基于多信息融合表示学习的关联用户挖掘算法

韩忠明[1,2] 郑晨烨[1] 段大高[1] 董 健[3]
(北京工商大学计算机与信息工程学院 北京 100048)[1]
(食品安全大数据技术北京市重点实验室 北京 100048)[2]
(信息网络安全公安部重点实验室公安部第三研究所 上海 200031)[3]

摘 要 随着互联网技术的迅速发展和普及,越来越多的用户开始通过社会网络进行各种信息的分享与交流。网络中同一用户可能申请多个不同账号进行信息发布,这些账号构成了网络中的关联用户。准确、有效地挖掘社会网络中的关联用户能够抑制网络中的虚假信息和不法行为,从而保证网络环境的安全性和公平性。现有的关联用户挖掘方法仅考虑了用户属性或用户关系信息,未对网络中含有的多类信息进行有效融合以及综合考虑。此外,大多数方法借鉴其他领域的方法进行研究,如去匿名化问题,这些方法不能准确解决关联用户挖掘问题。为此,文中针对网络关联用户挖掘问题,提出了基于多信息融合表示学习的关联用户挖掘算法(Associated Users Mining Algorithm based on Multi-information fusion Representation Learning,AUMA-MRL)。该算法使用网络表示学习的思想对网络中多种不同维度的信息(如用户属性、网络拓扑结构等)进行学习,并将学习得到的表示进行有效融合,从而得到多信息融合的节点嵌入。这些嵌入可以准确表征网络中的多类信息,基于习得的节点嵌入构造相似性向量,从而对网络中的关联用户进行挖掘。文中基于3个真实网络数据对所提算法进行验证,实验网络数据包括蛋白质网络PPI以及社交网络Flickr和Facebook,使用关联用户挖掘结果的精度和召回率作为性能评价指标对所提算法进行有效性验证。结果表明,与现有经典算法相比,所提算法的召回率平均提高了17.5%,能够对网络中的关联用户进行有效挖掘。

关键词 关联用户,社会网络安全,表示学习,用户嵌入
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.04.012

Associated Users Mining Algorithm Based on Multi-information Fusion Representation Learning

HAN Zhong-ming[1,2] ZHENG Chen-ye[1] DUAN Da-gao[1] DONG Jian[3]
(School of Computer and Information Engineering,Beijing Technology and Business University,Beijing 100048,China)[1]
(Beijing Key Laboratory of Food Safety Big Data Technology,Beijing 100048,China)[2]
(The Third Research Institute of The Ministry of Public Security,The Ministry of Public Security Key Laboratory of Information Network Security,Shanghai 200031,China)[3]

Abstract With rapid development and popularization of Internet technologies,more and more users have begun to share and exchange various information through social networks. The same user in the network may apply for multiple different accounts to distribute information,and these accounts constitute the associated users in the network. Effectively mining associated users in social networks can suppress false information and illegal behaviors in the network,and thus ensure the security and fairness of the network environment. Existing associated user mining methods only consider user attributes or user relationship information without merging multiple types of information contained in the network comprehensively. In addition,most methods draw lessons from the methods in other fields,such as de-anonymization,and they can't accurately solve the problem of associated user mining. In light of this,this paper proposed an associated user mining algorithm based on multi-information fusion representation learning(AUMA-MRL). In this algorithm,the idea of network representation learning is utilized to learn various dimensional information in the networks,such as user attributes,network topology,etc. Then the learned multi-information is effectively fused to obtain multi-information node embedding,which can accurately characterize multiple types of information in networks,and mine associated users in networks through similarity vectors between node embedding. The proposed algorithm was validated on three real networks namely protein network PPI and social network Flickr,Facebook. In the experiment,the accuracy and recall rate

到稿日期:2018-11-09 返修日期:2019-01-14 本文受国家自然科学基金(61170112)资助。
韩忠明(1972—),男,博士,教授,主要研究方向为社会网络、数据挖掘、大数据处理工作,E-mail:hanzm@th.btbu.edu.cn(通信作者);**郑晨烨**(1994—),女,硕士生,主要研究方向为社会网络、数据挖掘;**段大高**(1976—),男,博士,副教授,主要研究方向为社会计算、多媒体信息处理;**董 健**(1974—),男,博士,高级工程师,主要研究方向为网络数据挖掘、网络安全。

is selected as the performance evaluation indexes. The results show that the recall rate of proposed algorithm is increased by 17.5% on average compared with the existing classical algorithms, and it can effectively mine associated users in networks.

Keywords Associated users, Social networks security, Representation learning, Node embedding

1 引言

社会网络是由不同社会成员之间的互动和联系构成的虚拟社区。随着互联网的快速发展,各种社交网络平台逐渐深入人们生活和工作的方方面面,如 Twitter、豆瓣等,这些社交网络对人类世界具有至关重要的作用;但在信息交互的过程中也存在一些安全隐患。在社交网络中,存在同一用户实体在不同社会网络中注册不同账号的情况,这些不同的虚拟账号被称为马甲或关联用户。关联用户是许多社会热点事件传播的幕后推手,带来了众多的安全隐患,如马甲作弊、水军谣言传播等。如何有效识别关联用户成为社会网络内容安全研究的热点问题。

如图 1 所示,DonnaA 和 DoonaB 是真实社会的同一自然人分别在社会网络 A 和 B 中创建的两个不同账号,因此 DonnaA 和 DoonaB 构成了一对关联用户。将这组关联用户作为桥梁,可以将网络 A 和网络 B 进行连接融合,从而更深刻地理解多社会网络的信息传播模式,也能对跨社会网络的热点事件进行建模。

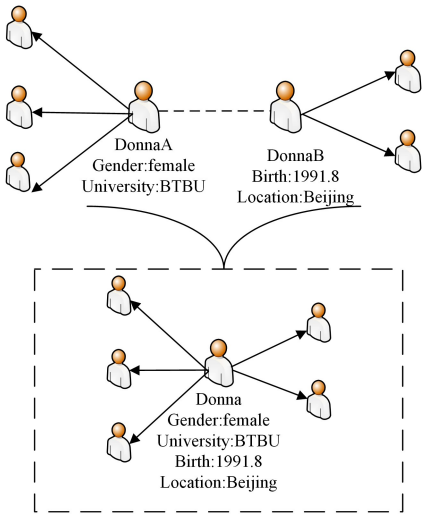

图 1 网络融合示意图

Fig. 1 Schematic diagram of network fusion

传统的关联用户挖掘方法大多通过度量单一社会网络中用户属性信息的相似度进行关联用户的识别。然而,社会网络用户属性的相似性、虚假性和不一致性,可能导致仅使用用户属性信息进行分析易受恶意用户的攻击。利用不同维度的用户行为、用户属性以及网络结构信息构建全面、准确的关联用户挖掘具有很大的挑战性。

大部分关联用户挖掘方法借鉴社会网络中“去匿名化”问题的解决方法,其在一定程度上与关联用户挖掘问题具有一定的相似性,但在实际应用场景中存在较大差别。“去匿名化”问题主要涉及的两个网络在某子网上具有高度相似性;而关联用户挖掘问题中涉及的两个网络的用户及用户关系的重叠度较低,大致在 60%左右[1]。因此,大多数去匿名化方法并不能很好地完成关联用户挖掘任务。文章针对关联用户挖掘问题,提出了一种无监督的学习算法,将社会网络包含的不同类别信息,包括用户属性(文本信息、用户标签等)和用户关系(网络结构、邻域结构)等,嵌入到低维向量空间,并基于空间中的节点嵌入构造了更准确、健壮的关联用户挖掘模型。该算法解决了用户信息缺失、网络平台数据不一致性的问题,提高了关联用户挖掘的安全性。

2 研究现状

2.1 关联用户挖掘

现有的关联用户挖掘方法大都借鉴其他领域的研究,如词共现[2]、实体匹配[3]、命名辨别[4-6]等。这些研究为社会网络关联用户挖掘奠定了很好的基础。关联用户挖掘研究方法主要可分为两类:基于用户属性的关联用户挖掘和基于用户关系的关联用户挖掘。

基于用户属性的关联用户挖掘是传统方法,该类方法使用用户昵称、头像等属性计算用户相似性,将相似性较大的用户对评定为关联用户。Liu 等[7]通过用户名分词及 n-gram 概率计算分词稀有性,构建了无监督的关联用户挖掘方法。Zafarani 等[8]从人类群体局限、个体内外在因素方面对用户命名行为进行特征建模,从而识别未知关联用户。Zhang 等[9]基于用户的内部特征和外部特征,采用朴素贝叶斯进行用户区分,从而得到网络中的关联用户。以上方法在实验数据集上取得了较好的结果,但在真实的大型社会网络中存在着很多具有相似属性信息但并不关联的用户,同时用户可能会提供虚假的信息,恶意用户极易伪造虚假用户,导致模型的健壮性较差。

基于用户关系的关联用户挖掘方法使用社会网络中稠密、可靠且可获取的信息进行关联用户挖掘,如网络拓扑结构等。根据是否需要先验节点,将该类方法分为两类:基于先验节点的关联用户挖掘方法和无先验节点的关联用户挖掘方法。基于先验节点的关联用户挖掘方法根据先验节点定义关联用户之间的相似度,将具有较高相似度的用户判定为关联用户,并通过迭代方法关注越来越多的用户。Narayanan 等[10]根据网络节点出度、入度以及未知关联用户与先验节点的连接情况,构建了用户相似度计算公式,并根据该公式计算用户间的相似度,发现网络关联用户。Zhou 等[11]通过度量真实社会网络的密度以及待融合网络的重叠度得到关联用户与非关联用户之间的概率关系,从而构建了关联用户挖掘模型。上述方法得到的关联用户的质量和数量依赖于先验节点的数量和质量。无先验节点的关联用户挖掘方法大多将问题转化为以下求解过程:以总体相似度最大为目标,建立关联用户挖掘模型及其求解方法。Fu 等[12]认为网络节点相似度可以由其邻居节点间的相似度决定,并采用二分图的一对一最

优匹配算法衡量目标节点的邻居节点的相似度。Signh 等[13]提出基于用户邻居构建用户相似性，根据阈值挖掘关联用户，该方法在蛋白质交互网络的检索中取得了较好的效果。由于不同社会网络在网络结构上的差异较大，存在大量度较小的节点，很难仅通过用户关系进行关联用户挖掘。

综上所述，合理融合用户属性和用户关系是构建准确、全面的关联用户模型的更优方法。但由于用户属性和用户关系的融合存在不一致性，因此综合用户属性和关系的关联用户挖掘方法还处于初步探索阶段。

2.2　网络节点嵌入

对社会网络数据中的有效信息进行挖掘是社会网络分析中的基础问题，其关键是要找到有效、简洁的网络数据表示，即如何将网络中的各类信息通过节点嵌入方法（网络表示学习方法）转化为低维稠密的实数向量。这些节点嵌入可以作为网络分析任务的输入，从而在时间和空间上高效地执行各种网络高级分析任务[14]。近年来，网络表示学习吸引了大量国内外研究者的关注，现在常用的节点嵌入模型主要基于深度神经网络及其变型。如 Wang 等[15]提出的 SDNE 算法使用 Laplace 矩阵监督一阶相似度建模，由无监督的深层自编码器对二阶相似度建模，将深层自编码器的中间层作为节点嵌入。此外，一些方法使用卷积神经网络进行节点嵌入学习，如 Kipf 等[16-17]提出的 GCN 算法针对节点分类问题构建了半监督的节点嵌入模型，来对网络拓扑结构和网络节点特征进行编码，从而学习得到网络节点嵌入。上述方法保留了较为完整、深层的网络节点信息，且在大规模网络的实际应用中取得了良好的效果。

然而上述方法也存在一些问题，如对于网络中含有多种不同类别的信息（节点属性信息、网络结构信息、文本信息等），上述方法只对网络中的某类信息进行学习，具有一定的片面性。此外，不同的节点嵌入方法设计了不同的优化目标，对网络任务的普适性较差。

3　融合用户属性和用户关系的关联用户挖掘算法

关联用户挖掘的目标是：在两个稀疏重叠的网络中发现准确、全面的关联用户。本文基于节点属性、邻域信息以及网络全局结构信息，提出了融合多信息的关联用户挖掘算法 AUMA-MRL。该算法主要分为以下两个步骤：

1）将待融合社交网络中的每个用户看作节点，使用节点嵌入方法分别习得各个节点的嵌入，该嵌入融合了网络的用户属性和用户关系信息；

2）根据用户节点嵌入计算两个待融合网络间用户对的相似向量，该相似向量表示用户之间不同维度的相似性。基于这些相似向量，构建关联用户挖掘算法。

AUMA-MRL 算法的核心是将待融合网络包含的多类信息进行融合并嵌入至低维的向量空间中，通过节点相似性建立两个向量空间的联系，构建关联节点挖掘算法，因此如何得到准确、有效的节点嵌入是本算法的关键。

通常，社交网络中拥有大量共同邻居的节点之间具有更高的相似性。本文通过对网络节点的邻域采样得到网络的局部拓扑结构。AUMA-MRL 首先对目标节点进行 K 阶邻域均匀采样，并设定采样窗口大小为 ω。如图 2(a)所示，以窗口大小 $\omega=2$ 为例对采样过程进行说明。图中虚线和实线之间的连接构成了目标节点的邻居序列，实线为采样到的节点，若目标节点的邻居总数小于 ω，则可重复采样。网络中每个节点都有 n 维特征向量描述节点属性信息（如节点文本信息、标签信息等），通过深度神经网络训练 K 个融合函数 $fusion_k$，$\forall k\in\{1,2,\cdots,K\}$，以习得不同深度邻域的节点属性特征分布，其中每层融合函数对该层采样的邻域信息进行融合，并对不同采样深度的局部邻域信息进行迭代传播。图 2(b)为图 2(a)对应的网络拓扑结构，该图展示了节点邻域信息的融合过程，其中虚线箭头为 $fusion_1$，实线箭头为 $fusion_2$。

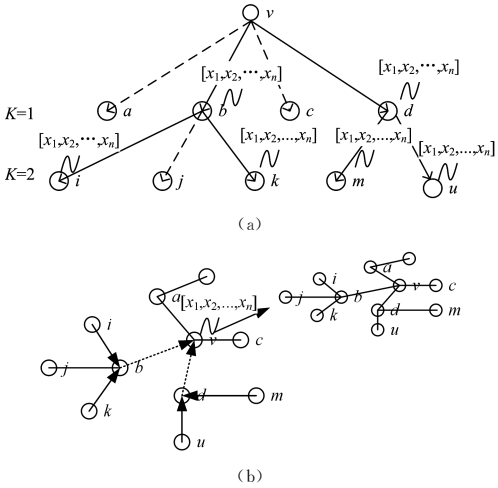

图 2　AUMA-MRL 的采样和融合过程

Fig. 2　Sampling and fusion processes of AUMA-MRL

我们选择池化函数进行节点邻域信息融合，目标节点 v 在邻域深度为 k 时，其邻域信息融合后可表示为 $h_{N(v)}^k=\max(\{\sigma(W_{\text{pool}}h_u^{k-1}+b),\forall u\in N(v)\})$，根据该函数可以将目标节点 v 的每个邻居节点的向量通过全连接神经网络独立传播，最后通过最大池化对 v 的邻域信息进行融合。

将上述邻域信息融合过程所得向量 $hk_{N(v)}$ 与 v 的当前向量表示 h_v^{k-1} 进行级联，通过非线性激活函数 σ 得到 h_v^k，其中 h_v^k 为 $k+1$ 阶邻域融合，提供节点的向量表示。上述邻域信息融合过程为网络中的每个节点习得一个表示向量 $z_v=h_v^K$，$\forall v\in V$，该过程通过节点邻域采样保存了节点邻域的拓扑结构信息，并通过邻域信息融合习得了目标节点的属性信息。

为了得到有效融合用户属性和用户关系的嵌入，使属性、结构相似的节点具有相似的嵌入表示，本文使用基于图的损失函数和梯度下降法来学习融合函数中的参数。基于图的损失函数如式(1)所示，其假设邻近节点具有相似的嵌入，离散的节点具有低相似度的嵌入。

$$L(z_v)=-\log(\sigma(z_v^{\mathrm{T}}z_u))-Q\cdot\mathrm{E}_{u_n\sim P_n(u)}\log(\sigma(-z_v^{\mathrm{T}}z_{u_n})) \tag{1}$$

其中，节点 u 出现在以节点 v 为起点的随机游走序列中，P_n 为负采样分布，Q 为负样本数量，z_v 是由节点 v 局部邻域中的特征生成的融合表示。

由于上述节点邻域信息融合过程只对目标节点的 K 阶邻域进行采样，且采样窗口固定，因此该过程在学习节点属性

信息的同时，间接保存了节点的邻域局部结构信息；但该过程没有保存节点在网络中的全局拓扑结构信息，即完整的用户关系。为了完整、有效地融合用户的属性和关系，本文将邻接矩阵 A 引入损失函数中。由于邻接矩阵 A 表示网络中节点之间的关系，因此该矩阵保存了完整的网络结构信息，即用户关系。邻接矩阵的定义如下：如果节点 i 和节点 j 之间存在链接，则 $A_{ij}=1$，否则 $A_{ij}=0$。通过最大化节点 v 的全局关系特征 A_v 与邻域特征 z_v 的相关性[18]对用户属性和用户关系信息进行融合，从而得到融合用户属性和用户关系的节点嵌入。如式(2)所示，其中邻接矩阵 A 中的每行 A_v 表示节点 v 的全局关系特征。

$$L(z_v)=-\log(\sigma(z_v^{\mathrm{T}}z_u))-Q\cdot \mathrm{E}_{u_n\sim P_n(u)}\log(\sigma(-z_v^{\mathrm{T}}z_{u_n}))-\mathrm{corr}(f_1(A;\theta_1),f_2(z_v;\theta_2)) \tag{2}$$

式(2)中，$f(x)=\theta x+b$，通过求解损失函数可以得到最终的节点嵌入。该嵌入融合了社会网络中用户的属性信息和用户关系信息，将这些信息表示为低维稠密的向量，为关联用户挖掘问题提供了良好的特征基础。上述节点嵌入的产生过程如图 3 所示。

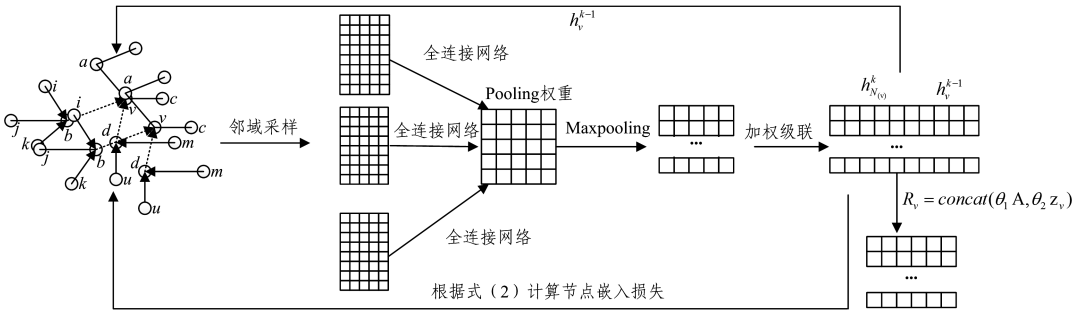

图 3 节点嵌入的产生过程

Fig. 3 Generation process of node embedding

由于同一自然人的社会关系和属性信息在不同社交网络平台中具有一定的相似性，因此本文通过网络间节点对的相似性来判断节点对是否互为关联用户。通过上文得到的节点嵌入可以直接进行节点对间相似性的度量，相似度计算公式如式(3)所示：

$$Sim_{ij}=\sqrt{(R_i^A-R_j^B)(R_i^A-R_j^B)^{\mathrm{T}}} \tag{3}$$

通过 R^A-R^B 计算得到网络间节点的相似性向量，并将节点间的相似度作为相似性向量中的一维。本文使用网络中少量用户账号的已知关联信息作为网络间节点对相似向量的标签，对已标记的节点对构造模型并进行参数训练，得到关联用户挖掘模型，并使用该模型判断无标签节点对是否为关联用户。给定集合 n，其中 n 是从 N_t 中抽取的有真实标签的数据$\{(x_{ij},y_{ij})\}$，x_{ij} 表示用户 i 和用户 j 之间的 D 维相似向量，$y_{ij}\in\{1,-1\}$ 表示两个用户在现实世界中是否为同一自然人。

AUMA-MRL 基于支持向量机[19](SVM)建立了用户对关联挖掘模型 f，用于判断待融合网络间的节点对是否属于同一自然人，如式(4)所示：

$$f(x)=w^{\mathrm{T}}x+b \tag{4}$$

$$L_{(w,b)}=\frac{\gamma_L}{2}\|w\|^2+C\sum\xi_{ij} \tag{5}$$

$$\text{s.t. } y_{ij}(w^{\mathrm{T}}x_{ij}+b)\geqslant 1-\xi_{ij},\xi_{ij}\geqslant 0$$

其中，模型参数 w 和 b 可以通过最小化式(5)中的目标函数得到。目标函数 $L_{(w,b)}$ 为二分类的标准结构损失最小化问题，其中，C 为对误分类的惩罚参数，ξ_{ij} 为保证模型非线性可分的松弛变量，b 为数据的偏差。本文将关联用户挖掘问题转化为基于节点相似向量构造二分类模型。通过分类决策函数 $f(x)$将网络间的节点对分为关联用户和非关联用户两类，从而实现不同网络平台间的关联用户挖掘任务。上述过程如图 4 所示，R_A 和 R_B 分别表示两个网络的节点嵌入矩阵，N_A 和 N_B 分别表示两个网络的节点个数。

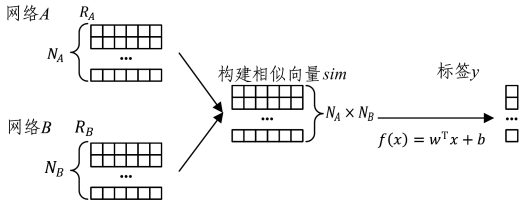

图 4 基于多信息融合表示学习的关联用户挖掘算法 AUMA-MRL

Fig. 4 Associated user mining algorithm based on multi-information fusion representation learning

算法 1 为 AUMA-MRL 算法的完整过程，对于给定的待融合网络集合$\{G_A(V,E),G_B(V,E)\}$，分别遍历其中每一个节点，并对节点的邻域信息进行采样和融合，从而得到邻域特征 Z_v(第 1－9 行)。使用参数 θ_1 和 θ_2 对邻域特征 Z 和全局特征 A 加权，得到完整的节点嵌入 R(第 11 行)。根据两个网络节点嵌入计算节点间的相似度，并根据先验知识构造模型的训练集(第 12－13 行)。寻找如式(4)所示的目标函数的最优参数，构建模型来对网络间的所有节点对进行关联用户挖掘。

算法 1 AUMA-MRL

输入：网络$\{G_A(V,E),G_B(V,E)\}$；节点特征$\{x_v,\forall v\in V\}$；采样深度 K；邻接矩阵 A_A，A_B

输出：关联用户标签 y_{ij}

1. for G_i in $\{G_1(V,E),G_2(V,E)\}$
2. $h_v^0=x_v,\forall v\in V$
3. for k in range(K):
4. for $v\in V$:
5. $h_{N(v)}^k=\max(\{\sigma(W_{pool}\,h_u^{k-1}+b),\forall u\in N(v)\})$
6. $h_v^k=\sigma(W_k\cdot \mathrm{concat}(h_v^{k-1},h_{N(v)}^k))$
7. end for
8. $h_v^k=h_v^k/\|h_v^k\|_2,\forall v\in V$
9. end for

10. $Z_v = h_v^K$

11. $R^i: R_v^K = \text{concat}(\theta_1 \cdot A_v, \theta_2 \cdot Z_v)$

12. end for

13. $S_{ij} = \text{concat}((R_i^A - R_j^B), \sqrt{(R_i^A - R_j^B)(R_i^A - R_j^B)^T})$, $S_{ij} \in S$ base R_i^A and R_j^B

14. Select the candidate pair set n and Initialize y_{ij} for training pairs

15. While the stopping criterion is not reached do

16. find w, b by max Eq. (4)

17. end while

18. for $S_{ij} \in S$:

19. $y_{ij} = w^T S_{ij} + b$

20. end for

AUMA-MRL 算法分为两个部分：首先通过节点嵌入方法习得网络的节点嵌入，然后基于节点嵌入构造节点对的相似向量，并使用二分类模型区分关联用户和非关联用户。综上，AUMA-MRL 算法的复杂度由两个部分组成，第一部分是获取节点嵌入，该阶段的时间复杂度为 $Time \sim O(\sum_N \sum_{l=1}^K \omega_l \cdot d^2 \cdot k)$，其中 N 为网络节点总数，K 为对网络节点进行采样的深度，ω 为节点邻域采样窗口的大小，d 为节点属性特征的维度，k 为习得节点嵌入的维度；第二部分是节点对相似向量分类，该阶段的时间复杂度为 $Time \sim O(M \cdot N_S)$，其中 M 为节点对相似性向量的维度，N_S 为支持向量机算法中的支持向量的数量。因此，AUMA-MRL 算法的时间复杂度为 $Time \sim O(\sum_N \sum_{l=1}^K \omega_l d^2 \cdot k + M \cdot N_S)$。

4　实验分析

为了验证 AUMA-MRL 算法在关联用户挖掘任务下的适用性和有效性，在 3 个真实的公开数据集上进行关联用户挖掘实验。3 个数据集的数据统计如表 1 所列。PPI 为生物蛋白质网络，该网络包含文本信息和节点类别信息，实验使用关联用户挖掘算法在蛋白质网络中进行蛋白质检索任务。Flickr 是著名的照片共享网站，该网站的用户构成了社会网络的人际关系，且该网站提供了用户的标签信息。Facebook 数据是使用 Facebook APP 通过调查参与者收集的，包含用户的多种属性信息。

表 1　数据集信息

Table 1　Dataset information

数据集	节点数	边数	群聚系数
PPI	14755	222109	0.1772
Flickr	80513	5899882	0.1652
Facebook	4039	88234	0.6055

实验分别从 3 个网络中抽取了重叠度分别为 33%，45%，60% 和 80% 的多组重叠网络。网络重叠度使用 $\frac{|X \cap Y|}{|X \cup Y|}$ 度量，其中 X 和 Y 分别表示两个网络的节点集合。因此，当网络有 1/2 的节点相同时，节点重叠度约为 33%。实验中，我们根据一部分用户提供的帐号关联信息从网络间节点对中抽取 20% 的节点对来构造实验的训练集。为了合理评价本文模型，选择目前关联用户挖掘效果较好的 NS 算法和 Grh 算法进行对比，NS 算法和 Grh 算法选用了度值较大的节点作为种子节点，即从网络 top25% 节点度值中选取 10% * N 个节点作为种子节点，其中 N 为网络节点总数。首先使用精度和召回率对关联用户挖掘结果进行评估。对比了当网络重叠度为 60% 时，不同算法在构造的 3 组待融合网络中进行关联用户挖掘的召回率。可以看出，AUMA-MRL 算法在 3 个数据集上都取得了最佳效果，证明了融合用户属性和用户关系信息比仅使用用户关系进行关联用户挖掘的效果更好，AUMA-MRL 算法可以有效地挖掘网络中的关联用户。

表 2　重叠度为 60% 时算法召回率的对比

Table 2　Comparison of algorithm recall rate when overlap is 60%

算法	PPI	Flickr	Facebook
NS	0.3144	0.3250	0.2300
Grh	0.3997	0.1194	0.1157
AUMA-MRL	0.4195	0.3937	0.4291

为了验证所提算法的健壮性，实验分别构造了重叠度为 33%，45%，60% 和 80% 的网络进行对比实验。图 5(a) 对比了两种算法在不同网络重叠度下完成关联用户挖掘任务的精度。从中看出，AUMA-MRL 算法的精度高于 NS 算法。由于网络重叠度的增加使得节点嵌入包含的相似信息更加丰富，因此关联用户的精度会随着网络重叠度的增加而增加。图 5(b) 展示了两种算法在不同网络重叠度下的召回率，结果表明 AUMA-MRL 算法的召回率略高于 NS 算法的召回率，且随着重叠度的增加，召回率不断提高。待融合网络中关联用户的比例比未关联用户的比例低，预测时对提高未关联用户对的召回率的作用较小，因此召回率比精度略低。

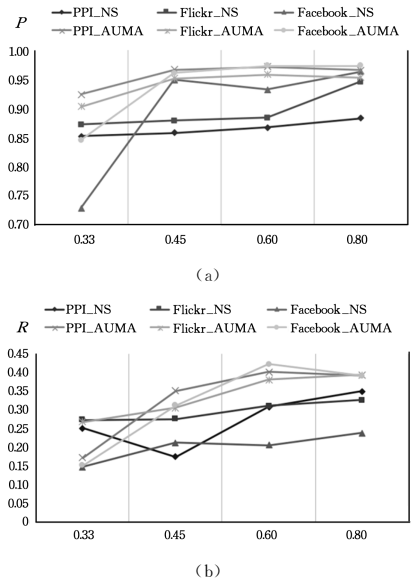

图 5　AUMA-MRL 和 NS 在不同网络重叠度下精度和召回率的对比

Fig. 5　Comparison of precision and recall rate of AUMA-MRL and NS under different network overlap degrees

由上述实验结果可知，本文所提的基于多信息融合表示学习的关联用户挖掘算法 AUMA-MRL 在不同网络重叠度下均具有良好的性能。此外，由于 AUMA-MRL 算法的节点嵌入是通过邻域采样得到的，对于网络中的新增节点，本算法

可以快速得到新节点嵌入以及新节点与网络中其余节点间的相似性向量，从而对网络新增节点关联用户进行快速挖掘，增强了网络关联用户挖掘算法的健壮性。

结束语 基于多信息融合表示学习的关联用户挖掘算法AUMA-MRL，通过网络节点嵌入方法将用户属性和用户关系进行融合，构建了关联用户挖掘模型。该模型可以避免恶意用户的攻击，提升了关联用户挖掘模型的精度和召回率。

由于社会网络中数据量大且用户属性具有相似性、稀疏性、虚假性和不一致性，面向社会网络融合的关联用户挖掘方法面临很多挑战：1)随着先验信息获取难度的增加，如何在无先验或者极少先验的情况下精确地挖掘关联用户，是当前关联用户挖掘的重要研究内容；2)当今社会网络用户规模达到了几千万甚至几亿，现有的许多关联用户挖掘方法由于计算复杂度问题已经不再适用，如何对海量数据下的社会网络进行关联用户挖掘，将是一个重要的研究方向。

参 考 文 献

[1] ZHOU X P，LIANG X，ZHAO J C，et al. A Survey of Related User Mining Methods for Social Network[J]. Journal of Software，2017，28(6)：1565-1583.(in Chinese)
周小平，梁循，赵吉超，等. 面向社会网络融合的关联用户挖掘方法综述[J]. 软件学报，2017，28(6)：1565-1583.

[2] CAI J，STRUBE M. End-to-end coreference resolution via hypergraph partitioning[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics，2010：143-151.

[3] WANG J，LI G，YU J X，et al. Entity matching：How similar is similar[J] Proceedings of the VLDB Endowment，2011，4(10)：622-633.

[4] KALASHNIKOVD V，CHEN Z Q，MEHROTRA S，et al. Web People Search via Connection Analysis[J]. IEEE Transactions on Knowledge and Data Engineering，2008，20(11)：1550-1565.

[5] QIAN Y，HU Y，CUI J，et al. Combining machine learning and human judgment in author disambiguation[C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM，2011：1241-1246.

[6] TANG J，FONG A C M，WANG B，et al. A Unified Probabilistic Framework for Name Disambiguation in Digital Library[J]. IEEE Transactions on Knowledge and Data Engineering，2012，24(6)：975-987.

[7] LIU J，ZHANG F，SONG X，et al. What's in a name? an unsupervised approach to link users across communities[C]// ACM International Conference on Web Search and Data Mining. ACM，2013：495-504.

[8] ZAFARANI R，LIU H. Connecting users across social media sites：a behavioral-modeling approach[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2013：41-49.

[9] ZHANG H，KAN M Y，LIU Y，et al. Online Social Network Profile Linkage[M]// Information Retrieval Technology. Springer International Publishing，2014：197-208.

[10] NARAYANAN A，SHMATIKOV V. De-anonymizing Social Networks[C]// Security and Privacy IEEE Symposium. IEEE，2009：173-187.

[11] ZHOU X，LIANG X，ZHANG H，et al. Cross-Platform Identification of Anonymous Identical Users in Multiple Social Media Networks[J]. IEEE Transactions on Knowledge and Data Engineering，2015，28(2)：411-424.

[12] FU H，ZHANG A，XIE X. Effective social graph deanonymization based on graph structure and descriptive information[C]// ACM Transactions on Intelligent Systems and Technology (TIST)，2015，6(4)：1-29.

[13] SINGH R，XU J B，BERGER B. Global alignment of multiple protein interaction networks with application to functional orthology detection[J]. Proceedings of the National Academy of Sciences of the United States of America，2008，105(35)：12763-12768.

[14] CAI H Y，ZHENG V W，CHANG K. A Comprehensive Survey of Graph Embedding：Problems，Techniques and Applications[J]. IEEE Transactions on Knowledge and Data Engineering，2018，30(9)：1616-1637.

[15] WANG D，CUI P，ZHU W. Structural Deep Network Embedding[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2016：1225-1234.

[16] KIPF T N，WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[J]. arXiv preprint arXiv：1609.02907，2016.

[17] HAMILTON W，YING Z，LESKOVEC J. Inductive representation learning on large graphs[M]// Advances in Neural Information Processing Systems. Bertin：Springer，2017：1024-1034.

[18] ANDREW G，ARORA R，BILMES J，et al. Deep canonical correlation analysis[C]// International Conference on Machine Learning. JMLR. org，2013：1247-1255.

[19] BURGES C J C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery，1998，2(2)：121-167.