基于用户评分差异性和相关性的协同过滤推荐算法

王劲松 蔡朝晖 李永凯 刘树波
(武汉大学计算机学院 武汉 430072)

摘 要 传统的协同过滤相似性度量方法主要考虑用户评分之间的相似性,缺少对评分差异性的考虑。文中将用户评分关系分为差异部分和相关部分,提出了一种基于用户评分差异性和相关性的相似性度量方法。该方法在非极其稀疏数据集下有较好的推荐效果。针对该方法在稀疏数据集下存在推荐不准确的问题,采用预填充方法对其进行改进。实验表明,该方法在预填充后的推荐精度得到明显提高。

关键词 协同过滤推荐,差异性,相关性,预填充
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.05.032

Collaborative Filtering Recommendation Algorithm Based on Difference and Correlation of Users' Ratings

WANG Jing-song CAI Zhao-hui LI Yong-kai LIU Shu-bo
(School of Computer, Wuhan University, Wuhan 430072, China)

Abstract The traditional similarity measurement in collaborative filtering mainly pays attention to the similarity between users' ratings, lacking the consideration of difference of users' ratings. This paper divided the relationship of users' ratings into differential part and correlated part, and proposed a similarity measurement based on the difference and the correlation of users' ratings on the non-sparse dataset. In order to solve the problem that the algorithm's recommendation is not accurate in spare dataset, this paper improved this algorithm by prefilling the vacancy of rating matrix. Experiment results show that this algorithm can significantly improve the accuracy of recommendation after prefilling the rating matrix.

Keywords Collaborative filtering recommendation, Difference, Correlation, Prefilling

1 引言

随着信息技术的迅猛发展,网络信息资源量呈现指数增长趋势。面对海量的信息资源,用户需要耗费大量的时间和精力寻找所需的物品,且常会迷失在海量信息中。个性化推荐算法通过分析不同用户的兴趣点,可以帮助用户快速地定位到自己感兴趣的内容,解决互联网中信息过载[1]和用户喜好之间的矛盾。目前,个性化推荐算法已被广泛应用到电子商务[2]、电影和视频网站、数字图书馆[3]、新闻网站[4]、个性化音乐网站等系统中。

协同过滤是迄今为止应用得最成功的个性化推荐技术[5-6],其思想是根据用户对物品或内容的偏好,发现用户、物品或内容的相关性,然后基于这种关联性进行推荐。例如,在基于用户的推荐系统中,首先需要根据用户的评分信息发现用户之间的相似性,然后基于相似性来预测目标用户对未评分物品或内容的评分。

在基于用户的协同过滤推荐算法中,用户之间的相似性度量是否准确直接关系到整个推荐系统的效果。传统的相似性度量方法包括余弦相似度、修正余弦相似度、相关相似度以及上述算法的改进方法[7]。在本质上,传统相似度算法只能反映用户之间的兴趣方向的相似性或相关性,不能反映用户之间对同一资源的评分差异性。因此,传统相似度算法很难全面地体现用户之间的相似性,从而导致推荐系统质量下降。针对上述问题,本文提出了一种基于用户评分的差异性与相关性(Difference and Correlation,DC)的相似性度量方法。该方法既考虑了用户评分之间的相关部分,还考虑了差异部分。由于数据集的稀疏度对DC算法的精度有很大影响,因此本文通过预填充DC算法(Prefilled Difference and Correlation,PDC)来解决DC算法在稀疏数据集下存在的性能问题。

2 相关工作

传统基于用户的协同过滤算法为目标用户进行推荐时需要经过数据表述、寻找 K 最近邻居、产生推荐数据集这3个步骤[8],其具体描述如下。

(1)数据表述

数据表述主要是收集代表用户兴趣的信息,如用户评分、

到稿日期:2017-03-20 返修日期:2017-06-10 本文受国家自然科学基金(41671443)资助。
王劲松(1990-),男,硕士,主要研究方向为推荐系统、本体,E-mail:2014202110087@whu.edu.cn;**蔡朝晖**(1968-),女,博士,副教授,主要研究方向为分布式信息处理,E-mail:zhcai@whu.edu.cn(通信作者);**李永凯**(1988-),男,博士,主要研究方向为信息安全,E-mail:2013102110070@whu.edu.cn;**刘树波**(1970-),男,博士,教授,主要研究方向为嵌入式系统、多媒体及其安全,E-mail:liu.shubo@whu.edu.cn。

购买信息等。若在评分系统中，给定用户集合 $U=\{u_1,u_2,\cdots,u_m\}$ 和项目集合 $I=\{i_1,i_2,\cdots,i_n\}$，则用户兴趣模型可以表示为 $m\times n$ 的用户项目评分矩阵 R，如式(1)所示：

$$R=\begin{bmatrix} R_{1,1} & R_{1,2} & \cdots & R_{1,n} \\ R_{2,1} & R_{2,2} & \cdots & R_{2,n} \\ \vdots & \vdots & & \vdots \\ R_{m,1} & R_{m,2} & \cdots & R_{m,n} \end{bmatrix} \tag{1}$$

其中，$R_{u,i}$ 表示用户 u 对项目 i 的评分，$R_{u,i}=0$ 表示用户 u 未对项目 i 进行评分；m 表示用户数；n 表示项目数。

(2)寻找 K 最近邻居

根据收集到的用户兴趣度信息，通过相似度算法计算用户之间的相似度。按照相似度从大到小的顺序排列，取前 K 个与目标用户最相似的邻居。

(3)产生推荐数据集

相似用户对同一项目的兴趣度也相同，因此可以利用目标用户的 K 最近邻居评分值来预测目标用户对未评分项目的评分值。根据用户对所有未评分项目的预测值，可以得到用户的推荐集。假设目标用户 u 的最近邻居集合为 N_u，用户 u 的未评分项目 i 的预测评分为 $r_{u,i}$，则可以通过用户 u 的邻居 N_u 对项目 i 的加权平均值来逼近用户 u 的评分值[8]，如式(2)所示：

$$r_{u,i}=\overline{R_u}+\frac{\sum\limits_{v\in N_u} sim(u,v)\times(R_{v,i}-\overline{R_v})}{\sum\limits_{v\in N_u}|sim(u,v)|} \tag{2}$$

其中，$\overline{R_u}$ 和 $\overline{R_v}$ 分别表示用户 u 和用户 v 对已评分项目的平均评分，$sim(u,v)$ 表示用户 u 和用户 v 的相似度。依据式(2)获取用户 u 对所有未评分项目的预测值，从预测值中获取最大的 N 个预测值作为推荐集。

从上述 3 个步骤可以看出，在协同过滤过程中寻找目标用户的 K 最近邻居是推荐系统的关键。合适的相似度算法直接决定着推荐系统的质量。

2.1　传统相似性度量方法

传统相似性度量方法主要有余弦相似性(Cosine)、改进的修正余弦相似性(Adjusted Cosine)和相关相似性(Pearson)算法。

(1)余弦相似性

余弦相似性通常将每一个用户的评分信息看作一个 n 维向量。如果用户对某一项目未进行评分，则将其评分设置为 0。用户之间的相似性通过两个用户向量的余弦夹角来衡量。设用户 u 和用户 v 的评分向量分别为 u 和 v，则用户 u 与用户 v 之间的相似度如式(3)所示：

$$\begin{aligned} sim(u,v)&=\cos(u,v)\\ &=e_u\cdot e_v\\ &=\frac{u\cdot v}{|u|\times|v|}\\ &=\frac{\sum\limits_{c=1}^{n}R_{u,c}R_{v,c}}{\sqrt{\sum\limits_{c=1}^{n}R_{u,c}^2}\sqrt{\sum\limits_{c=1}^{n}R_{v,c}^2}} \end{aligned} \tag{3}$$

其中，向量 e_u 和 e_v 分别为向量 u 和 v 的单位向量，$R_{u,c}$ 和 $R_{v,c}$ 分别为用户 u 和用户 v 对项目 c 的评分。

(2)修正余弦相似性

余弦相似性度量方法忽略了不同用户之间的评价尺度问题，修正余弦相似性度量方法通过减去用户对项目的平均评分来修正用户之间的评价尺度。假设 I_{uv} 表示用户 u 和用户 v 共同已评分的项目集合，I_u 和 I_v 分别表示用户 u 和用户 v 已评分的项目集合，则用户 u 和用户 v 之间的相似性如式(4)所示：

$$sim(u,v)=\frac{\sum\limits_{c\in I_{uv}}(R_{u,c}-\overline{R_u})(R_{v,c}-\overline{R_v})}{\sqrt{\sum\limits_{c\in I_u}(R_{u,c}-\overline{R_u})^2}\sqrt{\sum\limits_{c\in I_v}(R_{v,c}-\overline{R_v})^2}} \tag{4}$$

其中，$\overline{R_u}$ 和 $\overline{R_v}$ 分别表示用户 u 和用户 v 对所有已评分项目的平均评分。

(3)相关相似性

相关相似性从统计学的角度讨论数据之间的相关性，Pearson 相似性算法只考虑用户共同评分过的项目，一般适用于两个变量之间存在定距关系。若一个用户的评分总是高于另一个用户的评分，则两个用户之间的相关性较高。用户 u 和用户 v 之间的相关相似性如式(5)所示：

$$sim(u,v)=\frac{\sum\limits_{c\in I_{uv}}(R_{u,c}-\overline{R_u})(R_{v,c}-\overline{R_v})}{\sqrt{\sum\limits_{c\in I_{uv}}(R_{u,c}-\overline{R_u})^2}\sqrt{\sum\limits_{c\in I_{uv}}(R_{v,c}-\overline{R_v})^2}} \tag{5}$$

2.2　传统相似性度量方法分析

在传统相似度算法中不同算法存在不同的弊端。用户项目评分矩阵如表 1 所列，由余弦相似度计算公式得出 u_1 和 u_3 的相似度高于 u_1 和 u_2 的相似度，但从实际评分来看，u_1 与 u_2 都喜欢项目 i_1 和 i_2，而 u_3 不喜欢 i_1 和 i_2，由此可知 u_1 与 u_2 比 u_1 与 u_3 更加相似。u_1 与 u_4 的相似度高于 u_1 与 u_5 的相似度，与实际情况不符合，因为从共同评分项目的角度来看，u_1，u_4，u_5 在项目 i_2 和 i_3 上的评分相同，无法比较 u_1 与谁更相似。在相关相似度性的计算方式下，u_3 与 u_6 的相似度为 -1，但实际上它们有着较大的相似度。特别地，修正余弦相似度和相关相似度都无法计算 u_7 与其他用户的相似度，这是因为 u_7 的平均评分为 4，在计算相似度时，用户评分减去 4 后会使计算公式的分母为 0。

表 1　用户项目评分矩阵

Table 1　User-item rating matrix

	i_1	i_2	i_3	i_4	i_5
u_1	4	5	4		
u_2	4	4	1		
u_3	1	2	1	2	1
u_4		5	4	2	1
u_5		5	4	5	4
u_6	2	1	2	1	2
u_7	4	4	4	4	4

余弦相似度算法通过用户评分向量之间的夹角余弦值进行评估，只从用户兴趣的方向来考虑用户之间的相似度，而没有考虑用户之间的评分差异性对用户兴趣相似度的影响。式(3)中的余弦相似度也可以表示为向量 u 和 v 单位向量 e_u 和 e_v 上的夹角余弦值，与具体的评分高低无关。修正余弦相似度在余弦相似度算法的基础上修正不同用户之间的评价尺度，一定程度上缓解了评分差异性，但并没有从每一个评分项目上体现这种差异性。就相关相似性而言，当一个用户都评高分而另一用户都评低分时，相关相似性就不能正确表现用

户之间的相似性，其原因在于相关相似性只考虑了整体上的差异性，同样没有考虑每一个评分项目上的差异性。从公式角度来看，相关相似性与修正余弦相似性相同，都只考虑了用户之间的评价尺度上的整体差异性。因此，传统相似性度量方法没有针对每个项目来考虑评分差异性问题。

针对上述问题，研究者提出了不同的解决方法。文献[9]根据用户兴趣相似度和兴趣向量的欧几里德距离的倒数关系来计算用户相似度。文献[10]通过归一化欧氏距离来消除不同评分维度之间存在的不同评价尺度。文献[11]提出了一种在用户共同评分项目上的平均平方差(Mean Square Difference)，并通过设置一定误差阈值来过滤掉不相似用户。但上述方式只考虑了用户评分的差异性，没有考虑用户评分的相关性。文献[12]在传统相似性度量方法上进行改进，主要通过设置项目评分差值阈值来判断用户是否属于同一局部相似性。为了体现用户之间评分的差异性，文献[13]定义了一种评分差异度，作为传统相似性度量方法的权重因子。文献[12-13]都引入用户评分差异性来优化传统的相似性度量算法。

3 基于差异性与相关性的相似性度量

通过上述分析，本文将用户之间的评分关系分为差异关系和相关关系，提出了一种基于用户评分差异性与相关性(Difference and Correlation，DC)的相似度计算方法。此方法在一定条件下可以提高相似度算法的准确度。同时，针对 DC 算法在稀疏数据集下存在的弊端，提出了一种改进算法，即基于预填充的差异性与相关性(Prefilled Difference and Correlation，PDC)相似度算法。

3.1 基于差异性与相关性的相似性度量方法

定义 1(用户评分差异关系) 将用户的评分信息看作一个 n 维的向量，将用户在某一项目上的评分差异看作向量在这一维度上的评分差值的绝对值。两个用户总体上的评分差异关系可以描述为两个向量的相减关系，定量描述为 $(u-v)^2$，其中向量 u 和 v 分别表示用户 u 和用户 v 的评分向量。

定义 2(用户评分相关关系) 将用户在每一项目上的评分相似度看作向量在这一维度上的评分乘积。将两个用户总体上的评分相关关系描述为两个向量的点积，定量描述为 $u \cdot v$，其中向量 u 和 v 分别表示用户 u 和用户 v 的评分向量。

用户的评分关系可以分为用户评分差异关系和用户评分相关关系。用户评分差异关系反映的是用户对每一项目的评分的不同之处，如定义 1 所示，传统相似性度量方法无法对此进行体现。用户对同一项目的评分越不相同，用户之间应该越不相似。同理，用户评分相关关系体现了用户对每一个项目的评分的相同之处，如定义 2 所示，用户评分相关关系可以定量表示为 $u \cdot v = e_u \cdot e_v \times |u| \times |v|$，其中 $e_u \cdot e_v$ 表示用户的兴趣方向，$|u| \times |v|$ 则表示用户的评分大小。

用户评分差异关系和用户评分相关关系从两个不同的方面体现了用户的评分关系，因此相似性的度量应该充分考虑用户评分之间的差异关系和相关关系。差异关系越大，则相似性越小；相关关系越大，则相似性越大。同时，考虑到相似性处于 0～1 之间，用户相似性度量如式(6)所示：

$$sim(u,v)=\frac{u \cdot v}{(u-v)^2+u \cdot v} \tag{6}$$

其中，向量 u 和 v 表示用户 u 和用户 v 的评分信息。

在实际情况中，用户的评分是十分稀疏的。用户评分向量中存在大量的未评分项目，处理未评分项目的方式将对相似性度量产生很大的影响。设置未评分项目的评分值为 0，从表达式 $(u-v)^2$ 中可知用户评分差异关系可扩大到 $\max(u^2,v^2)$，从表达式 $u \cdot v$ 中可知用户相关关系将减小到 0，此时的相似性度量与实际值的偏差较大。

考虑到未评分项目的问题，本文根据用户是否对项目进行评分为式(6)提供两种处理方式。

1)DC 算法

针对未评分项目的问题，如果只考虑用户之间的共同评分项，则可以消除未评分项目对相似性度量的影响，如式(7)所示：

$$sim(u,v)=\frac{\sum\limits_{c \in I_u \cap I_v} \min(R_{u,c},R_{v,c})^2}{\sum\limits_{c \in I_u \cap I_v}(R_{u,c}-R_{v,c})^2+\sum\limits_{c \in I_u \cap I_v} \min(R_{u,c},R_{v,c})^2} \tag{7}$$

其中，I_u 和 I_v 分别表示用户 u 和用户 v 评分过的项目集合，$R_{u,c}$ 和 $R_{v,c}$ 分别为用户 u 和用户 v 对项目 c 的评分。

2)PDC 算法

式(7)虽然消除了未评分项目的默认评分 0 值对相似性度量的影响，但是这种方法丢弃了许多非共同评分信息，使得相似性度量时可参考的评分非常稀少，从而影响度量的准确度。其次，在稀疏数据集中，用户共同评分项目更加稀少，这将严重影响相似性的度量。因此，可以考虑用户之间已评分项目的并集，这时相似性度量可参考的评分信息较多。但是这种方法存在处理用户评分缺失的问题。

处理用户评分缺失问题的方法可以看作是对用户评分矩阵的一种预填充方法。选取恰当的预填充方法将会明显改善式(6)在差异性和相关性上的计算误差。预填充是解决数据稀疏性问题的常用方法，其中平均值填充是最简单的。基于评分预测的填充方法也是一种常用的填充方法，即将系统第一次预测的评分作为未评分项目的评分进行填充。文献[14]和文献[15]分别提出了基于信任度传播的填充方法和基于云模型的评分矩阵填充方法。文献[16-17]采用基于项目分类信息的协同过滤算法进行预填充。文献[18]通过建立信任模型对评分矩阵进行预填充。

因此，本文借用传统预填充方法来改善基于差异性与相关性的相似性度量方法。假设 $r_{u,i}$ 表示用户 u 在评分项目 i 上的填充值，预填充后的评分矩阵 R' 如式(8)所示：

$$R'=\begin{cases} R_{u,i}, & \text{if user } u \text{ rated item } i \\ r_{u,i}, & \text{if user } u \text{ not rated item } i \end{cases} \tag{8}$$

基于预填充的评分矩阵，考虑到用户已评分项目的并集，可将式(6)改写为：

$$sim(u,v)=\frac{\sum\limits_{c \in I_u \cup I_v} \min(R'_{u,c},R'_{v,c})^2}{\sum\limits_{c \in I_u \cup I_v}(R'_{u,c}-R'_{v,c})^2+\sum\limits_{c \in I_u \cup I_v} \min(R'_{u,c},R'_{v,c})^2} \tag{9}$$

3.2 算法描述

DC 和 PDC 是对式(7)进行不同处理的方法，PDC 方法如算法 1 所示。

算法 1　基于预填充差异性与相关性的相似性推荐算法

输入：用户评分矩阵 R，预填充后的用户评分矩阵 R′，推荐用户 u，最近邻居个数 K，推荐项目数 N

输出：推荐项目集 I

1. Begin

2. Set I=∅；相似度集合 S=∅；预测集合 RP=∅；从矩阵 R 中获取用户集合 $U=\{u_1,u_2,\cdots,u_m\}$，项目集合 $I=\{i_1,i_2,\cdots,i_n\}$；用户 u 的评分项目集合为 I_u；用户项目平均评分集合为 $\{\overline{R_1},\overline{R_2},\cdots\overline{R_m}\}$；

3. For each $v\in U-\{u\}$：

4. 采用式(9)计算用户之间的相似度；

5. $S=S\cup\{sim(u,v)\}$；

6. End For

7. $N_u=TOP(S,K)//N_u$ 为用户 u 的 K 个最近邻居

8. For each $i\in I-I_u$：

9. $r_{u,i}=\overline{R_u}+\dfrac{\sum\limits_{v\in N_u}sim(u,v)\times(R_{v,i}-\overline{R_v})}{\sum\limits_{v\in N_u}|sim(u,v)|}$；

10. $RP=RP\cup\{r_{u,i}\}$；

11. End For

12. For each $i\in I-I_u$：

13. IF $i\in TOP(RP,N)$

14. $I=I\cup\{i\}$

15. End IF

16. End For

17. End

4　实验及分析

4.1　数据集

本文实验采用 MovieLens 站点提供的数据集。MovieLens 建立于 1997 年，其数据来源于电影推荐网站，被广泛用于研究推荐系统。目前，该站点的用户已超过 71000 人，用户已评分的电影超过 10000 部。MovieLens 提供了 3 种不同数量级的数据。本文采用 943 个用户对 1682 部电影做出的 10 万条评分数据作为实验的数据集。本文将数据集分为训练集和测试集，其中 80％的数据作为训练集，20％的数据作为测试集。

4.2　评估标准

本文采用平均绝度偏差(Mean Absolute Error，MAE)作为推荐系统的评价指标。假设预测用户的评分集合为 $\{p_1,p_2,\cdots,p_N\}$，实际对应的用户评分为 $\{q_1,q_2,\cdots,q_N\}$，则 *MAE* 如式(10)所示：

$$MAE=\frac{\sum\limits_{i=1}^{N}|p_i-q_i|}{N}\tag{10}$$

4.3　实验方案

本文将从 3 个方面进行实验。由于 DC 相似度算法只考虑用户共同评分项目，因此数据稀疏性对相似度具有一定影响。首先，对不同稀疏数据集下的 DC 算法性能进行分析；其次，由于 DC 相似度算法考虑了用户之间的差异性，因此通过实验分析不同算法在不同评分差异的数据集下的关系；最后，对改进的 DC 算法(PDC)与其他算法进行分析。

1)实验 1

为了获取不同稀疏度的数据集，本文将用户按照用户评分次数排序，分别取前 100，200，…，800，900，940 个用户表示不同稀疏度的训练集。实验分别在不同训练集下比较 Cosine，Pearson，DC 算法的性能，比较结果如图 1 和图 2 所示。

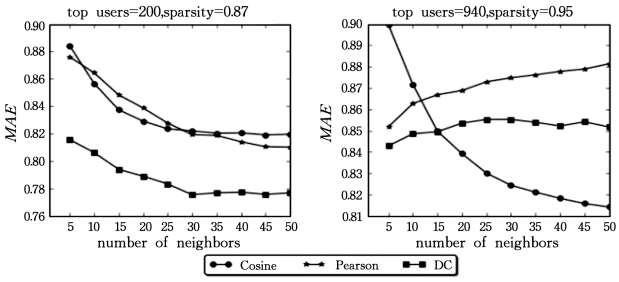

图 1　固定稀疏性时不同邻居用户数下的 *MAE* 值比较

Fig. 1　Comparison of *MAE* with different number of neighbors and fixed sparsity

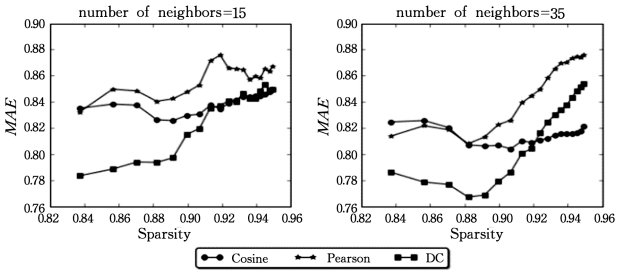

图 2　固定邻居用户数时不同稀疏性下的 *MAE* 值比较

Fig. 2　Comparison of *MAE* with fixed number of neighbors and different sparsity

由图 1 和图 2 可知，稀疏性对 Pearson 和 DC 的影响较大，主要原因在于这两种算法在进行相似度计算时都选取了用户共同评分项。从整体上来看，DC 相比于 Pearson 在性能上有很大提升，而与 Cosine 相比在稀疏度小于 0.92 时有明显优势。

2)实验 2

用户对项目的评分一般服从正态分布，正态分布的方差大小代表了数据集中的数据分布的差异性大小。为了获取不同差异的数据集，本文将项目按照项目评分方差进行排序，分别取方差最大的前 100，200，…，1600 个项目表示不同评分差异的数据集。为了较好地反映评分误差对相似性度量的影响，首先选取稀疏度为 0.91 的数据集，再将数据集转换成不同评分差异的数据集。实验分别在不同训练集下比较 Cosine，Pearson，DC 算法的性能，比较结果如图 3 和图 4 所示。

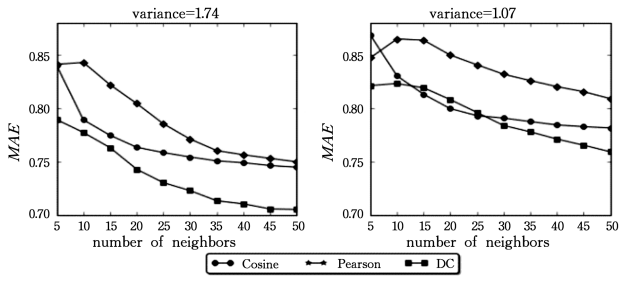

图 3　固定方差时不同邻居用户数下的 *MAE* 值比较

Fig. 3　Comparison of *MAE* with fixed standard deviation and different number of neighbors

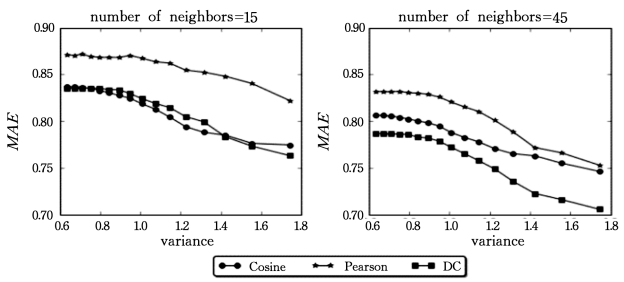

图 4 固定邻居数时不同方差下的 *MAE* 值比较

Fig. 4 Comparison of *MAE* with fixed number of neighbors and different standard deviation

由图 3 和图 4 可知，Pearson 在不同数据集差异下的表现基本相同，而 Cosine 和 DC 算法的变化较大，这是因为 Pearson 算法只考虑了用户评分的相关性，而没有考虑差异性。另外，相比于 Cosine 算法，DC 算法对数据差异性更加敏感，随着数据方差的增大，邻居个数的增多，DC 算法更显优势。

3)实验 3

实验 1 和实验 2 分别从稀疏性和差异性说明了 DC 算法的特性，这意味着改善这两个特性将会改进 DC 算法的性能。评分差异性是数据集的固有属性，在不同的特定领域，差异性表现不同。本文提出的 PDC 算法通过预填充的方法来改善 DC 算法的稀疏性问题。实验分别采用平均值和分类信息预测对 PDC 算法进行预填充，并将其与传统的相似度算法、DC 算法以及预填充后的 Cosine，Pearson 和 PDC 算法进行比较分析。

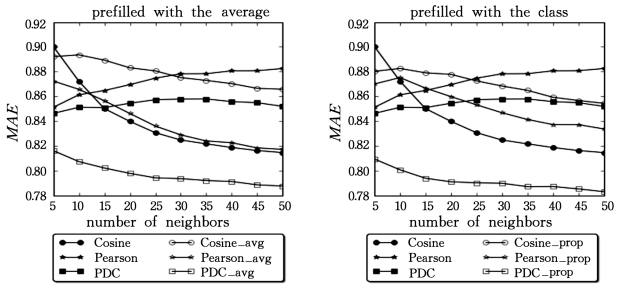

图 5 PDC 算法预填充性能对比

Fig. 5 Comparison of prefilled PDC algorithm with others

图 5 给出了 PDC 算法与其他算法的对比结果，图 5(a)为 PDC 算法采用平均值填充的效果对比，图 5(b)为 PDC 算法采用分类预测方法填充的效果对比。其中 Cosine_avg，Pearson_avg 和 PDC_avg 分别表示 Cosine，Pearson 和 PDC 采用平均值填充后的算法；Cosine_prop，Pearson_prop 和 PDC_prop 分别表示 Cosine，Pearson 和 PDC 采用分类预测方法填充的算法。从图 5 可知，两种预填充方法对 PDC 算法的影响基本相同。Cosine 算法经过预填充后的性能急剧下滑，其原因在于预填充在 Cosine 算法的执行过程中添加了一些不可靠评分数据。但是对于 Pearson 和 DC 算法而言，预填充评分减少了相似度计算误差，使其性能均有较大改善。从整体来看，预填充后的 DC 算法(PDC)的性能增长明显，比其他算法的 *MAE* 值降低了 0.04 以上，性能优越。

结束语 本文提出了一种基于用户评分相关性与差异性(DC)的相似度算法。该算法在稀疏性较小、评分差异性较大的数据集上有较好的性能。针对 DC 算法在稀疏数据集中的缺点，本文采用预填充方法对其进行改进(PDC)。实验表明，PDC 能够很好地适应数据稀疏性问题，并且相对传统算法有着更好的推荐质量。下一步将针对 DC 算法和 Cosine 算法对稀疏度的不同表现，对两种算法的融合模型进行研究，提出一种适应数据稀疏度变化的相似度算法。

参 考 文 献

[1] EPPLER M J, MENGIS J. The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines [J]. The Information Society, 2004, 38(5): 325-344.

[2] SCHAFER J B, KONSTAN J, RIEDL J. Recommender systems in e-commerce[C]// Proceedings of the 1st ACM conference on Electronic commerce. ACM, 1999: 158-166.

[3] JAYAWARDANA C, HEWAGAMAGE K P, HIRAKAWA M. A Personalized Information Environment for Digital Libraries [J]. Information Technology & Libraries, 2000, 20(4): 185-196.

[4] KONSTAN J A, MILLER B N, MALTZ D, et al. GroupLens: applying collaborative filtering to Usenet news [J]. Communications of the Acm, 2000, 40(3): 77-87.

[5] LINDEN G, SMITH B, YORK J. Amazon. com Recommendations: Item-to-Item Collaborative Filtering [J]. IEEE Internet Computing, 2003, 7(1): 76-80.

[6] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]// Fourteenth Conference on Uncertainty in Artificial Intelligence. 2013: 43-52.

[7] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(17): 734-749.

[8] SU X, KHOSHGOFTAAR T M. A survey of collaborative filtering techniques [J]. Advances in Artificial Intelligence, 2009, 2009(12): 4.

[9] SINGH A, YADAV A, RANA A. K-means with Three different Distance Metrics [J]. International Journal of Computer Applications, 2013, 67(10): 13-17.

[10] CHENG X H, GAO Y. Collaborative Filtering Recommendation Based on Optimization Euclidean Distance [J]. Computer and Modernization, 2015(3): 37-40. (in Chinese)
陈小辉，高燕. 基于优化欧氏距离的协同过滤推荐[J]. 计算机与现代化，2015(3)：37-40.

[11] SHARDANAND U. Social Information Filtering for Music Recommendation[J]. Massachusetts Institute of Technology, 1994: 74-81.

[12] WU F Q, HE L, XIA W W, et al. A recommendation algorithm based on users' partial similarity [J]. Journal of Computer Applications, 2008, 28(8): 1981-1985. (in Chinese)
吴发青，贺樑，夏薇薇，等. 一种基于用户兴趣局部相似性的推荐算法[J]. 计算机应用，2008，28(8)：1981-1985.

[13] DANG B, JIANG J L. Collaborative filtering recommendation al-

gorithm based on score difference level and user preference [J]. Journal of Computer Applications, 2016, 36(4): 1050-1053. (in Chinese)
党博，姜久雷. 基于评分差异度和用户偏好的协同过滤算法[J]. 计算机应用，2016，36(4)：1050-1053.

[14] JANG S, YANG J, KIM D K. Minimum MSE design for multi-user MIMO relay [J]. IEEE Communications Letters, 2010, 14(9): 812-814.

[15] ELDAR Y C. Universal Weighted MSE Improvement of the Least-Squares Estimator [J]. IEEE Transactions on Signal Processing, 2008, 56(5): 1788-1800.

[16] LI C, LIANG C Y, DONG K. A Collaborative filtering recommendation algorithm based on item category similarity [J]. Journal of Hefei University of Technology (Nature Sicence), 2008, 31(3): 360-363. (in Chinese)
李聪，梁昌勇，董珂. 基于项目类别相似性的协同过滤推荐算法[J]. 合肥工业大学学报(自然科学版)，2008，31(3)：360-363.

[17] XU Y, ZHANG D. Accelerating the kernel-method-based feature extraction procedure from the viewpoint of numerical approximation [J]. Neural Computing and Applications, 2011, 20(7): 1087-1096.

[18] YANG X Y, YU J, TURGENI B, et al. Collaborative Filtering Recommendation Model Based on Trust Model Filling [J]. Computer Engineering, 2015(5): 6-13. (in Chinese)
杨兴耀，于炯，吐尔根·依布拉音，等. 基于信任模型填充的协同过滤推荐模型[J]. 计算机工程，2015(5)：6-13.

(上接第 179 页)

函数 P-集合在动态数据(信息)规律挖掘、动态数据(信息)规律识别方面获得了应用。本文改进 P-集合的理论模型，提出能被工程应用的 P-数据模型，给出 P-数据模型的结构与生成方式。利用内 P-数据模型与内 P-数据推理交叉，给出内 P-数据模型与数据智能获取-风险识别应用。P-数据模型是研究动态数据分析、动态数据筛选、动态数据风险估计的新模型、新方法。

参 考 文 献

[1] SHI K Q. P-sets [J]. Journal of Shandong University (Natural Science), 2008, 43(11): 78-84. (in Chinese)
史开泉. P-集合[J]. 山东大学学报(理学版)，2008，43(11)：78-84.

[2] SHI K Q. P-sets and its applications[J]. An International Journal Advances in Systems Science and Applications, 2009, 9(2): 209-219.

[3] SHI K Q. P-sets and its applied characteristics[J]. Computer Science, 2010, 37(8): 1-8. (in Chinese)
史开泉. P-集合与它的应用特征[J]. 计算机科学，2010，37(8)：1-8.

[4] SHI K Q. P-information law intelligent fusion and soft information image intelligent generation[J]. Journal of Shandong University (Natural Science), 2014, 49(4): 1-17. (in Chinese)
史开泉. P-信息规律智能融合与软信息图像智能生成[J]. 山东大学学报(理学版)，2014，49(4)：1-17.

[5] SHI K Q. P-reasoning and P-reasoning discovery- identification of information [J]. Computer Science, 2011, 38(7): 1-9. (in Chinese)
史开泉. P-推理与 P-推理信息发现-辨识[J]. 计算机科学，2011，38(7)：1-9.

[6] SHI K Q. P-sets, inverse P-sets and the intelligent fusion-filter identification of information [J]. Computer Science, 2012, 39(4): 1-13. (in Chinese)
史开泉. P-集合，逆 P-集合与信息智能融合-过滤辨识[J]. 计算机科学，2012，39(4)：1-13.

[7] LI Y Y, LIN H K, SHI K Q. Characteristics of data discrete interval and data discovery- application [J]. Systems Engineering and Electronics, 2011, 33(10): 2258-2262. (in Chinese)
李豫颖，林宏康，史开泉. 数据离散区间特征与数据发现-辨识[J]. 系统工程与电子技术，2011，33(10)：2258-2262.

[8] LI Y Y, ZHANG L, SHI K Q. Generation and recovery of compressed data and redundant data [J]. Quantitative Logic and Soft Computing, 2010, 1(2): 661-672.

[9] LI Y Y, LIN H K. Application of data identification on $(\bar{F}, F)$-data discrete rectangle region[J]. Journal of Shandong University (Natural Science), 2011, 46(3): 46-51. (in Chinese)
李豫颖，林宏康. $(\bar{F}, F)$-数据离散矩形区域在数据辨识中的应用[J]. 山东大学学报(理学版)，2011，46(3)：46-51.

[10] FAN C X, LIN H K. P-sets and the reasoning identification of disaster information [J]. An International Journal of Convergence Information Technology, 2012, 7(1): 337-345.

[11] LIN H K, FAN C X. The dual form P-reasoning and identification of unknown attribute [J]. International Journal of Digital Content Technology and its Applications, 2012, 6(1): 121-131.

[12] SHI K Q, LI X H. Camouflaged information identification and its application [J]. Advances in Systems Science and Applications, 2010, 10(2): 157-167.

[13] ZHANG L, CUI Y Q. Outer P-sets and data internal recovery [J]. An International Journal Advances in Systems Science and Applications, 2010, 10(2): 189-199.

[14] ZHANG L, REN X F. P-sets and its $(\bar{f}, f)$-heredity[J]. Quantitative Logic and Soft Computing, 2010, 1(2): 735-743.

[15] ZHANG G Y, LI E Z. Information gene and identification of its information Knock-out/knock-in [J]. Advances in Systems Science and Applications, 2010, 10(2): 308-315.

[16] YU X Q. Generation of iterative $\bar{F}$-internal embedding information and its heredity discovery- application [J]. Systems Engineering and Electronics, 2011, 33(12): 2691-2695. (in Chinese)
于秀清. 迭代 $\bar{F}$-内嵌入信息生成与它的遗传发现-应用[J]. 系统工程与电子技术，2011，33(12)：2691-2695.

[17] ZHANG G Y, ZHOU H Y, SHI K Q. P-sets and the recovery-identification of double P-data [J]. Systems Engineering and Electronics, 2010, 32(9): 1919-1924. (in Chinese)
张冠宇，周厚勇，史开泉. P-集合与双 P-数据发现-辨识[J]. 系统工程与电子技术，2010，32(9)：1919-1924.

[18] SHI K Q. Function P-sets [J]. International Journal of Machine Learning and Cybernetics, 2011, 2(4): 281-288.

[19] SHI K Q. P-sets, Function P-sets [J]. Journal of Shandong University (Natural Science), 2011, 46(2): 62-69. (in Chinese)
史开泉. P-集合，函数 P-集合[J]. 山东大学学报(理学版)，2011，46(2)：62-69.