文章编号：1003-0077(2018)05-0080-09

基于信任关系和词相关关系的冷启动用户词特征重建

高亨德[1]，王智强[1]，李 茹[1,2,3]

(1. 山西大学 计算机与信息技术学院，山西 太原 030006；
2. 山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006；
3. 山西大学大数据挖掘与智能技术协同创新中心，山西 太原 030006)

摘 要：文本是社交媒体用户的重要信息之一，从文本中获取用户的词特征是实现用户主题建模、兴趣挖掘及个性化推荐等任务的基础。然而社交媒体中存在许多用户(冷启动用户)只含有少量甚至缺乏文本信息，为此该文提出一种融合用户信任关系及词相关关系的词特征重建方法。该方法通过对用户信任关系矩阵、词相关关系矩阵和用户词特征矩阵进行联合概率矩阵分解来实现对冷启动用户的词特征重建。在新浪微博和 Twitter 的四组数据集上的实验结果表明，该文所提出的冷启动用户词特征重建算法能够取得较好的词特征重建结果。

关键词：社交媒体；词特征；概率矩阵分解；冷启动

中图分类号：TP391 **文献标识码**：A

Word Feature for Cold Start Users Based on Trust Relationships and Word Correlation

GAO Hengde[1], WANG Zhiqiang[1], LI Ru[1,2,3]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;
2. Key Laboratory of Computation Intelligence & Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China;
3. Collaborative Innovation Center of Big Data Mining& Intelligent Technology, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: The user's word feature obtained from the text is the basis for achieving the task of user theme modeling, interest mining, and personalized recommendation. To derive the word feature for cold start users who contain scarcely texts, this paper presents a method of merging the trust relations of the user and the word correlation, Specifically, we combine the user's trust relation matrix, words correlation matrix and the feature word matrix via probabilistic matrix factorization. The experimental results on 4 data sets from Sina microblogging and twitter show that the proposed algorithm achieves better results.

Key words: social media; words feature; probabilistic matrix factorization; cold start

0 引言

近年来，社交媒体逐渐成为了用户分享和传播信息的主要平台。用户可以通过在社交媒体平台上发布视频、图片或文本等来分享和传播信息。其中，文本成为了社交媒体用户表达观点、获取和分享信息的重要内容之一。从文本中可以抽取出用户丰富的词特征信息，对于许多数据挖掘任务如用户主题建模、兴趣挖掘、用户画像及个性化推荐等具有重要作用。Jiang 等人[1]利用上下文信息的词特征对用户进行微博推荐，张晨逸等人[2]利用微博词特征对微博进行主题挖掘，高明等人[3]对热门微博进行实时个性化推荐，程南昌等人[4]利用微博短文本，进行短文本倾向性分析，付博等人[5]利用人工标注微博文本进行用户的消费意图识别，另外一些评测任务中也用到用户历史文本信息。

在社交媒体中，文本作为用户信息的重要方面，

收稿日期：2017-10-20 **定稿日期**：2017-11-28

基金项目：国家 863 计划(2015AA015407)；国家自然科学基金(61772324)

挖掘用户的丰富历史文本信息(如微博/Tweets)是支撑许多社交媒体应用如用户画像、社会化推荐等的重要前提。然而,社交媒体中同时存在大量用户,他们未发表过或者发表过很少量的文本信息,面向此类用户的文本挖掘任务则无法展开。针对此问题,本文试图从词特征层面来重建缺乏历史文本信息的用户词特征,从而为面向社交媒体用户的挖掘任务奠定基础。本文将上述所指的缺乏文本信息的用户称为冷启动用户,本文的具体研究问题称之为冷启动用户词特征重建。

具体而言,本文将冷启动用户的词特征重建问题形式化为:给定 m 个用户的集合 $\boldsymbol{S}=\{u_1,\cdots,u_m\}$ 与 n 个特征词的集合 $\boldsymbol{K}=\{w_1,\cdots,w_n\}$,其中有 m_1 个用户具有丰富的历史文本信息,此类用户能够直接通过现有的词特征提取方法(如 tf、tf-idf 等)获取每个用户的特征词及其特征权重。而用户集合 S 中还包含 m_2 个用户在社交媒体中缺少历史文本信息,无法建立其特征词及特征权重,此类用户即为冷启动用户。本文研究的目标就是针对此类冷启动用户,通过借助社交媒体环境中的其他可用信息重建此类用户的特征词及其特征权重。图 1 为用户—词特征矩阵 $\boldsymbol{R}=[R_{ij}]_{m\times n}$,矩阵的前 m_1 行对应上文所述的 m_1 个用户已建立的词特征向量,而其余 m_2 个冷启动用户的词特征向量为空,需要重建,且 $m=m_1+m_2$。

	w_1	w_2	w_3	w_4	…		w_n
u_1							
u_2							
⋮							
u_{m_1}	?	?	?	?	?	?	?
u_{m_1+1}	?	?	?	?	?	?	?
u_{m_1+2}	?	?	?	?	?	?	?
⋮							
$u_{m_1+m_2}$	?	?	?	?	?	?	?

图 1　用户—词特征矩阵

与该冷启动用户的词特征重建问题类似的研究已有很多,特别是在协同过滤推荐领域中,与推荐领域中的冷启动推荐问题类似,可以借鉴解决推荐系统冷启动问题的算法来解决本问题。

在推荐领域中给定一个用户—商品评分矩阵,我们需要基于历史评分信息对矩阵中的未知评分进行填充。但有些用户缺少或只有少量的历史评分信息从而很难依据这些历史信息对其进行评分的填充,这就是协同过滤推荐系统中面临的冷启动用户推荐问题。类似的,本文也可看作对信息的填充问题,但同时,用户特征词的维度高且词特征权重取值不同于评分的取值(1～5),词特征(tf、tf-idf)的值可取范围更广,因此本文所研究的词特征重建问题相对协同过滤推荐系统中的冷启动问题更具挑战性。

已有解决冷启动问题的方法主要有如下三种类型。

(1) 引入辅助信息的推荐

该类方法主要是结合外部数据,如属性数据、文本数据、标签数据等来帮助解决冷启动前提下的推荐问题。例如,于洪等[6]提出充分利用用户时间权重与标签、项目属性、时间等信息,获得个性化推荐评分,实现个性化推荐,解决新项目冷启动问题。Zhang 等[7-8]提出将标签信息应用到推荐算法中,并以三分图的形式来描述用户、项目与标签三者之间的关系,以解决推荐算法的冷启动问题。

Zhang 等[9]结合不同的上下文信息构建预测模型,然后通过协同策略使不同模型之间相互学习,以此解决推荐系统中地址冷启动问题。Wang 等[10]引入非拓扑信息并建立非拓扑信息与拓扑信息的连接来预测最终用户与现有用户之间连接的可能性,来解决冷启动问题。高玉凯等[11]结合用户在其他系统的消费信息,学习用户的潜在特征,然后使用迭代决策树算法训练更优的用户偏好,达到解决冷启动问题的目的。

(2) 基于用户间信任关系的推荐

由于信任关系的可靠性,准确挖掘用户间的信任关系,发现用户的信任用户,是基于信任关系方法的关键问题。Jamali 等[12]将用户的信任传播算法引入矩阵分解方法中,更精确地发现用户在社交网络中的信任用户。郭磊等[13]提出利用信任关系的强度来进一步提高算法的性能,印桂生等[14]提出利用受限的信任关系来约束用户的信任关系矩阵,解决推荐系统中的冷启动问题。Wang 等[15]利用用户间的相似值作为对用户的社会信任关系的约束,通过给用户的信任关系赋予不同权重,选出信任关系较强的信任用户,以此提高算法准确率。

(3) 使用混合方法的推荐

该方法在确定和冷启动用户相似的用户之后,使用混合算法计算相似性或者产生预测评分。如 Wang 等人[16]为解决用户冷启动问题,提出一个使用混合方法的推荐框架。首先,结合用户上下文信

息对用户进行分类，然后，根据分类结果，动态地选择合适的推荐算法，完成推荐。郭等[17]结合用户社会网络数据得出用户信任关系矩阵，然后利用推荐对象间的关联关系进行混合计算，生成共享的用户和推荐对象潜在特征空间，使其同时考虑用户社会关系和推荐对象间的关联关系，完成推荐。

从以上解决冷启动问题的相关研究可看出，如何利用好辅助信息和对用户信任关系权重进行计算，是解决用户冷启动推荐问题的关键。如在推荐时，利用社交媒体中用户间社交关系和用户属性等辅助信息，并约束用户之间的信任度。而以上介绍的这些利用社交关系进行推荐的算法，虽然能很好地对推荐的社会化过程进行建模，但它们在推荐过程中只是单纯地从社交关系的角度对用户信任进行计算，而忽略了推荐对象间的关联关系，Wang 等[15]从评分相似和信任的角度进行了建模，虽然使用用户相似度但没有考虑推荐对象间的关联关系。郭等[17]综合考虑了信任矩阵与推荐对象间的关联关系，但忽略了用户相似度对信任矩阵的影响，且由于用户所发表的特征词之间具有密切的词义或用法上的相关关系，有区别于新项目冷启动问题，不能将其视为独立的个体，故已有推荐对象间关联关系算法不能直接用于解决本文所提出的问题。为此，本文面向冷启动用户的词特征重建研究，提出一种结合用户信任关系和词相关关系的冷启动用户词特征重建方法。该方法中除了利用已有的词特征信息外，还利用用户信任关系和词相关关系信息，并将三种信息通过一种联合概率矩阵分解的方法进行融合，最终实现面向冷启动用户的词特征重建。

本文方法借用矩阵分解方法将用户信任关系矩阵、用户词特征矩阵及词相关关系矩阵这三个矩阵进行联合分解，充分利用这三方面的信息，在低维特征空间上得到用户的隐含特征矩阵及词特征的隐含特征矩阵。方法中的内容我们将在后文中分小节进行详细介绍。

1 基于用户信任关系和词相关关系的词特征重建算法

为了便于本文方法的描述，以下给出本文方法中所用的主要符号及其解释，如表1所示。

表1 符号表示

符 号	解 释
$U=\{u_1,u_2,\cdots,u_m\}$	用户集合
$V=\{v_1,v_2,\cdots,v_n\}$	特征词集合
$\boldsymbol{R}=[r_{ij}]_{m\times n}$	用户词频矩阵，描述 m 个用户的 n 个特征词的词频
$\boldsymbol{W}=[w_{ij}]_{n\times n}$	词相关关系矩阵，描述 n 个特征词之间的相关度，其中 $W_{ij}\in(0,1]$ 表示特征词 t 对特征词 j 的词相关度权重
$\boldsymbol{T}=[t_{ik}]_{m\times m}$	用户信任关系矩阵，描述 m 个用户间的信任关系，其中 $t_{ik}\in(0,1]$ 表示用户 i 对用户 k 的信任权重
$\boldsymbol{U}\in^{l\times m}$	用户特征矩阵，由 l 维列向量描述用户主要特征
$\boldsymbol{Q}\in^{l\times m}$	信任特征矩阵，由 l 维列向量描述用户信任其他用户的主要因素
$\boldsymbol{V}\in^{l\times m}$	词特征矩阵，由 l 维列向量描述不同特征词的主要特征

1.1 信任关系与词相关关系矩阵的构建

本文在 Ma[18]等提出的 SoRec 方法基础上，结合社交网络特点对信任关系权重的计算进行改进，在用户信任关系矩阵构建时，由于在社交网络中用户更倾向于信任其所关注的用户，所以本文只关注单向的用户关注网络，而不关注用户的被关注网络。因此，本文将用户的被关注连边去掉，在用户社交关系图中，假设用户 i 关注用户 j，则 $D_{ij}=1$，$D_{ji}=0$。且加入用户之间的间接信任关系，比如，关注同一个四六级英语老师的两个用户，有可能都在学习四六级相关知识，那么他们之间的经验就可以相互学习并进行推荐。本文将间接信任关系定义，如式(1)所示。

$$T_{ij}=\frac{1}{d}t_{ij} \tag{1}$$

其中，d 是根据宽度优先搜索算法得出的用户 i 和用户 j 的最短路径，当用户 i 到用户 j 的传播路径越长时，用户 i 对用户 j 表现出的局部信任越小。这一点在现实生活中也可以得到验证，即当两个用户越亲密时，他们之间的信任关系也越强烈。t_{ij} 表示算法搜索的总步数。相似度的计算方法有很多，最简单的是欧几里德距离，其他常见的方法有相关相似性(皮尔逊相关系数)、余弦相似性和修正的余弦相似性等。本文使用词向量的夹角余弦来衡量用户相似度，如式(2)所示。

$$D_{ij} = \cos(\text{word}_{u_i}, \text{word}_{u_j}) \tag{2}$$

其中，word_{u_i} 与 word_{u_j} 表示用户 i 与用户 j 的所有特征词的词向量相加的向量值。

本文使用式(3)将用户信任关系和用户相似度相结合构建更加精确的用户信任关系矩阵。构建新的用户信任度矩阵分段函数，如式(3)所示。

$$\begin{cases} 0 & \text{用户 } i \text{ 与用户 } j \text{ 无直接或间接关注} \\ (1-\alpha)\boldsymbol{T}_{ij} + \alpha \boldsymbol{W}_{ij} & \text{用户 } i \text{ 间接关注用户 } j \\ 1 & \text{用户 } i \text{ 直接关注用户 } j \end{cases} \tag{3}$$

用户发表文本除了受信任用户影响，还受词相关关系的影响. 例如，文本中出现“篮球”时，“乔丹”“科比”等出现的概率很大。我们采用如下方法构建词相关关系矩阵。其中，两个特征词之间相关性权重的计算方法为：使用 Word Embedding[19] 得到全部特征词的词向量，遍历每个用户的特征词，将两词的相似度作为特征词之间的相关性权重，将两个特征词 k 与 j 之间的相关性权重定义为 w_{ij}。

1.2　联合概率矩阵分解

Ma 等[17] 提出基于联合概率矩阵分解(UPMF)方法，并把该方法应用于广告推荐领域。本文把 UPMF 方法首次应用于解决冷启动用户词特征重建问题上，它结合三方面的信息进行矩阵分解。在四个数据集上的结果表明，本文算法在解决冷启动用户词特征重建问题上有更高的准确率。

用户词频矩阵 $\boldsymbol{R}$ 的条件概率分布可以定义为[20]：

$$P(\boldsymbol{R} \mid \boldsymbol{U}, \boldsymbol{V}, \delta_R^2) = \prod_{i=1}^{\mathrm{m}} \prod_{j=1}^{n} [N(r_{ij} \mid g(\boldsymbol{U}_i^{\mathrm{T}} \boldsymbol{V}_j), \delta_R^2)]^{\boldsymbol{I}_{ij}^R} \tag{4}$$

其中，$N(x|\mu, \delta^2)$ 表示 x 服从均值为 μ，方差为 δ^2 的高斯分布；$\boldsymbol{I}_{ij}^R$ 是一个指示函数，如果用户 u_i 的特征词 w_i 词频大于 0 则其值为 1，反之为 0。r_{ij} 表示用户 i 的特征词 j 的词频数。函数 $g(x)=1/(1+\exp(-x))$ 的目的是将预测结果限定在区间 $[0,1]$ 之间。假设 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 同样服从均值为 0 的高斯先验分布[21]，则

$$P(\boldsymbol{U} \mid \delta_U^2) = \prod_{i=1}^{m} (\boldsymbol{U}_i \mid 0, \delta_U^2 \boldsymbol{I}) \tag{5}$$

$$P(\boldsymbol{V} \mid \delta_V^2) = \prod_{j=1}^{n} (\boldsymbol{V}_j \mid 0, \delta_V^2 \boldsymbol{I}) \tag{6}$$

为了分析用户间的信任关系和词相关关系是否会影响用户的文本特征词，本文使用共享的用户特征空间将用户间的信任关系与用户词相关信息结合在一起，通过对这两部分信息进行联合概率分解，识别出在词特征上比较相近并且具有社会关系的用户以帮助用户进行词特征的重建。使用的概率图模型如图 2 所示，其中，W_{tj} 表示词相关关系矩阵中元素，R_{ij} 表示用户-特征词矩阵中元素，T_{ik} 表示用户信任关系矩阵中元素，V_t、V_j 表示词特征矩阵中元素，U_i 表示用户特征矩阵中元素，Q_k 表示用户信任特征矩阵中元素。

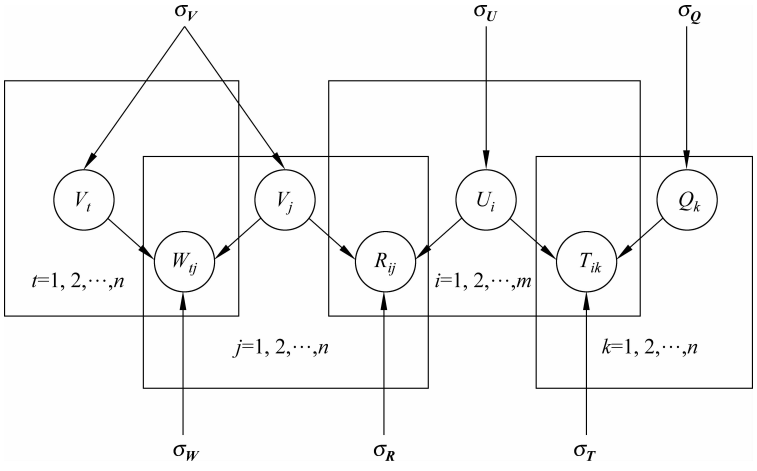


图 2　概率图模型

用户信任关系矩阵表示成用户特征矩阵和信任特征矩阵内积的形式；用户词特征矩阵表示成用户特征矩阵和词特征矩阵内积的形式，词相关关系矩阵表示成不同词特征矩阵内积的形式。

考虑用户间信任关系和词相关关系，经过贝叶斯推断，可以得到 $\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{Q}$ 的后验概率分布如下：

$$\ln P(\boldsymbol{U},\boldsymbol{V},\boldsymbol{Q}\mid\boldsymbol{R},\boldsymbol{W},\boldsymbol{T},\delta_U^2,\delta_V^2,\delta_Q^2,\delta_R^2,\delta_W^2,\delta_T^2)$$
$$=-\frac{1}{2\delta_R^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{I}_{ij}^R(r_{ij}-g(\boldsymbol{U}_i^T\boldsymbol{V}_j))^2-$$
$$\frac{1}{2\delta_T^2}\sum_{i=1}^{m}\sum_{k=1}^{m}\boldsymbol{I}_{ij}^T(d_{ik}-g(\boldsymbol{U}_i^T\boldsymbol{Q}_k))^2-$$
$$\frac{1}{2\delta_W^2}\sum_{t=1}^{n}\sum_{j=1}^{n}\boldsymbol{I}_{tj}^W(w_{tj}-g(\boldsymbol{V}_t^T\boldsymbol{V}_j))^2-$$
$$\frac{1}{2\delta_U^2}\sum_{i=1}^{m}\boldsymbol{U}_i^T\boldsymbol{U}_i-\frac{1}{2\delta_V^2}\sum_{j=1}^{n}\boldsymbol{V}_j^T\boldsymbol{V}_j-$$
$$\frac{1}{2\delta_Q^2}\sum_{k=1}^{m}\boldsymbol{Q}_k^T\boldsymbol{Q}_k-\frac{1}{2}(\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{I}_{ij}^R)\ln\delta_R^2-$$
$$\frac{1}{2}(\sum_{i=1}^{m}\sum_{k=1}^{m}\boldsymbol{I}_{ik}^T)\ln\delta_T^2-\frac{1}{2}(\sum_{t=1}^{n}\sum_{j=1}^{n}\boldsymbol{I}_{tj}^W)\ln\delta_S^2-$$
$$\frac{1}{2}(ml\ln\delta_U^2+nl\ln\delta_V^2+ml\ln\delta_Q^2)+S \tag{7}$$

联合用户信任关系矩阵和词相关关系矩阵的分解可得既满足用户信任关系又满足词相关关系约束的用户特征矩阵，进而由用户特征矩阵和词特征矩阵的内积得到用户词频矩阵中的缺失词频项。其中，S 是和参数无关的常量，求参数固定时 $\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{Q}$ 的极大后验概率，相当于最小化如下误差平方和函数：

$$E(\boldsymbol{R},\boldsymbol{W},\boldsymbol{T},\boldsymbol{U},\boldsymbol{V},\boldsymbol{Q})$$
$$=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{I}_{ij}^R(r_{ij}-g(\boldsymbol{U}_i^T\boldsymbol{V}_j))^2+$$
$$\frac{\lambda_W}{2}\sum_{t=1}^{n}\sum_{j=1}^{n}\boldsymbol{I}_{tj}^W(w_{tj}-g(\boldsymbol{V}_t^T\boldsymbol{V}_j))^2+$$
$$\frac{\lambda_T}{2}\sum_{i=1}^{m}\sum_{k=1}^{m}\boldsymbol{I}_{ik}^T(t_{ik}-g(\boldsymbol{U}_i^T\boldsymbol{Q}_k))^2+$$
$$\frac{\lambda_U}{2}\|\boldsymbol{U}\|_F^2+\frac{\lambda_V}{2}\|\boldsymbol{V}\|_F^2+\frac{\lambda_Q}{2}\|\boldsymbol{Q}\|_F^2 \tag{8}$$

其中 $\lambda_T=\delta_R^2/\delta_T^2$，$\lambda_W=\delta_R^2/\delta_W^2$，$\lambda_U=\delta_R^2/\delta_U^2$，$\lambda_V=\delta_R^2/\delta_V^2$，$\lambda_Q=\delta_R^2/\delta_Q^2$，$\|\cdot\|_F^2$ 表示 Frobenius 范数。

对于式(8)所示的目标函数，我们对 $\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{Q}$ 进行随机初始化，然后在 $\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{Q}$ 上，采用梯度下降法求解最小值，将函数逐步进行迭代，直到达到局部最小值。对目标矩阵 $\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{Q}$ 分别求梯度，如式(9)～式(11)所示。

$$\frac{\partial\boldsymbol{E}}{\partial\boldsymbol{U}_i}=\sum_{j=1}^{n}\boldsymbol{I}_{ij}^Rg'\boldsymbol{U}_i^T\boldsymbol{V}_j(g(\boldsymbol{U}_i^T\boldsymbol{V}_j)-r_{ij})\boldsymbol{V}_j+$$
$$\lambda_T\sum_{k=1}^{m}\boldsymbol{I}_{ik}^Tg'\boldsymbol{U}_i^T\boldsymbol{Q}_k(g(\boldsymbol{U}_i^T\boldsymbol{Q}_k)-t_{ik})\boldsymbol{Q}_k+\lambda_U\boldsymbol{U}_i \tag{9}$$

$$\frac{\partial\boldsymbol{E}}{\partial\boldsymbol{V}_j}=\sum_{i=1}^{m}\boldsymbol{I}_{ij}^Rg'\boldsymbol{U}_i^T\boldsymbol{V}_j(g(\boldsymbol{U}_i^T\boldsymbol{V}_j)-r_{ij})\boldsymbol{U}_i+$$
$$\lambda_W\sum_{t=1}^{n}\boldsymbol{I}_{tj}^Wg'\boldsymbol{V}_t^T\boldsymbol{V}_j(g(\boldsymbol{V}_t^T\boldsymbol{V}_j)-s_{tj})\boldsymbol{V}_t+\lambda_V\boldsymbol{V}_j \tag{10}$$

$$\frac{\partial\boldsymbol{E}}{\partial\boldsymbol{Q}_k}=\lambda_T\sum_{i=1}^{m}\boldsymbol{I}_{ik}^Tg'\boldsymbol{U}_i^T\boldsymbol{Q}_k(g(\boldsymbol{U}_i^T\boldsymbol{Q}_k)-t_{ik})\boldsymbol{U}_i+\lambda_Q\boldsymbol{Q}_k \tag{11}$$

1.3 算法描述

下面以微博用户为例对本文算法进行描述：

输入：用户社会关系矩阵，用户词频矩阵。

输出：微博冷启动用户的用户词频矩阵。

Step1 根据余弦相似度公式计算得到用户相似度矩阵。

Step2 使用宽度优先搜索算法遍历用户社会关系矩阵得到用户之间的直接和间接信任关系 T_{ij}，结合用户之间相似度，根据式(1)计算用户之间的信任关系权重，从而得出用户信任关系矩阵 $\boldsymbol{T}$。需要注意的是，为了简化计算，本文将步数设定在三步以内。

Step3 使用式(2)计算词之间的相关性权重，根据用户词频矩阵得到词相关关系矩阵 $\boldsymbol{W}$。

Step4 将用户信任关系矩阵 $\boldsymbol{T}$ 和词相关关系矩阵 $\boldsymbol{W}$ 进行联合概率矩阵分解，通过梯度下降求得用户特征矩阵 $\boldsymbol{U}$，信任特征矩阵 $\boldsymbol{Q}$ 和词特征矩阵 $\boldsymbol{V}$。

Step5 根据用户特征矩阵 U 和词特征矩阵 V 重建微博冷启动用户词频矩阵，从而重建冷启动用户的词特征。

2 实验结果与分析

2.1 实验数据

本文数据集来源于 Zhang 等[22]提供的新浪微博数据集和 Twitter 数据集。为了满足实验需求，本文分别从这两个数据集中抽取出两个子集用于实验。实验数据中的用户既包含一定规模文本又包含部分社交关系。其中，用户网络抽取方式为：在大数据集中，随机选取一个满足如下约束条件的用户，抽取和其有连边的全部用户，然后抽取和这些用户有连别的用户，逐层抽取，最终得到所需的连通子集。

对子集中全部用户的约束条件如下：

(1) 用户发送和转发微博总数超过 100 条。

(2) 每个用户至少有一条连边，即保证本文所抽取的用户社交关系子图为连通子图。

随后，针对抽取出的用户所发表的社交媒体文本数据进行预处理，抽取出用户特征词。用户特征词抽取的详细步骤如下：

Step1　使用停用词表去掉停用词、标点符号、非中文和非英文字符、中文单字。另外，针对微博文本，本文将停用词表加入了“转发”“分享”“微博”等无意义但出现频率大的词。

Step2　针对社交媒体中文本的特殊性，本文去除社交媒体文本中常见的表情文本和一些网络中特有的符号如“23333”“T_T”等文本，因为文本中的这些符号虽然代表了用户发文时的状态和情绪，但本文的目的在于重建词特征，为后续的文本挖掘任务做铺垫，这些词不具有实在意义，且词频较大，可能为后续挖掘任务增加噪声。

Step3　将词语进行繁体转简体，并将处理完的文本进行分词处理。

Step4　统计词频，去掉词频数小于 5 的词。

Step5　构建用户词频矩阵，使用 tf-idf 计算每个用户的词权重，为了防止矩阵维度过大，本文选取每个用户的 tf-idf 权重排名为前 20 的词作为该用户特征词。

最终得到的数据集包含用户的社交关系数据及用户的特征词及词频数据，数据集的基本特征信息如表 2 所示。

表 2　数据集的基本特征信息

数据集	用户数	边数	特征词数
Sina1	977	3 182	6 011
Sina2	1 378	5 877	3 740
Twitter1	288	2 101	919
Twitter2	337	1 551	1 877

为了验证算法的准确性，将每个数据集分为训练集和测试集，训练集用来学习或训练推荐方法中的相关参数，测试集用来验证推荐的准确性。本文按 9∶1 的比例将数据随机地分为训练集和测试集。将测试集中的用户作为冷启动用户，将训练集中对应的用户词频全部置为 0，然后使用处理后的训练集和测试集进行实验。

2.2　比较方法

为了验证用户间的信任关系和词相关关系在推荐过程中所起到的作用，以及它们对推荐结果产生的影响，在实验中我们选择了五种矩阵分解或其改进算法作为比较算法，分别为 PMF、SoRec、SocialMF、PMFUI 和 TS_MF。

在论文中我们引入概率矩阵分解方法[22] PMF (probabilistic matrix factorization)作为基本比较方法之一。PMF 方法通过对用户-商品的评分矩阵进行分解，得出用户和推荐商品的低维潜在特征矩阵，然后通过随机梯度下降法得出最优的潜在特征矩阵，完成对未知评分的填充，但该方法只利用了用户的评分矩阵信息来对用户和推荐对象的潜在特征进行计算，推荐结果并不是很精确。

Ma 等[18]在 PMF 算法的基础上提出 SoRec 方法，引入用户的社会关系信息。该方法通过对用户社会关系分解学习得出用户社交行为的低维潜在特征信息，并将用户社会关系信息和用户评分信息进行联合分解，识别出在评分上比较相近并且具有社会关系的用户来进行推荐，相较于 PMF，该方法在推荐准确率上有了较大提高。

在 SoRec 算法基础上，Ma[23] 等又提出 SocialMF 算法，加入用户的信任传播，进一步优化用户信任矩阵，使算法能够选出信任度更高的用户，借此提高算法的推荐准确率。但这两种方法只利用了用户社会关系和用户评分这两方面的信息，而未考虑推荐对象间的关系。

在 SoRec 方法基础上，郭等[17] 提出 PMFUI (prob- abilistic matrix factorization with user and item relations)算法，该算法在已有的社会化推荐算法基础上，将推荐对象间的关联关系用于约束共享的用户和推荐对象潜在特征空间的求解，使其同时考虑用户社会关系和推荐对象间的关联关系，从而进一步提高推荐算法的准确率。

Wang 等[15]在利用信任关系算法的基础上，提出 TS_MF 算法，该算法结合用户间的相似关系对用户的社会信任关系增加不同权重，增强对用户信任邻居的计算能力。通过对用户相似度约束的用户信任关系矩阵的分解，得到更精确的信任用户，以此提高算法准确率。

2.3　评价指标

为了评价冷启动用户的词特征重建结果，实验

中借鉴了推荐领域中广泛使用的平均绝对误差MAE(mean absolute error)和均方根误差RMSE(root mean squared error)这两种指标。它们在本文中反映的是冷启动用户的预测词频与实际词频的贴近程度,MAE与RMSE的值越小,表示方法的预测结果越好。计算公式如下:

$$\mathrm{RMSE}=\sqrt{\frac{1}{T}\sum_{i,j}(R_{i,j}-\widehat{R_{i,j}})^2} \tag{12}$$

$$\mathrm{MAE}=\frac{1}{T}\sum_{i,j}|R_{i,j}-\widehat{R_{i,j}}| \tag{13}$$

其中$\widehat{R_{i,j}}$表示由算法预测的用户i的特征词j的词频数,$R_{i,j}$表示用户i的特征词j的真实词频数。

2.4 实验结果

在实验过程中,我们在训练集上尝试不同参数值,然后在测试集上验证结果。经过反复测试,我们发现,实验中的参数设置为:$\lambda_u=\lambda_v=\lambda_a=0.01$时,算法耗时最小,故将以上三个参数设置为0.01。表3给出了潜在特征向量为10的情况下的实验结果。

表3 不同方法的结果比较

方法	RMSE				MAE			
	Sina1	Sina2	Twitter1	Twitter2	Sina1	Sina2	Twitter1	Twitter2
PMF	1.014 5	1.115 1	0.762 8	0.672 1	0.982 8	0.985 1	0.745 1	0.647 8
SoRec	1.006 5	1.104 5	0.745 2	0.665 6	0.978 4	0.984 7	0.731 6	0.642 4
SocialMF	0.996 1	1.091 4	0.746 3	0.664 1	0.976 1	0.983 1	0.732 8	0.641 1
PMFUI	0.997 6	1.094 7	0.741 5	0.653 8	0.972 1	0.978 6	0.731 3	0.646 3
TS_MF	0.993 1	1.094 5	0.722 5	0.655 7	0.971 3	0.981 2	0.728 3	0.654 1
本文方法	**0.976 1**	**1.064 5**	**0.705 6**	**0.640 1**	**0.931 5**	**0.922 4**	**0.696 8**	**0.638 1**

从表中可以看出,相比于其他方法,本文方法在RMSE和MAE指标下取得了较好的结果。而PMFUI与SocialMF的结果相近且优于PMF较多,表明结合用户信任关系与结合词间关系都对结果有较大影响,而PMFUI结果优于SocialMF较少,表明词间关系对结果的影响较小。而TS_MF结果优于PMFUI,表明结合相似度和用户间信任关系的方法对改善实验结果有影响。而本文方法综合考虑了用户间信任关系与词相关关系及用户相似度,对实验结果有较大提升。

2.5 参数影响

图3给出了推荐结果随潜在特征矩阵维度l的变化情况。

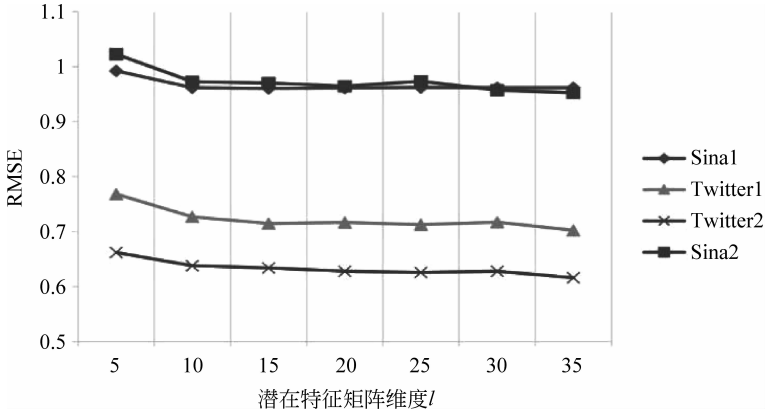


图3 潜在特征矩阵维度l对算法影响

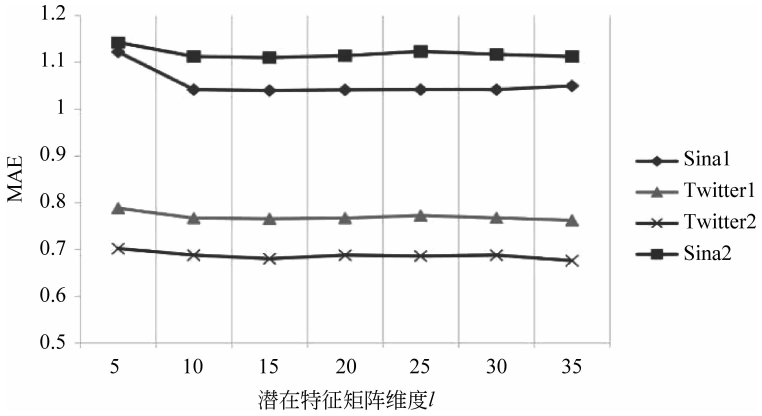


图 3　（续）

图 3 中的两个子图分别表示在四个数据集中，当潜在特征矩阵维度 l 变化时，本文算法对推荐指标 RMSE 值和 MAE 值的影响。由图 3 可知，随着潜在特征矩阵维度 l 增加，对四个数据集而言，RMSE 值和 MAE 值都逐渐减少，之后逐渐趋于稳定，即增加潜在特征矩阵维度可以提高算法的准确率，但同时增大潜在特征矩阵维度会降低算法计算效率，且加大计算开销，通过仔细观察，发现在 l 取值为 $[0,15]$ 时，随着潜在特征矩阵维度的增加 RMSE 值和 MAE 值减小 0.1 左右，而当 l 超过 15 时，RMSE 值和 MAE 值减小幅度不到 0.01，综合算法效率和准确率考虑，本文取 $l=15$ 为最佳维度。

图 4 中的两个子图分别表示四个数据集中，当参数 α 变化时，算法对推荐指标 RMSE 值和 MAE 值的影响。

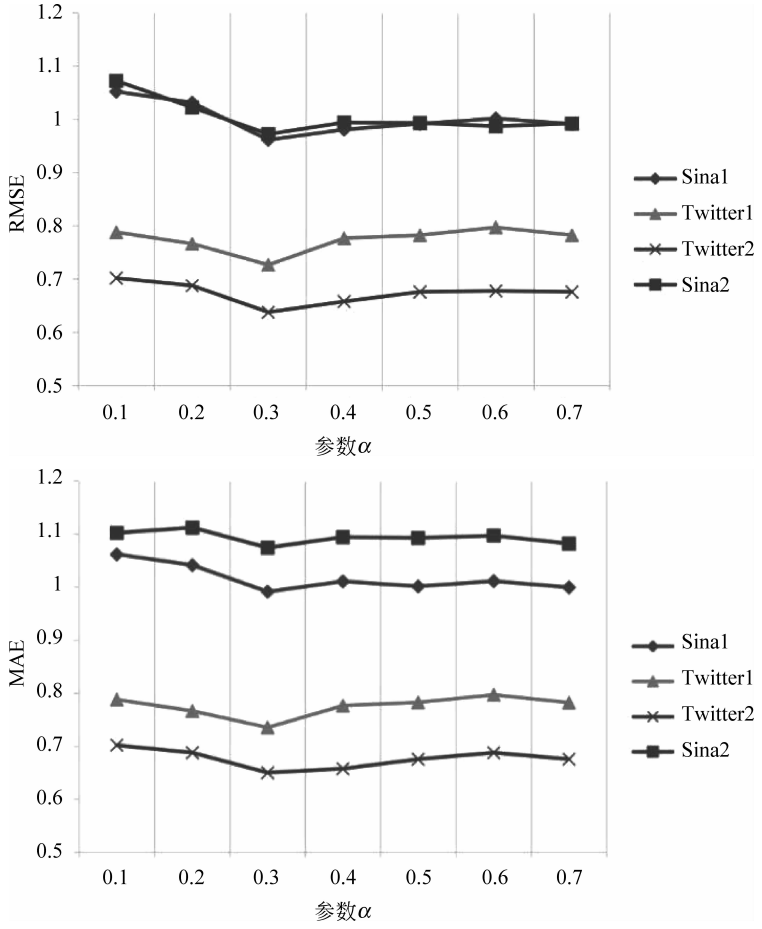


图 4　参数 α 对算法影响

由图4可知，随着参数α增加，推荐误差先减少后增加，之后逐渐趋于稳定，参数α越大表示用户相似度的重要性越大，反之用户社交关系的重要性越大。当$\alpha=0.3$时，四个数据集上的推荐效果均达到最优，另外当α从0.2增加到0.3时算法的RMSE值和MAE值的降低幅度达到0.05左右，推荐效果提升最明显。

3 总结

本文提出一种融合用户信任关系及词相关关系的词特征重建方法，通过对用户信任关系矩阵、用户词频矩阵和词特征相关关系矩阵的联合概率分解，为冷启动用户的词特征进行重建，为冷启动用户的词特征重建研究提供了新思路。未来我们将结合已有的语言知识库如HowNet[24]，Chinese FrameNet[25]等，来提高冷启动用户的词特征重建准确率。

参考文献

[1] Jiang M, Cui P, Liu R, et al. Social contextual recommendation[C]//Proceedings of the ACM International Conference on Information and Knowledge Management. ACM, 2012: 45-54.

[2] 张晨逸，孙建伶，丁轶群. 基于MB-LDA模型的微博主题挖掘[J]. 计算机研究与发展，2011，48(10)：1795-1802.

[3] 高明，金澈清，钱卫宁，等. 面向微博系统的实时个性化推荐[J]. 计算机学报，2014(4)：963-975.

[4] 程南昌，侯敏，滕永林. 基于文本特征的短文本倾向性分析研究[J]. 中文信息学报，2015，29(2)：163-169.

[5] 付博，陈毅恒，邵艳秋，等. 基于用户自然标注的微博文本的消费意图识别[J]. 中文信息学报，2017，31(4)：208-215.

[6] 于洪，李俊华. 一种解决新项目冷启动问题的推荐算法[J]. 软件学报，2015，26(6)：1395-1408.

[7] Zhang Z K, Liu C, Zhang Y C, et al. Solving the cold-start problem in recommender systems with social tags [J]. 2010, 92(2): 28002-28007.

[8] Zi-Ke Zhang, Tao Zhou, Yi-Cheng Zhang. Tag-aware recommender systems: A state-of-the-art survey[J]. 计算机科学技术学报(英文版)，2011，26(5)：767-777.

[9] Zhang M, Tang J, Zhang X, et al. Addressing cold start in recommender systems: A semi-supervised co-training algorithm[C]//Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 2014: 73-82.

[10] Wang Z, Liang J, Li R, et al. An approach to cold-start link prediction: Establishing connections between Non-topological and topological information [J]. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(11): 2857-2870.

[11] 高玉凯，王新华，郭磊，等. 一种基于协同矩阵分解的用户冷启动推荐算法[J]. 计算机研究与发展，2017(8)：1813-1823.

[12] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks[C]//Proceedings of the ACM Conference on Recommender Systems. ACM, 2010: 135-142

[13] 郭磊，马军，陈竹敏. 一种信任关系强度敏感的社会化推荐算法[J]. 计算机研究与发展，2013，50(9)：1805-1813.

[14] 印桂生，张亚楠，董宇欣，等. 基于受限信任关系和概率分解矩阵的推荐[J]. 电子学报，2014，42(5)：904-911.

[15] Wang M, Ma J. A novel recommendation approach based on users' weighted trust relations and the rating similarities[J]. Soft Computing, 2016, 20(10): 3981-3990.

[16] Wang J H, Chen Y H. A distributed hybrid recommendation Frame work to Address the new-user cold-start problem[C]//Proceedings of the Ubiquitous Intelligence and Computing and 2015 IEEE, Intl Conf on Autonomic and Trusted Computing and 2015 IEEE, Intl Conf on Scalable Computing and Communications and ITS Associated Workshops. IEEE, 2016: 1686-1691.

[17] 郭磊，马军，陈竹敏，等. 一种结合推荐对象间关联关系的社会化推荐算法[J]. 计算机学报，2014，37(1)：219-228.

[18] Ma H, Yang H, Lyu M R, et al. SoRec: Social recommendation using probabilistic matrix factorization [C]//Proceedings of the ACM Conference on Information and Knowledge Management, 2008: 931-940.

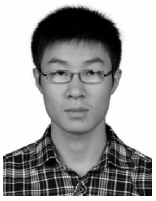
[19] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of International Conference on Neural Information Processing Systems. Curran Associates Inc, 2013: 3111-3119.

[20] Yin D, Hong L, Davison B D. Structural link analysis and prediction in microblogs[C]//Proceedings of ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October. DBLP, 2011: 1163-1168.

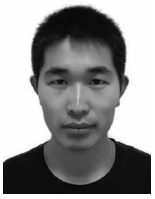
[21] Dueck D, Frey B J. Probabilistic sparse matrix factorization[R]. University of Toronto Technical Report Psi, 2004.

(下转第96页)

熊玲(1997—),学士,主要研究领域为自然语言处理,信息抽取。
E-mail: lingxiong97@gmail.com



王潇斌(1991—),硕士,主要研究领域为自然语言处理,信息抽取。
E-mail: czwangxiaobin@foxmail.com



徐增壮(1993—),硕士,主要研究领域为自然语言处理,信息抽取。
E-mail: nedxuwork@gmail.com

(上接第 88 页)

[22] Zhang J, Liu B, Tang J, et al. Social influence locality for modelingretweeting behaviors[C]//Proceedings of the International Joint Conference on Artificial Intelligence. AAAI Press, 2013: 2761-2767.

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

[24] Li L F, Fan X Z, Li H Q. Domain-specific QA driven by computation of semantic similarity[J]. Journal of Beijing Institute of Technology, 2005, 25(11): 958-962.

[25] Ru L, Wang Z, Li S, et al. Chinese sentence similarity computing based on frame semantic parsing[J]. Journal of Computer Research & Development, 2013, 50(8): 1728-1736.

高亨德(1990—),硕士,主要研究领域为社会媒体数据挖掘、自然语言处理。
E-mail: 476908322@qq.com

李茹(1963—),教授,博士,博士生导师,主要研究领域为中文信息处理、信息检索。
E-mail: liru@sxu.edu.cn

王智强(1987—),博士研究生,主要研究领域为社会媒体数据挖掘、自然语言处理。
E-mail: zhiq.wang@163.com