基于耦合相似度的矩阵分解推荐方法

郭梦娇 孙劲光 孟祥福
(辽宁工程技术大学电子与信息工程学院 葫芦岛 125100)

摘 要 随着因特网和信息技术的高速发展,信息过载现象越来越严重。推荐系统能够给个人和商家(例如电子商务和零售商)提供个性化的推荐。数据稀疏性和分数预测质量问题被公认为是现存推荐系统中的主要挑战。当前绝大多数推荐系统技术都依赖于协同过滤方法,它主要利用用户-项目评分矩阵来表示用户和项目之间的关系。一些研究利用附加信息来提高推荐准确性,但是,绝大多数现存的引入项目之间关系的方法并不能很好地用于预测和推荐,因为其假设项目属性之间是独立同分布的,而实际上项目(或用户)的属性之间是存在耦合关系的。由此提出了基于属性耦合关系的矩阵分解模型,它能有效地刻画项目之间的耦合相关性,从而更加合理地预测用户对项目的评分。实验结果表明,所提出的模型在热启动和冷启动的推荐准确性方面均优于传统的推荐算法。

关键词 推荐系统,相似度,矩阵分解,冷启动,预测
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.4.050

Coupling Similarity-based Matrix Factorization Technique for Recommendation

GUO Meng-jiao SUN Jing-guang MENG Xiang-fu
(College of Electronic and Information Engineering, Liaoning Technical University, Huludao 125100, China)

Abstract With the rapid development of Internet and information technology, information overload becomes more and more seriously. Recommender system can provide personalized recommendations to both individual users and businesses (such as e-commerce and retail enterprises). The data sparsity and prediction quality are recognized as the key challenges in the existing recommender systems. Most of the existing recommender systems depend on collaborating filtering (CF) method, which mainly uses the user-item rating matrix to represent the relationship between users and items. Several researches consider utilizing extra information to improve the accuracy. However, most of the existing methods usually fail to provide accurate information for predicting recommendations, as there is an assumption that the relationship between attributes of items is independent and identically distributed, while, there are often several kinds of coupling relationships or connections existing among items or users in real applications. This paper incorporated the coupling relationship analysis to capture under-discovered relationships of items and aimed to make the ratings more reasonable. This paper proposed a coupled attribute-based matrix factorization model, which can capture the coupling correlations between items effectively. The experimental evaluations demonstrate the proposed algorithms outperform the state-of-the-art algorithms in the warm start and cold start settings.

Keywords Recommender systems, Similarity, Matrix factorization, Cold-start, Predicting

1 引言

推荐系统是一种解决信息飞速增长问题的重要途径,它主要应用信息过滤技术向用户提供满足其兴趣和偏好的信息[10]。然而,目前大多数推荐系统的推荐准确性并不高,主要是由大量没有评分的项目或者新用户的加入导致的冷启动和数据稀疏性问题造成的。由于新项目通常缺少用户评价,而现有推荐系统又需要根据项目的多个用户评价才能将其较为准确地推荐给其他用户,因此,如果项目的用户评价缺失或用户评价较少,将导致推荐系统的准确性不高,这就是推荐系统的冷启动问题。由于新项目通常是用户较为感兴趣的资源,而它们的用户评价又较少,因此本文将重点研究项目的冷启动问题。

为了解决冷启动问题,研究人员提出使用附加信息改善推荐质量的方法,附加信息一般包括标签[5]和地点[12]等。同时,也有文献考虑了社交关系(用户与用户之间的某种联系)[14,15],但是在许多场合下社交关系信息并不存在。因此,本文将通过对项目的属性建立关联规则,进而建立项目之间的关系,从而解决向用户合理推荐项目的问题。以往关于上下文感知方法的文献大都是假设属性之间服从独立同分布关系,传统的推荐系统通常忽略了属性间的关联关系(其中包含显式或者隐式关联关系)。而在实际应用中,属性之间通常存在耦合关系(例如同一个属性内的内耦合关系和不同属性间的间耦合关系)。因此,需要同时考虑项目属性的耦合关系和

到稿日期:2015-03-27 返修日期:2015-06-15 本文受国家青年科学基金(61003162),辽宁省高等学校杰出青年学者成长计划(LJQ 201303038)资助。

郭梦娇(1988—),女,硕士生,主要研究方向为推荐系统,E-mail:gmj23678111@hotmail.com;孙劲光(1962—),女,博士,教授,主要研究方向为图形理论与技术、图像工程、数据库原理与系统;孟祥福(1981—),男,博士,副教授,主要研究方向为Web数据库和XML数据个性化柔性查询。

用户评分矩阵来提高推荐准确性。项目的属性能够被用来解决评分稀少情况下的冷启动问题。基于本文思想，图 1 例举了一个电影推荐问题。

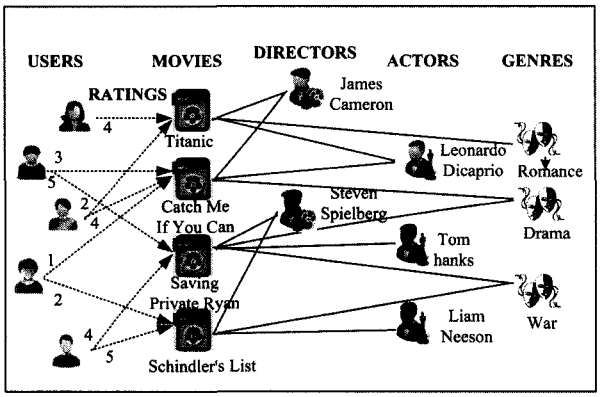


图 1　电影数据集的例子

在该问题中，向用户推荐的项目是电影。用 E 表示同类项目构成的集合，集合中的元素即为各种不同的影片。例如，电影的导演、演员以及基因构成电影的属性，不同的导演、不同的演员及各种不同的类别构成了相应属性的属性“值”。用户观看的每部影片所提供的信息就是其所有相关的属性值，通过属性值的相似性可以建立起不同项目（即影片）间的联系；同时，通过统计数据可以得到各种属性之间的关系，进而也可以推出项目（影片）间的关联。将这两种关系耦合在一起所提供的信息作为附加信息，进而提高对用户的评价能力。

以往推荐策略所面临的主要问题是：协同过滤方法主要依赖于用户评分矩阵，而其中大多数的信息是缺失的。隐因子模型中的矩阵分解技术将评分矩阵分解成低维的项目特征矩阵和用户特征矩阵，然后在相应的隐因子空间中通过重构的低维矩阵预测用户对项目的评分。矩阵分解方法主要用于同时将所有的项目结构进行评估。

本文在预测评分模型中引入耦合属性相似度度量方法，使得项目的耦合能够有效地解释现存的评分尺度问题。因此，其相关内容可应用在推荐系统中。本文主要贡献如下：

1）利用项目耦合关系反映了项目之间的隐式关联关系，从而有效解决了数据的稀疏性所带来的问题；

2）将用户的主观偏好评分尺度与项目的耦合关系相结合，推导出的隐式关系引入到矩阵分解学习模型中，从而提高推荐的准确性；

3）在两组真实数据集上进行了实验，验证了本文所提方法在解决冷启动方面优于传统的推荐算法。

由于以往推荐系统中项目相似度度量的不足及其所面临的数据稀疏导致的冷启动问题，本文提出了基于耦合相似度的矩阵分解方法，该方法能够相对准确地刻画用户对项目喜好的预测模型。实验结果表明，新算法能有效提高推荐质量。

2　耦合关系分析

大多数推荐系统中运用的相似度度量方法主要是根据用户对项目的历史评分，但这些评分往往是稀少且没有将项目的分类属性关系融入其中。应用最广泛的相似度度量方法是皮尔逊相关系数法，它以变量之间存在线性关系为前提[2]，而实际上通常属性之间所具有的关系并不是线性的，而耦合相似度度量方法能相对准确地刻画此种非线性关系。

本文主要利用分类属性的信息来揭示耦合属性相似度，从而挖掘隐式关系。耦合属性相似度是由内耦合相似度和外耦合相似度构成，它能够更为全面地捕获项目间的关联关系[3,23]。

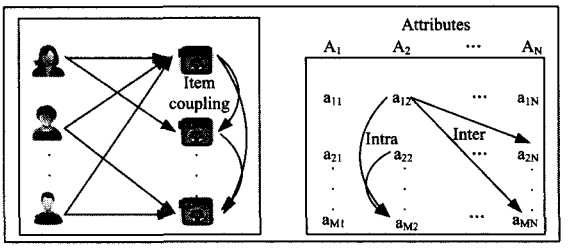


图 2　项目耦合关系

2.1　属性的内耦合相似度

对于给定的项目类集合 O，A_j 为项目的某一属性，属性 A_j 的内耦合相似度（IaAVS）是指该属性的两个属性值之间的一种关联程度值[3,23]。属性内耦合相似度定义为：

$$\delta_j^{Ia}(x,y)=\frac{|g_j(x)|\cdot|g_j(y)|}{|g_j(x)|+|g_j(y)|+|g_j(x)|\cdot|g_j(y)|} \tag{1}$$

其中，$g_j(x)$ 和 $g_j(y)$ 表示集合 O 中属性 A_j 的属性值为 x 和 y 的子集合，$|g_j(x)|$ 和 $|g_j(y)|$ 表示对应子集合的大小，即统计频数。

2.2　属性间耦合相似度

属性的内耦合相似度仅考虑了同一属性下的属性值之间在分布上的相关关系，而属性的间耦合相似度考虑了不同属性之间的关联度，进而挖掘其中的隐含关联关系。对于任意两个项目，在属性 A_j 下属性值为 x 和 y 的属性间耦合相似度（IeAVS）定义如下：

$$\delta_j^{Ie}(x,y)=\sum_{k=1,k\neq j}^{N}\gamma_k\delta_{j|k}(x,y) \tag{2}$$

其中，γ_k 代表属性 A_k 的权重参数，$\gamma_k\in[0,1]$，$\sum_{k=1,k\neq j}^{N}\gamma_k=1$。

本文采用香农熵来描述属性的权重。由于取得可靠的主观权重是比较困难的，因此使用客观权重赋值方法。客观权重估计方法被广泛应用于多属性决策领域[11]。熵的概念是一条信息中的信息量并作为一种统计度量。香农熵是由概率理论来表述不确定信息的度量。熵权重是指在特定条件下解释信息彼此之间有多大的影响程度。熵权重的计算过程如下：

$$p_{ik}=\frac{|g_{ik}(x)|}{M},i=1,\cdots,m,k=1,\cdots,n$$

p_{ik} 代表项目集合中第 k 个属性下第 i 个属性值的出现频率，M 表示项目类集合 O 的项目个数，h_k 表示第 k 个属性熵值，计算过程如下：

$$h_k=-h_0\sum_{i=1}^{m}p_{ik}\ln p_{ik},i=1,\cdots,m,k=1,\cdots,n$$

其中，h_0 是对熵值的约束，其值等于 $(\ln M)^{-1}$，当 $p_{ik}=0$ 时，$p_{ik}\ln p_{ik}$ 的值为 0。

$d_k=1-h_k,k=1,\cdots,n,d_k$ 表示多样化程度值。

$\gamma_k=\frac{d_k}{\sum_{k=1}^{n}d_k},k=1,\cdots,n,\gamma_k$ 表示属性 k 的权重。

$\delta_{j|k}(x,y)$ 表示属性 A_j 的属性值为 x 和 y 的属性间耦合相似度。$\delta_{j|k}(x,y)$ 的定义如下：

$$\delta_{j|k}(x,y)=\sum_{w\in\cap}\min\{P_{k|j}(w|x),P_{k|j}(w|y)\} \tag{3}$$

其中，$\cap$ 表示在属性 j 的属性值分别取 x 和 y 的情况下，属性 k 的属性值均为 w 的交集。$P_{k|j}(w|x)$ 是属性 j 的属性值为 x 的条件下属性 k 取值为 w 的条件概率：

$$P_{k|j}(w|x)=\frac{|g_k(w)\cap g_j(x)|}{|g_j(x)|} \tag{4}$$

通过对属性内耦合相似度（IaAVS）和属性间耦合相似度

(IeAVS)的分析，属性 A_j 的属性值为 x 和 y 之间的耦合相似度(CAVS)为：

$$\delta_j^A(x,y)=\delta_j^{Ia}(x,y)*\delta_j^{Ie}(x,y) \tag{5}$$

2.3 耦合属性相似度

项目 o_i 和项目 o_j 之间的耦合相似度(CIS)由属性内耦合相似度 $\delta_j^{Ia}(x,y)$和属性间耦合相似度 $\delta_j^{Ie}(x,y)$的乘积得出，即：

$$CIS(o_i,o_j)=\sum_{j=1}^{N}\delta_j^A(x,y) \tag{6}$$

基于以上讨论，耦合相似度度量同时考虑了属性的内耦合相似度和间耦合相似度，能够较准确地挖掘项目之间复杂的隐关系。

3 耦合相似度矩阵模型

传统的矩阵分解技术是上下文非感知的且单纯依靠用户评分矩阵来进行分数预测的，它将高维的稀疏的评分矩阵映射为隐因子空间中低维度的用户和项目矩阵，然后通过对应用户特征向量和项目特征向量的内积来预测缺失的评分。这种方法在预测全局信息时相对有效，但会导致局部的信息(小范围内项目之间的强关联信息)的丢失。

本文提出融合项目间耦合属性相似度与传统的矩阵分解模型来进行推荐的方法，其充分挖掘了已有数据和未知数据之间的潜在关系。值得注意的是，尽管项目间存在差别，但是每个项目与其近邻之间从某些方面必然存在一些共同特性，这间接反映了项目特性的传递。项目的特征向量和它的近邻特征向量在相应的空间中会比较近似[15,16]。由此，我们采用整体结构和局部信息全面地提高预测模型的精确性。

由于耦合相似度刻画的是两两对象之间的关系[3,23]，因此本文构建的相似度矩阵是对称的。设 W_{ij} 表示项目 i 和项目 j 规范化的相似度，可以表示为：

$$W_{ij}=\frac{CIS(o_i,o_j)}{\sum_{j\in\tau(i)}CIS(o_i,o_j)} \tag{7}$$

预测评分 R_{ui} 中项目的特征向量来自于项目本身和它的近邻的组合。

$$r_{ui}\approx\hat{r}_{ui}=\alpha p_u^{\mathrm{T}}q_i+(1-\alpha)\sum_{j\in\tau(i)}W_{ij}p_u^{\mathrm{T}}q_j \tag{8}$$

其中，参数 α 是在预测评分时隐因子 q_i 和其近邻之间的平衡因子。参数 α 表示项目依赖其本身和近邻的程度，$\tau(i)$表示项目 o_i 的 k 个最相似的近邻项目。

为了避免预测评分超过范围要求，本文把原始评分 R_{ui} 通过函数 $f(x)=x/R_{\max}$ 映射到[0,1]范围内，$R_{\max}$ 表示评分的最大尺度。采用文献[20]提出的 logistic 函数 $g(x)=1/(1+\exp(-x))$使得预测分数控制在[0,1]范围内。本文提出的以项目为导向的方法能够通过耦合属性相似度提供一个合理的解释性。本模型使用以下优化算法进行训练：

$$L(p,q)=\min_{p^*,q^*}\frac{1}{2}\sum_{(u,i)\in E}I_{ui}(R_{ui}-g(\alpha p_u^{\mathrm{T}}q_i+(1-\alpha)\sum_{j\in\tau(i)}W_{ij}p_u^{\mathrm{T}}q_j))^2+\frac{\lambda_p}{2}\|p_u\|_F^2+\frac{\lambda_q}{2}\|q_i\|_F^2 \tag{9}$$

其中，I_{ui} 是指示函数，当其值等于 1 时表示用户 u 评价过项目 i，当其值等于 0 时表示用户 u 没有评价过项目 i。$\|\cdot\|$ 是 Frobenius 范数，正则化参数为 $\lambda_p,\lambda_q>0$，$\frac{\lambda_p}{2}\|p_u\|_F^2+\frac{\lambda_q}{2}\|q_i\|$ 称为罚项，其作用是防止过拟合。凸优化的优化问题的解决可以通过解决最小二乘问题而得出，通过随机梯度下降方法依次进行迭代，由目标函数的局部最小找到其最速下降方向从而取得模型参数 p_u 和 q_i。

$$\frac{\partial L}{\partial p_u}=\sum_{i=1}^{n}I_{ui}g'(\alpha p_u^{\mathrm{T}}q_i+(1-\alpha)\sum_{j\in\tau(i)}W_{ij}p_u^{\mathrm{T}}q_j)\times(g(\alpha p_u^{\mathrm{T}}q_i+(1-\alpha)\sum_{j\in\tau(i)}W_{ij}p_u^{\mathrm{T}}q_j)-r_{ui})\times(\alpha q_i+(1-\alpha)\sum_{j\in\tau(i)}W_{ij}q_j)+\lambda_p p_u \tag{10}$$

$$\frac{\partial L}{\partial q_i}=\alpha\sum_{u=1}^{m}I_{ui}g'(\alpha p_u^{\mathrm{T}}q_i+(1-\alpha)\sum_{j\in\tau(i)}W_{ij}p_u^{\mathrm{T}}q_j)p_u\times(g(\alpha p_u^{\mathrm{T}}q_i+(1-\alpha)\sum_{j\in\tau(i)}W_{ij}p_u^{\mathrm{T}}q_j)-r_{ui})+(1-\alpha)\sum_{k\in N(i)}\sum_{u=1}^{m}I_{uk}g'(\alpha p_u^{\mathrm{T}}q_k+(1-\alpha)\sum_{j\in\tau(k)}W_{kj}p_u^{\mathrm{T}}q_j)\times(g(\alpha p_u^{\mathrm{T}}q_k+(1-\alpha)\sum_{j\in\tau(k)}W_{kj}p_u^{\mathrm{T}}q_j)-r_{uk})W_{ki}p_u+\lambda_q q_i \tag{11}$$

其中，$g'(x)=\exp(x)/(1+\exp(-x))^2$ 是 $g(x)$的导数。$N(i)$是与项目 o_i 最相似的 k 个近邻。

对应的隐因子适当地根据逆梯度方向做相应调整，其学习过程的公式如下所示：

$$p_u^{(k+1)}\leftarrow p_u^{(k)}-\eta\frac{\partial L}{\partial p_u^{(k)}} \tag{12}$$

$$q_i^{(k+1)}\leftarrow q_i^{(k)}-\eta\frac{\partial L}{\partial q_i^{(k)}} \tag{13}$$

算法 1 基于项目耦合相似度的矩阵分解模型算法的伪代码

Input：用户评分矩阵 E，项目集合 O，迭代次数 iter，用户的正则化参数 λ_p，项目的正则化参数 λ_q，平衡因子 α，学习率 η

Output：用户的隐特征模型 P 和项目的隐特征模型 Q

第一步：//计算项目之间的相似度

1. for each $o_i\in O$ do
2. for each $o_j\in O$ & $o_i\neq o_j$ do
3. 根据式(6)计算对象 o_i 和 o_j 之间的耦合相似度 CIS(o_i,o_j)
4. end for
5. end for

第二步：//迭代学习

6. 赋予 P(0)和 Q(0)任意值，将其初始化并且 iter← 0
7. repeat
8. for u=1 to M and i=1 to N do
9. if $I_{ui}\neq 0$ then
10. 根据式(10)和式(11)计算目标函数 L 对 p_u 和 q_i 的导数
11. 由式(12)和式(13)对 p_u,q_i 沿逆梯度做调整
12. end if
13. end for
14. iter++
15. 直到在测试集上的误差率趋于平稳、增加或者 iter<maxIter
16. return P,Q

4 实验结果与分析

4.1 数据集

为了证明本文提出的基于项目耦合相似度的矩阵分解方法的精确度，用 5 折交叉验证的方法来对数据集进行训练和测试。我们随机地将数据样本分成 5 份，然后迭代地将其中 4 份作为训练集，其余部分作为测试集。

本文采用两个被业界广泛应用的推荐数据集对提出的模型进行实验验证，数据集分别是 MovieLens100K (ML-100K)和 MovieLens1M (ML-1M) (http://www.grouplens.org)。这两个数据集由美国明尼苏达大学计算机科学与工程学院的 GroupLens 研究组提供。

ML-100K 包含 943 名用户对 1682 部电影的十万条评分数据，评分范围是 1－5，ML-1M 中包含了 6040 名用户对 3900 部电影的评分。另外，所有的数据集中的用户至少评价过 20 部电影，它们的数据稀疏程度分别是 0.9369 和0.9553。数据集中除了具有历史评分还提供了额外的关于电影属性的

信息，其中包含电影基因和发行年份，因此此数据集对于基于项目的推荐特别有意义。

4.2 评价标准

本文针对所提模型采用的评价指标是平均绝对误差(*MAE*)和均方根误差(*RMSE*)。*MAE* 和 *RMSE* 分别定义如下：

$$MAE=\frac{\sum_{(u,i)\in r_{test}}|r_{ui}-\hat{r}_{ui}|}{|r_{test}|}$$

$$RMSE=\sqrt{\frac{\sum_{(u,i)\in r_{test}}(r_{ui}-\hat{r}_{ui})^2}{|r_{test}|}}$$

其中，r_{ui} 是真实评分，$\hat{r}_{ui}$ 表示用户 u 对项目 i 评分的预测，r_{test} 是测试集中的用户对项目的评分。*MAE* 和 *RMSE* 的值可以通过计算预测的项目评分与用户对项目的实际评分之间的偏差得到，显然，值越小证明该算法的推荐精度越高。

4.3 与其他典型方法的效果比较

为了说明本文所提基于项目的耦合相似度矩阵分解(CISMF)预测模型的准确性，将其与以下具有代表性的经典模型进行比较。

1)正则化的奇异值分解模型(RSVD)：此模型作为一个经典的基准模型运用正则化的奇异值分解方法[17]。

2)非负矩阵分解模型(NMF)：此模型[8]规定在学习过程中隐因子的更新只能是非负的。

3)概率矩阵分解模型(PMF)：此模型作为一个知名的传统推荐模型由文献[20]提出。

4)贝叶斯概率矩阵分解模型(BPMF)：此模型有效地利用了马尔科夫蒙特卡罗方法并由文献[21]提出。

5)内容加强矩阵分解模型(CBMF)：此模型将相关内容信息直接加入到矩阵分解模型中[16]。

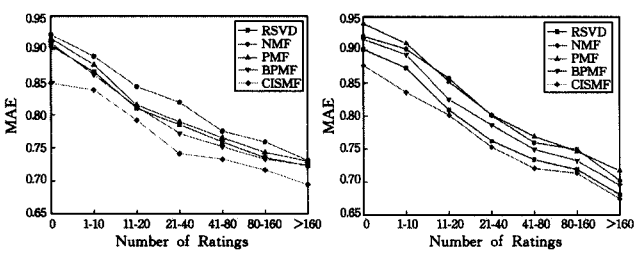
表 1 在 MovieLens 数据集上的结果的比较

Metrics	MovieLens100K						MovieLens1M					
	RSVD	NMF	PMF	BPMF	CBMF	CISMF	RSVD	NMF	PMF	BPMF	CBMF	CISMF
MAE	0.7433	0.7724	0.7522	0.7465	0.7315	**0.7279**	0.6885	0.7286	0.7306	0.7023	0.6932	**0.6814**
RMSE	0.9473	0.9874	0.9667	0.9533	0.9297	**0.9268**	0.8670	0.9203	0.9234	0.8907	0.8630	**0.8592**

本文所提方法的实验参数设置为：$\alpha=0.6$，*Top-k* $=10$，$\lambda_p=\lambda_q=0.001$，$d=5$。表 1 总结了以上模型和本文所提模型在测试集上的结果，可以发现最后一列的结果(即推荐准确性)优于其他方法。原因是本文考虑了 item 之间在属性层面上的耦合关系，现实中这种耦合关系会影响到用户评分。因此在推荐算法中把 item 之间的耦合关系考虑进去，会有效提高推荐算法的准确性。此外，由于维数的增大可能会在学习过程中引入噪声，在此只给出维数为 5 的情况。

4.4 冷启动项目的分析

推荐系统领域的一个亟需解决的问题是冷启动问题。由于当前很少有解决项目评分稀少问题的相关方法，因此将本文中提出的方法与其他对比方法针对不同程度的冷启动项目进行比较。项目根据用户的评分数量情况被分成了 7 组，分别是“0”，“1－10”，“11－20”，“21－40”，“41－80”，“81－160”和“>160”。



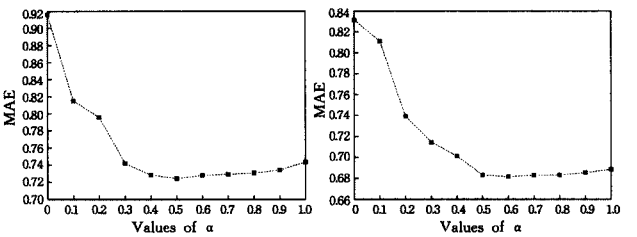
(a)MovieLens100K (b)MovieLens1M

图 3 不同项目的对比

图 3 展示了拥有不同数量评分项目的分类。如图 3 所示，基于项目的耦合矩阵分解(CISMF)方法相对于其他方法取得了较好的效果，从而证明了考虑耦合关系的有效性。

4.5 参数 α 的分析

参数 α 控制项目本身的重要程度和其近邻的贡献程度，本实验将它们融合用于预测评分。通过在 0 到 1 范围内调整参数 α 来探究本文所提模型预测结果的变化趋势。当参数 $\alpha=1$ 时，意味着只依赖于评分矩阵进行分数预测，此时模型等同于 RSVD[17]。当参数 $\alpha=0$ 时，意味着只依赖近邻进行评分预测，不考虑其本身。



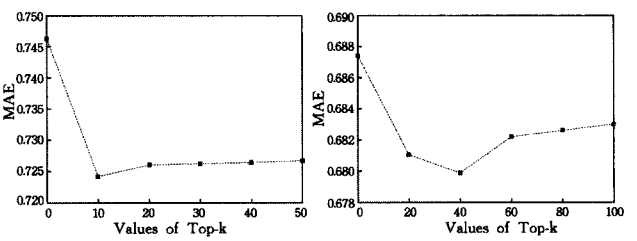
(a)MovieLens100K (b)MovieLens1M

图 4 参数 α 的影响

图 4 展示了融合不同比重的近邻项目的耦合相似度矩阵分解(CISMF)模型表现。起初随着 α 值的增加，*MAE* 的值下降(预测准确性上升)，但当 α 超过一定的阈值后，*MAE* 的值又开始上升(预测准确性下降)，MovieLens 100K 在 $\alpha=0.5$ 时和 MovieLens 1M 在 $\alpha=0.6$ 时分别取得最优结果，由此证明基于耦合关系近邻的特征对于本模型是有价值的。

4.6 近邻的大小的分析

Top-k 的大小表示相似项目的选取个数，它也是影响模型表现的重要因素之一。我们将 MovieLens100K 数据集近邻的选择范围设为 10 到 50，以 10 为步长，参数 $\alpha=0.5$，将 MovieLens1M 数据集近邻的选择范围设为 20 到 100，以 20 为步长，参数 $\alpha=0.6$。



(a)MovieLens100K (b)MovieLens1M

图 5 近邻数量的影响

图 5(a)所示为 MovieLens100K 数据集近邻规模的选取对 *MAE* 的评测效果的影响。从图中可以观察到在 *Top-k* 取值从 10 到 50 的过程中，当 *Top-k* $=10$ 时，误差达到了最小

值，随后误差增大。图 5(b)所示为 MovieLens1M 数据集近邻大小对 *MAE* 效果的影响。*Top-k* 的选取范围为 20～100，随着 *Top-k* 取值的增长，当 *Top-k* 的取值大于 40 后，*MAE* 的值没有呈现出明显的变化。通过分析可以证明，数量相当的近邻能够提供足够的信息，过多的近邻可能带来相对较多的不相关信息，从而使得精度下降。

结束语 本文旨在解决当前研究新项目和用户对其评分较少的项目的冷启动问题的局限。由于项目属性信息能够改善预测评分的准确性，本文将耦合相似度度量方法与矩阵分解技术相结合来优化推荐系统的预测能力。通过分析和实验，证明了耦合相似关系为寻找相似项目提供了较准确的信息。

在接下来的工作中，需要收集更多带有属性信息的相关数据集，以进一步改进算法。同时，由于本文没有探究冷启动用户的表现，因此考虑加入社交关系来丰富推荐框架以及更有效的算法都有待进一步研究。

参 考 文 献

[1] Balabanovic M, Shoham Y. Fab: Content-based collaborative filtering [J]. Communications of the ACM, 1997, 40(3): 66-72

[2] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering [C]// Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. 1998: 43-52

[3] Cao L B, Ou Y, Yu P S. Coupled behavior analysis with applications [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(8): 1378-1392

[4] Gantner Z, Drumond L, Freudenthaler C, et al. Learning attribute-to-feature mappings for cold-start recommendations [C]// Proceedings of the 10th International Conference on Data Mining. 2010: 176-185

[5] Jaschke R, Marinho L, Hotho A, et al. Tag recommendations in folksonomies [C]// Proceedings of the 11th Conference on European Conference on Principles and Practice of Knowledge Discovery in Databases. 2007: 506-514

[6] Hotho A, Jaschke R, Schmitz C, et al. FolkRank: A Ranking Algorithm for Folksonomies [J]. LWA, 2006, 1: 111-114

[7] Koren Y. Collaborative filtering with temporal dynamics [J]. Communications of the ACM, 2010, 53(4): 89-97

[8] Lee D D, Seung H S. Algorithms for non-negative matrix factorization [C]// Proceedings of the 14th Conference on Advances in Neural Information Processing Systems. 2001: 556-562

[9] Middleton E, Shadbolt R, De Roure C. Ontological user profiling in recommender systems [J]. ACM Transactions on Information Systems, 2004, 22(1): 54-88

[10] Linden G, Smith B, York J. Amazon. com Recommendations: Item-to-item Collaborative Filtering [J]. IEEE Internet Computing, 2003, 7(1): 76-80

[11] Lotfi H, Fallahnejad R. Imprecise shannon's entropy and multi attribute decision making [J]. Entropy, 2010, 12(1): 53-62

[12] Levandoski J, Sarwat M, Eldawy A, et al. LARS: A Location-Aware Recommender System [C]// Proceedings of the 28th Conference on Data Engineering. 2012: 450-461

[13] McAuley J, Leskovec J. From amateurs to connoisseurs: Modeling the evolution of user expertise though online reviews [C]// Proceedings of the 22th International Conference on World Wide Web. 2013: 897-908

[14] Ma H, King I, Lyu M R. Learning to recommend with social trust ensemble [C]// Proceedings of the 32th Conference on Research and Development in Information Retrieval. 2009: 203-210

[15] Ma H, Zhou D, Liu C. Recommender system with social regularization [C]// Proceedings of the 4th Conference on Web Search and Data Mining. 2011: 287-296

[16] Nguyen J J, Zhu M. Content-boosted matrix factorization techniques for recommender systems [J]. Statistical Analysis and Data Mining, 2013, 6(4): 286-301

[17] Paterek A. Improving regularized singular value decomposition for collaborative filtering[C]// Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining. 2007: 5-8

[18] Sarwar B, Karypis G, Riedl J. Item-based collaborative filtering recommendation algorithms [C]// Proceedings of the 10th International Conference on World Wide Web. 2001: 285-295

[19] Sarwar M, Karypis G, Konstan J, et al. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering [C]// Proceedings of the 5th International Conference on Computer and Information Technology. 2002

[20] Salakhutdinov R, Mnih A. Probabilistic matrix facotorization [C]// Proceedings of the 20th Conference on Neural Information Processing Systems Foundation. 2007: 1257-1264

[21] Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using markov chain monte carlo [C]// Proceedings of the 25th Conference on International Conference on Machine Learning. 2008: 880-887

[22] Wang J, De Vries A P, Reinders M J T. Unifying user-based and item-based collaborative ltering approaches by similarityfusion [C]// Proceedings of the 29th Conference on Research and Development in Information Retrieval. 2006: 501-508

[23] Song Y, Cao L B, Wu X, et al. Coupled behavior analysis for capturing coupling relationships in group-based market manipulations [C]// Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining. 2012: 976-984

(上接第 223 页)

[13] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2): 191-203

[14] W Ren-xia, Y Xiao-ya, S Xiao-ke. A Weighted Fuzzy Clustering Algorithm for Data Stream[C]// International Colloquium on Computing, Communication, Control, and Management, 2008 (CCCM'08). 2008: 360-364

[15] Jaworski M, Duda P, Pietruczuk L. On fuzzy clustering of data streams with concept drift[C]// Artificial Intelligence and Soft Computing. Springer Berlin Heidelberg, 2012: 82-91

[16] Jiawei H, Micheline K, Jian P. Data Minging: Concepts and Techniques(Third Edition)[M]. San Francisco: Morgan Kaufmann Publishers, 2012: 323-350

[17] Shi Feng, Wang Hui, Yu Lei, et al. Matlab Intelligent Algorithm: Analysis of 30 Cases [M]. Beijing: Beihang University Press, 2011: 188-196(in Chinese)
史峰，王辉，郁磊，等. Matlab 智能算法：30 个案例分析[M]. 北京：北京航天航空大学出版社，2011：188-196

[18] David A. UCI Machine Learning Repository[OL]. http://archive. ics. uci. edu/ml/datasets. html