

基于用户属性聚类与矩阵填充的景点推荐算法

刘荣权,袁仕芳,赵锦珍,杨伟杰

(五邑大学 数学与计算科学学院,广东 江门 529020)

摘要:随着互联网和旅游业的发展,可以选择的旅游景点越来越多。在海量的景点信息中,景点的选择成为旅客出行的一个重要问题。该文采用改进的协同过滤算法,给每个旅客推荐合适的旅游景点,以解决他们出行难的问题。首先对传统的协同过滤算法进行改进,即对用户属性进行二分聚类;再利用奇异值分解算法填充稀疏的用户评分矩阵,得到多个聚类类别的中心和一个填充完整用户评分矩阵;然后计算出目标用户各属性到各个聚类中心的欧氏距离,将其分到距离最小的类别;再利用 Pearson 相似度方法和填充完整的用户评分矩阵计算出目标用户与同一类别中其他用户的相似度;最后结合相似度,用 Top - N 推荐方法将预测景点评分进行降序排序,并推荐给目标用户,从而提高推荐算法的精准度。实验结果表明,该算法比传统协同过滤算法的推荐质量有显著提高。

关键词:景点;用户属性;数据稀疏;聚类; Top - N

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2020)11-0200-05

doi:10.3969/j.issn.1673-629X.2020.11.037

Tourist Spot Recommendation Algorithm Based on User Attribute Clustering and Matrix Filling

LIU Rong-quan, YUAN Shi-fang, ZHAO Jin-zhen, YANG Wei-jie

(School of Mathematics and Computational Science, Wuyi University, Jiangmen 529020, China)

Abstract: With the development of the Internet and tourism, more and more tourist attractions are available. Among the massive information, the choice of tourist attractions has become an important issue for travelers. We adopt an improved collaborative filtering algorithm to recommend suitable tourist attractions for each traveler to solve their travel difficulties. We firstly improve the traditional collaborative filtering algorithm, which is binary clustering of user attributes, and fill the sparse user rating matrix by the SVD algorithm to obtain the centers of multiple clustering categories and a complete user rating matrix. Then we calculate the Euclidean distance from each attribute of the target user to each cluster center which is divided into the category with the smallest distance, and then calculate the similarity between the target user and other users in the same category by using Pearson similarity method and filling the complete user rating matrix. Finally combining similarity, we use the Top - N recommendation method to sort the predicted attraction scores in descending order and recommend them to the target users for improving the accuracy of the recommendation algorithm. Experiment shows that the proposed algorithm has significantly improved the recommendation quality compared with traditional collaborative filtering algorithms.

Key words: tourist spots; user attributes; data sparsity; clustering; Top - N

0 引言

在当今数据化时代,人们出行需要从海量景点信息中选取自己喜欢的景点,导致旅游用户出行前有选择困难。为了解决这个问题,许多旅游推荐系统会根据用户的行为特征,给用户推荐一些可能喜欢的景点,帮助用户做出选择。近几年来,基于用户的协同过滤算法和基于项目的协同过滤算法^[1]在这方面起到了

重要的作用。目前有很多学者不断改进传统的协同过滤算法,以提高旅游推荐的质量^[2-4]。

传统协同过滤算法主要是根据评分矩阵计算推荐结果,评分矩阵的好坏直接影响推荐的质量。评分矩阵最大的不足是数据稀疏^[5],因为大部分用户去过的旅游景点数较少,只会对海量景点中的少数景点做出评分,从而做出的评分数量很少,使得评分矩阵出现大

收稿日期:2019-12-03

修回日期:2020-04-05

基金项目:广东省自然科学基金项目(2015A030313646);2018年五邑大学教学质量工程与教学改革项目(JX2018024);五邑大学创空间大学生创新创业项目(18KSX02)

作者简介:刘荣权(1997-),男,研究方向为数据处理;通讯作者:袁仕芳(1972-),男,博士,教授,硕导,研究方向为数据处理和数值分析。

量的缺失值。

针对上述问题,该文尝试基于用户的属性对用户进行聚类,计算目标用户与同一类中的其他用户的相似度,这样不仅减少了计算量,而且使所得近邻用户集更准确。然后用奇异值分解算法填充评分矩阵,从而在计算预测评分时减少数据稀疏带来的影响,提高预测评分的准确性。

1 用户属性聚类

目前传统 k -means 算法^[6]和二分 k -means 聚类算法^[7]在用户属性聚类算法中有重要应用。传统 k -means 算法:随机选取初始簇心、无法确定类别个数造成聚类误差较大,文献[8]证明了二分 k -means 聚类减小初始随机选取簇心的影响,降低聚类误差。传统的推荐算法是在单一的评分矩阵中给目标用户推荐项目,如果数据过于稀疏,将导致目标用户的近邻用户集不准确。本节主要讨论将二分 k -means 聚类算法应用到用户属性聚类中,因为同一类用户的特征越相似,则它们的相似度越高,这使得目标用户的相近邻用户集越准确。

1.1 旅游用户属性预处理

用户属性是指用户的一些基本特征。在文献[9]中提到,影响用户出行的因素主要有职业、性别、年龄、出行时段、出游陪同、人均消费和出发天数。以文献[9]为基础,该文选取用户性别、注册年份、天数、出发时间、陪同、人均消费和玩法作为用户属性,构成数据集。对数据集进行量化,玩法量化主要计算每个用户的玩法的词频和与总玩法的词频和之比。量化如表 1 所示。

表 1 用户属性量化

属性	区间选项
用户性别	1 = “男”; 2 = “女”
用户注册年份	1 = “2001 ~ 2005 年” 2 = “2006 ~ 2010 年” 3 = “2011 ~ 2015 年” 4 = “2016 ~ 2019 年”
游玩天数	1 = “1 天”; 2 = “2 天”; 3 = “3 天”; 4 = “4 天”; 5 = “其他”
出发时间(月份)	1 = “1 ~ 3 月”; 2 = “4 ~ 6 月”; 3 = “7 ~ 9 月”; 4 = “10 ~ 12 月”
出游陪同	1 = “和朋友”; 2 = “亲子”; 3 = “一个人”; 4 = “情侣”; 5 = “夫妻”; 6 = “和父母”

续表 1

属性	区间选项
人均消费	1 = “0 ~ 500 元”; 2 = “500 ~ 1000 元”; 3 = “1000 ~ 1500 元”; 4 = “1500 ~ 2000 元”; 5 = “其他”
玩法	1 = “0 ~ 0.3”; 2 = “0.3 ~ 0.35”; 3 = “0.35 ~ 0.4”; 4 = “0.4 ~ 0.45”; 5 = “0.45 ~ 0.5”

1.2 二分 k -means 聚类

定义用户集 $\{x_i | i = 1, 2, \dots, n\}$ 和用户属性集 $\{a_{ij} | j = 1, 2, \dots, 7\}$ (例如用 a_{i1} 可表示用户 i 的性别属性值,后面的以此类推),对每一个用户属性集进行归一化处理,消除指标之间的量纲的影响,处理方法如下:

$$a_{ij}^* = \frac{a_{ij} - \min(a_j)}{\max(a_j) - \min(a_j)}$$

其中, $\max(a_j)$ 、 $\min(a_j)$ 是用户属性 j 的最大值、最小值。

k -means 聚类算法描述如下:

输入:用户属性集合,聚类数 K 。

输出: K 个聚类结果。

(1) 初始化 K 个聚类中心。在每个属性的范围内选择 K 个随机值,生成 K 个聚类中心,以保证每个聚类中心仍处于原数据集内。定义第 i 个聚类中心为:

$$c_i = \{c_{i1}, c_{i2}, \dots, c_{i7}\}, i \in \{i | 1 \leq i \leq K, i \in N\}$$

$$c_{ij} = [\max(a_j) - \min(a_j)] \cdot \text{rand} + \min(a_j)$$

$$\{j | 1 \leq j \leq 7\}$$

其中,rand 是 (0,1) 之间的随机数, c_{ij} 是聚类中心 c_i 第 j 个用户属性的随机值。

(2) 计算所有样本点到 K 个聚类中心的欧氏距离,并把样本点归类到距离最小的类中。计算公式如下:

$$d(x_i, c_j) = \sqrt{(a_{i1} - c_{j1})^2 + (a_{i2} - c_{j2})^2 + \dots + (a_{i7} - c_{j7})^2}$$

其中, x_i 是第 i 个用户的用户属性向量, c_j 是第 j 类的聚类中心。

(3) 经过步骤(2)得到 K 个类别,计算每个类别各个属性的均值 avg,并判断:如果 avg _{j} 与原聚类中心 c_j 相等,则目标函数收敛,迭代结束;否则,将各类的属性平均值 avg 作为新的聚类中心,继续执行步骤(2)、

(3),直到迭代结束,其中 $\text{avg}_j = \frac{\sum_{i \in C_j} [a_{i1}, a_{i2}, \dots, a_{i7}]}{|C_j|}$, $|C_j|$ 表示第 j 类中用户的数量, a_{i1} 表示用户 i 的第一个属性值(a_{i2} 等以此类推)。

当聚类数 $K = 2$ 时, k -means 算法变为二分 k -means 算法, 它的主要聚类过程是 k -means 算法。算法一般采用误差平方和作为目标函数, 不断迭代优化聚类结果, 当目标函数收敛, 迭代结束。目标函数如下:

$$\text{error} = \sum_{i=1}^k \sum_{x \in C_i} \|x - c_i\|^2$$

其中, error 为所有用户与对应聚类中心的误差平方和, c_i 为第 i 类的聚类中心, x 为第 i 类的样本点。

二分 k -means 聚类算法描述如下:

输入: 用户属性集合, 聚类数 K 。

输出: K 个聚类结果。

(1) 初始时, 建立类表(用于存储二分聚类结果), 将所有用户视为同一类 c 并存入类表。

(2) 当类表中只有一类 c 时, 将此类 c 中所有用户属性代入上述 k -means 聚类算法($K = 2$), 得到两个新类 c_1, c_2 , 存入类表并取代 c 。

当类表中有两类及以上时, 以聚类后所得误差平方和最小为目标, 遍历当前类表中的所有类, 执行上述 k -means 聚类算法; 当遍历完毕时, 若某类满足上述目标, 使其成为目标类并执行二分聚类, 将所得两个新类取代目标类。

(3) 当 K 达到输入要求时, 算法停止, 否则继续执行步骤(2)。

2 景点推荐算法

上一节对用户属性进行二分 k -means 聚类, 使同一类中的用户具有更高的相似度; 在计算目标用户与同一类其他用户的相似度时, 不仅减少了计算量, 而且使所得近邻用户集更准确。但在实际中, 用户之间共同评分的景点数较少, 使评分矩阵稀疏程度偏大。本节采用文献[10]中的矩阵奇异值分解(SVD)对用户景点评分矩阵进行填补, 并基于用户的协同过滤算法, 预测目标用户景点评分。

2.1 传统的协同过滤算法

传统的协同过滤算法包括基于用户的协同过滤算法和基于项目的协同过滤算法^[11]。由于大部分景点缺失描述信息, 也没有相应描述标签, 使景点之间的相似度不易计算。若采用基于项目的协同过滤算法进行推荐, 则难以实施。因此该文采用基于用户的协同过滤算法。

它的主要内容如下:

(1) 构建用户景点评分矩阵。

收集 m 个用户对 n 个景点的评分数据, 定义用户集 $\{x | x_1, x_2, \dots, x_m\}$, 景点集 $\{y | y_1, y_2, \dots, y_n\}$, 构成大小 $m \times n$ 的评分矩阵, 如图1所示。

	y_1	y_i	y_n
x_1	$S(x_1, y_1)$	\dots	$S(x_1, y_n)$
\dots	\dots	\dots	\dots
x_i	$S(x_i, y_1)$	\dots	$S(x_i, y_n)$
\dots	\dots	\dots	\dots
x_m	$S(x_m, y_1)$	\dots	$S(x_m, y_n)$

图1 用户-景点评分矩阵

其中, $S(x_i, y_j)$ 为用户 x_i 对景点 y_j 的评分; 规定用户未评价景点的评分为-1, 依据用户对景点的喜爱程度, 将评分设为五分制。

(2) 用户相似度。

在推荐系统中, 计算相似度的方法有 Jaccard 相关系数^[12]、余弦相似度和 Pearson 相关系数等等^[13], 其中 Pearson 相关系数应用较为广泛。该文使用 Pearson 相关系数计算用户相似度, 公式如下:

$$\text{sim}(x_i, x_j) = \frac{\sum_{n_r \in N} (n_r^i - \bar{x}_i)(n_r^j - \bar{x}_j)}{\sqrt{\sum_{n_r \in N} (n_r^i - \bar{x}_i)^2 \sum_{n_r \in N} (n_r^j - \bar{x}_j)^2}}$$

其中, $N \setminus \{n_1, n_2, \dots, n_n\}$ 为目标用户与其他用户共同评分的景点集合, n_r^i, n_r^j 分别是用户 x_i, x_j 对景点 r 的评分, \bar{x}_i, \bar{x}_j 分别是用户 x_i, x_j 所有已评分景点的评分均值。

(3) 景点评分预测。

该文使用 Top- N 推荐为每个用户推荐个性化的景点列表。通过用户相似度计算用户 x_i 近邻用户集 $X_i = \{x_1, x_2, \dots, x_n\}$; 设 $\tilde{N} = \{\tilde{n}_1, \tilde{n}_2, \dots, \tilde{n}_n\}$ 是目标用户 x_i 为没评价而其近邻用户集 X_i 已经评价的景点集合, 计算用户 x_i 与 \tilde{N} 中每个景点的预测评分, 将所得结果降序排序构成 Top- N 推荐列表, 向用户展示。计算用户 x_i 对未评分景点 \tilde{n}_1 的预测评分公式^[14]如下:

$$R(x_i, \tilde{n}_1) = \frac{\sum_{x_j \in X_i} \text{sim}(x_i, x_j) \times (R(x_j, \tilde{n}_1) - \bar{R}_{x_j})}{\sum_{x_j \in X_i} \text{sim}(x_i, x_j)}$$

其中, $\bar{R}_{x_i}, \bar{R}_{x_j}$ 分别是用户 x_i 、近邻用户 x_j 对其已游览景点评分的均值, $R(x_j, \tilde{n}_1)$ 是用户 x_j 对景点 \tilde{n}_1 的评分。

2.2 SVD 算法填充矩阵

在推荐系统中,用户景点评分矩阵的数据稀疏性严重影响推荐算法的预测准确度。对于数据稀疏,文献[15]的解决办法有:简单的数据填补、对用户进行聚类和数据降维。

本节采用 SVD 算法^[16]对原数据进行降维分解,最大程度地保留原数据的信息,填充原数据的缺失值,得到完整的评分矩阵。填充过程描述如下:

(1)初始时,计算每一列景点评分的均值 $\bar{r}_j (1 \leq j \leq n)$ 和每一行用户评分的均值 $\bar{r}_i (1 \leq i \leq m)$,用 \bar{r}_j 填充每一列空缺值,每一行元素减去对应用户评分的均值 \bar{r}_i ,得到一个 $m \times n$ 的标准化矩阵 R' ,对 R' 进行奇异值分解,分解如下:

$$R' = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$$

在 $\Sigma_{m \times n}$ 中,对角元素称为奇异值,它包含着 R' 中重要的信息,这也是数据降维的关键。

(2)规定阈值为 95%。当前 k 个奇异值的累计贡献率达到 95%,可认为这前 k 个奇异值保存足够的特征。

(3)对于矩阵 $\Sigma_{m \times n}$,只保留 k 个对角元素得到新的对角矩阵 $\Sigma_{k \times k}$;对矩阵 U, V , 同理可得新的矩阵 $U_{m \times k}$ 和 $V_{k \times n}^T$ 。

(4)计算 $\sqrt{\Sigma_k}$,再作 $U_k \cdot \sqrt{\Sigma_k}$ 和 $\sqrt{\Sigma_k} \cdot V_k^T$ 运算。

(5)对缺失值进行预测。计算用户 x 在景点 i 缺失值的预测评分,可以通过下式求解:

$$M_{x,i} = \bar{R}_x + U_k \times U_k \sqrt{\Sigma_k}(x) \times \sqrt{\Sigma_k} V_k^T(i)$$

其中, \bar{R}_x 是用户 x 的景点评分均值。

2.3 总体的算法流程

算法实现流程如图 2 所示。

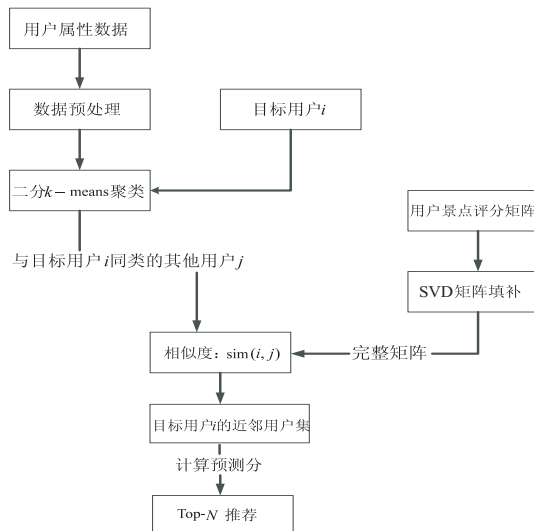


图 2 算法实现流程

3 数据实验与分析

3.1 实验环境

实验环境为 Windows7、64 位操作系统,所用编程语言为 Python。

3.2 实验数据

该研究在携程网站(www.ctrip.com),以“深圳”为关键字,收集用户和评分数据。共获得 1 551 个用户数据以及用户对 49 个景点的评分数据。分析可知,评分矩阵的稀疏度为 92.5%。随机选取 80% 的实验数据作为训练集,20% 作为测试集。

3.3 评测指标

本节采用平均绝对误差(MAE)^[17]评定预测评分的精准度。对于测试集而言,其 MAE 的值越小,表明推荐结果越好。给定测试集中某用户 u 和某景点 i ,令 $R(u, i)$ 为用户 u 对景点 i 的实际评分,而 $\hat{R}(u, i)$ 是用户 u 对景点 i 的预测评分。

MAE 的定义如下:

$$MAE = \frac{\sum_{i \in N} |R(u, i) - \hat{R}(u, i)|}{|N|}$$

其中, $|N|$ 是未评分的景点数量。

3.4 实验结果与分析

初始时,二分 k -means 聚类数 K 是一个不确定值。以二分聚类结果的误差平方和最小为目标,循环运算二分 k -means 聚类算法得到最优的 K 值。从图 3 中可见,当聚类数为 5 时,误差平方和最小。综上该文选取聚类 $K = 5$ 。

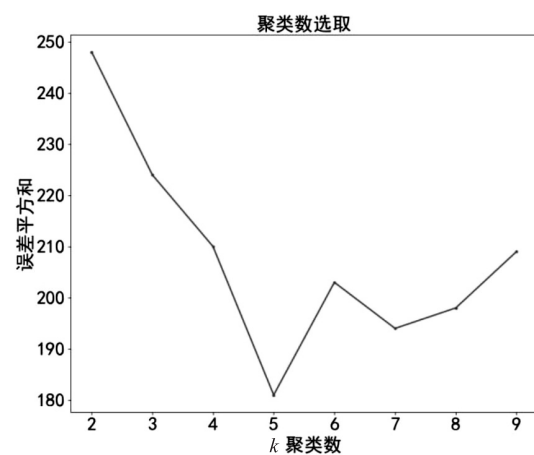


图 3 最佳 K 聚类数

对于多维聚类,利用 Python 的 TSNE 进行数据可视化,将五维数据的聚类结果展示到二维空间,如图 4 所示。可见原数据聚类成五类,同一类中聚拢程度高,不同类间分界明显,聚类效果符合预期目标。

为了评测该算法的推荐质量,将基于用户的协同过滤算法、经聚类优化后的协同过滤算法与文中算法

进行比较。三个算法在不同的近邻用户数量下,计算 MAE 并比较,结果如图 5 所示。

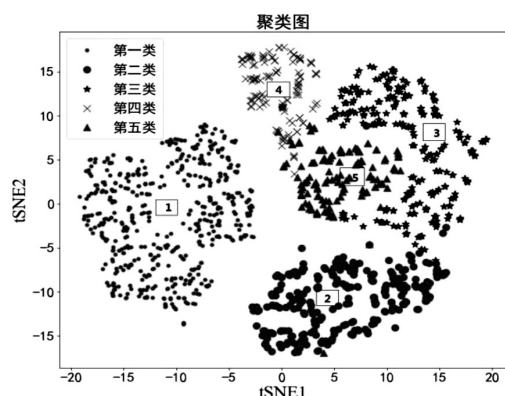


图 4 聚类结果可视化

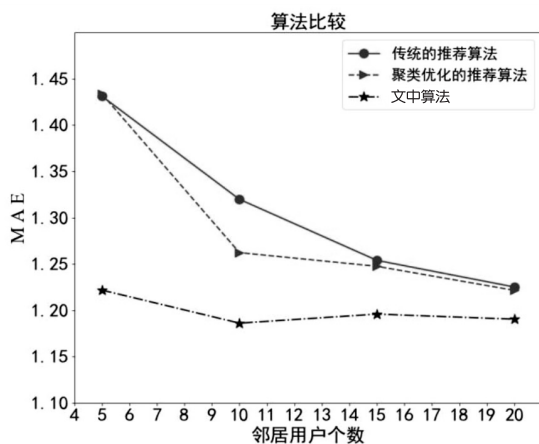


图 5 不同算法比较

由图 5 可知,在不同的近邻用户数量下,文中算法所得 MAE 值均比基于用户的协同过滤算法、聚类后的协同过滤算法的小,说明文中算法的景点预测评分与真实评分之间误差最小,推荐质量最优。基于用户的协同过滤算法使用较稀疏的评分矩阵计算相似度,所得 MAE 值最大;对用户属性进行聚类的协同过滤算法在一定程度上使目标用户的近邻用户集更准确,所得 MAE 值较小;而该文对用户进行属性聚类并填充稀疏的评分矩阵,减小数据的稀疏度,所得 MAE 值在三者中最小。

4 结束语

针对旅游数据稀疏性偏大的问题,对用户属性进行二分 k -means 聚类,对评分矩阵进行填充,得到完整的用户评分矩阵;在目标用户所在的类别中,对目标用户计算 Pearson 相似度,找到目标用户的近邻用户集;计算目标用户对未游览景点的预测评分,构成 Top- N 推荐列表,向目标用户展示。实验表明,该算

法的推荐质量优于传统推荐算法。

参考文献:

- [1] 常亮,曹玉婷,孙文平,等. 旅游推荐系统研究综述[J]. 计算机科学,2017,44(10):1-6.
- [2] 王志强,益民,李芳. 基于多方面评分的景点协同推荐算法[J]. 山东大学学报:工学版,2016,46(6):54-61.
- [3] 吴军. 基于协同过滤的个性化旅游推荐系统的研究与实现[D]. 北京:北京交通大学,2017.
- [4] 徐旋旋. 个性化旅游景点推荐研究[D]. 天津:天津理工大学,2017.
- [5] 姜维,庞秀丽. 面向数据稀疏问题的个性化组合推荐研究[J]. 计算机工程与应用,2012,48(21):21-25.
- [6] WAGSTAFF K, CARDIE C, ROGERS S, et al. Constrained k -means clustering with background knowledge[C]//Proceedings of the eighteenth international conference on machine learning. San Francisco: Morgan Kaufmann, 2001: 577-584.
- [7] KASHEF R, KAMEL M S. Enhanced bisecting k -means clustering using intermediate cooperation[J]. Pattern Recognition, 2009, 42(11): 2557-2569.
- [8] 陈贤宇,李有强,吕苗苗,等. 基于二分法的 K -means 算法的实现[J]. 无线电通信技术, 2017, 43(6): 37-40.
- [9] 葛学峰. 旅游目的地选择意向影响因素研究[D]. 大连:大连理工大学,2012.
- [10] 夏培勇. 个性化推荐技术中的协同过滤算法研究[D]. 青岛:中国海洋大学,2011.
- [11] LINDEN G, SMITH B, YORK J. Amazon.com recommendations: item-to-item collaborative filtering[J]. IEEE Internet computing, 2003(1): 76-80.
- [12] PATEREK A. Improving regularized singular value decomposition for collaborative filtering[C]//Proceedings of KDD cup and workshop. California: ACM, 2007: 5-8.
- [13] 马宏伟,张光卫,李鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统, 2009, 30(7): 1282-1288.
- [14] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2008: 471-480.
- [15] HUANG Z, CHEN H, ZENG D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering[J]. ACM Transactions on Information Systems, 2004, 22(1): 116-142.
- [16] 刘晴晴,罗永龙,汪逸飞,等. 基于 SVD 填充的混合推荐算法[J]. 计算机科学, 2019, 46(6A): 468-472.
- [17] 项亮. 推荐系统实践[M]. 北京:人民邮电出版社,2012.