

# 基于时序模型和矩阵分解的推荐算法\*

蔡海尼<sup>1</sup>, 牛冰慧<sup>1</sup>, 文俊浩<sup>1</sup>, 王喜宾<sup>2</sup>

(1. 重庆大学 软件学院, 重庆 400044; 2. 重庆邮电大学 软件工程学院, 重庆 400065)

**摘要:** 时间序列数据是一种数据属性随时间变化的高维数据类型,反映了用户兴趣的动态变化。基于时序数据的推荐系统利用用户的行为时间提高推荐的准确性,但不适用于大规模数据集的推荐任务。矩阵分解方法是处理高维数据集时常用的降维方法。为此,提出一种基于时序模型和矩阵分解的推荐算法。基于该方法,首先利用矩阵分解提取原始时序数据的特征,然后通过时序模型挖掘特征的趋势,最后根据预测的特征得到预测结果并进行推荐。实验结果表明,所提出的算法与已有的推荐算法相比,在均方根误差(root mean square error, RMSE)和平均准确率(mean average precision, MAP)两个指标上均有较好表现,且适用于大规模数据的推荐任务。

**关键词:** 推荐算法; 概率矩阵分解; 时序行为; 行为预测

**中图分类号:** TP391.9

**文献标志码:** A

**文章编号:** 1001-3695(2018)06-1624-04

doi:10.3969/j.issn.1001-3695.2018.06.005

## Recommender algorithm based on time series model and matrix factorization

Cai Haini<sup>1</sup>, Niu Binghui<sup>1</sup>, Wen Junhao<sup>1</sup>, Wang Xibin<sup>2</sup>

(1. School of Software Engineering, Chongqing University, Chongqing 400044, China; 2. School of Software Engineering, Chongqing University of Posts & Telecommunications, Chongqing 400065, China)

**Abstract:** Time series data is a high-dimensional data type in which the attributes change over time, and it reflects the dynamic change of user interest. Recommender system based on the time series data is able to improve the efficiency of recommendation, but it is not suitable for large scale data sets, matrix factorization is a commonly used method to reduce the dimension of high dimensional data set. This paper proposed a method to combine the time-series model with matrix factorization. Firstly, this method extracted the characteristics of the original time series data by using the matrix factorization, then it extracted the trend of the feature by the sequential model. Finally, it obtained the prediction results according to the predicted characteristics. The experiments demonstrate that the proposed method performs better on root mean square error (RMSE) and mean average precision (MAP) than the conventional recommender algorithms, and it is also appropriate for the prediction task of large-scale data.

**Key words:** recommender algorithms; probabilistic matrix factorization (PMF); sequential behavior; behavior prediction

## 0 引言

推荐系统(recommender system, RS)能够从大规模的数据中快速找到用户感兴趣的信息,是个性化服务研究领域的一个重要分支。推荐算法是推荐系统中的核心内容。传统的推荐算法通过研究用户与项目之间的二元关系,识别和预测用户偏好,提供个性化服务。目前,推荐算法在电子商务(如阿里巴巴、Netflix、Amazon等)、搜索引擎(如Google、百度等)以及互联网广告、移动服务等众多应用领域内取得了较大进展。然而传统的推荐算法,如协同过滤<sup>[1]</sup>、基于内容的过滤等主要针对推荐任务中存在的数据稀疏性<sup>[2,3]</sup>和冷启动<sup>[4]</sup>等问题进行研究,当信息规模越来越大时,推荐的质量和效率就遇到了很大的挑战。随着数据量的爆炸式增长和用户需求的升级,推荐算法面临一系列的挑战,如稀疏性问题、用户兴趣漂移问题等。传统的推荐算法如协同过滤等不能根据用户兴趣偏好的变化快速作出反映,只能生成静态的用户偏好模型,当用户偏好发生很快变化时,推荐结果准确度会急剧下降。

时间序列数据(time series data)是一种重要的高维数据类型,时序数据的属性随时间变化。为了解决传统推荐算法中用户兴趣漂移问题,如何在推荐系统中有效地利用时序数据开始受到研究人员的关注。比如,用户近一个经常听流行音乐,而之前的大部分时间经常听古典音乐,那么可以认为用户的偏好产生了漂移,由古典音乐变为流行音乐。传统的推荐算法通常会根据用户的大部分偏好为用户推荐古典音乐,而利用时间序列数据进行推荐时,推荐系统可以学习到用户的兴趣变化,为用户推荐流行音乐。因此,结合时序数据的推荐系统可以帮助用户从海量动态变化的资源中获取真正满足自身需要的内容,具有重要的研究意义和应用价值,近年来成为推荐系统研究领域的活跃分支。现有的利用时序数据的推荐方法主要包括两类:一类是在传统的推荐系统中以时间权重的形式加入时间特性<sup>[5]</sup>,但是随着数据规模的越来越大以及用户需求的增加,这种方法的算法复杂度较高,并且对于时间权重使用,无法有效利用时序数据规律变化的特征;为解决这个问题,研究人员提出另一类通过张量分解利用时间信息的方法<sup>[6,7]</sup>,这类方法将时间作为单独的一维数据进行数据建模,但是通常具有很

**收稿日期:** 2017-02-14; **修回日期:** 2017-04-09 **基金项目:** 国家自然科学基金资助项目(61672117); 重庆市自然科学基金资助项目(cstc2014jcyjA40054); 重庆市教育委员会科学技术研究项目(KJ1600437)

**作者简介:** 蔡海尼(1972-),女,海南文昌人,副教授,博士,主要研究方向为服务计算(hainicai@cqu.edu.cn); 牛冰慧(1993-),女,山东淄博人,硕士研究生,主要研究方向为推荐算法、数据挖掘; 文俊浩(1969-),男,河南漯河人,教授,博士,主要研究方向为服务计算、数据挖掘; 王喜宾(1985-),男,河南洛阳人,讲师,博士,主要研究方向为机器学习、数据挖掘、个性化推荐。

高的计算复杂度。针对用户—项目—时间张量分解的复杂度问题,本文以时序模型与矩阵分解模型相结合的方法减小算法的复杂度。本文方法首先提取用户和项目的特征;然后通过时序模型挖掘时序数据特征的趋势,提高推荐结果的有效性。

本文提出一种基于时序模型和矩阵分解的推荐算法。主要内容包括:

a) 针对每一时间段的用户—项目二元关系进行矩阵分解(matrix factorization, MF), 得到用户和项目的特征矩阵。

b) 将用户和项目特征矩阵的时序数据代入多维时序模型中, 得到未来时间段的预测特征矩阵, 计算未来时间段中用户对项目的行为预测值, 选择预测值 Top- $N$  项目进行推荐。

c) 通过实验与现有方法进行比较, 证明该方法可有效提高预测准确率, 并且预测精度会随着特征向量维度和时序数量的增加而增加。

## 1 相关工作

### 1.1 概率矩阵分解

概率矩阵分解(PMF)是 Salakhutdinov 等人<sup>[8]</sup>提出的一种矩阵分解模型, 该模型可以有效地获取用户、项目的潜在特征矩阵, 起到矩阵降维的作用。该模型从概率的角度解释了矩阵分解原理, 将用户对项目的喜好程度作为一系列组合问题。PMF 模型通过推导出两个满足高斯分布的项目和用户的低维潜在特征矩阵来降低计算复杂度, 在大规模数据集上具有较高的可扩展性和准确性。对于一个带高斯噪声的概率线性模型, 定义观测评分的条件分布为

$$p(R|U, V, \sigma^2) = \prod_{i=1}^N \prod_{j=1}^M [N(R_{ij}|U_i^T V_j, \sigma^2)]^{I_{ij}} \quad (1)$$

对于评分矩阵  $R$ ,  $M$  和  $N$  分别表示项目集合  $V$  的数量和用户集合  $U$  的数量, 矩阵分解得到的特征向量有  $K$  个特征值;  $N(x|\mu, \sigma^2)$  表示均值为  $\mu$ 、方差为  $\sigma^2$  的高斯分布的概率密度;  $I_{ij}$  是指示函数, 当用户  $i$  对于项目  $j$  有评分时为 1, 反之则为 0。优化最大化项目和用户特征的后验概率可以表示为

$$E = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^M I_{ij} (R_{ij} - U_i^T V_j)^2 + \frac{\lambda_U}{2} \sum_{i=1}^N \|U_i\|_{\text{Fro}}^2 + \frac{\lambda_V}{2} \sum_{j=1}^M \|V_j\|_{\text{Fro}}^2 \quad (2)$$

其中: Fro 表示斐波那契范数, 并且  $\lambda_U = \sigma^2 / \sigma_U^2$ ,  $\lambda_V = \sigma^2 / \sigma_V^2$ 。PMF 模型可以看做是奇异值分解(singular value decomposition, SVD)模型的概率扩展。如果所有的评分都被观测到, 式(2)就变成先验方差趋于无穷的 SVD。PMF 可以结合传统推荐算法应用于推荐系统中, 如协同过滤算法<sup>[9,10]</sup>。文献[11]在 TrustSVD 算法的基础上, 加入用户隐式信任信息, 提高推荐准确率。文献[12]结合用户、广告和网页三者的信息进行广告推荐, 利用联合概率矩阵分解解决冷启动问题。本文将利用 PMF 模型进行矩阵分解, 得到每一时间段的用户和项目特征矩阵, 对特征矩阵进行趋势预测。

### 1.2 时间序列模型

时序数据反映了用户行为在时间序列上的变化情况。对于推荐预测任务来说, 其核心就是推测目标用户或对象的未来行为趋势, 因此使用时序数据可以有效利用用户行为在时间序列上的趋势, 进而得到推荐预测。时序数据的趋势变化是按时间顺序的、海量的和潜在无限的。由于时序数据的这些特点, 如何准确快速地发现时序数据的变化成为研究的一个难点, 通常研究人员通过使用或者改进现有的时序模型来对数据进行处理。常用的时序模型有两种:

a) 自回归移动平均模型(auto regression moving average,

ARMA)。该模型由 AR 和 MA 两部分组成, AR 代表  $p$  阶自回归过程, MA 代表  $q$  阶移动平均过程, 常用于拟合平稳序列。公式如下:

$$z_t = \varphi_1 z_{t-1} + \varphi_2 z_{t-2} + \cdots + \varphi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \quad (3)$$

其中:  $z_t$  是时间序列中第  $t$  个元素;  $a_t$  是影响因子;  $\varphi$  和  $\theta$  是需要拟合的模型参数。

b) 差分自回归移动平均模型(auto regression integrated moving average, ARIMA)。ARIMA 是 Box 等人于 20 世纪 70 年代提出的一种著名时间序列预测方法。该方法在 ARMA 模型的基础上增加差分操作, 用来将不平稳的时间序列数据稳定在零点附近, 多数非平稳序列可以通过一次或多次差分后成为平稳序列。ARIMA 模型对于很多单变量的时序数据可以作出准确预测。文献[13]将 ARIMA 与 GARCH 结合, 提出一种可以获取服务质量属性波动的预测方法。文献[14]利用一个预测间隔内的交通流量与前几个间隔的交通流量的相关性, 结合 ARIMA 等方法提出一个新的贝叶斯组合方法(Bayesian combination method, BCM), 有效地提高单一预测因素的预测精度。文献[15]对比了 ARIMA、人工神经网络、灰色预测模型在预测流行色彩趋势中的效果, 实验发现单一的 ARIMA 方法并不能很好地预测色彩流行趋势。

本文将用户特征矩阵和项目特征矩阵中的每一单变量的时序数据代入时间序列模型, 求出该变量的下期值, 按序组合每一个单变量的预测值, 即求出未来时间段的用户和项目特征矩阵。

### 1.3 基于时序信息的推荐算法

基于时序信息的推荐算法是在推荐算法中利用时序信息提高推荐效率的一种方法。该算法通常将时序信息加入到常用的推荐模型中, 使模型可以学习到数据的动态变化, 解决了用户兴趣漂移的问题<sup>[16]</sup>。文献[17]提出的 TimeSVD++ 算法在用户(项目)的特征向量中加入时间信息, 提高了传统 SVD 方法的推荐效果。文献[7]建立用户—项目—时间三维张量, 利用张量分解的方式研究用户行为的动态变化。文献[18]通过将时间序列分段聚类, 将每个分类作为一个隐状态得到状态转移概率矩阵, 建立关于分段的隐马尔可夫模型, 对用户的未来行为状态进行预测。文献[19]提出的基于多维时间序列分析的推荐系统利用项目标签通过主题模型建模, 将每个项目表示成若干隐含主题构成的一个概率分布, 研究时间序列中概率分布的变化, 预测用户的行为。与上述相关工作不同, 本文通过将时序模型与矩阵分解相结合的方式, 研究时间序列中用户和项目特征向量的变化。该方法不仅解决了用户兴趣漂移的问题, 同时也起到了缓解数据稀疏性和数据降维的效果。

## 2 基于时序模型和矩阵分解的推荐算法

### 2.1 算法设计思想

本节将详细介绍本文所提出的基于时序模型和矩阵分解的推荐算法, 描述算法设计思想, 分析并推导出目标函数。

传统的概率矩阵分解模型的核心是学习用户和项目的特征向量, 根据特征向量预测未知项的评分。但是在本文场景中, 用户每个时间段的行为已知, 需要预测未来时间段中用户的行为, 而用户的行为在时间上往往是有规律的, 利用这些规律可以挖掘出用户潜在兴趣偏好的漂移情况。图 1 给出某用户的音乐收听行为的时序信息。数据显示该用户在四个时间段中对  $M_1 \sim M_4$  四首音乐的兴趣偏好存在以下规律: 对于  $M_1$

和  $M_4$  两首音乐的兴趣呈递减趋势,对于  $M_2$  和  $M_3$  两首音乐的兴趣呈递增趋势。因此在未来时间  $T_5$  中,为用户推荐  $M_2$  和  $M_3$  两首音乐更符合用户偏好。在实际数据集中,数据量通常远大于图 1 的例子,并且实际数据集通常是高度稀疏的,本文很难通过直接观察的方式分析出数据的变化趋势,因此利用原始数据进行数据预测的方法是不可行的。结合时序数据隐含行为规律的特点和概率矩阵分解在处理大规模数据集上的良好表现,本文采用时序模型与矩阵分解相结合的方法提高推荐效率。

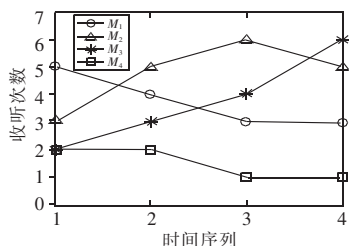


图 1 用户音乐收听次数的时序信息

给定时间序列  $T = \{t_1, t_2, \dots, t_p, \dots, t_P\}$ ,  $P$  表示时间序列的长度,本文方法包括以下两个步骤:

a) 对某一时间段  $t_p$  的音乐收听矩阵进行概率矩阵分解,得到当前时间段的用户特征矩阵  $U_p$  和项目特征矩阵  $V_p$ ,每个特征向量的维度为  $K$ ,令用户  $U$  和项目  $V$  的特征向量满足零均值的球形高斯先验:

$$p(U_p | \sigma_{U_p}^2) = \prod_{i=1}^N N(U_{ip} | 0, \sigma_{U_p}^2, I) \quad (4)$$

$$p(V_p | \sigma_{V_p}^2) = \prod_{j=1}^M N(V_{jp} | 0, \sigma_{V_p}^2, I) \quad (5)$$

b) 得到每一时间段的特征向量后,结合 ARIMA 模型得到在  $P$  时段中每个特征向量的时序模型,预测  $P+1$  时段的用户  $i$  的特征向量为  $U_{iP+1}$ ,  $P+1$  时段的项目  $j$  的特征向量为  $V_{jP+1}$ ,原特征向量序列通过一次差分拟合,根据特征向量的时序数据拟合 ARMA 模型参数,然后将参数代入 ARMA 模型预测下期值。

$$U_{iP+1} = \varphi_{i1} U_{iP} + \varphi_{i2} U_{iP-1} + \dots + \varphi_{ix} U_{iP-x} + a_{iP} - \theta_{i1} a_{iP-1} - \dots - \theta_{iy} a_{iP-y} \quad (6)$$

$$V_{jP+1} = \varphi_{j1} V_{jP} + \varphi_{j2} V_{jP-1} + \dots + \varphi_{jx} V_{jP-x} + a_{jP} - \theta_{j1} a_{jP-1} - \dots - \theta_{jy} a_{jP-y} \quad (7)$$

预测得到  $P+1$  时段的用户特征矩阵  $U_{P+1}$  和项目特征矩阵  $V_{P+1}$  后,通过矩阵分解的原始函数即可计算得到  $P+1$  时间段的预测收听数矩阵  $R_{P+1}$ 。对于某一用户  $U_i$ ,选择预测收听数最大的 TOP- $N$  首音乐为其进行推荐。

## 2.2 算法复杂度分析

本算法中主要的资源消耗包括两部分,第一部分是对每个时间段的收听矩阵进行概率矩阵分解。概率矩阵分解使用梯度下降法逼近近似值。对于  $T$  个时间段的收听矩阵的概率矩阵分解,时间复杂度为  $O(nT)$ ,  $n$  是收听矩阵中非零元素的个数。第二部分是对时序特征矩阵求解 ARIMA 模型参数,其中 ARMA 模型是线性的。因此本文算法的时间复杂度呈线性增长,算法适用于大规模数据。

## 3 实验结果及分析

### 3.1 实验数据集

本文提出的基于时序模型和矩阵分解的推荐算法主要是对用户的时序行为进行预测推荐。考虑到对算法性能和适应性的测试,实验主要围绕以下几个问题展开:

- a) 本文方法与对比方法对比推荐效果。
- b) 本文方法与对比方法对比模型训练时间。
- c) 不同特征向量维度  $K$  对本文方法的影响。
- d) 不同时间序列长度  $T$  对本文方法的影响。

为研究上述问题,本文从阿里音乐中抓取数据作为实验数据集。阿里音乐拥有数百万的曲库资源和每天千万的用户活跃度,有数亿人次的用户行为,并记录了行为发生的时间,因此适合本文实验研究,称之为 AlibabaMusic 数据集。该数据集包括 2015 年 3 月~8 月的部分歌曲艺人数据,以及与之相关的用户行为历史记录。首先本文对原始数据集进行预处理,将在统计时间内的活跃度小于 5 的用户筛选掉,保留活跃度较高的用户。为验证算法在不同时间序列长度  $T$  上的表现,按照用户行为时间对预处理后的数据集进行筛选,得到 AlibabaMusic-1 ( $T=30$ )、AlibabaMusic-2 ( $T=60$ ) 和 AlibabaMusic-3 ( $T=150$ )。数据情况如表 1 所示。

表 1 AlibabaMusic 数据集的各项统计信息

属性	AlibabaMusic	AlibabaMusic-1	AlibabaMusic-2	AlibabaMusic-3
用户数量	51 658	34 873	41 481	48 334
音乐数量	8 652	5 899	6 773	7 856
时间范围	03.01 ~ 08.31	07.01 ~ 07.31	06.01 ~ 07.31	03.01 ~ 07.31

### 3.2 算法及参数设定

本文选取了以下四种方法作为对比算法:

- a) 概率矩阵分解模型 (PMF)。文献 [8] 提出的概率矩阵分解模型,它没有考虑用户行为的时序信息。
- b) 差分自回归移动平均模型 (ARIMA)。Box 等人提出的时间序列预测方法,该方法直接对原始时序数据进行回归预测。
- c) 结合时序信息的张量分解算法 (BPTF)。文献 [7] 提出的一种结合时序信息的张量分解推荐算法。该算法在 User-Item 评分矩阵的基础上,将时间项作为第三维,对三维张量进行张量分解,提高传统二维矩阵分解的推荐效果。
- d) 针对时间序列聚类的隐马尔可夫模型 (CHMM)。文献 [18] 提出的一种基于聚类的隐马尔可夫模型。该模型假设时间序列是由一系列隐状态组成。时间序列整体来看就是在不同隐状态之间进行转换,通过阶段线性表示将时序数据分段,根据数据的时序性和相似性进行聚类作为隐状态,通过状态转移概率矩阵建立隐马尔可夫模型。该方法对于时间序列多步预测有较好的预测精度。

在实验中,为控制模型计算的复杂度,本文设定概率矩阵分解中最大迭代次数为 100,并且在迭代更新的过程中,设置自适应学习速率  $lRate = 0.001$ 。为了验证本文方法的特性,本文方法除了与对比算法进行对比实验外,还将通过实验说明不同的特征向量维度对方法的影响。选取特征数为  $K=5, K=10, K=20$ 。需要说明的是,ARIMA 和 CHMM 方法的推荐效果与  $K$  无关,只与时序长度  $T$  有关。不同时间序列长度  $T$  对本文方法的影响也会通过实验进行说明,具体的分组方式如表 1 所示。

### 3.3 评价指标

本文采用均方根误差和平均准确率对预测结果的有效性进行评估。RMSE 通过计算预测评分与实际评分之间的偏差来度量预测的准确性,是最常用的一种推荐效果评估方法。本文采用音乐收听数量作为用户对项目的评分。MAP 是预测准确率的平均值,模型预测的项目排位越靠前,MAP 的值就越大,它可以反映模型在全部相关项目上的预测性能。

设  $N$  个项目的预测评分为  $\{p_1, p_2, \dots, p_N\}$ , 实际评分为  $\{r_1, r_2, \dots, r_N\}$ , 则 RMSE 的计算可以表示为

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (p_i - r_i)^2}{N}} \quad (8)$$

设  $N$  个项目的预测排序为  $\{d_1, d_2, \dots, d_N\}$ , 实际排序为  $\{s_1, s_2, \dots, s_N\}$ , 则 MAP 的计算可以表示为

$$MAP = \frac{1}{N} \sum_{i=1}^N \frac{s_i}{d_i} \quad (9)$$

### 3.4 实验结果分析

#### 3.4.1 算法推荐效果对比

实验首先比较了本文算法和对比算法在相同条件下的推荐效果。在实验中,设特征向量维度  $K=5, 10, 20$ 。针对不同长度的时间序列,在 AlibabaMusic-1 ( $T=30$ )、AlibabaMusic-2 ( $T=60$ ) 和 AlibabaMusic-3 ( $T=150$ ) 上进行实验。算法的其他参数设置为使各算法最优时的相应值。

图2~7是不同特征向量维度和不同时序长度下,不同算法在 RMSE 和 MAP 上的对比结果。本文提出的方法用 PMF-TS 表示。

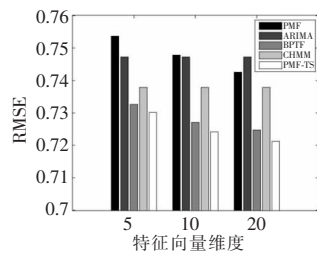


图2 AlibabaMusic-1 数据集的 RMSE 结果

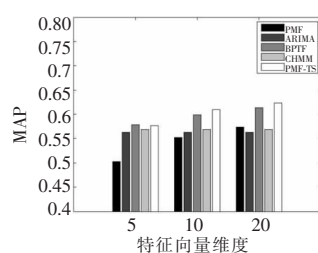


图3 AlibabaMusic-1 数据集的 MAP 结果

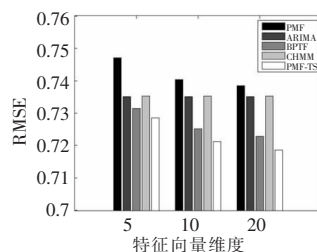


图4 AlibabaMusic-2 数据集的 RMSE 结果

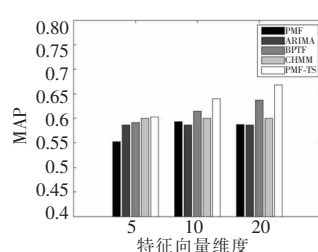


图5 AlibabaMusic-2 数据集的 MAP 结果

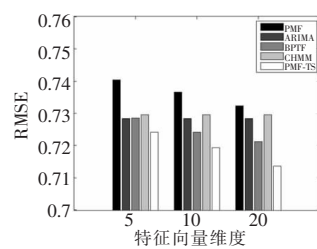


图6 AlibabaMusic-3 数据集的 RMSE 结果

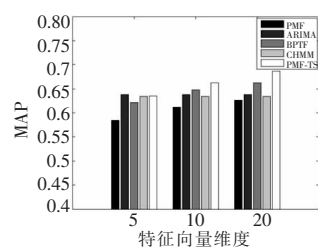


图7 AlibabaMusic-3 数据集的 MAP 结果

从实验结果可以看出:

a)除 ARIMA 和 CHMM 这两种不考虑特征向量维度  $K$  的方法之外,随着  $K$  的增加,其他算法的推荐精度都有不同程度的提高,但是  $K$  的增大会影响模型的运行时间。

b)随着时间序列长度  $T$  的增加,数据集的数据量变大,算法的推荐效果也随着数据量的增加而提高。

c)PMF 方法推荐效果最差,这是因为 PMF 方法并没有考虑时间因素的影响,用户的兴趣偏移降低了算法的推荐精度。ARIMA 方法的效果比 PMF 方法略好,这主要是因为考虑到了用户兴趣对时间的偏移,但是 ARIMA 方法受数据集稀疏性的影响较大,所以推荐效果受到了限制。

d)BPTF 和 CHMM 方法比 PMF 和 ARIMA 精度有所提高,

这是由于它们都考虑到了时间因素的影响,BPTF 方法利用张量分解模型,在数据量逐渐增大时可有效提高推荐列表的质量。

e)本文方法比 BPTF 和 CHMM 有了进一步的提高。由于本文方法在计算时只需要简单的用户行为时间信息,有效控制了计算复杂度,对大规模数据来说有更大的优势。

实验结果表明,结合时间信息的 BPTF、CHMM 和 PMF-TS 方法比基本的 PMF 方法有了很大的提高,本文方法相比另外几种方法在推荐精度和推荐列表的质量上效果更好,充分说明了时间因素对解决用户兴趣偏移的问题具有有效性和合理性,本文方法只需要比较容易获取的用户行为时间信息,因此算法的实际应用场景更广泛,对大规模数据的矩阵分解模型的扩展给出了新思路。

#### 3.4.2 算法运行时间对比

在 2.2 节中分析了本文方法的时间复杂度。在本实验中,表 2 给出了各种算法的运行时间。特征向量维度为  $K=10$ 。从表中可以看到,本文方法比基础的 PMF 和 ARIMA 方法耗时多,这主要是因为考虑了时间因素以及特征向量维度的影响。但总体来说,由于本文方法的时间复杂度是线性增长的,所以本文方法比 BPTF 和 CHMM 方法运行速度快;BPTF 方法由于使用张量分解方法,大大增加了时间复杂度;CHMM 方法使用分段聚类,耗时比本文方法多。此外考虑到随着时间序列长度的增加,用户数量和项目数量随之增加,BPTF 方法的耗时快速增长,而本文方法有效控制了算法运行时间,符合大规模数据场景下的计算要求。

表2 各种算法的运行时间比较

数据集	PMF	ARIMA	BPTF	CHMM	PMF-TS
AlibabaMusic-1	1.5	5.1	15.6	9.7	6.7
AlibabaMusic-2	1.6	6.4	17.8	11.3	8.3
AlibabaMusic-3	1.8	8.3	21.1	15.0	10.2

## 4 结束语

本文利用用户收听音乐行为的时序信息提出时序模型与矩阵分解相结合的推荐算法。首先对现有的基于矩阵分解和基于时序数据的推荐算法进行了分析总结,然后针对两类方法各自的缺陷提出改进方法。由于与标签、位置信息等相比用户行为时间信息更容易获取,并且概率矩阵分解对高维矩阵和矩阵稀疏性问题有良好的推荐精度,所以本文通过结合两类方法的优势提高推荐效率。

在真实阿里音乐数据集上的实验表明,本文方法比现有算法有更好的推荐效果,在 RMSE 和 MAP 数值指标上均有较好表现,并且可以有效控制算法的时间复杂度,因此更适合大数据环境下的个性化推荐任务。此外,本文提出的方法可以适用于多类矩阵分解和张量分解方法。在接下来的工作中,本文将利用社交网络关系等将该方法应用于张量分解算法中,对用户—项目—上下文的张量进行张量分解,结合其他数据信息进一步提高推荐效果。

### 参考文献:

- [1] 郑修猛,陈福才,柯丽虹. 基于信任和概率矩阵分解的协同推荐算法研究[J]. 计算机应用研究,2016,33(11):3240-3244.
- [2] Sharifi Z, Rezghi M, Nasiri M. A new algorithm for solving data sparsity problem based on non negative matrix factorization in recommender systems[C]//Proc of International Conference on Computer and Knowledge Engineering. Piscataway, NJ: IEEE Press, 2014:56-61.
- [3] Hu Yan, Peng Qimin, Hu Xiaohui, et al. Time aware and data sparsity tolerant Web service recommendation based on improved collaborative filtering[J]. IEEE Trans on Services Computing, 2015, 8(5): 782-794.



是异常轨迹,是因为 TOD-TIED 方法不仅仅除了能检测出形状异常的轨迹,还能够检测出形状正常但是其他特征表现异常的轨迹。如轨迹运动的路径正常,但是轨迹移动的速度异常的轨迹。由于轨迹信息熵距离计算采用加权多特征距离,是从多个角度判断轨迹间相似度,进而能够发掘出轨迹其他特征(如速度等特征)存在异常的隐藏的异常轨迹。

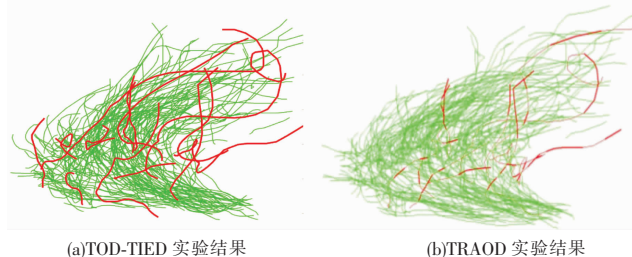


图6 TOD-TIED 与 TRAOD 的实验结果对比

另外,在算法参数选取上,本文提出的 TOD-TIED 方法利用异常数据判定的  $k\sigma$  准则选取异常轨迹阈值,在异常检测阶段无须设置参数,灵活性较高,并且能够全面地了解轨迹运动分布情况,而 TRAOD 需要提前进行相关参数设置,适用性低,且若参数设置不合理将会影响到异常检测效果。因此,本文提出的 TOD-TIED 要比 TRAOD 算法更具有现实意义。

#### 4 结束语

本文研究异常轨迹检测,针对异常轨迹多特征检测算法检测效率低的问题,提出了 TOD-TIED。该方法在轨迹段聚类阶段,轨迹段间距离度量应用了加权多特征距离,实验结果表明本文的轨迹段聚类方法发掘出的聚类更具有实际意义。在异常轨迹检测阶段,定义了轨迹信息熵以及采用  $k\sigma$  准则确定异常轨迹阈值对异常轨迹进行检测。检测结果表明,本文提出 TOD-TIED 算法提高了异常轨迹检测效果,是一种有效的异常轨迹检测算法。由于聚类阶段时间复杂度较高,所以,下一步工作是优化聚类过程,提高聚类的效率。

#### 参考文献:

- [1] 毛嘉莉,金澈清,章志刚,等. 轨迹大数据异常检测:研究进展及系统框架[J]. 软件学报,2017,28(1):17-34.
- [2] Yu Yanwei, Cao Lei, Rundensteiner E A, et al. Detecting moving object outliers in massive-scale trajectory streams[C]//Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014:422-431.
- [3] Liu Siyuan, Ni L M, Krishnan R. Fraud detection from taxis' driving behaviors[J]. IEEE Trans on Vehicular Technology, 2014, 63(1):464-472.
- [4] Chen Zaiben, Shen Hengtao, Zhou Xiaofang. Discovering popular routes from trajectories[C]//Proc of IEEE International Conference on Data Engineering. Washington DC: IEEE Computer Society, 2011: 900-911.
- [5] Costa G, Manco G, Masciari E. Dealing with trajectory streams by clustering and mathematical transforms[J]. Journal of Intelligent Information Systems, 2014, 42(1):155-177.
- [6] Deng Ze, Hu Yangyang, Zhu Mao, et al. A scalable and fast OPTICS for clustering trajectory big data[J]. Cluster Computing, 2015, 18(2):549-562.
- [7] Lee J G, Han Jiawei, Li Xiaolei. Trajectory outlier detection: a partition-and-detect framework[C]//Proc of IEEE International Conference on Data Engineering. Piscataway, NJ: IEEE Press, 2008:140-149.
- [8] 刘良旭, 乐嘉锦, 乔少杰, 等. 基于轨迹点局部异常度的异常点检测算法[J]. 计算机学报, 2011, 34(10):1966-1975.
- [9] 韩博洋, 汪兆洋, 金蓓弘. 一种基于轨迹大数据离线挖掘与在线实时监测的出租车异常轨迹检测算法[J]. 中国科学技术大学学报, 2016, 46(3):247-252.
- [10] Lee J G, Han Jiawei, Whang K Y. Trajectory clustering: a partition-and-group framework[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2007:593-604.
- [11] 唐梦梦, 吉根林, 赵斌. 利用 MapReduce 的异常轨迹检测并行算法[J]. 地球信息科学学报, 2015, 17(5):523-530.
- [12] Chang Cheng, Zhou Baoyao. Multi-granularity visualization of trajectory clusters using sub-trajectory clustering[C]//Proc of IEEE International Conference on Data Mining Workshops. Washington DC: IEEE Computer Society, 2009:577-582.
- [13] 袁冠, 夏士雄, 张磊, 等. 基于结构相似度的轨迹聚类算法[J]. 通信学报, 2011, 32(9):103-110.
- [14] Jiang Shengyi, Li Qinghua, Li Kenli, et al. GLOF: a new approach for mining local outlier[C]//Proc of International Conference on Machine Learning and Cybernetics. Piscataway, NJ: IEEE Press, 2003: 157-162.
- [15] Kafsi M, Grossglauser M, Thiran P. Traveling salesman in reverse: conditional Markov entropy for trajectory segmentation[C]//Proc of IEEE International Conference on Data Mining. Piscataway, NJ: IEEE Press, 2015:201-210.
- [16] 田尧, 秦永彬, 许道云, 等. 基于双信任机制的 TrustSVD 算法[J]. 计算机科学与探索, 2015, 9(11):1391-1397.
- [17] 涂丹丹, 舒承椿, 余海燕. 基于联合概率矩阵分解的上下文广告推荐算法[J]. 软件学报, 2013, 24(3):454-464.
- [18] Amin A, Colman A, Grunke L. An approach to forecasting QoS attributes of Web services based on ARIMA and GARCH models[C]//Proc of IEEE International Conference on Web Services. Washington DC: IEEE Computer Society, 2012:74-81.
- [19] Wang Jian, Deng Wei, Guo Yuntao. New Bayesian combination method for short-term traffic flow forecasting[J]. Transportation Research Part C: Emerging Technologies, 2014, 43(6):79-94.
- [20] Yu Yong, Hui C L, Choi T M. An empirical study of intelligent expert systems on forecasting of fashion color trend[J]. Expert Systems with Applications, 2012, 39(4):4383-4389.
- [21] 孙光福, 吴乐, 刘洪, 等. 基于时序行为的协同过滤推荐算法[J]. 软件学报, 2013, 24(11):2721-2733.
- [22] Koren Y. Collaborative filtering with temporal dynamics[C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009:89-97.
- [23] 章登义, 欧阳黥霏, 吴文李. 针对时间序列多步预测的聚类隐马尔可夫模型[J]. 电子学报, 2014, 42(12):2359-2364.
- [24] 王守涛. 一种基于多维时间序列分析的音乐推荐系统研究与实现[D]. 南京: 南京大学, 2014.
- [25] Tu Shitao, Zhu Lanjuan. A bandit method using probabilistic matrix factorization in recommendation[J]. Journal of Shanghai Jiaotong University: Science, 2015, 20(5):535-539.
- [26] Song Yang, Zhuang Ziming, Li Huajing, et al. Real-time automatic tag recommendation[C]//Proc of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2008:515-522.
- [27] 王守辉, 于洪涛, 黄瑞阳, 等. 基于模体演化的时序链路预测方法[J]. 自动化学报, 2016, 42(5):735-745.
- [28] Xiong Liang, Chen Xi, Huang T K, et al. Temporal collaborative filtering with Bayesian probabilistic tensor factorization[C]//Proc of SIAM International Conference on Data Mining. 2010:211-222.
- [29] Salakhutdinov B R, Mnih A. Probabilistic matrix factorization[C]//Proc of the 20th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2015: 1257-1264.
- [30] Bokde D, Girase S, Mukhopadhyay D. Matrix factorization model in collaborative filtering algorithms: a survey[J]. Procedia Computer Science, 2015, 49(1):136-146.
- [31] Li Gai, Ou Weihua. Pairwise probabilistic matrix factorization for implicit feedback collaborative filtering[J]. Neurocomputing, 2016, 204(9):17-25.