \boldsymbol{h}_{2*}^m\end{bmatrix}$，均为 K 维. 如算法 1 中式(1)，对于不含序关系的特征（如是否大床），直接判断对应维度值是否相同. 对维度中存在序关系的特征（如房间的价格），其首先求得两向量间的绝对值差值，并利用所有用户该维度的均值对该差值进行归一化处理，最后求得负指数即为该维度的偏好分. 若两向量对应维度间差值越小，其偏好分越大，表示用户更偏好该母基础房型. 利用历史均值进行归一化处理可以缓解差值不均匀分布造成的影响，使偏好分更加可信.

由此可得偏好分矩阵 $\boldsymbol{S}_n^m$ 如表 1 所示. 由于每个用户有多条历史信息，故用式(3)进行融合，得到最终结果如表 2 所示.

表 1　偏好分矩阵 $\boldsymbol{S}_n^m$

Tab. 1　Data of $\boldsymbol{S}_n^m$

母基础房型	用户最近 Y 年内的信息	偏好分向量
$\boldsymbol{I}_1$	$\boldsymbol{h}_{1*}^m$	$[\mathrm{sim}_1 \quad \mathrm{sim}_2 \quad \cdots \quad \mathrm{sim}_K \quad \mathrm{sim\_fianl}]_1$
$\boldsymbol{I}_1$	$\boldsymbol{h}_{2*}^m$	$[\mathrm{sim}_1 \quad \mathrm{sim}_2 \quad \cdots \quad \mathrm{sim}_K \quad \mathrm{sim\_fianl}]_2$
$\boldsymbol{I}_2$	$\boldsymbol{h}_{1*}^m$	$[\mathrm{sim}_1 \quad \mathrm{sim}_2 \quad \cdots \quad \mathrm{sim}_K \quad \mathrm{sim\_fianl}]_3$
$\boldsymbol{I}_2$	$\boldsymbol{h}_{2*}^m$	$[\mathrm{sim}_1 \quad \mathrm{sim}_2 \quad \cdots \quad \mathrm{sim}_K \quad \mathrm{sim\_fianl}]_4$

表 2　最终的偏好分矩阵 $\boldsymbol{S}_n^m$

Tab. 2　Finl data of $\boldsymbol{S}_n^m$

母基础房型	用户最近 Y 年内的信息	偏好分向量
$\boldsymbol{I}_1$	$\boldsymbol{H}^m$	$[\mathrm{sim}_1 \quad \mathrm{sim}_2 \quad \cdots \quad \mathrm{sim}_K \quad \mathrm{sim\_fianl}]_a$
$\boldsymbol{I}_2$	$\boldsymbol{H}^m$	$[\mathrm{sim}_1 \quad \mathrm{sim}_2 \quad \cdots \quad \mathrm{sim}_K \quad \mathrm{sim\_fianl}]_b$

2.3　物品排序

本文分别利用线性、非线性 2 种分类器方法实现对物品的排序，具体为经典算法逻辑回归(LR)[23]与目前效果最佳的 XGBoost[24]. 逻辑回归是建立在线性回归模型之上的分类问题，其在回归模型之后利用 sigmod 函数来将预测值进行归一化. Xgboost 属于集成学习，该算法在优化公式中引入正则项，并对损失函数进行二阶泰勒展开，从而快速求得模型的最优解.

最后按照预测分对所有待推荐物品进行排序，排序越靠前代表用户越可能喜欢该物品，即为推荐物品. 针对子房型推荐问题，当用户从手机 APP 端点进某酒店页面时，将该酒店中排在首位的子房型在猜您喜欢板块展示，作为对该用户推荐的子房型.

3　实验

3.1　数据集

本文的数据来自于某 OTA 公司酒店部门从 2017 年 1 月 6 日～2017 年 1 月 10 日，手机 APP 端非取消订单的详情页信息. 根据交叉验证的思想，将数据集按时间划分为训练集、验证集和测试集. 其中训练集和验证集用于模型参数的调整，然后再将训练集和验证集同时作为训练集在最佳参数设置下进行训练，得到用于预测的模型. 数据集的具体划分

为:2017-01-06～2017-01-08 的数据为训练集;2017-01-09 的数据作为验证集. 为了提升训练速度,本文对训练集和验证集的负样本进行了下采样,处理后训练集的正负样本比例为 1∶2,共包含 1 539 290 条记录. 取 2017-01-10 的数据作为测试集,其正负样本比例为 1∶37,共包含 5 386 783 条记录. 表 3 所示为数据集统计信息.

表 3 数据集统计信息

Tab. 3 Some statistics of the dataset

数据集	时间	正负样本比例	数据量
训练集	2017-01-06～2017-01-08	1∶2	1 131 194
验证集	2017-01-09	1∶2	408 096
测试集	2017-01-10	1∶37	5 386 783

3.2 评估指标

本文采用推荐系统中常用的 Top-K 命中率(Hit Rate)来进行评估,Top-K Hit Rate $\eta(K)$指系统每次推荐给用户 K 个产品,用户最终的选择在这 K 个产品里的次数占系统推荐总次数的比例. 用 Z 表示系统推荐次数,W 表示命中次数,计算公式如下:

$$\eta(K) = W/Z$$

3.3 参数设定

本文主要对 XGBoost 模型进行调节使其达到最优,并且在一定区间上以一定步长进行遍历的方法在训练集和验证集上进行寻参,寻参的结果为:最大树深(max_depth)为 11;学习速率(learning_rate)为 0.06;迭代次数(n_estimator)为 500;采样率(Subsample)为 0.8.

3.4 实验结果与分析

实验 1(特征比较) 维度的选择和构造是模型学习中重要的一部分,本文侧重于个性化推荐方向,从用户维度角度出发构造了一系列偏好度维度. 对 4 组不同的维度进行实验:维度 1 中只有酒店/子房型最基本的信息;维度 2 在维度 1 的基础上加入了用户信息;维度 3 则在维度 1 的基础上添加了本文新构造的偏好度维度;维度 4 则同时加入了用户和偏好度 2 种信息. 实验结果如图 3 所示.

由图 3 可知,4 组维度设定中,由于维度 1 不含任何用户维度,其各项均为最低,$\eta(1)$ 只有 37.462%. 在基本信息中加入用户维度后命中率会有提升,但是提升相对来说不是很大. 维度 3 在维度 2 的基础上加入了偏好度维度,其 $\eta(1)$相比于维度 1 提升了将近 5%,有着很明显的效果. 当同时加入用户维度和偏好度,其结果相较于维度 1 有着极大的提升,但是相对于维度 3 则提升有限. 这说明本实验中构造的用户偏好度特征能够很好地表示用户的偏好信息,对推荐精度的提升有着重要的作用. 同时也说明,只是一味地堆加维度对精度提升的作用十分有限,构造能体现出更多信息的有效维度才能得到更好的结果.

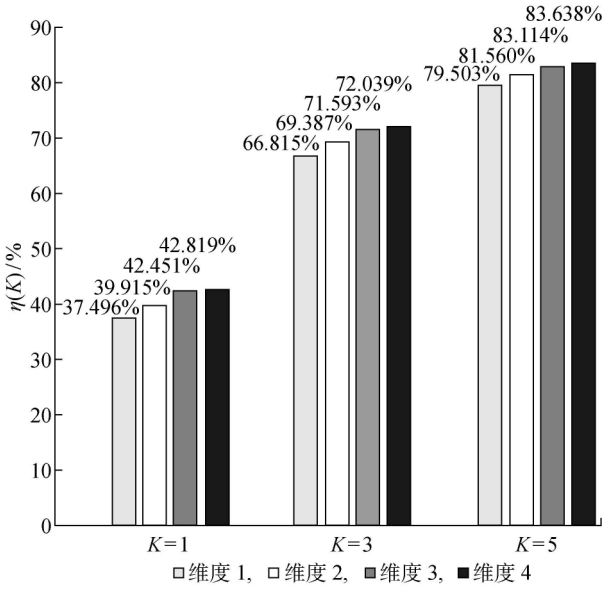

图 3 不同维度设定的实验结果

Fig. 3 Results of different feature settings

此外,由于本实验借助 XGBoost 将物品进行排序,实验时间较短,且模型适用于所有用户,能够满足子房型推荐实时性的要求.

实验 2(方法比较) 本实验将逻辑回归和 XGBoost 方法在包含及不包含偏好度的 2 种维度下进行子房型推荐,并对新提出的偏好度特征构造方法是否具有通用性进行验证.

本实验的设置:① XGBoost 维度与实验 1 中的维度 2 和维度 4 相同. 考虑到逻辑回归对维度的敏感性,利用特征选择的前向搜索方法在归一化处理后的维度 2 和维度 4 中分别挑选出最优的维度子集,作为其维度设置. ② XGBoost 的参数与实验 1 中一致;逻辑回归设置,正则化项为 L2,训练误差取 0.000 1. 图 4 所示为不同方法的实验结果.

由图 4 可以看出,本文提出根据用户及物品间的潜在相关性构造偏好度向量对逻辑回归以及 XGBoost 2 种分类器方法的结果均有 3%左右的提升. 说明新构造的偏好度特征对线性、非线性分类方法均有效果,具有通用性.

除此之外还可得知,子房型推荐是一个复杂的

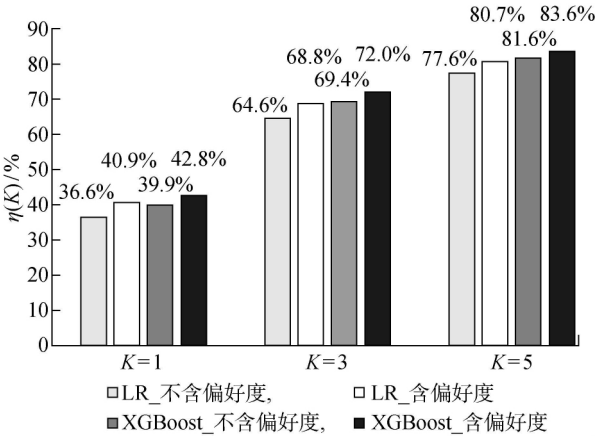

图 4　不同方法的实验结果

Fig. 4　Results of different algorithms

问题，分类器方法对该问题的学习有着一定的优势，能取得较好的结果，且非线性分类器相对来说有着更好的学习能力.

实验 3(训练集比较)　训练集的时长(一定程度上也反映了历史信息的时长)也会影响模型的效果. 在这个实验中，选取不同时间长度的训练集进行模型训练，考察模型在训练集上的效果变化.

本实验的设置：① 共 5 组对照实验，每组实验的训练集对应不同的时长，分别是测试集日期的前 1，2，…，5 d；② 5 组实验参数设置与实验 2 中 XGBoost 保持一致. 表 4 所示为不同训练集的实验结果.

表 4　不同训练集的实验结果

Tab. 4　Results of different training sets

时长/d	时间	$\eta(1)$/%
1	2017-01-09	42.229
2	2017-01-08～2017-01-09	42.441
3	2017-01-07～2017-01-09	42.696
4	2017-01-06～2017-01-09	42.819
5	2017-01-05～2017-01-09	42.758

由表 4 可知，当训练集时长为 1 d 时，其精度是比较低的；随着训练集长度的增加，精度会有一定的提升；但是到训练集时长为 5 d 时，开始出现下降. 当训练集时长较少时，训练数据也较少，模型不能充分地学习；当训练集时长过长时，有些数据已不能代表用户的偏好，会对结果造成一定的负面影响.

训练集时长在一定程度上反映了历史信息的时长，本实验说明用户的历史信息不是越久越好，太久之前的历史信息有时不能代表该用户当前的兴趣爱好，反而会带来一些负面的影响.

4　结语

针对传统推荐算法的不足，本文提出一种基于偏好度特征构造的个性化推荐算法，利用用户的历史信息，构造用户与物品间的偏好度特征. 在实际应用中，本文从减少用户决策时间和提高用户满意度的角度出发，首次在酒店信息展示板块提出对用户进行个性化子房型推荐. 实验部分的结果显示偏好度特征相比于单纯维度的叠加有了 5% 以上的提升，说明了新算法的有效性. 将新维度在多种方法中进行实验，验证了其通用性. 此外，新算法借助机器学习领域的分类器算法，使得其能够满足实时性的要求.

随着个性化推荐的不断发展，推荐也早已不局限于浅层的用户和物品挖掘，但对用户行为的挖掘仍是一个相对困难的问题；另外，深度学习已经证明在图像及自然语言处理领域取得了较好的效果，在推荐系统领域中的应用也越来越多. 而本文提出的基于内容的子房型推荐系统存在着数据较稀疏、冷启动等问题，日后可尝试将本文提出的方法与深度学习相结合来攻克这些难题.

参考文献：

[1] SARWAR B, KARYPIS G, KONSTAN J, *et al*. Analysis of recommendation algorithms for e-commerce[C]// **ACM Conference on Electronic Commerce**. Minneapolis: ACM, 2000:158-167.

[2] PERUGINI S, GONCALVES M A. Recommendation and personalization: A survey[J]. **Journal of Intelligent Information Systems**, 2002, 23(2):107-143.

[3] BATET M, MORENO A, ISERN D. Turist@: Agent-based personalised recommendation of tourist activities [J]. **Expert Systems with Applications**, 2012, 39(8): 7319-7329.

[4] GAVALAS D, KENTERIS M. A web-based pervasive recommendation system for mobile tourist guides [J]. **Personal & Ubiquitous Computing**, 2011, 15(7): 759-770.

[5] VINODHINI S, RAJALAKSHMI V, GOVINDARAJULU B. Building personalised recommendation system with big data and Hadoop mapreduce[J]. **Metabolism Clinical & Experimental**, 2009, 58(1): 38.

[6] VERMA J P, PATEL B, PATEL A. Big data analysis: Recommendation system with Hadoop framework [C]// **IEEE International Conference on Computational Intelligence & Communication Technology**. Ghazi-

abad: IEEE, 2015: 92-97.

[7] ZHANG K, WANG K, WANG X, *et al*. Hotel recommendation based on user preference analysis[C]// **IEEE International Conference on Data Engineering Workshops**. Seoul: IEEE, 2015:134-138.

[8] GAO H, LI W. A hotel recommendation system based on collaborative filtering and Rankboost algorithm[C]// **International Conference on Multimedia & Information Technology**. Kaifeng: IEEE, 2010: 317-320.

[9] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. **Communications of the ACM**, 1974, 18(11): 613-620.

[10] AHN J W, BRUSILOVSKY P, GRADY J, *et al*. Open user profiles for adaptive news systems: Help or harm?[C]// **International Conference on World Wide Web, WWW 2007**. Banff: Springer International Publishing, 2007: 11-20.

[11] LI ZN, DREW M S, LIU J. Content-based retrieval in digital libraries[M]. Switzerland: Springer International Publishing, 2014: 93-95.

[12] **刘健, 毕强, 刘庆旭, 等. 数字文献资源内容服务推荐研究——基于本体规则推理和语义相似度计算[J]. 现代图书情报技术**, 2016, 32(9): 70-77.
LIU Jian, BI Qiang, LIU Qingxu, *et al*. New content recommendation service of digital literature[J]. **New Technology of Library and Information Service**, 2016, 32(9): 70-77.

[13] **江周峰, 杨俊, 鄂海红. 结合社会化标签的基于内容的推荐算法[J]. 软件**, 2015, 36(1): 1-5.
JIANG Zhoufeng, YANG Jun, E Haihong. A content-based recommendation algorithm with social tagging[J]. **Software**, 2015, 36(1): 1-5.

[14] WANG M, ZHENG Y, QIU M, *et al*. Research on schedule-based user recommendation model based on improved *K*-means algorithm[J]. **Journal of Computational Methods in Sciences & Engineering**, 2016, 16(3): 1-10.

[15] **闫东东, 李红强. 一种改进的基于内容的个性化推荐模型[J]. 软件导刊**, 2016, 15(4): 11-13.
YAN Dongdong, LI Hongqiang. An improved content-based personalized recommendation model[J]. **Software Guide**, 2016, 15(4): 11-13.

[16] **李聪, 梁昌勇, 马丽. 基于领域最近邻的协同过滤推荐算法[J]. 计算机研究与发展**, 2008, 45(9): 1532-1538.
LI Cong, LIANG Changyong, MA Li. A collaborative filtering recommendation algorithm based on domain nearest neighbor[J]. **Journal of Computer Research and Development**, 2008, 45(9): 1532-1538.

[17] **黄创光, 印鉴, 汪静, 等. 不确定近邻的协同过滤推荐算法[J]. 计算机学报**, 2010, 33(8): 1369-1377.
HUANG Chuangguang, YIN Jian, WANG Jing, *et al*. Uncertain neighbors' collaborative filtering recommendation algorithm[J]. **Chinese Journal of Computers**, 2010, 33(8): 1369-1377.

[18] WANG Z, LI Q, LIU Y, *et al*. Online personalized recommendation based on streaming implicit user feedback[C]// **Asia-Pacific Web Conference**. Guangzhou, China: Springer International Publishing, 2015: 720-731.

[19] **王琳琳. 基于协同过滤的在线学习个性化推荐技术研究[J]. 微型电脑应用**, 2017, 33(5): 49-51.
WANG Linlin. Research on personalized recommendation technology of online learning based on collaborative filtering [J]. **Microcomputer Applications**, 2017, 33(5): 49-51.

[20] **应毅, 刘亚军, 陈诚. 基于云计算技术的个性化推荐系统[J]. 计算机工程与应用**, 2015, 51(13): 111-117.
YING Yi, LIU Yajun, CHEN Cheng. Personalization recommender system based on cloud-computing technology[J]. **Computer Engineering and Applications**, 2015, 51(13): 111-117.

[21] SALAKHUTDINOV R, MNIH A, HINTON G. Restricted Boltzmann machines for collaborative filtering[C]// **Proceedings of the 24th International Conference on Machine Learning**. Corvallis: ACM, 2007: 791-798.

[22] HE X, LIAO L, ZHANG H, *et al*. Neural collaborative filtering[C]// **Proceedings of the 26th International Conference on World Wide Web**. Perth: Springer International Publishing, 2017: 173-182.

[23] HOSMER D W, LEMESBOW S. Goodness of fit tests for the multiple logistic regression model[J]. **Communications in Statistics**, 1980, 9(10): 1043-1069.

[24] CHEN T, HE T. Higgs boson discovery with boosted trees[C]// **NIPS 2014 Workshop on High-energy Physics and Machine Learning**. Montreal: MIT Press, 2014: 69-80.