任务推荐中考虑任务关联度与时间因素的改进OCCF方法

王 刚 王含茹 胡 可 贺曦冉
(合肥工业大学管理学院 合肥 230009)

摘 要 随着众包系统的兴起,人们对众包系统的关注逐渐增多。基于众包系统中的任务推荐,研究者大多将用户对任务的行为数据转化为评分,但没有考虑任务关联关系以及用户兴趣变化对推荐结果的影响。为此,提出一种考虑任务关联度与时间因素的改进OCCF方法,以对任务进行推荐。一方面,在负例抽取阶段引入兴趣遗忘函数,并根据用户活跃度抽取一定数量的负例;另一方面,在概率矩阵分解阶段融合任务相似度信息以进行分解。将所提出的方法应用于众包系统的任务推荐中,利用威客任务中国的数据集进行了实验。实验结果表明,与主流方法相比,所提方法取得了更好的结果,能有效地提高推荐质量。

关键词 任务推荐,推荐系统,OCCF,时间因素,用户兴趣变化
中图法分类号 TP391.3 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.07.030

Improved OCCF Method Considering Task Relevance and Time for Task Recommendation

WANG Gang WANG Han-ru HU Ke HE Xi-ran
(School of Management, Hefei University of Technology, Hefei 230009, China)

Abstract With the development of crowdsourcing system, researchers pay more attention to the crowdsourcing system. Based on the task recommendation of crowdsourcing, most of research scholars convert the behavior data into rate data, without considering the relationship between tasks or the influence caused by the change of user interest on the recommendation results. Therefore, this paper proposed an improved OCCF method considering the task relevance and the time factor to recommend task. On the one hand, this paper introduced a forgetting function when extracting the negative cases, and extracted a certain number of negative cases according to users' activity. On the other hand, it merged the similarity information of tasks in the probability matrix factorization phase. The proposed method was further applied to recommend tasks in the crowdsourcing system. This paper used the data set of Taskcn to conduct experiments. The experimental results show that the proposed method achieves better results, and effectively improves the quality of recommendation compared with the mainstream methods.

Keywords Task recommendation, Recommendation system, OCCF, Time factor, Changes of user interest

1 引言

众包,是指将一项任务以公开呼叫的形式外包给网络上的一些群体或个人,以降低生产成本[1]。近年来,众包系统已经吸引了广泛关注,一些流行的众包系统包括Amazon Mechanical Turk(或MTurk)、CrowdFlower、猪八戒、威客任务中国等。在众包系统中,雇主提交任务请求,用户选择参加并完成任务,并且雇主仅对成功完成任务且满意的用户支付报酬。但众包系统普遍存在一个问题,即雇主数量庞大且每天发布的任务太多,而用户感兴趣的任务只占其中很小的比例,用户很难找到合适的任务。因此需要研究者对用户进行任务推荐。

众包中的任务推荐之所以重要是因为:1)可以提高用户的工作质量。如果用户做是自己感兴趣的任务,则任务完成度会更好。2)可以为众包平台留住人才。如果用户经常找到合适的任务,并成功完成,则会增强用户自信心并吸引该用户继续做更多业务。在个性化推荐中,协同过滤推荐算法是最早被提出的推荐算法,也是目前研究最多、应用最广的推荐技术[2]。协同过滤方法按其处理数据的不同可分为两类:1)处理显式的数据,如用户的评分;2)处理隐式数据,即不能明确表示用户偏好的数据,如网页点击、收藏、购买等,对于此类问题,Pan等将其定义为单类协同过滤问题[3]。单类问题通常用0-1矩阵$R_{m\times n}$表示(m为用户数,n为物品数)。$R_{ij}=1$代表用户i点击(或收藏、购买)了该物品j,表明用户i对物品j的明确喜好。而用户没进行操作的$R_{ij}=0$却可能存在两个原因:1)用户看到了物品j但是不喜欢,所以未进行操作;2)由于物品过多,用户虽然喜欢,但是并没有看到该物品。因此,单类协同过滤往往存在两方面缺陷:1)在训练时,由于只有正样本,缺少负样本,导致最终对未知值的预测结果都是正向的,没有区分性;2)相比于评分数据,单类协同过滤的数据

到稿日期:2017-01-19 返修日期:2017-04-15 本文受国家自然科学基金(71471054,91646111),安徽省自然科学基金(1608085MG150)资助。
王 刚(1980-),男,副研究员,主要研究方向为商务智能与商务分析,E-mail:wgedison@gmail.com(通信作者);王含茹(1994-),女,硕士生,主要研究方向为社会化推荐;胡 可(1996-),男,主要研究方向为任务推荐;贺曦冉(1994-),女,硕士生,主要研究方向为社会化推荐。

往往更为稀疏，会造成分解后的矩阵对特征的提取不够充分[4]。因此，对于单类协同过滤而言，从负例抽取和解决数据高度稀疏性两方面进行改进是一个重要思路。对于已有的任务推荐研究，研究者们大多将用户对任务的行为数据转化为评分数据来进行推荐，而将其考虑为单类协同过滤问题的研究还很少。

目前，许多学者对单类协同过滤问题进行了相关的研究。一方面研究解决负例抽取问题。Zong 等人运用低秩逼近（LRA）技术，把点击数据当作正例，将其他数据都当作负例[5]；Pan 等人提出通过均匀采样、偏重用户采样、偏重项目采样等方式，从未知值中抽取出一个与正例规模类似的集合作为负例[6]；Wang 等人[7]运用 PMF 技术，把点击数据当作正例，将其他数据都当作负例；Hu 等人[8]在设置一个相对较小的权重的前提下，将所有用户的未操作项都当作负例；Paterek[9]针对该问题采用了 SVD 技术；而 Rendle 等人[10]则使用基于 KNN 的协同过滤算法来解决该类问题；Chen 等人[11]通过将随机抽样得到的物品假定为负例来进行推荐。另一方面研究解决数据高度稀疏性问题。如 Li 等人[12]将丰富的用户搜索查询历史、购买和浏览活动等用户信息嵌入单类协同过滤模型；Wang 等人[7]基于传统的矩阵分解模型添加了推荐对象的文本描述信息；Kaya 等人[13]基于传统的最近邻模型添加了属于特定领域的社交网络信息。以上这些方法虽然在一定程度上优化了单类协同过滤，但都没有关注到用户的兴趣变化以及任务关联度对推荐结果产生的影响。

本文在已有研究的基础上，提出了一种考虑任务关联度与时间因素的改进 OCCF 方法（One Class Collaborative Filtering with Time and Task Relevance，OCCF-TTR）。一方面，针对缺少负例的问题，在负例抽取阶段综合考虑了任务相似度与时间因素对负例抽取的影响，即用户未参加的任务与其参加过的任务是否具有相似性，以及用户对参加过的任务的兴趣是否随时间的推移发生了变化。如果任务具有很高的相似性，但其是用户很久以前参加过的，则认为当前用户的兴趣已发生了变化，通过遗忘函数降低其关联度，增加其作为负例的可能性。在此基础上，基于用户活跃度确定要抽取的负例数量，用户越活跃，则认为其在更多情况下看见了任务但不想参加，因此对活跃度高的用户应该抽取更多的负例。另一方面，针对数据高度稀疏的问题，在矩阵分解阶段融入任务相似度信息，构建任务-任务相似度矩阵，将任务-任务相似度矩阵融入用户-任务矩阵中，实施概率矩阵分解，从而得到用户与任务的潜在特征矩阵，最终对结果进行更精确化的预测。为了验证所提方法的有效性，本文在真实的威客任务中国数据集上进行了实验，实验结果表明，与传统的几种方法相比，本文提出的方法在 *MAP*，*MAR* 和 *MAF* 3 个评价指标下均取得了较好的实验结果，提高了推荐质量。

2　问题定义

首先对任务推荐场景进行说明。在众包系统中，雇主在网站上发布悬赏任务，用户选择参加感兴趣的任务。由于任务数量太多，用户无法看到所有的任务，因此，对用户进行任务推荐是当前众包系统的一大关键问题。任务推荐可以提高系统中任务的完成效率，在帮助用户更快地找到其真正感兴趣的任务的同时也能帮助雇主获得更高质量的任务作品。但是，任务推荐不同于传统的商品推荐，用户完成任务的过程中不会对任务给予评级来表示他们对每项任务的青睐程度，因此它缺乏用户的明确偏好。我们在推荐的过程中只能明确正例，即用户参加过的任务，而无法判断用户未参加的任务是其不喜欢的负例还是喜欢但是被遗漏的正例。

本文将此问题形式化定义为 OCCF 的推荐问题。假设推荐系统中存在 M 个用户 $U=\{u_1,u_2,\cdots,u_i,\cdots,u_M\}$，$N$ 个发布的任务 $V=\{v_1,v_2,\cdots,v_k,\cdots,v_j,\cdots,v_N\}$，从用户的角度来说可将任务集 V 中的任务分为两个部分：1）用户已参加过的任务；2）用户未参加过的任务。用户-任务矩阵 $R=\{R_{i,j}\}_{M\times N}$，表示用户参加任务的历史，其中若用户 u_i 曾经参加过任务 v_j，则 $R_{i,j}=1$，说明用户确实对该任务感兴趣，否则 $R_{i,j}$ 为 0，此时无法辨别用户对该任务不感兴趣还是感兴趣但没有看到。任务与任务之间的相似度矩阵可以表示为 $S=\{S_{j,k}\}_{N\times N}$，则 $S_{j,k}$ 的值为任务 v_j 和任务 v_k 描述文本内容之间的相似程度。基于矩阵分解的 OCCF 方法利用矩阵分解模型学习用户和任务的潜在特征向量，然后基于此特征向量预测未知的用户行为。

3　考虑任务关联度与时间因素的单类协同过滤方法

对于本文中的任务推荐问题，我们从负例抽取和概率矩阵分解两个方面对原有方法进行改进。从负例抽取的角度来说，由于只有正样本，缺乏负样本，在进行训练时只有正向的激励，缺乏反向的惩罚，导致训练结果对缺失值的预测都是正向的，结果的区分度低。我们仅能够明确用户已参加的任务确实为用户真正感兴趣的任务，但并不能确定用户没有参加的任务是用户真正不感兴趣的还是实际感兴趣但被遗漏的任务。因此，如何从用户没有参加的任务中准确且合理地选择出负样本进行训练是一个关键问题。同时，在概率矩阵分解的过程中，考虑到任务潜在特征矩阵要受其相似任务的影响，融入任务相似度信息，若两个任务之间的相似度越高，则这两个任务的潜在特征向量应该越相似，若在此基础上，进行联合概率矩阵分解，预测用户对任务的偏好。

3.1　考虑任务关联度与时间因素的负例抽取方法

目前，已有研究主要将所有用户未选择的任务都当作负样本来进行训练（All Missing as Negative，AMAN），将所有用户未选择的项都当作未知项而不进行训练（All Missing as Unknown，AMAU），从所有用户未选择的项中等可能地随机抽取一些项作为负样本进行训练。本文在现有研究的基础上，提出了一种考虑任务关联度与时间因素的负例抽取方法。具体来讲，就是在负例抽取过程中计算任务相似度时融入用户参加任务的时间因素。在用户-任务矩阵 $R=\{R_{i,j}\}_{M\times N}$ 中，1 代表用户实际参加的任务，而未知的值由负例抽取添加且添加到矩阵中的值 $R_{i,j}$ 在 0 到 1 间取值，抽取的负例与用户历史参加过的任务相似度越小，时间间隔越长，则用户真正不想参加的可能性就越大，其 $R_{i,j}$ 值也就越接近 0。

首先，单独计算用户未参加任务 v_j 与已参加的某项任务 v_k 之间的相似度 $s(k,j)$，在此基础上融入用户参加任务 v_k 的时间因素。由于用户的兴趣随时间的推移不断变化，且在较短的一段时间内用户的兴趣是相对稳定的，用户近期参加过

的任务对推荐该用户未来可能想要参加的任务具有比较重要的作用，而早期的参加记录对生成推荐的影响相对较小，因此，我们在比较任务相似度的同时融入时间因素，以提高推荐结果的准确性。于是，在任务相似度计算过程中，引入用户 u_i 对已参加任务 v_k 的非线性遗忘函数[14]：

$$h(t_{i,k})=(1-\theta)+\theta\left(\frac{t_{now}-t_{i,k}}{t_{now}-t_{start}}\right)^2 \tag{1}$$

其中，$0\leqslant\theta\leqslant1$，$0<h(t_{i,k})<1$，$t_{i,k}$ 为用户 u_i 参加任务 v_k 的时间，t_{start} 为用户最早一次参加某项任务的时间，t_{now} 为系统当前的时间，$\frac{t_{now}-t_{i,k}}{t_{now}-t_{start}}$ 表示用户 u_i 参加任务 v_k 的时间与系统当前时间间隔在其整个任务周期（从开始做任务的时间到系统当前时间的时间间隔）中所占的比例。对于用户 u_i 而言，若参加某项任务的时间和当前时间间隔越长，则对该项任务遗忘得越多。θ 为遗忘系数，其值反映了遗忘速度，即 θ 值越大，遗忘速度越快，用户的兴趣变化也就越快。当 $\theta=0$ 时，未对用户已参加过的任务进行非线性遗忘处理；当 $0<\theta<1$ 时，对用户已参加过的任务进行部分非线性遗忘处理；当 $\theta=1$ 时，对用户已参加过的任务进行完全非线性遗忘处理。θ 值的设定应结合推荐系统中的用户兴趣变化的速度，若要兴趣变化快，则 θ 值应设得大一些；否则，反之。基于遗忘函数计算用户 u_i 未参加过的任务 v_j 与用户 u_i 每个已参加过的任务的相似度 $s(k,j)\cdot[1-h(t_{i,k})]$，其中 $s(k,j)$ 为任务 v_k 与任务 v_j 的余弦相似度。于是，计算用户 u_i 未参加的任务 v_j 与用户 u_i 已参加任务集的相似度 $D(i,j)$，则有：

$$D(i,j)=\frac{\sum_{k=1}^{m}s(k,j)\cdot[1-h(t_{i,k})]}{m} \tag{2}$$

其中，m 为用户 u_i 已参加任务的个数，$D(i,j)$ 在单个相似度基础上计算总体加权平均，以此来调节任务 v_j 与用户 u_i 已参加任务集的相似度，提高相似度预测的准确性。基于以上算法，将与用户 u_i 已参加任务集的相似度较小的任务作为负例，而负例的选取数目则根据用户的活跃度来决定。采用公式 $N_i=\beta\cdot\sum R_i$ 来计算，其中 β 表示负正例比例，$\sum R_i$ 表示用户 u_i 已参加的所有任务总数，对于活跃的用户来说，其未参加的任务是将用户在多数情况看见了但未参加，因此对活跃度高的用户应该抽取更多的负例。

基于以上分析，得到本文考虑任务关联度与时间因素的负例抽取算法，如算法 1 所示。

算法 1 考虑任务关联度与时间因素的负例抽取算法

输入：用户-项目矩阵 $R_{i,j}$，用户参加任务的时间 $t_{i,k}$，任务描述文本 w，遗忘系数 θ，权重系数 α，负例抽取比例 β

输出：添加完负例的用户-项目矩阵 $R_{i,j}$

1. For i=1,2,…,M
2. 根据用户-项目矩阵 $R_{i,j}$ 中用户 u_i 已选择的正例，确定 u_i 应抽取的负例数：
   $N_i=\beta\cdot\sum R_i$
3. 初始化一个负例候选列表 list={0}；
4. For j=1,2,…,N
5. 利用 tf-idf 生成各任务关键词的词频向量；
6. If(用户 u_i 未参加过任务 v_j)
7. For(用户 u_i 已参加任务集)
8. 计算任务 v_j 与用户 u_i 已参加任务的余弦相似度 $s(k,j)=\frac{x\cdot y}{\|x\|^2\times\|y\|^2}$，其中 x 为用户 u_i 已参加任务的词频向量，y 为用户 u_i 未参加任务 v_j 的词频向量；
9. 利用式(1)计算用户 u_i 对每个已参加任务 v_k 的兴趣遗忘函数；
10. 计算用户 u_i 未参加过任务 v_j 与用户 u_i 每个已参加过任务 v_k 的相似度：
    $s(k,j)\cdot[1-h(t_{i,k})]$
11. End for
12. 利用式(2)计算 v_j 与用户 u_i 整个已参加任务集的相似度，其中，m 为用户 u_i 已参加任务的个数；
13. End if
14. End for
15. 根据相似度 D(i,j)的值，按照从小到大的顺序选取 N_i 个负例，将 $R_{i,j}$ 中的对应位置置换为 D(i,j)；
16. End for

3.2 考虑任务关联度与时间因素的改进 OCCF 方法

为了对评分信息进行有效建模，引入概率矩阵分解方法（Probabilistic Matrix Factorization，PMF）作为基本的推荐框架。PMF 方法通过对用户-任务矩阵进行分解来推导两个分别表示用户和任务的低维潜在特征矩阵，这些特征是刻画用户和任务的关键因素。本文在原有框架的基础上，融入了任务的相似度信息。具体来说，就是在已添加负例的前提下，根据各任务的描述文本得到各任务的关键词，计算任务与任务之间的文本相似度，据此构造任务-任务相似度矩阵。显然，若两个任务之间的相似度越高，则这两个任务的潜在特征向量应该越相似。之后，将任务-任务关系矩阵融合到用户-任务矩阵中，实施概率矩阵分解，得到用户和任务的潜在特征矩阵。其概率图模型如图 1 所示，其中 U_i 表示用户在潜在特征空间的分布向量，V_j 表示任务在潜在特征空间的分布向量，$S_{j,r}$ 表示任务 V_j 的与任务 V_r 的文本相似度，$D(j)$ 表示 V_j 的相似任务的集合。

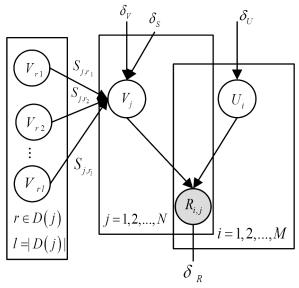

图 1 融合任务关联信息的概率矩阵分解图

Fig. 1 Probabilistic matrix factorization model of fusing task relevance information

如图 1 所示，本文提出的方法在对用户的评分矩阵进行分解以推导出用户和任务的低维潜在特征矩阵的同时，考虑到社会网络中任务的潜在特征矩阵要受其相似任务的影响，将任务相似度关系与用户-任务矩阵这两种信息进行有效结合，实施概率矩阵分解，得到用户潜在特征矩阵 $U\in R^{D\times M}$，任务潜在特征矩阵 $V\in R^{D\times N}$，使 U^TV 的值尽可能逼近用户-任务矩阵 R。其中，U_i 表示用户 u_i 的 D 维特征向量，V_j 表示任务 v_j 的 D 维特征向量。根据以上定义，用户以前参加任务行为的条件概率定义如下：

$$P(R|U,V,\sigma_R^2)=\prod_{i=1}^{M}\prod_{j=1}^{N}[N(R_{i,j}|g(U_i^{\mathrm{T}}V_j),\sigma_R^2)]^{I_{i,j}^R} \quad (3)$$

其中,$N(x|\mu,\sigma^2)$表示均值为 μ、方差为 σ^2 的高斯分布;$I_{i,j}^R$ 是指示函数,如果用户 u_i 参加过任务 v_j,则 $I_{i,j}^R=1$,否则 $I_{i,j}^R=0$;$g(x)=1/(1+\exp(-x))$,作用是将 $U_i^{\mathrm{T}}V_j$ 的值映射在[0,1]区间内。

另外,为了防止过拟合,本文假设 U_i 和 V_j 均服从均值为 0 的高斯分布且相互独立,其中任务的特征向量不仅要服从高斯分布,而且要受到其相似任务的特征向量的影响,即:

$$P(V|S,\sigma_V^2,\sigma_S^2)=\prod_{j=1}^{N}N(V_j|0,\sigma_V^2I)\times\prod_{j=1}^{N}N(V_j|\sum_{r\in D(j)}S_{j,r}V_r,\sigma_S^2I) \quad (4)$$

$$P(U|\sigma_U^2)=\prod_{i=1}^{M}N(U_i|0,\sigma_U^2I) \quad (5)$$

经过贝叶斯推断可以得到 U 和 V 的后验概率分布后如下:

$$\begin{aligned}P(U,V|R,S,\sigma_R^2,\sigma_U^2,\sigma_V^2,\sigma_S^2)&\propto P(R|U,V,\sigma_R^2)P(V|S,\sigma_V^2,\sigma_S^2)P(U|\sigma_U^2)\\&=\prod_{i=1}^{M}\prod_{j=1}^{N}[N(R_{i,j}|g(U_i^{\mathrm{T}}V_j),\sigma_R^2)]^{I_{i,j}^R}\times\prod_{j=1}^{N}N(V_j|0,\sigma_V^2I)\times\\&\quad\prod_{j=1}^{N}N(V_j|\sum_{r\in D(j)}S_{j,r}V_r,\sigma_S^2I)\times\prod_{i=1}^{M}N(U_i|0,\sigma_U^2I)\end{aligned} \quad (6)$$

为了方便求解,对通过式(6)得到的后验概率进行对数处理:

$$\begin{aligned}&\ln P(U,V|R,S,\sigma_U^2,\sigma_V^2,\sigma_R^2,\sigma_S^2)\\&=-\frac{1}{2\sigma_R^2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{i,j}^R(R_{i,j}-g(U_i^{\mathrm{T}}V_j))^2-\frac{1}{2\sigma_S^2}\sum_{j=1}^{N}(V_j-\sum_{r\in D(j)}S_{j,r}V_r)^{\mathrm{T}}(V_j-\sum_{r\in D(j)}S_{j,r}V_r)-\frac{1}{2\sigma_V^2}\sum_{j=1}^{N}V_j^{\mathrm{T}}V_j-\\&\quad\frac{1}{2\sigma_U^2}\sum_{i=1}^{M}U_i^{\mathrm{T}}U_i-\sum_{i=1}^{M}\sum_{j=1}^{N}I_{i,j}^R\ln\sigma_R^2-D\sum_{i=1}^{M}\ln\sigma_U^2-D\sum_{j=1}^{N}\ln\sigma_V^2+P\end{aligned} \quad (7)$$

其中,D 表示特征向量的维度,P 是不依赖参数的常量。求参数固定时 U 和 V 的极大后验概率,相当于最小化以下误差平方和函数:

$$\begin{aligned}E(U,V,R,S)=&\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{i,j}^R(R_{i,j}-g(U_i^{\mathrm{T}}V_j))^2+\frac{\lambda_s}{2}\sum_{j=1}^{N}\\&(V_j-\sum_{r\in D(j)}S_{j,r}V_r)^{\mathrm{T}}(V_j-\sum_{r\in D(j)}S_{j,r}V_r)+\\&\frac{\lambda_v}{2}\sum_{j=1}^{N}V_j^{\mathrm{T}}V_j+\frac{\lambda_u}{2}\sum_{i=1}^{M}U_i^{\mathrm{T}}U_i\end{aligned} \quad (8)$$

其中,$\lambda s=\frac{\sigma_R^2}{\sigma_S^2}$,$\lambda_V=\frac{\sigma_R^2}{\sigma_V^2}$,$\lambda_u=\frac{\sigma_R^2}{\sigma_U^2}$,反映各个矩阵对目标函数的影响程度的大小。对于式(8)所示的目标函数,通过在 U_i,V_j 使用梯度下降的方法进行求解,使其达到局部极小值:

$$\frac{\partial E}{\partial U_i}=\sum_{j=1}^{N}I_{i,j}^R(g(U_i^{\mathrm{T}}V_j)-R_{i,j})g'(U_i^{\mathrm{T}}V_j)V_j+\lambda_uU_i \quad (9)$$

$$\begin{aligned}\frac{\partial E}{\partial V_j}=&\sum_{i=1}^{M}I_{i,j}^R(g(U_i^{\mathrm{T}}V_j)-R_{i,j})g'(U_i^{\mathrm{T}}V_j)U_i+\lambda_s(V_j-\sum_{r\in D(j)}S_{j,r}V_r)-\lambda s\sum_{r\in D(j)}S_{j,r}(V_j-\sum_{r\in D(j)}S_{j,r}V_r)+\lambda_vV_j\end{aligned} \quad (10)$$

基于以上分析,本文提出的融合社会化信息的改进单类协同过滤方法的详细过程如算法 2 所示。

算法 2　改进的 OCCF 算法

输入:矩阵 R 和 S,潜在特征维数 D,正则化参数 λ_u,λ_V,λ_S,学习率 α,最大迭代次数 I

输出:用户和任务的潜在因子矩阵 U 和 V

1. 根据用户活跃度、任务相似度以及时间半衰期进行负例抽取,并将其添加到用户-项目矩阵 R 中;
2. 初始化 U 和 V,生成随机矩阵 U 和 V;
3. For iter=1,2,…,I do:
4.　For each ⟨i,j⟩∈R:
5.　　根据式(9)所求梯度更新 $U_i=U_i-\alpha\frac{\partial E}{\partial U_i}$;
6.　　根据式(10)所求梯度更新 $V_j=V_j-\alpha\frac{\partial E}{\partial V_j}$;
8.　End for
9. End for

3.3　时间复杂度分析

本文提出的算法的时间复杂度主要体现在目标函数与对应的梯度下降计算中。对于任务关联度矩阵的建立,由于数据中共包含 N 个任务,计算两个任务相似度的时间复杂度为 $O(N^2)$。在联合概率矩阵分解阶段,假设平均每个用户参与过的任务数目为 $\bar{t}$,每个任务平均有 $\bar{r}$ 个用户参与。因此在联合概率矩阵分解的求解阶段,每一次迭代过程中梯度计算的时间复杂度为 $O(M\bar{t}K+N\bar{r}K+Nl^2K)$。可以看出,时间复杂度与矩阵中非零数据的数量呈线性关系,故本文提出的融合任务相似度信息的概率矩阵分解可以扩展到更大数据集的应用中,具有良好的可扩展性。

4　实验设计

本节首先介绍实验所用数据集、评价标准以及对比方法,然后给出本文所提方法与其他方法的对比实验结果,并对实验结果进行分析。

4.1　实验数据

为了验证本文提出的考虑任务关联度与时间因素的改进 OCCF 方法(OCCF_TTR)在任务推荐中的有效性,本文使用从威客任务中国[1)]抓取的数据作为实验数据集。威客任务中国是一个任务众包平台。在该网站中,雇主发布任务请求,同时对任务进行详细的描述,以便用户可以辨别自己是否感兴趣,用户可以选择其感兴趣的任务参加并完成,雇主会选择成功完成任务的用户并支付报酬。综上所述,威客任务中国数据集比较适合本文的实验研究。首先,抓取的数据主要是威客任务中国网站中所有的悬赏任务;然后,逐一抓取该悬赏任务的描述文本、参加该任务的所有用户以及任务最终悬赏的用户;接着,抓取该网站中所有的用户,即威客;最后,对抓取的数据进行一定程度的处理,去除没做过任务的用户,剔除无人问津的任务,我们认为其中参加任务次数少于 10 的用户的数据量太少不能作为有效数据,将其剔除。最终得到经过处理的含有 2700 个用户、50029 个任务、7280415 次用户参与任务记录的数据。

4.2　评价标准

本文采用了常用的 *MAP*,*MAR* 和 *MAF* 3 种指标来评价推荐结果的好坏。*MAP* 为平均准确率的均值,*MAR* 为平均召回率的均值,*MAF* 为平均 F-measure 的均值,其中,*MAF* 为 *MAP* 与 *MAR* 的调和平均数。

$$MAP=\frac{1}{|U|}\sum_{u\in U}(\frac{1}{T}\sum_{1\leqslant j\leqslant N}precision(j)\times rec(j)) \quad (11)$$

$$MAR=\frac{1}{|U|}\sum_{u\in U}(\frac{1}{T}\sum_{1\leqslant j\leqslant N}recall(j)\times rec(j)) \quad (12)$$

1) http://www.taskcn.com

$$MAF=\frac{2\times MAP\times MAR}{MAP+MAR} \tag{13}$$

其中，$precision(j)$为 top-j 的准确率。$recall(j)$为 top-j 的召回率。若 V_j 命中，则 $rec(j)=1$；否则，$rec(j)=0$。N 是推荐个数，T 是测试集中用户感兴趣的任务总数。

4.3 对比方法及参数设置

为了验证本文所提方法的有效性，选择了 6 种方法作为对比方法。1）SVD，忽视所有未选择项目，只对正例使用 SVD 方法进行建模[15]；2）PMF，忽视所有未选择项目，只对正例使用 PMF 方法进行建模[16]；3）PMF-Task，忽视所有未选择项目，在传统的 PMF 方法的基础上融入任务关联度信息；4）OCCF-AMAN，将所有未选择项目作为负例，然后使用 PMF 方法进行建模[17]；5）OCCF-RMAN，从用户未选项目中随机抽取一定数量的负例，然后使用 PMF 方法进行建模；6）OCCF-Time，根据用户活跃度确定抽取数量，并根据时间因素确定抽取可能性，以抽取负例，然后使用 PMF 方法进行建模。

在实验过程中，随机选择 80％的实验数据集作为训练集，20％的实验数据作为测试集。同时，为了保证实验结果的可靠性，进行了 10 次实验并取 10 次实验的平均值作为最终结果。另外，将参数设置为 $\lambda_U=\lambda_V=0.001$，$\lambda_S=0.05$，经过试错法反复实验以选择推荐任务个数 n、特征向量的维度 D、负例比例以及兴趣遗忘系数 θ。

5 实验结果与分析

5.1 实验结果

根据上述实验设计，本文方法与对比算法的推荐结果如表 1 所列，其中黑体为最好的结果。

表 1 本文方法与对比算法的推荐结果比较（$n=20$，$D=20$）

Table 1 Comparison results between proposed method and comparitive methods ($n=20$, $D=20$)

模型	*MAP*	*MAR*	*MAF*
SVD	0.16	0.26	0.20
PMF	0.19	0.33	0.24
PMF-Task	0.23	0.36	0.28
OCCF-AMAN	0.25	0.42	0.31
OCCF-RMAN	0.20	0.34	0.25
OCCF-Time	0.26	0.45	0.33
OCCF-TTR	**0.29**	**0.47**	**0.36**

由表 1 可知，本文提出的方法在 *MAP*，*MAR*，*MAF* 3 个评价指标上均优于其他 6 种推荐方法，说明了本文提出的方法在任务推荐应用中是有效的。在 SVD，PMF 和 PMF-Task 3 种方法的比较中，PMF-Task 在 3 个评价指标上均表现较好；在 PMF，OCCF-AMAN 和 OCCF-RMAN，OCCF-Time 4 种方法的比较中，OCCF-Time 在 3 个评价指标上也均表现较好。这一结果分别表明，融入任务关联度的概率矩阵分解与融入时间因素的负例抽取均有利于更高质量的推荐，进一步验证了本文提出的方法 OCCF-TTR 的有效性。

5.2 结果分析与参数讨论

在本文所提出的 OCCF-TTR 中，推荐任务个数 n、用户、任务的潜在特征维数 D、负例抽取比例 β 以及兴趣遗忘系数 θ 都会对推荐结果产生重要影响。因此，接下来进一步研究如何选择适当的参数来进行单类协同过滤推荐。以下实验均在威客任务中国数据集上进行。

1）推荐任务个数 n 对方法的影响

对任务推荐来说，我们的最终目的是在推荐列表中尽可能多地给用户展示其感兴趣的任务，因此，确定一个最佳的推荐任务个数 n 具有重要意义。由图 2 可以看出，随着 n 的增加，推荐效果 *MAF* 先越来越大，但当大于 20 后，推荐系统的 *MAF* 反而降低。同时，图 2 也表明了本文提出的方法在不同的推荐个数下都能取得较好的结果。

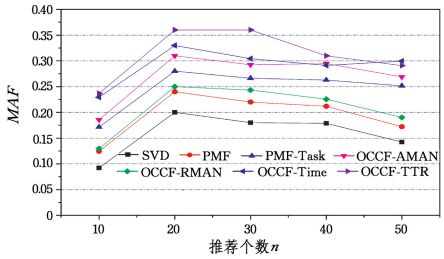

图 2 推荐个数 n 对方法的影响

Fig. 2 Influence of recommendation number n on different methods

2）潜在特征维数 D 对方法的影响

用户与任务的潜在特征维数 D 的大小对推荐结果有重要影响。特征维数过小则无法完全表达用户、任务的隐性特征，过大则会增加计算复杂度，造成过拟合。由图 3 可以看出，随着 D 的缓慢增加，推荐结果的 *MAF* 不断增大，但当特征维数大于 20 后，*MAF* 的增长速率慢慢降低，虽然此时继续增加特征维数 D 会使结果改善，但效果并不显著，且会使模型过拟合以及增加计算复杂度，这会对推荐结果产生不良影响。

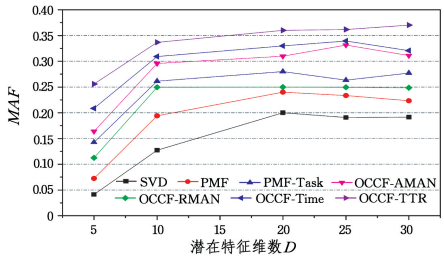

图 3 潜在特征维数 D 对方法的影响

Fig. 3 Influence of latent feature dimension D on different methods

3）负例比例 β 对方法的影响

针对本文提出的方法，负例比例 β 是影响负例抽取方法的重要因素，其值过大则会导致抽取过多的负例，使训练结果趋于负向。从图 4 中可以看出，当 β 在较小的范围里变化时，随着 β 的增大，推荐结果的 *MAF* 不断增大；当 β 接近 15 时，*MAF* 趋于最大；当 β 大于 15 之后，MAF 开始呈逐渐降低的趋势。

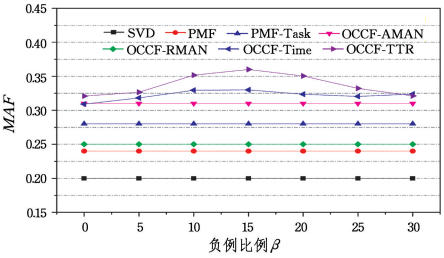

图 4 负例比例 β 对方法的影响

Fig. 4 Influence of negative instance rate β on different methods

4)兴趣遗忘系数 θ 对方法的影响

基于时间因素抽取负例强调了用户近期访问数据对推荐结果的重要性,而遗忘系数 θ 则是其中确定用户兴趣变化的重要影响因素,θ 值过大则会导致样本中本应是正例的项被当作负例,造成训练结果出现偏差。由图 5 可以看出,当 θ 较小时,随着 θ 的增大,推荐结果的 MAF 不断增大,当 θ 在 0.6 附近时,MAF 趋于最大,大于 0.6 之后便又开始降低;无论 θ 的取值如何,本文所提方法都要优于其他几种比较方法,从而说明了本文方法的有效性。

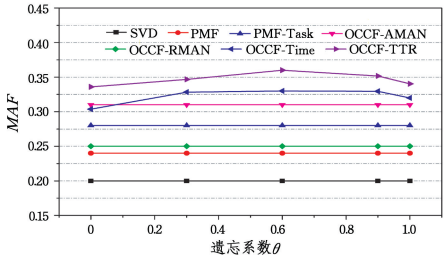

图 5　兴趣遗忘系数 θ 对方法的影响

Fig. 5　Influence of interest forgetting coefficient θ on different methods

结束语　本文针对众包系统中的任务推荐问题,提出了一种考虑任务关联度与时间因素的改进 OCCF 方法。该方法融入时间因素来改进任务相似度之后,根据用户活跃度抽取一定数量的负例,在样本添加负例的基础上,融合任务关联度信息进行概率矩阵分解。本文基于以上方法,利用威客任务中国的数据集进行实验,实验结果表明,本文方法较传统方法具有更高的推荐精度和良好的可扩展性。下一步工作将研究引入更能准确反映用户兴趣变化的遗忘函数,考虑为每个用户的兴趣遗忘速度设定不同的权重、冷启动等其他一些问题,以进一步改善推荐效果。

参 考 文 献

[1] YUEN M C,KING I,LEUNG K S. TaskRec:A Task Recommendation Framework in Crowdsourcing Systems[J]. Neural Processing Letters,2015,41(2):223-238.

[2] LI G,LI L. One-class Collaborative Filtering Based on Matrix Factorization [J]. Application Research of Computers,2012,29(5):1662-1665. (in Chinese)
李改,李磊. 基于矩阵分解的单类协同过滤推荐算法[J]. 计算机应用研究,2012,29(5):1662-1665.

[3] PAN R,SCHOLZ M. Mind the gaps:weighting the unknown in large-scale one-class collaborative filtering[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris,France,2009:667-676.

[4] WANG P,JING L P. Improved One-Class Collaborative Filtering for Recommendation System [J]. Journal of Frontiers of Computer Science and Technology,2014,8(10):1231-1238. (in Chinese)
王鹏,景丽萍. 改进的单类协同过滤推荐方法[J]. 计算机科学与探索,2014,8(10):1231-1238.

[5] ZONG W G,KIM J H,LOGANATHAN G V. A New Heuristic Optimization Algorithm: Harmony Search [J]. Simulation Transactions of the Society for Modeling & Simulation International,2001,76(2):60-68.

[6] PAN R,ZHOU Y,CAO B,et al. One-Class Collaborative Filtering[C]// Eighth IEEE International Conference on Data Mining. IEEE,2008:502-511.

[7] WANG C,BLEI D M. Collaborative topic modeling for recommending scientific articles[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,San Diego,Ca,Usa,August. 2011:448-456.

[8] HU Y,KOREN Y,VOLINSKY C. Collaborative Filtering for Implicit Feedback Datasets[C]// Eighth IEEE International Conference on Data Mining. IEEE,2008:263-272.

[9] PATEREK A. Improving regularized singular value decomposition for collaborative filtering[OL]. http://www.researchgate.net/publication/228629951_Improving_-_regularized_singular_value_decomposition_for_collaborative_fittering.

[10] RENDLE S,FREUDENTHALER C,GANTNER Z,et al. BPR:Bayesian personalized ranking from implicit feedback[C]// Conference on Uncertainty in Artificial Intelligence. AUAI Press,2009:452-461.

[11] CHEN K,CHEN T,ZHENG G,et al. Collaborative personalized tweet recommendation[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,2012:661-670.

[12] LI Y,HU J,ZHAI C X,et al. Improving one-class collaborative filtering by incorporating rich user information[C]// ACM International Conference on Information and Knowledge Management. ACM,2010:959-968.

[13] KAYA H,ALPASLAN F N. Using Social Networks to Solve Data Sparsity Problem in One-Class Collaborative Filtering[C]// Seventh International Conference on Information Technology:New Generations. IEEE,2010:249-252.

[14] ZHU G W,ZHOU L. Hybrid Recommendation Study Based on Forgetting Function and Domain Nearest Neighbor [J]. Journal of Management Sciences in China,2012,15(5):55-64. (in Chinese)
朱国玮,周利. 基于遗忘函数和领域最近邻的混合推荐研究[J]. 管理科学学报,2012,15(5):55-64.

[15] WU Y,LIN S P. SVD Recommendation Model Based on Positive and Negative Feedback Matrices [J]. Computer System Application,2015,24(6):14-18. (in Chinese)
吴扬,林世平. 基于正负反馈矩阵的 SVD 推荐模型[J]. 计算机系统应用,2015,24(6):14-18.

[16] TU D D,SHU C C,YU H Y. Using Unified Probabilistic Matrix Factorization for Contextual Advertisement Recommendation [J]. Journal of Software,2013,24(3):454-464. (in Chinese)
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.

[17] PAPPAS N,POPESCU-BELIS A. Adaptive sentiment-aware one-class collaborative filtering[J]. Expert Systems with Applications,2016,43(C):23-41.