计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190500130

基于极端评分行为的相似度计算

冯晨娇[1,2]　梁吉业[1]　宋　鹏[3]　王智强[1]
1 山西大学计算智能与中文信息处理教育部重点实验室　太原 030006
2 山西财经大学应用数学学院　太原 030006
3 山西大学经济与管理学院　太原 030006
(fengcj@sxufe.edu.cn)

摘　要　随着互联网技术的迅猛发展,互联网信息急剧增长,信息过载问题愈发凸显。面对海量的互联网信息,用户往往需要耗费大量的时间来搜索所需的信息或产品,而搜索的解往往受到制约。为解决信息过载问题,推荐系统应运而生。推荐系统根据用户的历史行为推测其需求、兴趣等,将用户感兴趣的信息、产品等推荐给用户。作为推荐领域中一类重要的推荐方法,基于记忆的协同过滤方法通常依据用户或产品的近邻信息来构造评分预测函数,其核心在于准确度量用户或产品之间的相似度。传统的相似度量,如皮尔逊、余弦及秩相关系数等,通常只考虑了用户之间的线性关系;而启发式相似度如基于3个特殊因子的PIP相似度及其改进方法,则只刻画了用户之间的非线性关系。事实上,在推荐系统中,就用户之间的相似关系而言,仅用线性或是非线性函数来度量均是不准确的。为了更为精细地刻画用户之间的相似程度,文中提出了基于非线性函数的用户极端评分行为的相似程度度量指数,通过将该指数融入传统的线性相关系数,构造了一个考虑极端评分行为的新的相似度。为验证该方法的有效性,基于Ml(100k)和Ml-latest-small两个数据集,将其与传统相似度以及启发式相似度进行比较,结果显示基于极端评分行为相似度的协同过滤方法在MAE和RMSE指标上能够获得更好的表现。

关键词:推荐系统;协同过滤;基于记忆的协同过滤;极端评分行为;相似度
中图法分类号　TP182

New Similarity Measure Based on Extremely Rating Behavior

FENG Chen-jiao[1,2],LIANG Ji-ye[1],SONG Peng[3] and WANG Zhi-qiang[1]
1 Key Laboratory of Computation Intelligence & Chinese Information Processing (Shanxi University),Ministry of Education,Taiyuan 030006,China
2 College of Applied Mathematics,Shanxi University of Finance and Economics,Taiyuan 030006,China
3 School of Economics and Management,Shanxi University,Taiyuan 030006,China

Abstract　With the rapid development of Internet technology,drastic Internet information explosion makes information overload as an increasingly serious problem. Faced with the massive Internet information,users consume a lot of time to search for information or products,but the search solution is constrained. The recommender systems is hence proposed to address the problem of information overload. The recommender systems use users' historical behaviors to speculate their needs,interests,etc.,and recommend the information and products users may be interested in. As an important type of recommendation approach,the memory-based collaborative filtering methods establish the rating prediction function based on neighbor information of the user or product. The essence of the function is to precisely measure the similarity between users or products. The traditional similarity measures such as Pearson,Cosin and Spearman rank correlation coefficients,only take into account the linear relationship between users,while the heuristic similarities,such as the PIP measurement based on three special factors and its improved version,only depict the non-liner relationship between users. Indeed,in the recommender systems,it is neither the linear relation nor the non-linear relation is good for measuring the similarity between users. In order to describe the similarity among users more finely,this

到稿日期:2019-05-23　返修日期:2019-08-12　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61876103);山西省重点研发计划重点项目(201603D111014);山西省回国留学人员科研资助项目(2017-005);山西省1331工程项目
This work was supported by the National Natural Science Foundation of China (61876103),Projects of Key Research and Development Plan of Shanxi Province(201603D111014),Research Project Supported by Shanxi Scholarship Council of China (2017-005) and 1331 Engineering Project of Shanxi Province, China.
通信作者:梁吉业(ljy@sxu.edu.cn)

paper proposed a similarity measure index of the correlation level considering the extreme rating behaviors based on a nonli-near function. By integrating this index with the traditional linear correlation coefficients, this paper constructed a novel similarity measure. Comparative experiments were conducted to test the practicability and validity of the proposed approach on Ml(100k) and Ml-latest-small datasets. The results demonstrate that the proposed method performs better judged by indicators of MAE and RMSE.

Keywords Recommender systems, Collaborative filtering, Memory-based collaborative filtering, Extremely rating behavior, Similarity

1 引言

互联网技术的迅猛发展，特别是移动终端（如移动电话、平板等）的广泛使用，使得互联网信息急剧增长，信息过载问题愈发凸显。面对海量的互联网信息，用户往往需要耗费大量的时间来寻找所需的信息或产品，最终还可能一无所获。为解决信息过载问题，推荐系统应运而生[1-6]。推荐系统是用于处理信息过载问题并根据历史信息为用户给出合理推荐的一种智能系统。近年来，推荐系统被广泛应用于各个领域，如亚马逊书籍推荐系统[1)]、Netflix 电影推荐系统[2)]、Jester 的笑话推荐系统[3)]和 Facebook 朋友推荐系统[4)]等。随着这些应用系统的有效推广，协同过滤方法作为推荐系统的一类典型技术，受到了学术界和工业界的重点关注[7-12]。

作为协同过滤中的一类重要推荐方法，基于记忆的协同过滤方法[13-18]（又称基于近邻的协同过滤方法）通过使用目标用户或产品的近邻相关信息来构造预测函数，以实现目标用户对未知产品的评分预测。其基本假设是用户的历史偏好具有延续性，即未来目标用户的偏好将会持续不变。因此，在基于近邻的协同过滤方法中，目标用户对未知产品的偏好预测通常是基于目标用户的近邻用户对该产品的评分进行加权平均。具体地，给定用户集合 $U=\{u_1,u_2,\cdots,u_m\}$、产品集合 $V=\{v_1,v_2,\cdots,v_n\}$ 以及用户和产品之间的评分矩阵 $\boldsymbol{R}=(r_{ij})_{m\times n}$。其中，$u_i$ 代表第 i 个用户；v_j 代表第 j 个产品；r_{ij} 为第 i 个用户对第 j 个产品的打分，通常采用 5 分制，1 分代表最弱的偏好，5 分代表最强的偏好。在基于记忆的协同过滤方法中，目标用户对未知项目的预测是通过输入评分矩阵 $\boldsymbol{R}$ 来构造相似度，并以其为准绳寻找与目标用户有共同偏好的近邻，然后将这些近邻对该项目的评分的加权平均作为预测结果。通常，相似度在这里又充当权重。显然，系统中任意两个用户的相似度扮演了选择近邻的依据和预测函数的权重两种角色。因此，如何准确度量用户或产品之间的相似度，是基于近邻协同过滤方法中最为关键的问题。

在实际的评分数据中，用户-产品之间的评分矩阵往往表现出显著的稀疏性，与之对应，用户间存在共同评分的产品数量往往较少，而统计学中样本量的不足将导致传统的相似度度量方法（如余弦相似度（Cosine Similarity，COS）[19]和皮尔逊相关系数（Pearson Correlation Coefficient，PCC）[20]等）难以准确计算用户或产品之间的相似度，进而影响最终的评分预测精度。为改进稀疏环境下用户/产品之间的相似度度量，许多相似度方法被提出，包括：改进传统的相似度，如将评分区分为正类和负类的被限制的皮尔逊相关系数（Constrained Pearson Correlation Coefficient，CPCC）[21]；考虑了评分项目自身重要性的频率权重皮尔逊相似度（Frequency-Weighted Pearson Correlation Coefficient，WFPCC）[13]；将评分排序后按序关系计算的秩相关系数（Spearman Rank Correlation Coefficient，SRCC）[22]等。在启发式相似度中，有考虑领域知识而构造的 3 个因素乘积的 PIP 相似度[23]；用 3 个连续函数表示这 3 个因素，同时考虑用户均值和方差的 NHSM 相似度[24]等。这些方法虽然可以一定程度上提高用户/产品之间相似度的度量准确性，但是仍然只考虑了单一的关系。例如，在传统相似度的改进方法中，CPCC，WFPCC，SRCC 等方法仍然仅考虑评分系统中用户的线性关系；而启发式相似度也只考虑了用户之间的非线性关系。事实上，在推荐系统中，就用户之间的相似关系而言，仅用线性或是非线性函数来度量均是不准确的。为准确建模用户间的复杂关系，本文提出了一种利用领域知识构造相似度的方法，该方法的思想主要体现在以下两个方面：

（1）用户之间共同的强烈偏好行为将加强用户之间的相似性，用非线性函数来刻画用户之间共同的极端评分行为的相似程度；

（2）将极端评分行为的相似程度作为权重融入传统的相关系数中，从而更加精细地度量用户之间的相关关系，有助于区分传统相似度无法区分的情况，能更全面地描述用户之间的关系。

本文第 2 节介绍了相关工作；第 3 节详细描述了考虑极端评分行为的相似度构造方法；第 4 节给出了实验与结果分析；最后总结全文并展望未来。

2 相关工作

用户/产品之间的相似度计算是基于近邻协同过滤方法的核心，已有基于近邻协同过滤的方法也主要集中于如何构造用户/产品之间的相似度，其主要包括针对传统相似度的改进方法和启发式的相似度方法。

在针对传统相似度的改进方面，Shardanand 等[21]提出了 CPCC，该方法用系统中值代替用户均值来中心化数据，能够将评分分为正类和负类，有利于数据的整体比较。Breese

1) https://www.amazon.com

2) https://www.netflix.com

3) https://www.thisisjester.com

4) https://www.facebook.com

等[13]为产品增加了权重因子，称其为频率权重皮尔逊相关系数 WFPCC，即对该产品打分的人数越多，认为该产品在度量两个用户相似度上的贡献就越小。Shardanand 等[21]提出了均方差(Mean Squared Difference，MSD)度量，其直接用共同评分差值的平方和的均值的倒数来计算，该度量简单直观，但误差较大。Kendall 等[22]提出了秩相关系数 SRCC，该系数基于评分排序来计算相似度，其优点在于不需要对数据进行标准化，但排序后有大量的共同顺序项，使得秩相关系数不准确。表 1 列出了常见的几种具有代表性的传统相似性度量及其改进方法。

表 1 传统相似性度量及其改进方法

Table 1 Traditional similarities and improved metrics

相似性度量	定义
PCC	$\rho^{PCC}_{(u_i,u_k)}=\frac{\sum_{j\in I_{ik}}(r_{ij}-\bar{r}_{u_i})(r_{kj}-\bar{r}_{u_k})}{\sqrt{\sum_{j\in I_{ik}}(r_{ij}-\bar{r}_{u_i})^2}\sqrt{\sum_{j\in I_{ik}}(r_{kj}-\bar{r}_{u_k})^2}}$
COS	$\rho^{COS}_{(u_i,u_k)}=\frac{\sum_{j\in I_{ik}} r_{ij}\cdot r_{kj}}{\sqrt{\sum_{j\in I_{ik}} r_{ij}^2}\sqrt{\sum_{j\in I_{ik}} r_{kj}^2}}$
CPCC	$\rho^{CPCC}_{(u_i,u_k)}=\frac{\sum_{j\in I_{ik}}(r_{ij}-\bar{r}_{\mathrm{med}})(r_{kj}-\bar{r}_{\mathrm{med}})}{\sqrt{\sum_{j\in I_{ik}}(r_{ij}-\bar{r}_{\mathrm{med}})^2}\sqrt{\sum_{j\in I_{ik}}(r_{kj}-\bar{r}_{\mathrm{med}})^2}}$
FWPCC	$\rho^{FWPCC}_{(u_i,u_k)}=\frac{\sum_{j\in I_{ik}}\lambda_i(r_{ij}-\bar{r}_{u_i})(r_{kj}-\bar{r}_{u_k})}{\sqrt{\sum_{j\in I_{ik}}\lambda_i(r_{ij}-\bar{r}_{u_i})^2}\sqrt{\sum_{j\in I_{ik}}\lambda_i(r_{kj}-\bar{r}_{u_k})^2}}$
MSD	$\rho^{MSD}_{(u_i,u_k)}=\frac{\lvert I_{ik}\rvert}{\sum_{j\in I_{ik}}(r_{ij}-r_{kj})^2}$
SRC	$\rho^{SRC}_{(u_i,u_k)}=\frac{\sum_{j\in I_{ik}}(h_{ij}-\bar{h}_{u_i})(h_{kj}-\bar{h}_{u_k})}{\sqrt{\sum_{j\in I_{ik}}(h_{ij}-\bar{h}_{u_i})^2}\sqrt{\sum_{j\in I_{ik}}(h_{kj}-\bar{h}_{u_k})^2}}$

表 1 中，$\bar{r}_{\mathrm{med}}$，$\bar{r}_{u_i}$，$\bar{r}_{u_k}$ 分别表示系统中值、用户 u_i 的所有评分项目的评分均值及用户 u_k 的所有评分项目的评分均值；λ_i 表示项目 v_j 的重要度；h_{ij} 表示用户 u_i 在项目 v_j 上的排序，$\bar{h}_{u_i}$ 表示用户 u_i 在所有给予评分的项目上的平均排序；I_i 表示用户 u_i 评分的项目集合，I_{ik} 表示用户 u_i 和用户 u_k 的共同评分项目集合。

在启发式相似度度量方面，Ahn 等[23]首次提出了启发式的相关系数，该方法在相似度的度量中添加了刻画用户行为的度量公式。文献[23]指出，两个用户对产品都给出 5 分或 1 分的相似性比用户都给出 3 分的相似性高，且对于用户自身来说，给出产品的极端评分比给出这个产品的大众评分更有意义。该文分别定义了两个用户共同评分的邻近函数(Proximity)、影响函数(Impact)及大众函数(Popularity)，并将其乘积之和作为两个用户的相似度，记为 PIP。

Liu 等[24]在此基础上提出了一种新的启发式相似度(NHSM)，该度量将 3 个函数改进为[0，1]范围内的连续取值，同时考虑了用户均值和方差的影响。值得注意的是，因为用户对之间的共同评分数目大部分是不相等的，所以上述相似度计算通常基于不同的样本容量；然而，在统计学上，不同的样本容量计算出的度量是不具有可比性的。为解决这一问题，一些基于权重的相似度被提出。直观上，两个用户共同评分的项目越多，相似度就越可靠；对应于统计学中的样本统计量估计，样本容量越大，估计值就越可靠。因此，当用户共同评分项目不同时，应该采用不同的惩罚函数。表 2 列出了当前具有代表性的权重函数。

表 2 具有代表性的权重函数

Table 2 Representative weight functions

权重	定义
①	$w^{\gamma}_{(u_i,u_k)}=\frac{\min\{\lvert I_{ik}\rvert,\gamma\}}{\gamma}$
②	$w^{\beta}_{(u_i,u_k)}=\frac{\lvert I_{ik}\rvert}{\lvert I_{ik}\rvert+\beta}$
③	$w^{J}_{(u_i,u_k)}=\frac{\lvert I_{ik}\rvert}{\lvert I_i\cup I_k\rvert}$
④	$w^{S}_{(u_i,u_k)}=\frac{1}{1+\exp(-\frac{\lvert I_{ik}\rvert}{2})}$

表 2 列出了 4 种权重函数，其中，$\lvert I_i\rvert$ 表示这个集合中元素的个数，即用户 u_i 评分的项目数；$\lvert I_{ik}\rvert$ 为用户 u_i 和用户 u_k 的共同评分项目数，β，γ 是参数。这 4 种函数都是关于 $\lvert I_{ik}\rvert$ 的函数值在[0，1]之间的增函数，其中函数①是连续函数[25]；函数②是连续光滑函数[26]；函数③是杰卡德相似性度量[3]，经常用于比较集合之间的相似性与差异性；函数④是 Sigmoid 函数[26]，是一种将变量映射到[0，1]之间的 S 型函数。将表 2 中的权重函数和表 1 中的相似度相结合，产生了一些新的相似度，如权重皮尔逊相关系数(the Weighted Pearson Correlation Coefficient，WPCC)、基于 Sigmoid 函数的皮尔逊相关系数(Sigmoid Function Based Pearson Correlation Coefficient，SPC)、基于杰卡德的均方差相似度(Jaccard Function Based Mean Squared Difference，JMSD)等。

3 极端评分行为的相似度

本节首先给出极端评分行为的相似程度度量指数，然后采用权重方式将其融合到传统的相关系数中，以期更为全面地刻画两个用户之间的相关关系。

正如文献[23]提出的启发式相似度一样，用户的极端评分行为能够体现用户之间的关系。本文除了给出文献[23]中提出的两个用户对项目给出的极端评分和用户对项目给出远超于大众评分的两种情况以外，添加了用户对项目评分远超于个人习惯性评分的情况，具体定义如下：

$$S^1_{(u_{ij},u_{kj})}=\frac{1}{1+\exp(-\lvert r_{ij}-\bar{r}_{\mathrm{med}}\rvert\lvert r_{kj}-\bar{r}_{\mathrm{med}}\rvert)}$$

$$S^2_{(u_{ij},u_{kj})}=\frac{1}{1+\exp(-\lvert r_{ij}-\bar{r}_{v_j}\rvert\lvert r_{kj}-\bar{r}_{v_j}\rvert)}$$

$$S^3_{(u_{ij},u_{kj})}=\frac{1}{1+\exp(-\lvert r_{ij}-\bar{r}_{u_i}\rvert\lvert r_{kj}-\bar{r}_{u_k}\rvert)}$$

其中，$\bar{r}_{v_j}$ 表示项目 v_j 所有评分的平均值；$S^1_{(u_{ij},u_{kj})}$ 表示用户 u_i 和用户 u_k 对系统中值的共同极端评分行为的度量；同理，

$S^2_{(u_{ij},u_{kj})}$ 和 $S^3_{(u_{ij},u_{kj})}$ 分别表示用户 u_i 和用户 u_k 远离大众评分和习惯性评分的度量。采用相同的形式表示 3 个函数是为了体现 3 种极端行为都很重要，没有主次之分。利用这 3 个函数定义用户 u_i 和用户 u_k 之间对项目 v_j 的极端行为的度量，即：

$$S_{(u_{ij},u_{kj})}=S^1_{(u_{ij},u_{kj})} * S^2_{(u_{ij},u_{kj})} * S^3_{(u_{ij},u_{kj})}$$

$S_{(u_{ij},u_{kj})}$ 是一个[0,1]之间的连续函数，具有良好的数学性质。同时，采用将上述 3 个函数相乘的形式是为了进一步加强极端行为和常规行为的区分。通过上述分析，本文构造了一个新的基于极端评分行为的相似度，记为 ρ^{ERB}_{ik}。

$$\rho^{\mathrm{ERB}}_{ik}=\frac{\sum_{j\in I_{ik}} S_{(u_{ij},u_{kj})} * (r_{ij}-\bar{r}_{\mathrm{med}})(r_{kj}-\bar{r}_{\mathrm{med}})}{\sqrt{\sum_{j\in I_{ik}} S_{(u_{ij},u_{kj})} * (r_{ij}-\bar{r}_{\mathrm{med}})^2}\sqrt{\sum_{j\in I_{ik}} S_{(u_{ij},u_{kj})} * (r_{kj}-\bar{r}_{\mathrm{med}})^2}}$$

相应地，结合表 2 中的任一权重函数可得到最终的用户 u_i 和用户 u_k 的相似度表达式：

$$\rho_{ik}=w^J_{(u_i,u_k)} * \rho^{\mathrm{ERB}}_{ik}$$

从表 2 所列的 4 个权重函数来看，杰卡德权重函数无需参数估计，且函数表达简洁，因此本文选取该函数进行实验分析。

4 实验与结果分析

本文选择了美国明尼苏达大学 GroupLens 研究项目中的两个数据集 Ml(100k)和 Ml-latest- small，这两个数据集的评分都是以 5 分制给出，即最高为 5 分，最低为 1 分。其中，Ml(100k)中有 943 个用户对 1 680 个项目给出的 100 000 个评分，密度为 6.3%；而 Ml-latest-small 中有 671 个用户对 9 125 个项目给出的 100 004 个评分，密度为 1.6%。需要强调的是，这两个数据集的评分间隔不同，在 Ml(100k)中以 1 为间隔，而在 Ml-latest-small 中以 0.5 为间隔。

4.1 评价指标

本文选择了平均绝对误差(Mean Absolute Error，MAE)和均方根误差(Root Mean Squared Error，RMSE)作为评价指标，并采用五折交叉验证方法来评价基于新的相似度的推荐精度。具体公式如下：

$$MAE=\frac{\sum_{r_{ij}\in R_{\mathrm{test}}} |r_{ij}-\hat{r}_{ij}|}{|R_{\mathrm{test}}|}$$

$$RMSE=\sqrt{\frac{1}{|R_{\mathrm{test}}|}\sum_{r_{ij}\in R_{\mathrm{test}}} (r_{ij}-\hat{r}_{ij})^2}$$

其中，R_{test} 表示五折交叉验证法中随机选择出的一折测试集，$|R_{\mathrm{test}}|$ 表示该集合中存在的评分数目。

4.2 实验设计及结果分析

实验分为 3 个部分。第一部分是为了验证本文提出的相似度比其他的相似度有更好的推荐精度，通过比较 6 种不同相似度的基于用户的协同过滤方法的 *MAE* 和 *RMSE* 来进行验证。第二部分是随机删除数据集中的一些评分，让其变得更加稀疏，仍然通过上述方法进行比较。第三部分是通过 4 张曲线图说明近邻参数对预测结果的影响。

首先，将本文提出的相似度与传统相似度(PCC，COS，CPCC，JMSD)和启发式相似度(PIP，NHSM)作为寻找近邻的度量用在基于用户的协同过滤方法中，通过预测精度进行比较分析。为了方便，将这些协同方法分别记为 UB-PCC，UB-COS，UB-CPCC，UB-JMSD，UB-PIP，UB-NHSM；将本文提出的极端评分行为相似度的协同过滤方法记为 UB-ERB。表 3 列出了上述协同过滤方法的 *MAE* 和 *RMSE* 结果。

表 3 基于不同相似度的推荐精度的比较

Table 3 Comparison of recommender accuracy with different similarities

方法	Ml(100k)		Ml-latest-small	
	MAE±*std*	*RMSE*±*std*	*MAE*±*std*	*RMSE*±*std*
UB-PCC	0.737 0±0.005 0**	0.942 3±0.006 4	0.692 1±0.005 8**	0.908 0±0.005 8
UB-COS	0.742 6±0.005 8	0.947 7±0.007 1	0.697 3±0.007 0	0.912 2±0.006 5
UB-CPCC	0.737 9±0.005 5	0.939 2±0.007 0**	0.703 6±0.001 5	0.902 8±0.006 1**
UB-JMSD	0.798 2±0.003 7	1.021 6±0.005 5	0.735 1±0.007 4	0.958 6±0.007 6
UB-PIP	0.747 0±0.004 1	0.958 8±0.004 1	0.704 4±0.005 8	0.923 4±0.005 4
UB-NHSM	0.756 9±0.003 8	0.967 8±0.004 3	0.707 7±0.006 6	0.926 8±0.006 5
UB-ERB	0.736 0±0.005 3*	0.936 6±0.006 7*	0.689 8±0.006 6*	0.901 9±0.006 6*

注：*std* 表示五折交叉验证法中的标准差，* 表示该值最优，** 表示该值次优

从表 3 可以看出，本文方法在 *MAE* 和 *RMSE* 指标中都是误差最小的；在这两个数据集中，UB-PCC 在 *MAE* 上、UB-CPCC 在 *RMSE* 上有较好的预测精度；而 UB-PIP 和 UB-NHSM 相对差一些，说明在这两个数据集中用户之间更多是线性关系。但是，我们加入极端行为权重后对精度有所改进，说明极端评分行为在用户之间的相似程度度量上是有贡献的。

其次，随机删除数据集中 20%，40%，60%，80%的评分，以将数据稀疏化，得到的稀疏数据集记为 Sub1，Sub2，Sub3，Sub4。通过在同一个数据集的不同稀疏度下对各种相似度方法进行比较，来验证本文方法具有一定的稳定性。从表 4 中可以看出，在数据集 Ml(100k)中，本文方法在不同的稀疏度下保持前两名，但是其他方法的波动较大。比如：UB-CPCC 方法在子集 Sub1 和 Sub2 中排名第二，但是在子集 Sub3 和 Sub4 中成为了第三名；UB-COS 方法在子集 Sub3 和 Sub4 中排名第一，但是在 Sub1 中排名第四，在 Sub2 中排名第三。从表 5 中可以看出，本文方法和 UB-CPCC 保持了前两名。这两种方法在这个数据集上都具有一定的稳定性，原因可能是该数据集以 0.5 波动，其波动幅度较小，极端评分不突出。但总体而言，本文方法在两个数据集上均表现出了较好的稳定性。

表 4 在 Ml(100k)数据集的不同稀疏度下推荐精度的比较

Table 4 Comparison of recommender accuracy with different densities on Ml(100k)

方法	Sub1		Sub2		Sub3		Sub4	
	MAE	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*
UB-PCC	0.7412	0.9462	0.7637	0.9745	0.7880	1.0053	0.8926	1.1408
UB-COS	0.7447	0.9490	0.7604	0.9680	0.7742*	0.9852*	0.8521*	1.0855*
UB-CPCC	0.7405**	0.9418**	0.7600**	0.9658**	0.7760	0.9871	0.8608	1.1013
UB-JMSD	0.7990	1.0222	0.8125	1.036	0.8193	1.0432	0.8874	1.1338
UB-PIP	0.7463	0.9569	0.7624	0.9747	0.7788	0.9884	0.8633	1.1293
UB-NHSM	0.7566	0.9655	0.7715	0.9832	0.7823	0.9985	0.8885	1.1337
UB-ERB	0.7380*	0.9403*	0.7578*	0.9651*	0.7754**	0.9870**	0.8603**	1.1010**

注：* 表示该值最优，** 表示该值次优

表 5 在 Ml-latest- small 数据集的不同稀疏度下推荐精度的比较

Table 5 Comparison of recommender accuracy with different densities on Ml-latest-small

方法	Sub1		Sub2		Sub3		Sub4	
	MAE	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*
UB-PCC	0.7034	0.9200	0.7267	0.9523	0.7604	0.9893	0.8143	1.0608
UB-COS	0.7054	0.9269	0.7211	0.9455	0.7465	0.9732	0.8060	1.0516
UB-CPCC	0.7009**	0.9203**	0.7128*	0.9408*	0.7414*	0.9658*	0.8019*	1.0458*
UB-JMSD	0.7461	0.9765	0.7572	0.9889	0.7682	0.9977	0.8094	1.0559
UB-PIP	0.7109	0.9355	0.7245	0.9511	0.7463	0.9752	0.8028	1.0472
UB-NHSM	0.7140	0.9381	0.7281	0.9548	0.7496	0.9795	0.8089	1.0564
UB-ERB	0.7002*	0.9200*	0.7185**	0.9412**	0.7420**	0.9669**	0.8027**	1.0470**

注：* 表示该值最优，** 表示该值次优

最后，对参数进行分析。在近邻推荐中，目标用户的近邻个数 k 是一个重要的参数。本文从 10 到 100，以 10 为间隔进行了不同近邻个数的实验，实验结果如图 1 和图 2 所示。从图中可以看出，当 $k=20$ 时，预测结果最好，这似乎违背了人们的常规认识。通常，我们认为近邻越多越好，但是这里的数据集是稀疏的，也就是说目标用户具有多个近邻，但是这些近邻未必对目标用户期望预测的项目有评分。此时，无评分的近邻对于评分预测而言并没有相应的贡献。

本文中所有的算法都是基于“近邻对期望评分的项目有评分”的假设。例如，当 $k=100$ 时，目标用户的 100 个近邻中只有少数的近邻用户对想要预测的项目有评分，因此实际参与运算的近邻并没有 100 个，并且这样会导致不同的目标用户实际参与运算的近邻个数是不一样的，导致最终的预测效果不佳。

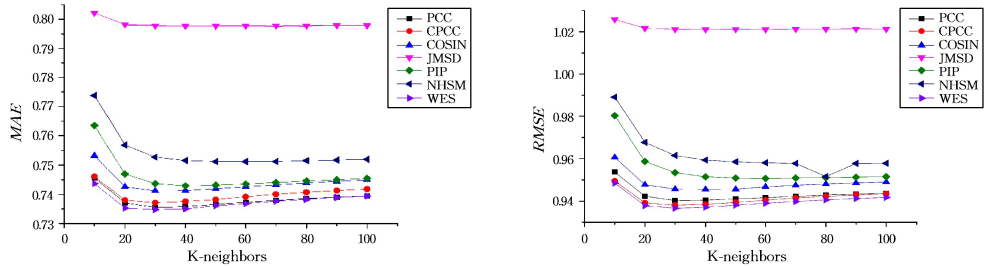

图 1 在 Ml(100k)数据集上近邻参数 k 对 *MAE* 和 *RMSE* 的影响

Fig. 1 Impact of neighbor parameter k on *MAE* and *RMSE* on Ml(100k)

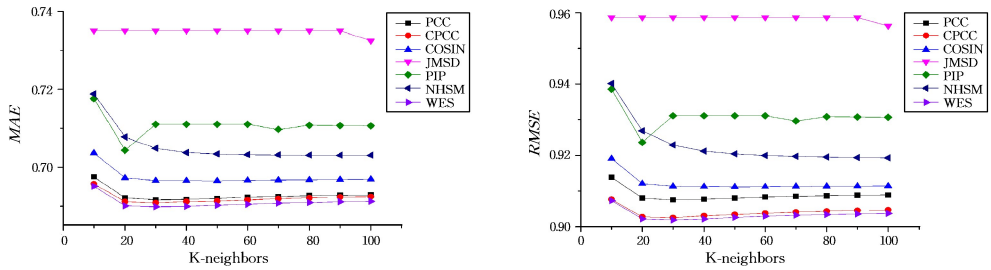

图 2 在 Ml-latest-small 数据集上近邻参数 k 对 *MAE* 和 *RMSE* 的影响

Fig. 2 Impact of neighbor parameter k on *MAE* and *RMSE* on Ml-latest-small

结束语 相似度作为协同过滤方法中的关键技术，在推荐系统中起着至关重要的作用。本文通过引入极端评分行为的相似程度度量指数，并将其融合到传统的线性相关系数中，提出了一种考虑用户极端评分行为的新的相似度。极端评分相似指数充分考虑了用户之间的 3 种极端行为，从而更加精准地刻画了用户之间的相似关系。实验结果不仅表明本文提出的相似度与协同过滤方法相结合可以得到更高的推荐精度，而且验证了本文方法的稳定性。未来可围绕相似度的高

效计算和多源信息融合计算等方面开展进一步的研究。

参 考 文 献

[1] GOLDBERG D, NICHOLS D, OKIB M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of ACM, 1992, 35(12): 61-70.

[2] RESNICK P, VARIAN H R. Recommender systems[J]. Communications of ACM, 1997, 40(3): 56-58.

[3] ZENEBE A, NORCIO A F. Representation: Similarity measures and aggregation methods using fuzzy sets for content-based recommender systems[J]. Fuzzy Sets and Systems, 2009, 160(1): 76-94.

[4] SCHAFER J B, KONSTAN J A, RIEDL J. E-commerce recommendation applications[J]. Data Mining and Knowledge Discovery, 2001, 5(1/2): 115-153.

[5] BOBADILLA J, ORTEGA F, HERNANDO A, et al. Recommender systems survey[J]. Knowledge-Based Systems, 2013, 46(1): 109-132.

[6] AAMIR M, BHUSRY M. Recommendation system: State of the art approach[J]. International Journal of Computer Applications, 2015, 120: 25-32.

[7] XIAO Y Y, ZHANG H Y. Friend recommendation method based on users' latent features in social networks[J]. Computer Science, 2018, 45(3): 220-254.

[8] ZHANG S, YAO L, SUN A, et al. Deep learning based recommender system: A survey and new perspectives [J]. ACM Computing Surveys, 2017, 1(1): 1-35.

[9] HANG L V, JIANG B T, LV S Y, et al. Survey on deep learning based recommender systems[J]. Chinese Journal of Computers, 2018, 41(7): 191-219.

[10] HSU C C, YEH M Y, LIN S D. A general framework for implicit and explicit social recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 14(8): 1-14.

[11] KATZMAN J, SHAHAM U, BATES J, et al. DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network[J]. Bmc Medical Research Methodology, 2016, 18(1): 24.

[12] QUADRANA M, CREMONESI P, JANNACH D. Sequence-aware recommender systems [J]. ACM Computing Surveys, 2018, 51(4): 373-374.

[13] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[J]. Uncertainty in Artificial Intelligence, 2013, 98(7): 43-52.

[14] SU X Y, KHOSHGOFTAAR T M. A survey of collaborative filtering techniques [J]. Advances in Artificial Intelligence, 2012, 2009(12): 1-19.

[15] SHI Y, LARSON M, HANJALIC A. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges[J]. ACM Computing Surveys, 2014, 47(1): 1-45.

[16] LEE S. Using entropy for similarity measures in collaborative filtering[J/OL]. Journal of Ambient Intelligence and Humanized Computing, 2019. https://doi.org/10.1007/s12652-019-01226-0.

[17] HE X, HE Z, SONG J, et al. NAIS: Neural attentive item similarity model for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(12): 2354-2366.

[18] LIAN D, GE Y, ZHANG F, et al. Scalable content-aware collaborative filtering for location recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1122-1135.

[19] SARWAR B M, KARYPIS G, KONSTAN J A, et al. Analysis of recommendation algorithms for e-commerce[C]// Proceedings of ACM E-Commerce. Minneapolis, Minn, USA, 2000: 158-167.

[20] RESNICK P, IACOVOU N, SUCHAK M, et al. Grouplens: An open architecture for collaborative filtering of netnews[C]// Proceedings of the ACM Conference on Computer Supported Cooperative Work. New York: ACM Press, 1994: 175-186.

[21] SHARDANAND U, MAES P. Social information filtering: algorithm for automating' word of mouth'[C]// Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems. New York: ACM Press, 1995: 210-217.

[22] KENDALL M G. Rank correlation methods[J]. British Journal of Psychology, 1990, 25(1): 86-91.

[23] AHN H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Sciences, 2008, 178(1): 37-51.

[24] LIU H, ZHENG H, MIAN A, et al. A new user similarity model to improve the accuracy of collaborative filtering[J]. Knowledge-Based Systems, 2014, 56(3): 156-166.

[25] HERLOCKER J L, KONSTAN J A, BORCHERS A, et al. An algorithmic framework for performing collaborative filtering [C]// Proceedings of the SIGIR '99 International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1999: 230-237.

[26] JAMALI M, ESTER M. TrustWalker: A random walk model for combining trust-based and item-based recommendation[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 397-406.

FENG Chen-jiao, born in 1977, doctorial student, lecturer, is member of China Computer Federation. Her main research interests include data mining, big data correlation analysis and recommender systems.

LIANG Ji-ye, born in 1962, Ph.D, professor, Ph.D supervisor, is member of China Computer Federation. His main research interests include granular computing, data mining and machine learning.