·先进计算与数据处理· 文章编号:1000-3428(2018)09-0045-07 文献标志码:A 中图分类号:TP391

基于融合表示学习的跨社交网络用户身份匹配

杨奕卓,于洪涛,黄瑞阳,刘正铭

(国家数字交换系统工程技术研究中心,郑州 450002)

摘 要: 针对现有跨社交网络用户身份匹配算法准确率较低与数据难以获取等问题,提出一种新的跨社交网络用户身份匹配算法。利用已知匹配的账号节点,通过网络融合算法使跨网络问题转化为单一网络问题,对用户名信息进行向量化表示,并与拓扑结构信息向量融合,运用网络表示学习技术,得到融合用户名和拓扑结构 2 种信息的账号节点向量,实现用户身份匹配。实验结果表明,该算法的平均 F1 值达到 79.7%,比传统的机器学习算法及现有 2 种基准算法高 7.3% ~28.8%,有效提升了用户身份匹配的效果。

关键词: 社交网络;用户身份匹配;用户名;信息融合;网络表示学习

中文引用格式: 杨奕卓,于洪涛,黄瑞阳,等. 基于融合表示学习的跨社交网络用户身份匹配[J]. 计算机工程,2018,44(9):45-51.

英文引用格式: YANG Yizhuo, YU Hongtao, HUANG Ruiyang, et al. Cross-social network user identity matching based on fusion representation learning[J]. Computer Engineering, 2018, 44(9):45-51.

Cross-social Network User Identity Matching Based on Fusion Representation Learning

YANG Yizhuo, YU Hongtao, HUANG Ruiyang, LIU Zhengming

(National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002, China)

【Abstract】 Aiming at the problems that the existing cross-social network user identity matching algorithm has low accuracy and difficult data acquisition, a new cross-social network user identity matching algorithm is proposed. Using the known matching account nodes, the network fusion algorithm is used to transform the cross-network problem into a single network problem, and the user name information is vectorized and integrated with the topology information vector, and the network representation learning technology is used to obtain the fusion user name and topology. The account node vector of the two types of information is structured to implement user identity matching. Experimental results show that the average F1 value of the algorithm is 79.7%, which is 7.3% ~28.8% higher than the traditional machine learning algorithm and the existing two benchmark algorithms, it effectively improves the user identity matching effect.

【Key words】 social network; user identity matching; user name; information fusion; network representation learning

DOI: 10.19678/j.issn.1000-3428.0050766

0 概述

随着互联网的普及和快速发展,社交网络应运而生,并且极大地方便和丰富了人们的生活。互联网技术的日益成熟,以及人们的需求日益多元化,使得社交网络软件呈现出了专业细分的趋势,并各自吸引了数量庞大的用户。据统计,Facebook(国外社交网络,其功能类似人人网)至今已拥有超过 14 亿用户,而在国内,仅在 2013 年,新浪微博用户就突破 5.56 亿[1]。由于不同的社交网络有着不同的功能定位,其所提供的服务类型也各不相同。国外对于社交网络的一项调查表明,42% 的用户拥有超过一个社交网络账号,有 93% 的 Instagram(国外一个以图片分享为特色的社交网站)用户在同时使用 Facebook[2],这说明单一的社交网络软件难以满足人们日常工作、生活、娱乐等需求,因此,很多用户会同时选择使用多个社交网络软件。

跨社交网络用户身份匹配就是根据所获取的社交网络数据,识别出多个网络中属于同一个用户的多个账号。这一问题在人们生活中有很重要的现实意义,包括商业网站的推荐系统、社交网络的好友推荐和通信录合并、网络舆论安全等领域都有着极强

基金项目: 国家自然科学基金创新群体项目(61521003)。

作者简介: 杨奕卓(1994—),男,硕士研究生,主研方向为网络大数据分析、社交网络分析;于洪涛(通信作者),研究员、博士;黄瑞阳,副研究员、博士;刘正铭,硕士研究生。

收稿日期: 2018-03-14 **修回日期:** 2018-04-28 **E-mail:** yizhuo_yang@foxmail.com

的研究价值和实际应用。目前针对跨社交网络用户身份匹配技术的研究主要从用户资料信息、用户行为信息以及用户拓扑结构信息 3 个方面入手。在社交网络中,拓扑结构即好友关系,很难进行伪造。另外,在线社交网络具有聚类系数高、节点度服从幂律分布、小世界特性等与复杂网络相类似的特性,可以利用研究复杂网络的方法来分析社交网络问题[3];同时用户名作为一个账号在网络中的标识,相比其他的用户资料也更容易获取。因此,本文选择结合拓扑结构信息和用户名信息进行跨社交网络用户身份匹配,并提出一种基于融合表示学习的跨社交网络用户身份匹配算法。

1 相关工作

当前结合拓扑结构信息和用户名属性信息进行跨社交网络用户身份匹配的研究以有监督方法为主。文献[4]运用拓扑结构和用户名属性进行身份匹配,主要基于拓扑结构信息,通过节点的共同邻居检测匹配的用户身份,同时利用用户名信息进一步提高了匹配效果。文献[5]从用户名属性出发,在身份匹配过程中提出一种基于超图的方法将拓扑结构信息,最后利用排序的方法完成身份匹配。此外,还有一些研究专注于拓扑结构或者用户名属性等。文献[6]基于网络中的重要节点提出了基于好友关系的用户身份匹配算法。文献[7]通过网络表示学习技术对节点进行分布式表示,并且还利用到深度学习技术来提取社交网络中的非线性关系。文献[8-9]则针对用户名属性进行用户身份匹配,分别提出了各自基于多特征融合的用户名字符串相似度算法。

以上方法都对身份匹配的准确率有一定提升,但还存在以下 3 个方面的问题:1)传统的网络节点相似性计算的方法对输入社交网络的数据量比较敏感。在大规模数据的情况下,会存在多个与同一个账号节点相似度高的节点,在匹配时会出现争议,即无法确定与之匹配的节点[10];2)传统的对用户名属性的处理通常将其看做普通字符串,并未考虑到用户名独有的特点,从而会导致原本相似度极高的2 个字符串被判定为不相似;3)利用网络表示学习挖掘拓扑结构信息的算法通常在 2 个网络各自进行分布式表示,再设计一种网络对齐算法[11]将一个向量空间投影到另一个中去[12],这种方法往往会造成信息丢失,从而降低最终匹配的准确率。

针对上述不足,本文提出一种融合拓扑结构和用户名 2 种信息并进行节点信息融合表示学习的算法。首先利用跨社交网络数据集中已知匹配的种子账号对进行网络融合,再将用户名信息根据其子字符串特征进行向量化表示,并与节点拓扑结构初始向量结合,利用网络表示学习技术得到每个节点的拓扑结构和用户名信息相融合的向量化表示,最后通过计算向量相似度,找到匹配的账号对。在多个现实社交网络上的实验结果表明,这种基于信息融合表示学习的跨社交网络用户身份匹配方法综合性能有了显著提高,从而验证了该方法的有效性。

2 用户身份匹配算法

出于表示和实验方便的考虑,本文算法仅以跨 2 个网络的用户身份匹配为例。对于跨多个网络的用户身份匹配问题,可以将其看作由多组跨 2 个网络的用户身份匹配问题组合而成,因此,本文所提供的方法可以很容易地扩展到多网络数据的情境中[1]。

2.1 相关定义

定义 1 定义社交网络 $G=(U,\varepsilon,P)$,其中,$U=\{u_1,u_2,\cdots,u_M\}$是网络中的节点集合,表示网络中所有的账号;$\varepsilon=\{e_{i,j}\}$, $i,j=1,2,\cdots,M(i\neq j)$是网络中的边集合,只有当网络中节点 i 和节点 j 存在好友关系时 $e_{i,j}$才存在;$P=\{p_i\}$, $i=1,2,\cdots,M$ 是网络中节点所对应的用户名集合,p_i 是网络中账号节点 i 的用户名。

定义 2(跨社交网络用户身份匹配) 给定 2 个社交网络 $G^X=(U^X,\varepsilon^X,P^X)$,$G^Y=(U^Y,\varepsilon^Y,P^Y)$和已知匹配的种子结点集 $A=\{(u_i^X,u_j^Y)\mid u_i^X\in U^X,u_j^Y\in U^Y\}$,其中,$(u_i^X,u_j^Y)$表示网络 X 中的账号 i 和网络 Y 中的账号 j 在现实世界中属于同一个用户。利用本文提出的算法,训练一个决策函数 $F:U^X\times U^Y\rightarrow\{0,1\}$ 使得:

$$F(u_i^X,u_j^Y)=\begin{cases}1, u_i^X \text{ 和 } u_j^Y \text{ 属于同一个用户}\\ 0, \text{ 其他}\end{cases}$$

2.2 网络融合算法

为提高跨网络用户身份匹配的准确性,首先要将网络 G^X 和 G^Y 利用已知匹配的种子结点集 A 进行融合。通常情况下在社交网络中存在如下的现象:用户 A 和 B 在网络 X 和 Y 中都有账号,即$F(u_A^X,u_A^Y)=F(u_B^X,u_B^Y)=1$,且 2 个用户在网络 X 中的账号是好友关系,即 $e_{A,B}^X\in\varepsilon^X$,则他们在网络 Y 中往往也存在好友关系,反之亦然。社交网站常常利用这种现象进行好友推荐,在识别出用户身份后为新入网的用户推荐可能认识的人。基于这个现象,本文可以通过网络融合算法,将网络中可能丢失的边进行扩展,图 1 所示为网络融合算法的示意图。

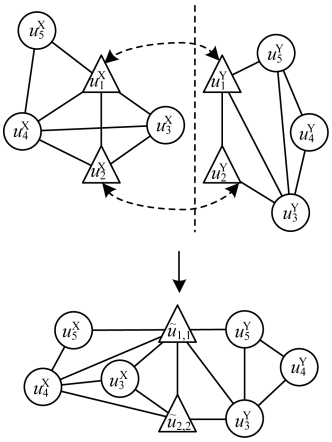


图 1　网络融合算法示意图

这种方法还可以避免现有方法在匹配阶段面临的向量空间对齐问题,提升算法的综合性能。传统方法先对网络 X 和 Y 进行表示学习,得到 2 个向量空间,再利用种子节点集 A 进行网络对齐。本文算法在初始阶段即将网络 X 和 Y 进行融合,相当于先进行网络对齐,再进行向量化表示,并且对拓扑结构进行了合理扩展。记 $N(u_i)=\{u_j \mid e_{i,j}\in\varepsilon\}$是一个社交网络中账号 u_i 的好友集合,则网络融合算法流程如下:

算法 1　跨社交网网络数据的网络融合算法

输入　网络 $G^X=(U^X,\varepsilon^X,P^X)$和 $G^Y=(U^Y,\varepsilon^Y,P^Y)$,已知匹配的种子结点集 A

输出　融合后的网络 $\tilde{G}=(\tilde{u},\tilde{\varepsilon},\tilde{p})$

1. for each(u_i^X,u_j^Y) in A

2. 删去节点 u_i^X 和 u_j^Y,添加新节点 $\tilde{u}_{i,j}$,令新节点对应的用户名 $\tilde{p}_{i,j}=\max(\text{len}(p_i^X,p_j^Y))$

3. for each u_j^X in $N(u_i^X)$

4. 将 u_j^x 与原来 u_i^x 节点之间的连边删去,在 u_j^x 与新节点 $\tilde{u}_{i,j}$之间添加连边

5. for each u_j^Y in $N(u_i^Y)$

6. 将 u_j^Y 与原来 u_i^Y 节点之间的连边删去,在 u_j^Y 与新节点 $\tilde{u}_{i,j}$之间添加连边

7. return 融合后的网络 $\tilde{G}=(\tilde{u},\tilde{\varepsilon},\tilde{p})$

算法 1 选取了原有 2 个账号节点中较长的用户名作为融合产生节点的用户名。由于方便记忆等原因,用户在 2 个社交网络中的用户名往往差异不大,选择较长的用户名可以最大程度保留账号的用户名属性信息。

2.3　用户名信息向量化

为了让用户名信息与节点的拓扑结构信息以向量的方式融合,本节提出对用户名信息进行向量化,向量中的每一个维度都表示用户名字符串的一个特征。在进行用户名信息向量化之前,先对用户名的字母大小写进行统一,然后去掉里面的特殊符号,只保留字母和数字作为用户名属性向量化的目标字符串[13]。随后在特征提取时,统计用户名字符串所包含的“n-gram”及其频数,来表示这个用户名。这里 n 的值不宜过大,因为用户名长度有限,并且过大的 n 会使“n-gram”特征维数过多。为了避免维数过多,取 $n=2$,此时理论上特征维数的最大值为 $(26+10)^2=1\,296$;最后利用 tf-idf 策略,计算每个特征的权值;最终实现用户名属性的向量化。表 1 是一个特征提取的示例,提取出所有“2-gram”出现的频数 *tf*,然后计算每个元素的逆文本频率 *idf*。

表 1　用户名的 2-gram 频数 *tf*

用户名	li	il	ia	an	ly	di	na
lilian	2	1	1	1	0	0	0
lily	1	1	0	0	1	0	0
diana	0	0	1	1	0	1	1

对每个“2-gram”特征 j,其 *idf* 值如下:

$$idf_j=\log_a\frac{|P|}{|j\in p_i\,|\,p_i\in P|} \tag{1}$$

其中,$|P|$为用户名总数,p_i 为 P 中的一个用户名。如对于特征“li”,其在 3 个用户名所组成的用户名集合 P 中的 *idf* 为 $idf_{\text{li}}=\log_a\frac{3}{2}=0.585$,同理得到其他几个特征的 *idf* 值,再与其频数相乘即可得到用户名向量的对应特征的值。例如:

$$\boldsymbol{p}_{\text{lilian}}=(2\times0.585,1\times0.585,1\times0.585,1\times0.585,0\times0.585,0\times0.585,0\times0.585)=(1.170,0.585,0.585,0.585,0,0,0)$$

同理有:

$$\boldsymbol{p}_{\text{lily}}=(0.585,0.585,0,0,1.585,0,0)$$

$$\boldsymbol{p}_{\text{diana}}=(0,0,0.585,0.585,0,1.585,1.585)$$

计算 3 个用户名的向量余弦相似度如表 2 所示,可见 lilian 和 lily 的相似度比 lilian 和 diana 的相似度要高,这也符合直观感受,说明了用户名属性向量化算法的合理性。

表 2　用户名之间的相似度

用户名	lilian	lily	diana
lilian	1	0.371	0.185
lily	0.371	1	0
diana	0.185	0	1

2.4　融合网络的表示学习技术

本节利用网络表示学习算法,在充分挖掘融合网络 $\tilde{G}$ 的拓扑结构信息的同时,将节点的用户名向量作为属性信息加入到表示学习算法的训练过程中,下文给出算法的具体技术。

2.4.1 拓扑结构信息表示学习

网络表示学习算法首先随机选择网络中的一个根节点,并从该节点出发,产生一个节点序列,以此根节点的向量表示来预测其“上下文”中的节点(即以根节点为中心,序列两侧的节点)。对于网络 $G=(U,\varepsilon,P)$ 中的节点 u_i,本文利用网络中的随机游走方法,得到窗口大小为 t 的一个随机游走节点序列 $S_i=\{u_{i-t},u_{i-t+1},\cdots,u_{i+t-1},u_{i+t}\}\backslash u_i$,通过最大化已知节点 u_i 的向量表示时节点序列 S_i 的条件概率,作为根节点对其“上下文”节点的预测,即 $\max\limits_{u_i}\log_a \Pr(\{u_{i-t},u_{i-t+1},\cdots,u_{i+t-1},u_{i+t}\}\backslash u_i|u_i)$,其中 u_i 是已知的节点 i 的拓扑结构向量表示。同时,假设节点之间相互独立,则有:

$$\Pr(\{u_{i-t},u_{i-t+1},\cdots,u_{i+t-1},u_{i+t}\}\backslash u_i|u_i)=\prod_{j=i-t,j\neq i}^{j=i+t}\Pr(u_j|u_i) \tag{2}$$

其中,$\Pr(u_j|u_i)$ 表示已知节点 i 的向量表示时,节点 j 在节点 i 的“上下文”序列中的条件概率。定义 $\Pr(u_j|u_i)$ 如下:

$$\Pr(u_j|u_i)=\frac{\exp(f(u_j,u_i))}{\sum\limits_{k=1}^{|\tilde{U}|}\exp(f(u_k,u_i))} \tag{3}$$

其中,$\tilde{U}$ 表示融合网络中的所有节点集合,$f(u_j,u_i)$ 是 2 个节点 u_i 和 u_j 的结构相似性得分,通常用 2 个节点向量的内积来表示,即 $f(u_j,u_i)=u_j'\cdot u_i$,其中对图中任意节点 i,其向量表示有 u_i 和 u_i',前者表示节点 i 本身的向量表示,后者表示节点 i 作为其他节点的“上下文”序列中节点时的向量表示,即:

$$\Pr(u_j|u_i)=\frac{\exp(u_j'\cdot u_i)}{\sum\limits_{k=1}^{|\tilde{U}|}\exp(u_k'\cdot u_i)} \tag{4}$$

由于利用内积衡量 2 个节点向量的相似程度不容易掌握真实社交网络中的非线性关系,因此本文采用多层感知机架构来定义 $f(u_j,u_i)$ 如下:

$$f(u_j,u_i)=u_j'\cdot g^{(n)}(W^{(n)}(\cdots g^{(1)}(W^{(1)}u_i+b^{(1)})\cdots)+b^{(n)}) \tag{5}$$

其中,n 是多层神经网络的隐藏层层数,$W^{(n)}$ 和 $b^{(n)}$ 分别是第 n 层网络的权值矩阵和偏置向量,$g^{(n)}$ 是第 n 层网络采用的激活函数。

2.4.2 融合信息的表示学习模型

为了将用户名信息融合到拓扑结构信息中,本文将账号用户名字符串转化成了向量 p_i,并将其与对应节点的拓扑结构向量 u_i 进行结合。由于在社交网络中,用户名信息和拓扑结构信息都能在一定程度上反映出节点的独特性,则在向量空间中,拓扑结构向量 u_i 和用户名向量 p_i 都应该对节点 i 的位置产生影响。因此,考虑在本文的多层感知机架构中,将 u_i 与 p_i 进行融合。模型框架如图 2 所示。

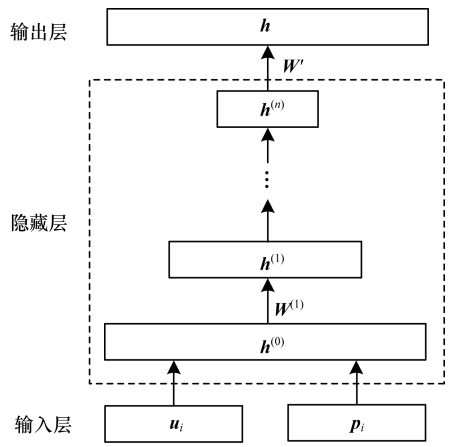


图 2 融合信息的表示学习模型

在图 2 中,$h^{(0)}=[u_i,\rho p_i]^{\mathrm{T}}$ 是隐藏层的输入向量,参数 $\rho\in\mathbb{R}$ 控制用户名属性在节点向量表示中的重要程度,W' 是由所有节点作为“上下文”时的向量组成的矩阵,节点的向量表示 $h_i^{(n)}$ 经由矩阵 W',最终得到当前节点的向量表示,并由此将式(5)写为:

$$f(u_j,u_i)=u_j'\cdot h_i^{(n)} \tag{6}$$

其中,$h^{(k)}=g^{(k)}(W^{(k)}h^{(k-1)}+b^{(k)})$,$k=1,2,\cdots,n$。所以,由式(3)、式(4)得:

$$\Pr(u_j|u_i)=\frac{\exp(u_j'\cdot h_i^{(n)})}{\sum\limits_{k=1}^{|\tilde{U}|}\exp(u_k'\cdot h_i^{(n)})} \tag{7}$$

同时为简便计算,以下用 $\Pr(u_j|u_i)$ 代替 $\Pr(u_j|u_i)$。

2.4.3 模型优化

模型中涉及到的参数 $\Theta=\{\Theta_h,W'\}$,其中,Θ_h 表示隐藏层中的参数,包括权值矩阵和偏置向量。对模型的优化过程就是最大化根节点 u_i 与其随机游走序列 S_i 中节点共现的条件概率,同时最小化与 S_i 之外节点共现的条件概率。对于考虑所有节点的全局网络来说,按下式对 Θ 进行优化:

$$\begin{aligned}\hat{\Theta}&=\arg\max_{\Theta}\prod_{i=1}^{|\tilde{U}|}\prod_{u_j\in S_i}\Pr(u_j|u_i)=\\&\arg\max_{\Theta}\sum_{u_i\in|\tilde{U}|}\sum_{u_j\in S_i}\log_a\Pr(u_j|u_i)=\\&\arg\max_{\Theta}\sum_{u_i\in|\tilde{U}|}\sum_{u_j\in S_i}\log_a\frac{\exp(u_j'\cdot h_i^{(n)})}{\sum\limits_{k=1}^{|\tilde{U}|}\exp(u_k'\cdot h_i^{(n)})}\end{aligned} \tag{8}$$

对式(8)中的对数似然函数求参数 Θ 的梯度:

$$\nabla\log_a\Pr(u_j|u_i)=\nabla f(u_j,u_i)-\sum_{u_j\in\tilde{U}}\Pr(u_j'|u_i)\nabla f(u_j',u_i) \tag{9}$$

其中,$f(u_j,u_i)$ 在式(6)中已经给出,由式(9)可以看出,参数的梯度分为 2 个部分,分别是给定 u_i 时的正例和全体负例的加和。本文采用一种负采样算法,通过采样得到的负例样本来估计全体负例的分布,即:

$$\nabla \log_a \Pr(\boldsymbol{u}_j \mid \boldsymbol{u}_i) \approx \nabla f(\boldsymbol{u}_j, \boldsymbol{u}_i) - \sum_{l=1}^{L} E_{v_l \propto P_n(v)} [\nabla f(\boldsymbol{u}'_l, \boldsymbol{u}_i)] \quad (10)$$

其中,第一项是正例,第二项是对负例的采样,$P_n(v) \propto d_v^{\frac{3}{4}}$[14],$d$ 是节点 v 的度,L 是采样数。

2.5　节点匹配及算法流程

得到节点向量表示后,需要计算节点之间的相似度。首先将网络中节点按其原来所属网络分为网络 G^X 中的节点和网络 G^Y 中的节点。然后计算 G^X 中的每一个节点向量与 G^Y 中所有节点向量的余弦相似度,同时设定一个相似度阈值 T,当 2 个节点向量的相似度最大且大于设定的阈值时,判定 2 个节点匹配。节点间相似度计算方法如下:

$$s(\boldsymbol{u}_i^X, \boldsymbol{u}_j^Y) = \frac{\boldsymbol{u}_i^X \cdot \boldsymbol{u}_j^Y}{|\boldsymbol{u}_i^X| \cdot |\boldsymbol{u}_j^Y|} \quad (11)$$

为了得到尽可能准确的匹配结果,本文在实验中设置了匹配阈值 T 和迭代次数 k,其中匹配阈值 T 较高,可以提高算法的准确率,同时适当选择迭代次数,可以弥补由于较高的匹配阈值导致的算法召回率过低的问题,最终运行算法 k 次后将每一次的匹配结果都作为最终结果。

综上,算法的整体流程如下:

算法 2　融合属性的跨社交网络用户身份匹配算法

输入　网络 $G^X = (u^X, \varepsilon^X, P^X)$ 和 $G^Y = (u^Y, \varepsilon^Y, P^Y)$,已知匹配的种子结点集 A,随机游走窗口大小 t,用户名属性的重要程度 ρ,相似度阈值 T,算法迭代次数 k

输出　匹配的节点对集合 M

1. 网络融合算法得到融合的网络 $\tilde{G} = (\tilde{u}, \tilde{\varepsilon}, \tilde{\rho})$
2. 用户名向量化表示
3. while k > 0
4. 以随机游走策略产生节点序列 S_i 的集合 S
5. for each S_i in S
6. for “上下文节点” in S_i
7. 基于式(8)更新 Θ 和中心节点向量 u_i
8. max = 0
9. for each u_i^X in G^X
10. for each u_i^Y in G^Y
11. 由式(11)计算 $s(u_i^X, u_j^Y)$
12. if $s(u_i^X, u_j^Y)$ > max
13. max = $s(u_i^X, u_j^Y)$
14. if max ⩾ T
15. 将(u_i^X, u_j^Y)加入匹配的节点对集合 M
16. k - -
17. end while
18. return M

3　实验结果与分析

3.1　实验数据集

为了验证本文提出算法的有效性,本文选取了目前较为流行的 Facebook、Twitter、Flickr、Last. fm 等社交网络中收集的账号信息进行跨社交网络用户身份匹配的实验。其中,Facebook 和 Twitter 中的数据是通过 Google + 获取的。Google + 的用户可以在其中添加自己在其他社交网络中的账号链接,从而使他人与自己在其他社交网络中都建立好友关系。本文通过 Google + 来收集用户信息,并在收集到的信息中选取包含 Facebook 和 Twitter 链接的 Google + 用户,再访问其 Facebook 和 Twitter 链接,利用社交网络提供的 API 接口,爬取用户所在社交网络的好友列表。需要说明的是,由于 Twitter 中不存在与 Facebook 类似的好友关系,而是一种关注与被关注关系[15],因此本文在获取的 Twitter 数据集中忽略了单向链接,只有当双方都关注了彼此时,才保留链接。Flickr 和 Last. fm 中的数据来源于文献[16],利用文献提供的匹配用户对和网络中的好友关系数据,提取出对于网络结构影响较大的大度节点组成实验数据。两组实验数据的具体信息如表 3 所示。

表 3　用户名信息向量化表示

社交网络	节点数	连边数	匹配账号对
Twitter	1 156	5 017	141
Facebook	878	9 219	
Last. fm	4 311	136 224	709
Flickr	9 348	135 378	

3.2　基准算法

本文实验采用 3 种基准算法来与提出的算法进行对比,包括一种有监督的机器学习算法:支持向量机(Support Vector Machine,SVM)算法,文献[8]提出的深度挖掘用户名信息中的各种特征来实现用户身份匹配的算法以及文献[7]提出的利用拓扑结构信息实现用户身份匹配的算法(PALE)。机器学习算法相当于把跨社交网络用户身份匹配问题看作一个二分类问题,判断一组账号对“匹配”或者“不匹配”;本文在实验中将用户名属性和拓扑结构属性分开,作为 2 个独立特征进行支持向量机的输入,设置训练时的正负样本比例为 1:1,输出为“0”或“1”表示是否匹配。文献[7]的方法通过提取用户名字符串中多个特征,包括平均最大匹配、最长公共子串相似度等,然后将这些特征转化为向量,最后计算这些向量之间的距离来判断用户名是否属于同一个人。PALE 算法在网络表示学习过程中没有采用传统的线性函数映射,而是利用多层感知机结构,使 2 个网络在进行表示的过程中不需要统一向量维度,并且可以学习网络结构中的非线性关系。

3.3 评价指标及实验参数设置

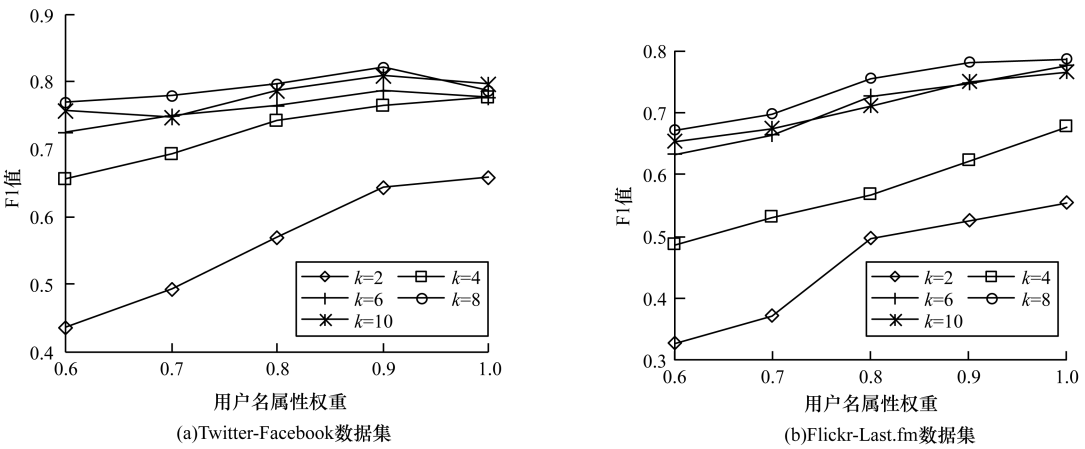
判断 2 个账号是否匹配的问题可以看作二分类问题,即给定 2 个用户账号时,若其属于同一个人,判定“匹配”,否则判定“不匹配”。因此,可以采用分类问题中的准确率 P、召回率 R 以及综合考虑准确率和召回率的 F1 值来作为衡量算法优劣的指标。这些指标的定义如下:$P = tp/(tp + fp)$,$R = tp/(tp + fn)$,$F1 = 2PR/(P + R)$,其中,tp 是真正例,即算法将实际匹配的账号对判断为“匹配”,fp 是假正例,即算法将实际不匹配的账号对判断为“匹配”,fn 是假负例,即算法将实际匹配的账号对判断为“不匹配”。F1 值是准确率与召回率的调和平均,综合反映了算法性能。

对于本文算法中的参数设定,控制用户名属性重要程度的变量 ρ 设定为 0.6 ~ 1.0 递增,迭代次数 k 分别设定为 2 ~ 10,匹配判定阈值 T 设定为 0.85,通过实验确定不同数据下的参数值。在与基准算法进行对比时,为了公平起见,实验过程中尽可能地保持参数设置一致,因此对其他基准算法的参数设定如下:网络表示学习阶段的节点拓扑结构向量维度统一设置为128维,训练率α_t统一设置为50% ,网络表示学习中随机游走的窗口大小 t 设置为 5,其余参数设定与原文保持一致。

3.4 实验结果

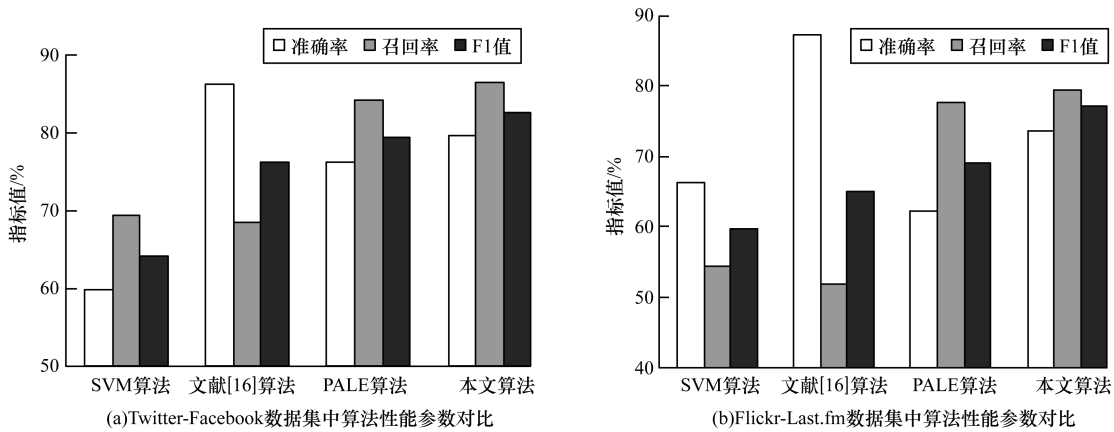
在实验中首先对控制用户名重要程度的变量 ρ 和迭代次数 k 在取不同值时 F1 值的变化进行了记录,如图 3(a)、图 3(b)所示分别是在两组数据集中进行的实验结果。从图 3 可以看出,随着迭代次数 k 的增大,F1 值呈现出先增大后减小的特点,这是由于迭代次数增加可以更充分地挖掘潜在匹配的节点,但如果迭代次数过多,会出现过拟合的问题,将原本不匹配的节点也视作匹配,因此迭代次数 k 不宜过小也不宜过大。另外,变量 ρ 的取值对实验结果也有一定程度的影响,总地来看,ρ 的取值在 0.9 ~ 1 之间时,实验效果较好,说明用户名属性在跨网络用户身份匹配任务中也扮演了比较重要的角色。因此,在后续对比实验中,本文算法中的变量 ρ 和迭代次数 k 分别设定为 1 和 8,同时为了使用户名向量与拓扑结构向量平衡,利用 tsne 降维技术将用户名向量维度也降为 128 维再与拓扑结构向量相结合。

3 种基准算法和本文算法在不同数据集中的性能参数对比情况分别如图 4(a)、图 4(b)所示。



(a)Twitter-Facebook数据集 (b)Flickr-Last.fm数据集

图 3 在不同数据集下 ρ 和 k 取值对算法综合性能的影响



(a)Twitter-Facebook数据集中算法性能参数对比 (b)Flickr-Last.fm数据集中算法性能参数对比

图 4 各算法性能参数对比

实验结果显示,本文算法相比其他 3 种算法效果更好。首先,对于传统的机器学习算法(以 SVM 为例),其对于数据的数量和质量有着极强的依赖性,并且对于跨网络用户身份匹配问题,当将其看作分类问题时,正例和负例是极其不均衡的:当2 个网络中的节点数都比较大时,会出现负例远多于正例的现象,尽管可以利用某些先验算法排除掉一些明显的负例,但仍无法从根本上解决正负样本不均衡的问题。其次,由实验结果可见,基于用户名信息的用户身份匹配算法有着极高的准确率 P,但受限于较低的召回率 R,其综合表现并不是很好。这是由于该算法对于"匹配"的判定比较"严苛",即用户在 2 个社交网络中的用户名相似度在极高的情况下才会认为 2 个用户匹配,一旦用户在2 个网络上注册的账号用户名差别较大,便会被算法认为不匹配,因此,会漏掉许多原本匹配的账号对。最后,对于利用拓扑结构进行网络表示学习的 PALE 算法,由于其没有挖掘除拓扑结构外的其他信息,因此准确率 P 并不高。本文算法综合考虑了社交网络中的拓扑结构信息和用户名信息,并将这两者结合起来判断账号是否匹配,使实验结果有了明显提升。从实验结果可以看出,本文算法在两组数据集中平均准确率达到 76.6% ,平均召回率达到83.0% ,平均 F1 值达到 79.7% ,其平均 F1 值比 SVM 算法提高 28.8% ,比文献[16]算法提高 12.7% ,比 PALE 算法提高 7.3% ,验证了本文提出的算法的有效性。同时,本文算法采用的拓扑结构信息和用户名信息在实际应用中获取的难度都不大,因此可以推广到实际应用中。

4　结束语

本文提出一种基于信息融合表示学习的跨网络用户身份匹配算法。该算法利用社交网络数据中的用户名信息和拓扑结构信息进行匹配,通过网络融合算法将跨网络问题转化为单一网络问题;运用网络表示学习技术,将网络中的节点拓扑结构信息向量化;同时使用用户名属性向量化方法,将用户名信息加入到节点向量化表示的过程中,使节点向量融合了拓扑结构信息和用户名信息,实现了在 Twitter-Facebook 和 Flickr-Last. fm 两组数据集中的跨网络用户身份匹配,并且取得了较好的效果。对于用户的挖掘账号信息,本文只针对用户名属性,未利用其他账号信息,存在一些拓扑结构相似度不高,用户名又不相似的匹配账号对在匹配过程中无法检出的情况。下一步将对社交网络账号的其他信息加以利用,找到一种统一所有账号信息的方法,以实现更高精度的跨网络用户身份匹配。

参考文献

[1] 尤　枫,曹天亮,卢　罡. 在线社交网络的自适应 UNI 采样方法[J]. 计算机工程,2017,43(4):200-206.

[2] SHU Kai,WANG Suhang,TANG Jiliang,et al. User identity linkage across online social networks: a review[J]. ACM SIGKDD Explorations Newsletter,2017,18(2):5-17.

[3] 罗由平,周召敏,李丽娟,等. 基于幂率分布的社交网络规律分析[J]. 计算机工程,2015,41(7):299-304.

[4] BUCCAFUIIR F, LAX G, NOCERA A, et al. Discovering links among social networks[J]. Lecture Notes in Computer Science,2012,7524:467-482.

[5] TAN Shulong, GUAN Ziyu, CAI Deng, et al. Mapping users across networks by manifold alignment on hypergraph[EB/OL]. [2018-03-21]. https://www.researchgate.net.

[6] ZHOU Xiaoping, LIANG Xun, ZHANG Haiyan, et al. Cross-platform identification of anonymous identical users in multiple social media networks[J]. IEEE Transactions on Knowledge and Data Engineering. 2016, 28(2):411-424.

[7] MAN Tong, SHEN Huawei, LIU Shenghua, et al. Predict anchor links across social networks via an embedding approach[C]//Proceedings of International Joint Conference on Artificial Intelligence. [S. 1.]: AAAI Press,2016:1823-1829.

[8] WANG Yubin, LIU Tingwen, TAN Qingfeng, et al. Identifying users across different sites using user names[J]. Procedia Computer Science,2016,80:376-385.

[9] LI Yongjun,PENG You,JI Wenli,et al. User identification based on display names across online social networks[J]. IEEE Access,2017(99):1.

[10] 吴　铮,于洪涛,黄瑞阳,等. 基于信息熵的跨社交网络用户身份识别方法[J]. 计算机应用,2017,37(8):2374-2380.

[11] LIU Li, WILLIAM C K, LI Xin, et al. Aligning users across social networks using network embedding[C]//Proceedings of International Joint Conference on Artificial Intelligence. Washington D. C., USA: IEEE Press,2016:1774-1780.

[12] 汪小帆,李　翔,陈关荣. 网络科学导论[M]. 北京:高等教育出版社,2013.

[13] 刘　东,吴泉源,韩伟红,等. 基于用户名特征的用户身份同一性判定方法[J]. 计算机学报,2015,38(10):2028-2040.

[14] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems,2013,26:3111-3119.

[15] 孟　波. 多社交网络用户身份识别算法研究[D]. 大连:大连理工大学,2015.

[16] ZHANG Yutao, TANG Jie, YANG Zhilin, et al. COSNET: connecting heterogeneous social networks with local and global consistency[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discorery and Data Mining. New York, USA: ACM Press,2015:1485-1494.

编辑　索书志