# 复杂网络视角下跨社交网络用户身份识别研究综述

邢 玲*，邓凯凯，吴红海，谢 萍

(河南科技大学信息工程学院 河南 洛阳 471023)

【摘要】社交网络是一种具有交互特性的复杂网络，利用复杂网络具有的网络特性可以链接不同社交网络中的节点，并分析节点之间存在的联系，结合相关的匹配算法可以有效地识别出用户在不同社交网络上的虚拟账号，有助于各大社交网络为用户提供更好的服务。该文对近十多年来数据挖掘领域中提出的跨社交网络用户身份识别技术进行了系统性地综述，详细阐述了 3 类用户身份识别技术相似度的计算方法和统一的识别框架，利用相关的评价指标对分类后的用户身份识别技术进行性能评估，最后展望了跨社交网络用户身份识别技术的未来研究方向。

**关 键 词** 跨社交网络； 复杂网络； 数据挖掘； 实体用户； 用户身份识别

**中图分类号** TP391 **文献标志码** A doi:10.12178/1001-0548.2019182

# Review of User Identification across Social Networks: The Complex Network Approach

XING Ling*, DENG Kai-kai, WU Hong-hai, and XIE Ping

(School of Information Engineering, Henan University of Science and Technology Luoyang Henan 471023)

**Abstract** Social network is a complex network with interaction characteristics. It can link nodes in different social networks by using the network characteristics of complex network, analyze the connections between nodes, and combine with the related matching algorithm to identify user's virtual accounts, which can help social networks to provide users with better services. This paper presents a systematic review on across social networks user identification techniques proposed in the field of data mining. Then the methods for calculating the similarity of the three types of user identification techniques and the unified identification framework are elaborated in detail. The relevant evaluation metrics are used to evaluate the classified user identification technique performances. Finally, the future research directions of across social networks user identification techniques are prospected based on the analysis of the research status.

**Key words** across social networks; complex network; data mining; entity user; user identification

近年来，数据挖掘技术得到了迅猛发展，促使人们对自然和社会现象的认知逐渐从宏观层面深入到微观层面。节点作为微观存在单元，是复杂网络的重要组成元素。社交网络是复杂网络中部分节点构成的社交服务平台。人们可以利用社交网络来满足自身的需求。由于社交网络所提供的服务存在差异性，人们会有选择性地参与到各个社交网络中。社交网络从不同的视角来刻画用户的实际生活状态，是真实世界在虚拟网络上的映射。由于社交网络之间数据的不互通和用户隐私保护的问题，用户的完整数据获取较难，导致很难形成一个完善的用户社交网络图。识别出用户在不同社交网络中的多重身份，能够最大限度地整合与完善用户信息，从而为用户提供更加便捷的服务。

跨社交网络用户身份识别的本质就是找出多个虚拟账号背后的实体用户，该问题的解决对很多领域都存在着重要的意义，主要体现在以下 4 个方面：

1) 用户信息完善：单一社交网络中用户数据有限，如果能够得识别出用户的多个社交账号，就可以更加全面的掌握用户信息。

收稿日期：2019－08－29；修回日期：2020－01－14

基金项目：国家自然科学基金 (61771185, 61772175, 61801171); 河南省高校科技创新团队支持计划 (21IRTSTHN015)

作者简介：邢玲 (1978－)，女，博士，教授，主要从事多媒体语义挖掘、社交计算和隐私保护等方面的研究. E-mail: xingling_my@163.com

2) 个性化服务推荐：分析单一社交网络中的用户数据不能够很好地实现个性化服务推荐。如果将多个社交网络的用户数据进行融合，充分利用用户产生的信息，则推荐效果将会显著提高[1-4]。

3) 数据挖掘：将具有链接性的多个社交账号进行数据挖掘，可以获取更多有研究价值的信息[5-6]。

4) 提供科研支撑：用户之间的关系可构成复杂网络[7]。复杂网络具有的特性在单个社交网络中被深入研究，当扩展到多个社交网络时，是否会产生新的特性需要进一步的研究。

这项技术给人们带来巨大收益的同时，也带来了泄露个人信息的危害。例如：恶意用户可以通过位置数据来推测正常用户的一些敏感信息[8-10]。只有最大限度的减小用户隐私泄露，才能保证人们愿意将自己的数据提交给网络应用，进而更大限度的满足人们日常生活的需求。

目前关于跨社交网络用户身份识别的相关研究已经取得了一系列重要的成果，本文在复杂网络视角下，分别介绍了跨社交网络用户身份识别的概念、模型、分类、相似度计算方法、基本框架等方面的技术，详细分析了现有方法在用户身份识别方面的性能评价，通过对比分析现有方法的优缺点，探讨了跨社交网络用户身份识别未来的研究方向。

# 1 跨社交网络用户身份识别技术

## 1.1 问题定义

跨社交网络用户身份识别技术是将不同网络平台上的社交账号进行整合的一个过程。复杂网络视角下的社交网络是一种利用多种连接或相互作用的模式而存在的一组人或群。用户身份识别主要目的就是将不同社交网络上具有同一性的虚拟账号进行识别整合。该技术可以简化为数学模型。用 $G\{V, E\}$表示一个社交网络，其中 $V$ 是用户集合，$E$ 是用户之间的关系集合，定义两个社交网络中好友关系网络为$G^X(V^X,E^X)$和$G^Y(V^Y,E^Y)$，挖掘出属于同一实体的账号，即匹配出用户对$(v_i^X,v_j^Y)$，如图 1 所示，用户身份识别所要解决的问题就是判定社交网络$G_A$上的账号$V_{Ai}$和$V_{Ak}$与社交网络$G_B$上的账号$V_{Bj}$是否为相同的自然人。

用户身份识别又称为用户身份解析、用户匹配、用户锚点链接。进行识别的前提是获取完整的数据集。首先需要通过爬虫[11] 获取社交网络上用户的数据信息。由于用户在不同的社交网络中产生的数据类型存在差异性，需要对用户的数据进行预处理来保证计算用户数据相似度计算过程的准确性。然后利用 simfunc() 函数对用户数据进行相似度计算可以获得候选匹配对。最终采用相关的匹配算法来对候选匹配对进行剪枝过滤，输出匹配结果。

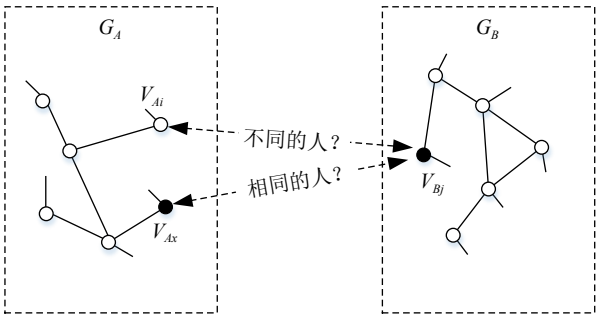



图 1 跨社交网络用户身份识别问题解析

由于用户在社交网络上产生的数据类型存在差异性，因此，跨社交网络用户身份识别技术大致可以分为 3 类：基于用户档案信息 (用户在注册社交账号时填写的个人信息)、基于网络拓扑结构 (用户社交所形成的好友关系，即节点之间的关系)、基于用户生成内容 (用户在社交过程中所发布的内容)。

## 1.2 基于用户档案信息

这类方法是将用户的用户名、生日和性别等档案信息转换成一个多维的向量形式，用来表征特定社交网络中的用户身份信息。多维向量可定义为$F_x=(a_1^x,a_2^x,...,a_n^x)$，其中$a_n^x$表示账号 $x$ 的第 $n$ 个属性。社交账号相似度向量可定义为$\boldsymbol{V}(F_A,F_B)=(v_1^{AB},v_2^{AB},...,v_n^{AB})$，其中$v_n^{AB}$表示 $A$，$B$ 账号之间第 $n$ 个属性的相似度，可由 simfunc() 函数计算，且$0\leqslant|v_n^{AB}|\leqslant 1$。然后为各个属性项分配合理的权重，最终衡量两个账号的相似度，即 $\text{similarity}(F_A,F_B)=\sum_1^n(w_n^{AB}v_n^{AB})$的值，来判定两个账号是否具有同一性。

## 1.3 基于网络拓扑结构

这类方法主要依靠用户的好友关系网络来识别未知用户。相关的方法中，利用复杂网络中节点的重要程度以及节点之间的匹配度来识别不同社交账号背后的实体用户[12]。识别过程中将用户账号等效成网络节点来计算网络节点间的相似度。

**定义 1** 实体用户指虚拟账号背后的真实用户。

**定义 2** 种子节点指被提前识别出来的网络节点。

**定义 3** 映射函数$\phi$表示虚拟账号$u_i$映射到实体用户 $p$，即$\phi(u_i)=p$。

输入：$n$ 个社交网络$G=\{G_1,G_2,\cdots,G_n\}$以及种子节点 (在无监督学习的算法中，不需要种子节点)。

输出：利用映射函数$\phi$得到的最终匹配结果，即$\phi(u_i^Y)=\phi(u_j^X)=p$。

### 1.4　基于用户生成内容

传统的用户身份识别忽略了用户生成内容的作用，而造成身份识别准确率低的问题。基于用户生成内容的研究工作将用户发表在各大社交网络上的内容作为文本数据集进行处理，挖掘与用户相关的行为信息来识别不同社交网络上用户的身份信息。由于用户在不同社交网络上发表的数据存在格式等方面的差异性，需要对用户的数据进行数据预处理。其中，对数据进行噪声处理，可以滤除文本内容中无意义的信息，减少信息检索和计算的时间。利用相关挖掘算法对文本数据进行分析，可以减少索引量，降低计算复杂度。随后，可以采用相关加权技术来判别某些词的重要程度，分析出能反应用户实际生活习惯的信息。最后，通过计算生成数据的相似度来判定用户是否匹配。

## 2　相似度计算

### 2.1　用户档案信息相似度计算

大多数用户档案信息的存储类型是字符串，因此，通过计算字符串序列相似度可以获取相关档案信息的相似度。常见的字符串有:

1) 编辑距离[13]：指计算两个字符串相等时所需的单个字符编辑步数，字符串$n_i$和$n_j$的相似度为：

$$\text{Simfunc}(n_i,n_j)=1-\frac{d(n_i,n_j)}{\max(|n_i|,|n_j|)} \tag{1}$$

式中，$|n_i|,|n_j|$表示字符的个数。

2) Dice 系数[14]：在计算字符串时，计算类型可分为多值属性和单值属性两类。字符串$l_i$和$l_j$的相似度计算公式为：

$$\text{Simfunc}(l_i,l_j)=2\frac{|l_i\cap l_j|}{|l_i|+|l_j|} \tag{2}$$

式中，$|l_i\cap l_j|$表示字符串的交集信息。例如：两个多值字符串“vivid music movie”和“music travel”交集信息为{“music”}，相似度为 2/5=0.4。

对于单值属性字符串，Dice 系数的计算方法如上式，只是交集信息有所不同。例如：单值字符串“johe”和“joh”，交集信息为“jo，oh”，相似度为 4/5=0.8。

3) Jaro 距离[15]：指利用公共字符个数与顺序的方法，字符串$s_i$和$s_j$的 Jaro 距离计算公式为：

$$d(s_i,s_j)=\frac{1}{3}\left(\frac{m}{|s_i|}+\frac{m}{|s_j|}+\frac{m-t}{m}\right) \tag{3}$$

式中，$m$ 是公共字符的个数；$t$ 是换位的个数。

可得相似度计算公式：

$$\text{Simfunc}(s_i,s_j)=1-\frac{d(s_i,s_j)}{\max(|s_i|,|s_j|)} \tag{4}$$

4) 最长公共子串[16]：该方法的思想就是找出两个字符串最长的公共子串，通过计算子串的相似度来判别用户的相似度。将“Dave Whelan”和“David J. Whelan”作为两个字符串$s_1$和$s_2$来计算最长公共子串的相似度。基于文献 [16] 的算法，在第一次迭代中$\text{lcs}(s_1,s_2)$是“Whelan”且$\text{Len}_{\text{lcs}}(1)=6$；在第二次迭代中，$\text{lcs}(s_1,s_2)$是“Dav”且$\text{Len}_{\text{lcs}}(2)=3$；在第三次迭代中，$\text{lcs}(s_1,s_2)$为空集。Len(David J. Whelan) = 13和Len(Dave Whelan) = 10。$\text{simfunc}_{\text{LCS}}$的计算如式 (5)，字符串$s_1$和$s_2$的相似度值为 0.25。

$$\text{simfunc}_{\text{LCS}}=\frac{\sum_{i=1}^{\text{cn}}\text{Len}_{\text{lcs}}(i)-\text{cn}+1}{\text{len}(s_1)+\text{len}(s_2)+\sum_{i=1}^{\text{cn}}\text{Len}_{\text{lcs}}(i)} \tag{5}$$

式中，$\text{Len}_{\text{lcs}}(i)$表示第 $i$ 次迭代子串的长度；cn表示迭代次数；$\text{len}(s_1)$和$\text{len}(s_2)$表示字符串$s_1$和$s_2$的长度。

此外，用户档案信息相似度计算的方法还有 MN 距离[17]、词频－逆文档频率 (TF-IDF)[18]、最长公共子序列[19] 和 Jensen-Shannon 距离[20] 等。

### 2.2　网络拓扑结构相似度计算

网络拓扑结构的相似性可以通过节点之间的相似性来表征。复杂网络中的“小世界网”的概念就可以很好的表述关于节点相似度计算的原理。社交网络中，大部分的好友关系网络可能都是“近程”的，利用节点之间的关系和共享的邻居节点数来计算网络节点之间的相似性。通常采用节点的邻居节点集合表示网络拓扑结构。已知节点$v_i^X,v_j^Y$分别来自社交网络 $X$ 和 $Y$，$\Gamma(v_i^X),\Gamma(v_j^Y)$分别表示节点$v_i^X,v_j^Y$的邻居节点集合。计算网络拓扑结构之间相似性的方法很多，其中最常用的是共同邻居指标 (CN)[21-22]，计算公式为：

$$S(v_i^X,v_j^Y)=|\Gamma(v_i^X)\cap\Gamma(v_j^Y)| \tag{6}$$

相似的相似性指标还有 Adamic-Adar 指标[23]、Jaccard 指标[24]、资源分配指标[25]、Salton 指标[26]、

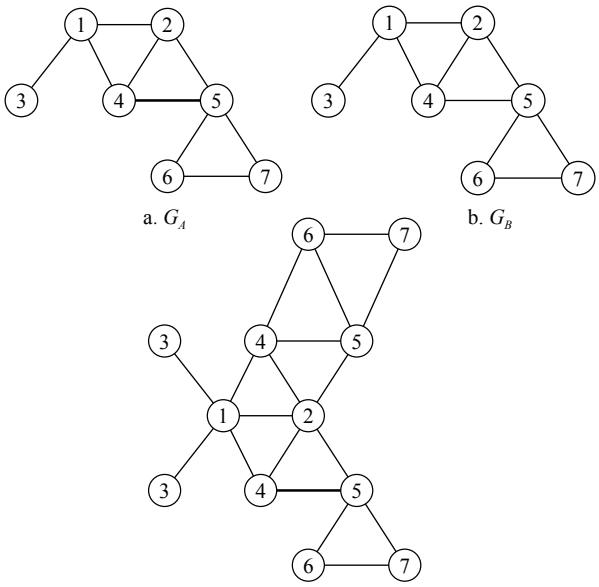
LHN-I 指标[27]、优先链接指标[28] 等。

除了上述的节点相似度计算方法，基于网络拓扑结构的节点相似度还有利用有监督学习的和无监督学习的方法：

1) 有监督学习的方法[29]：以图 2a 和图 2b 中的 $G_A$和 $G_B$ 为例 ($G_A$ 和 $G_B$ 为带有先验节点的社交网络)，如果一个节点只与节点 1 有关系，很明显其必须是节点 3。如果节点询问谁拥有节点 1 和 2 的节点集，那显然是节点 4。社交账号往往在不同的社交网络上有类似的朋友。因此，可以假设：如果给出了一些有效的先验种子节点 (有监督学习)，则可以推导出一组候选用户匹配对 (UMP)，以及账号在候选 UMP 中共享的已知的朋友越多，则它们属于同一个人的概率就越大。图 2c 给定了两个社交网络$G_A$和$G_B$，其具有先验种子节点$\mathrm{UMP}_{A\sim B}(1,1)$和$\mathrm{UMP}_{A\sim B}(2,1)$。可以发现节点$U_{A4}$和$U_{B4}$具有相同的节点集，然后可以根据先验节点集来计算节点$U_{A4}$和$U_{B4}$是否为一个新的用户匹配对。如果节点 4 被识别，可以添加到先验种子节点中，依次进行以下节点的匹配，其匹配度计算公式为：

$$M_{ij}=|F_{Ai}\cap F_{Bj}|+\frac{|F_{Ai}\cap F_{Bj}|}{\min(|F_{Ai}|,|F_{Bj}|)} \tag{7}$$

式中，$F_{Ai}$和$F_{Bj}$分别表示已识别的$U_{Ai}$和$U_{Bj}$的节点集。




a. $G_A$　　b. $G_B$

c. 带有先验种子节点的 $G_A$和 $G_B$

图 2　有监督学习方法的示意图

2) 无监督学习的方法[30]：顾名思义，该方法在进行用户身份识别的过程中不需要先验种子节点的支撑，即满足如下识别函数：

$$f(U_{Ai},U_{Bj}|p,G_A,G_B)=\begin{cases}1 & U_{Ai}=U_{Bj}\\ 0 & \text{其他}\end{cases} \tag{8}$$

式中，$p$=0 表示无先验种子节点。

通过将社交网络中每个用户的好友特征提取为好友特征向量，然后计算两个社交网络之间所有候选相同用户的相似度，其具体计算过程见文献 [30]。

### 2.3　用户生成内容相似度计算

用户生成内容基本都以文本信息的格式存在，计算用户生成内容相似性常用的方法有余弦相似度[31-32]、LDA 模型[33-34] 等。其中余弦相似度在计算过程中相对容易进行，余弦相似度的计算主要是通过计算向量之间的余弦值来判定用户生成内容之间的相似性，即需要想将文本信息转换成向量形式。例如，文本信息转换成的向量后为$\boldsymbol{A}=[A_1,A_2,\cdots,A_n]$，$\boldsymbol{B}=[B_1,B_2,\cdots,B_n]$。则余弦相似度的计算公式为：

$$\cos\theta=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{9}$$

另外，用户产生的内容可以从空间、时间以及文本方面来提取相应的特征，计算这些特征的相似度可以采用欧氏距离[35] 和 Jaccard 系数[36] 以及 2.1 节中使用的部分相似度计算方法。

## 3　基本框架

通过对用户数据进行相似度计算，可以获取账号之间的相似性关系。将甄选出来的候选匹配对在相关的匹配算法中做进一步的确认，就可获取最终的匹配结果。在社交网络上，对用户账号进行匹配的算法主要选用二部图最大权匹配算法[37]、稳定婚姻匹配[38] 等经典的算法。假定 $n$ 为待匹配账户的数目，则二部图最大权匹配算法的时间复杂度为 $O(n^3)$，当用户账号数目增多时，其时间复杂度会随之增加。而稳定婚姻匹配算法的时间复杂度为 $O(n^2)$，相比较而言，其时间复杂度相对较低，但用户账号的匹配精度有待提高。

为了进一步完善稳定婚姻匹配算法的性能，研究人员改进了这类算法。改进的算法采用逐步迭代地求取种子节点，其中迭代的过程可大致分为 3 步：账号选择、账号匹配、双向认证。通过迭代

计算可以验证虚拟账号是否匹配，获取最终的匹配结果集合。

现有的跨社交网络用户身份识别大多数都具有一个统一的匹配框架，如图 3 所示，用户身份识别的过程主要分为 5 个步骤：

1) 爬取用户在社交网络上产生的数据，划分不同用户的数据类型，并对数据进行泛化处理，形成用户数据信息集合 (包括用户档案信息、网络拓扑结构和用户生成内容)；

2) 分别采用章节 2 综述的相似度计算方法来给不同的用户信息进行相似度计算；

3) 用户不同信息类型的赋权对最终的识别结果至关重要，采用合理的权值分配来提高用户身份识别的性能；

4) 利用匹配算法以及步骤 2) 和 3) 中所得到的相似度和权值分配来获取用户匹配对；

5) 获取用户匹配结果后，为了更加准确地保证识别率，需要对用户账号进行剪枝过滤，以最大限度消除错误的匹配结果。

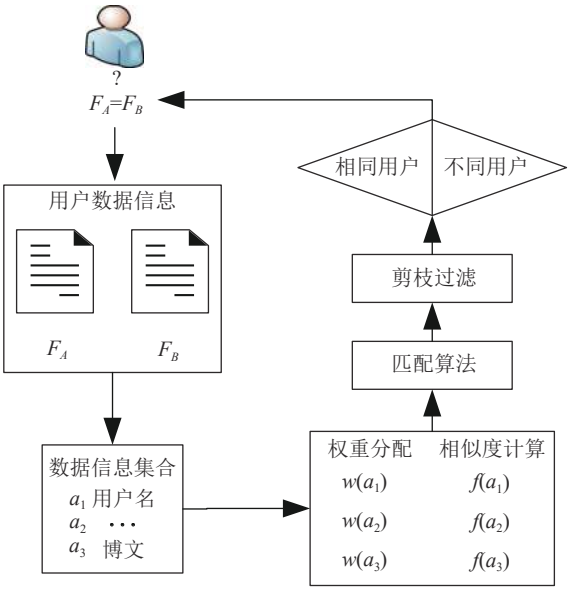



图 3　跨社交网络用户身份识别框架

# 4　跨社交网络用户身份识别技术

社交网络是复杂网络的重要组成部分，近年来得到了飞速的发展，且产生的用户数据种类具有多样性，但现有的用户身份识别还不足以完全适用于所有的社交网络。当社交网络中的用户具有多种匹配信息时，就可对这些信息进行整理分类来识别不同虚拟账号背后的实体用户。本文从 3 个方面来梳理研究现状。

## 4.1　基于用户档案信息的用户身份识别

用户在各大社交网络上注册账号时，通常需要进行个人信息的填写。档案信息很大程度上直观地反映了用户的身份，这些信息为跨社交网络用户身份识别提供了有力支撑。当用户档案信息能够被合理的获取和利用时，可以提高实体识别的准确率。现在已有很多研究在利用这类用户信息进行身份识别，大致可分为两类：基于单属性和基于多属性的用户身份识别。

### 4.1.1　基于单属性的用户身份识别

基于单属性项进行身份识别主要是指利用用户档案信息中的单个数据进行身份识别，其中应用最广泛的是用户名属性。

文献 [39] 最早提出将用户档案信息运用于身份识别研究中，利用用户名进行增加或移除其前缀与后缀的方法来完成用户名在社区之间的映射，进而判定用户的身份信息。为了进一步准确地验证用户名具有唯一性，研究人员通过对两个社交网络的用户数据进行训练提出了基于语言模型和马尔可夫链技术，并用 $n$-gram 概率来估计用户名的稀有性或共性，最终采用编辑距离的相似度计算方法来获取用户名属性的相似度，验证了直观的观察以及解释了在使用用户名链接公共配置文件的任务中具有的高准确性[40]。

经过对用户名具有的独特性进行分析，文献 [41] 利用用户名属性的唯一性为用户身份识别构建训练实例，将用户属性和用户进行的社交活动等相关内容进行结合来连接不同社交网络中的虚拟账号，可以有效的识别出虚拟账号背后的实体用户。用户名可以由不同的字符构成，研究人员通过定义一些合理的复杂特征来对用户选择用户名时的行为模式进行建模分析，通过分析用户的行为模式来判定用户是否具有同一性[42]。

由于用户名中含有冗余信息，文献 [43] 针对用户名属性特征进行了详细的研究，从用户的用户名中提取了上千种复杂特征，例如：数字特征、字符数字组合特征、日期特征等。将用户名特征等效成自信息向量，利用相关的相似度计算方法来分析向量之间的相似性，从而判定出相同的实体用户。

用户在命名用户名的过程中或多或少会存在一定的相似性。文献 [44] 对用户在不同社交网络上用户名命名存在的差异性进行分析，利用用户名中的冗余信息构建用户名特征，并使用有监督机器学习的方法对已识别的匹配账号进行进一步的确认，

有效地增加了实体识别的准确率。但用户身份识别仅仅依靠单属性进行分析，还不能完全满足现有的需求。用户的单属性信息具有高度的模仿性，而且单属性信息不足以完整地体现出用户的实际生活状态，存在鲁棒性差的问题，然而其具有的低计算复杂度特性也是不容忽视的。

#### 4.1.2 基于多属性的用户身份识别

用户名在一定程度上很好地反映了用户特征，但当识别的数据集过大时，单个属性会很容易出现高的重复性，这就造成识别准确率会随着用户数量增加而降低的负相关趋势。因此，这个关键性问题的解决就需要对用户的多个属性信息进行融合以寻求最优方案。

文献 [17] 最早提出将用户的多个属性项信息等效成 $n$ 维向量，对用户的各个属性采用不同的相似度计算方法，并为各个属性项进行相应的权重分配。它的缺点是属性与领域之间会存在紧耦合的关系，每次应用场景的改变都需要重新计算各个属性的权重值。研究人员还利用程序编程接口 (API) 对用户档案信息进行爬取，将用户档案信息表示成单词集合，通过计算分析单词之间的相似度来获取虚拟账号之间的相似度[45]。随后，文献 [46] 提出了一种基于 FOAF 词汇表的匹配架构，将用户档案数据转移到 FOAF 词汇表中，并使用判定算法来获取两个虚拟账号之间的相似度。由于在进行用户身份识别的过程中将电子邮件地址视为唯一的标识符，而这个属性信息不易被获取，因此，该算法存在普适性差的问题。

利用用户的多属性信息进行识别的过程中，权重的分配对识别结果的影响也不容忽视。文献 [47] 将用户名和用户标签进行结合，并采用简单的主观赋权法对二者进行赋权。当该方法应用到新的社交网络时，需要重新定义用户的属性权重，一定程度上增加了时间复杂度。随后，相关研究人员也提出一种基于主观导向的客观赋权方法，来对用户的多个属性进行相似性计算[48]，但该方法依赖大量的样本数据，在实际的应用中普遍性较差。

针对上述存在的问题，文献 [49] 提出了基于信息熵的客观赋权法，该方法可以利用用户属性的熵值来为各个属性项分配合理的权重。该方法在用户多属性分配权重方面具有高度的合理性，但所提方法需要对每个源账号都进行一次权重分配方案，当处理的用户规模较大时，权重分配方面的计算量将大大提高。相似的研究通过分析用户各个属性的后验概率来为各个属性进行赋权，该种方法明显区分了不同属性对用户身份识别的贡献程度，且不需要对每个源账号进行赋权，极大降低了计算量[50]。

在获取社交账号之间的相似度向量后，可以利用分类器对向量进行判决，识别出用户账号是否具有同一性。

文献 [51] 选取了 5 个社交网络上的用户多属性信息进行实体识别，分别利用决策树[52]、SVM[53] 等有监督分类模型进行对比训练。实验结果表明，贝叶斯分类模型在真阳性率 (TPR) 和伪阳性率 (FPR) 上和其他几种模型性能接近，但在运行时间上比其他模型要短的多，因此选用贝叶斯模型作为最终的分类结果，并对用户账号进行有监督的判决。

文献 [54] 综合考虑了用户在不同社交平台上的多属性信息，并结合均等评价模型和训练混合模型来融合多个属性项的相似度。文献 [55] 提出了一种基于重叠属性项的选择方法，与传统的随机选择法相比，该方法训练出来的模型具有很高的适用性。随着各大社交网络逐渐重视用户的隐私信息，部分用户属性信息不易获取。文献 [56] 提出了一种 UISN-UD 模型，利用用户名或显示名称[57] 中包含丰富的信息冗余来匹配用户，减少了用户身份识别过程中属性项的使用，降低了计算复杂度。最突出的优点在于基本不涉及个人隐私，并且具有高度可访问性，综合评价指标超过了 90%，为身份识别提供了一种新的思路。为了优化用户识别过程的性能，文献 [58] 设计了一种成对比较字符串匹配方法。该方法分析 LinkedIn 和 Facebook 上用户的多个档案属性，并使用相应的相似度计算方法分步计算不同属性的相似度值，每一步都与设定的阈值进行比较来删除错误的匹配对，从而获取最终的账号匹配对。

尽管上述方案可以实现良好的性能，但最大的挑战是用户数据的真实性和完整性，因为用户通常会因为隐私保护等问题不会提供其真实和完整的档案信息。因此，当档案信息的准确性不保证时，基于用户档案信息的方法对获得良好的匹配结果有一定的限制。

### 4.2 基于网络拓扑结构的用户身份识别

基于网络拓扑结构的跨社交网络用户身份识别实质上是将用户之间的好友关系等效成网络拓扑结构进行节点之间的相似度匹配。社交网络中的好友关系可以较易的通过开放的 API 来获取。这就使得

基于网络拓扑结构的实体识别得到了进一步的研究。

文献 [59] 最早提出了依靠网络拓扑结构来识别用户，从少量已知的种子节点出发，通过不断的迭代更新找出新的匹配节点，就可以实现在两个社交网络之间的用户身份识别，但身份识别的准确率和召回率都有待提高。随后，相关研究人员提出将用户档案信息与图相似性进行结合，实现了电子邮件网络向 Facebook 网络的映射[60]。但这样的映射关系存在一对一映射冲突的问题，对于最终的身份识别性能有一定的影响。针对以上有关映射方面的问题，文献 [61] 把文献 [60] 中存在的问题转化成了有向链路的预测问题，使得社交网络结构中存在的一对一映射冲突问题得到了解决，但提出的方法局限于匹配对的映射，应用场景受限。

通过将用户身份识别问题转换成数学问题会产生一定的积极影响，文献 [62] 将实体识别的问题转换成了严格的数学定义，认为现有的社交网络基本上都是通过一个潜在的用户图形结构采用概率生成的，并且网络拓扑结构中节点边的选择是近似概率的，且存在级联效应。利用这种数学思想，采用迭代计算来判定多个虚拟账号是否为同一实体用户。

在社交网络中进行节点之间相似度计算时会产生很大的计算量，文献 [63] 提出了超图[64] 的概念，对用户的好友关系进行建模，并设计了一个流形对齐框架，将各大社交网络中的用户映射到低维空间，很大程度上降低了运算过程的复杂度。

为了使得识别的性能更好，文献 [65] 考虑从能量模型的研究方向出发，提出一种基于能量模型 COSNET 的方法，使用能量级的方式来区分不同的用户属性和结构的匹配方式，并利用次梯度算法来训练能量模型，当获取的匹配结果达到最佳情况时能量最低。将社交账号匹配问题等效成了能量模型的求参问题，充分利用对偶问题分解的思想来提高社交账号匹配的性能。

基于网络拓扑结构的方法大多数采用有监督学习的方式进行用户身份识别。文献 [29] 利用不同社交网络中待匹配节点共有的种子节点数作为衡量节点间相似性大小的指标，并对评判依据匹配度进行改进，选择相似性较大的网络节点进行匹配。该方法的提出验证了基于网络拓扑结构可以更好的完成实体识别。提出的基于 FRUI 算法可以很好地应用于多个具有好友关系的在线社交平台，但用户的匹配结果过于依赖种子节点的质量和数量。针对上述问题，当种子节点无法获取时，即无监督学习。文献 [30] 提出了 FRUI-P 算法，通过借鉴随机游走的思路，使用了基于负采样技术的 CBOW 模型来学习网络向量。将社交网络中每个用户的好友特征提取为好友特征向量。在特征向量获取时，提出了 FFVM 模型来更合理的输出所有用户的特征向量。然后，利用相关的相似性计算方法[66] 更新用户节点之间的相似度来达到高准确率识别实体的目的。该方法优点在于无需知道种子节点 (先验知识) 就可以精确地识别社交账号线下的实体用户，并能够为现有的大部分基于有监督学习和半监督学习的方法提供可靠的先验知识用于用户身份识别。

随着相关研究工作的深入，文献 [67] 开发了一种基于图神经网络的用户身份识别框架，将社交网络形成的图形拓扑编码视为节点特征，此特征学习过程称为节点嵌入。嵌入的目的是将网络结构映射到低维节点空间，使得重建的基于学习节点特征社交网络接近原始社交网络。此外，他们还提出了一个深度图模型来学习一些大型社交网络的节点嵌入，并在社交网络中构建具有相同身份节点的非线性映射。这种半监督的学习方法有助于在实际案例中有效地识别用户。基于网络拓扑结构进行用户身份识别的方法主要是将用户之间的社交关系等效成网络节点图的形式，然后完成不同社交网络之间节点图匹配的任务。图匹配就是利用一种方法来计算图形之间的相似性大小，再结合相关的匹配算法，从海量的数据中获取最终的匹配结果。目前该方法已经应用到了很多领域[68]，并在社会安全领域方面帮助了一些机构很好的完成了一些特殊的任务。但由于社交网络具有异质性[69] 且部分用户的好友关系存在稀疏性，因此，基于网络拓扑结构的方法在识别用户的综合性能方面还有待提高。

### 4.3　基于用户生成内容的用户身份识别

用户生成内容是指用户在各大社交网络上产生的各类社交行为信息的总和。用户在社交网络上评论、转发、点赞等行为信息往往从一个侧面真实反映了用户的兴趣、爱好等[70]，无形中为用户打上了身份的标签。用户生成的行为数据进一步拓展了跨社交网络用户身份识别的信息维度。如果能够合理地加以利用，可为跨社交网络用户身份识别提供一条新的捷径。

文献 [71] 通过分析用户在生成内容上的书写风格来识别用户，验证了不同社交网络之间的可链接性，但在某种意义上存在用户隐私安全方面的问

题。为了提高社交网络之间的可链接性，文献 [72] 将用户档案信息、用户生成内容和其他的用户数据进行了整合，来进一步提高不同社交网络之间链接的准确性。用户产生的内容会在一段时间内形成一个明显的主题，文献 [73] 提出了一种动态核心兴趣映射算法 (DCIM)，该算法基于用户生成内容和自中心网络来考虑用户的拓扑和主题模型，主要用来分析用户的兴趣动态规律。然而，以上提出的方法或多或少地局限于某些特定的在线社交网络，这使得所提方法不能扩展到通用的在线社交网络上。

相关研究通过对用户发表的状态以及评论方面的信息进行了研究[74]，实现了多个社交网络上的用户身份识别，利用 Doc2Vec 将用户生成内容等效成向量形式，从而通过计算向量之间的相似性来进行实体识别。在用户发表内容的基础上，再辅助其他相关行为信息的挖掘，可以实现更好的身份识别性能。通过将用户产生的多个行为信息进行融合，文献 [61] 提出了一种 MNA 算法，从社交网络中获取了用户的地理位置、发表内容、签到时间、发表内容的主题和链接这 4 种信息。使用不同信息类型对应的特征值来训练 SVM 分类器，并计算用户不同社交网络中时间、空间等信息方面的相似性。为了保证一对一映射的限制条件，该算法优先将匹配分数较高的用户进行匹配，一旦匹配成功就不再考虑其他用户，一定程度上提高了匹配结果的准确率。但也不能保证完全无误差识别，因为用户的规模变大时，账号之间的相似度也会随之提高，对最终的匹配结果会有一定的影响。相似的研究通过利用用户在社交网络中生成的状态时间戳信息与移动设备生成的位置信息构建属于用户自己个性化的社交行为模式，以此来解决用户身份识别问题[75]，此类方法在文献 [76] 中也有所体现。

文献 [35] 提出了一种基于用户生成内容的用户身份识别模型 (U-UIM)。首先利用树状图的思想将识别模型表示出来，然后采用了几种算法分别测量了用户生成内容在空间、时间和内容维度上的相似性，并构成对应的特征向量。最后，利用有监督的机器学习算法在 3 个真实的数据集上验证综合性能的优劣。这项工作表明了当虚拟账号具有高度可访问的在线数据时精确匹配用户账号是可能的。

相关的研究还包括利用用户生成的轨迹信息进行身份识别，用户轨迹信息在一定程度上也反映了用户实际生活中的移动轨迹，可以作为体现用户社交产生的一个生成信息。因此，利用用户轨迹信息进行用户身份识别也渐渐引起了研究者们的注意。文献 [77] 通过分析用户轨迹中地理位置的共现频率来识别用户，提出了基于一种处理多源位置数据的算法，它的缺点是所用的参数过多，调整参数的过程非常繁琐，在算法计算的过程中，增加了身份识别的计算开销。相似的研究还将用户移动的轨迹范围划分成多个小网格，然后将用户的轨迹等效成若干小网格组成的序列，并采用 TF-IDF 模型将用户的轨迹数据转化成向量，最后利用余弦相似度来计算账号向量之间的相似度，进而判定用户账号的是否具有同一性[78]。用户在移动的过程中，每个地理位置都会产生相应的坐标，文献 [79] 将地理位置坐标表示成对应位置的语义词，使得用户轨迹构成一篇由语义位置组成的文本，然后用 LDA 模型获取用户的主题分布，最终利用 KL 散度得出用户轨迹之间的相似度来判决是否为同一用户。一般情况下，用户访问某个地点的次数会被统计，文献 [80] 假定用户在特定时间内访问某个地理位置的次数服从泊松分布，并用概率函数表示出两个账号具有同一性的概率形式，最终对目标函数进行优化获取最优的匹配结果。

部分研究工作还分别从用户轨迹产生的空间和时间两个维度信息的角度出发，将原始的多源时空数据转换成三部图的形式。通过计算三部图的最优划分，来获取最佳的用户匹配结果[81]。为了降低计算复杂度，文献 [82] 提出了一种基于最频繁分布 TOP-N(分布最频繁的 $N$ 个区域) 的识别解决方案。首先找到用户轨迹分布最频繁的前 $n$ 个区域，以降低计算复杂度。然后，利用基于概率偏差、角余弦和加权 Jaccard 相似度的方法计算两条轨迹的相似度，从而实现用户身份识别。

## 5 性能评估

跨社交网络用户身份识别在高识别度的基础上也应同时兼顾服务的可用性和计算开销等问题。本文从以下 3 个方面度量其性能。

1) 用户身份识别度，由识别性能的评价指标来反映，评价指标中的精确率、召回率、综合评价指标 (F1) 和 AUC 的值越高，说明用户身份识别度越好：

$$\text{精确率} = \frac{\text{tp}}{\text{tp}+\text{fp}} \tag{10}$$

$$\text{召回率} = \frac{\text{tp}}{\text{tp}+\text{fn}} \tag{11}$$

$$F1=\frac{2\times \text{precision}\times \text{recall}}{\text{prcision}+\text{recall}} \tag{12}$$

AUC 是 ROC 曲线下的面积，假阳性率 (FPR) 定义为 $X$ 轴，真阳性率 (TPR) 定义为 $Y$ 轴。由于用户身份识别的结果分为两类，即同一实体用户和不同的实体用户，AUC 可用于测量识别结果：

$$\text{TPR}=\frac{\text{tp}}{\text{tp}+\text{fn}} \tag{13}$$

$$\text{FPR}=\frac{\text{fp}}{\text{fp}+\text{tn}} \tag{14}$$

式中，tp 表示被正确匹配的实体数目；tn 表示账号未匹配上，且账号也不是同一个实体的数目；fp 表示被匹配上但不是同一实体的数目；fn 表示未被匹配上，但是同一实体的数目。

2) 服务的可用性，指社交网络为用户提供信息的便捷性和及时性，它反映了通过跨社交网络用户身份识别后用户获得的服务质量。

3) 计算开销，指跨社交网络用户身份识别的计算开销，包括预计算、运行时发生的存储和计算代价。存储代价主要发生在预计算时，预计算的代价在现有的算法中基本被忽略。运行时的计算代价一般根据算法的时间复杂度来度量计算开销。

本文对跨社交网络用户身份识别的研究现状进行了综述，并对目前该领域已有的研究成果进行了分类，介绍了 3 类跨社交网络用户身份识别技术。每类方法都有各自的特点，针对不同的应用需求，将它们的分析比较结果列在表 1 中，可以看出它们的适用范围、性能表现等不尽相同。通过对比分析，这 3 类识别技术都具有很高的识别度，但在计算开销、数据缺失和用户数据获取的难易程度上来说所表现的性能却不尽相同。基于网络拓扑结构和用户生成内容的实体识别技术在这 3 个方面表现要优于基于用户档案信息的实体识别技术。表 2 对跨社交网络用户身份识别作了进一步的对比分析，并给出了 3 类识别技术的优缺点以及典型代表方法。

**表 1　跨社交网络用户身份识别技术性能评估**

| 用户身份识别技术 | 实体识别度 | 计算开销 | 数据缺失 | 数据获取难易程度 |
|---|---|---|---|---|
| 基于用户档案信息 | 高 | 中 | 高 | 高 |
| 基于网络拓扑结构 | 高 | 高 | 低 | 低 |
| 基于用户生成内容 | 高 | 低 | 低 | 中 |

**表 2　跨社交网络用户身份识别技术对比分析**

| 用户身份识别技术 | 主要优点 | 主要缺点 | 代表方法 |
|---|---|---|---|
| 基于用户档案信息 | 实体识别度高，实现简单 | 数据缺失严重且易出现伪造 | FOAF匹配架构[46]、MADM[48]、UISN-UD模型[56] |
| 基于网络拓扑结构 | 数据较易获取且完整 | 存在网络异构性 | 去匿名化算法[59]、COSNET模型[65]、FRUI-P[30] |
| 基于用户生成内容 | 信息缺失性低且计算开销低 | 部分用户缺乏有效的行为信息 | 贝叶斯模型[71]、MNA算法[61]、U-UIM[35] |

## 6　未来研究方向

在大数据时代，获取信息的渠道越来越多，获取到的用户信息也越来越多样化。下面从用户数据权值分配、多维度数据融合和大规模用户身份识别 3 个方面来介绍未来的数据挖掘领域中用户身份识别的研究趋势。

### 6.1　用户数据权值分配

用户的不同数据类型会对用户身份识别度产生不同的影响。因此，合理的权值分配是必不可少的。在确定用户数据中各个数据项的权重系数时，传统的专家主观赋权法和客观赋权法会存在鲁棒性差和普适性较差的问题。研究人员通过变种熵值来对每个待识别源账号进行识别，然而当应用场景发生改变时，数据的权重系数也需要重新分配，因此会产生较大的计算量。

在信息论中，熵值的大小反映了信息的无序化程度，其值越小，则含有的信息量就越多。因此，可用信息熵来评价所用数据的有序性及有效性。熵值是通过计算用户数据概率得到的，为了使用户数据概率的描述更加准确，为各个数据分配更有效的权重。在文献[50] 信息熵的基础上进一步计算用户数据的后验概率，对提高用户身份识别准确率有一定积极作用。通过将用户数据的后验概率和信息熵结合，可以有效地为相关数据进行权值分配。当识别的用户数量发生变化时，其对应的后验概率是不变的，因此，可以大大降低识别过程中的计算量。如果在后验概率的基础上进一步计算用户数据项的权重系数，进行二级权重分配会不会产生更好的识别效果，这也将是我们下一步的研究工作。

### 6.2　基于多维度数据融合的用户身份识别

基于多维度数据融合的用户身份识别是指综合利用上述两种或三种用户数据类型进行识别用户。

针对一些特殊机构，需要高准确度识别用户的身份，这时利用单一维度信息进行实体识别就具有一定的局限性。相关研究在利用网络拓扑结构进行实体识别时，弥补了非好友关系的作用，提出了亲密度函数来判别好友关系和非好友关系对识别用户的重要程度，并采用一些匹配算法将用户档案信息和包含好友关系和非好友关系在内的链接关系进行统一，用来解决实体识别问题。

此外，相似的研究还综合考虑了网络拓扑结构、用户档案信息、用户生成内容之间的信息交互[83]，来实现社交网络的虚拟账号匹配。目前，融合多维度用户数据的跨社交网络用户身份识别的研究工作并很多，这方面的研究工作将是未来用户身份识别的重要组成部分。

此外，融合用户信息虽然可以提高用户身份识别的性能。然而，这样也给恶意攻击者提供了一条获取正常用户信息的途径，因此，从博弈论的角度来均衡好用户数据和隐私保护方面的问题也将是未来的研究热点。

### 6.3 大规模用户身份识别

现有的跨社交网络用户身份识别在待识别用户数量过大时，识别性能会随着用户数量的增加而呈现降低的负相关趋势。复杂网络中的社区发现可以有效地将社交网络中大规模用户分为不同的社区，利用社区之间的关系识别用户身份也将是未来最有潜力的研究方向。

## 7 结 束 语

本文在复杂网络的视角下，综述了近十多年来跨社交网络用户身份识别技术的研究现状。目前，用户身份识别的方法已经发展得比较成熟并在诸多领域中占有重要地位。这些方法可以帮助社交网络更好地为用户提供服务，并减少网络资源的消耗。本文首先对跨社交网络用户身份识别的概念和问题进行了阐述，然后从 3 个方面对现有的研究工作在模型、相似度计算方法、识别框架、研究现状以及性能评估等方面展开了比较和分析。最后，结合现有的研究工作对未来跨社交网络用户身份识别的研究方向进行了探讨。总之，跨社交网络用户身份识别属于大数据时代引领的新兴研究领域，仍然有许多关键性的问题需要进行深入细致的研究。

## 参 考 文 献

[1] LIU X Z, XIA T, YU Y Y, et al. Cross social media recommendation[C]//The International AAAI Conference on Web and Social Media. [S.l.]: AAAI, 2016: 1-10.

[2] ZAFARANI R, TANG L, LIU H. User identification across social media[J]. ACM Transactions on Knowledge Discovery from Data, 2015, 10(2): 1-30.

[3] LIU K, ZHANG L M, ZHOU L J. Survey of deep learning applied in information recommendation system[J]. Journal of Chinese Computer Systems, 2019, 40(4): 738-743.

[4] 陈玲姣, 蔡世民, 张千明, 等. 基于信任关系的资源分配推荐算法改进研究[J]. 电子科技大学学报, 2019, 48(3): 449-455.
CHEN Ling-jiao, CAI Shi-min, ZHANG Qian-ming, et al. Improved research on resource-allocation recommendation algorithm based on trust relationship[J]. Journal of University of Electronic Science and Technology of China, 2019, 48(3): 449-455.

[5] ZHONG E H, FAN W, YANG Q. User behavior learning and transfer in composite social networks[J]. ACM Transactions on Knowledge Discovery from Data, 2014, 8(1): 1-32.

[6] SHEN W, WANG J Y, LUO P, et al. Linking named entities in Tweets with knowledge base via user interest modeling[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2013: 68-76.

[7] 邵鹏, 胡平. 复杂网络特殊用户对群体观点演化的影响[J]. 电子科技大学学报, 2019, 48(4): 604-612.
SHAO Peng, HU Ping. The influence mechanism of special members on opinion evolution of group members in complex network[J]. Journal of University of Electronic Science and Technology of China, 2019, 48(4): 604-612.

[8] WICKER S B. The loss of location privacy in the cellular age[J]. Communications of the ACM, 2012, 55(8): 60-68.

[9] NARAYANAN A, SHMATIKOV V. Robust de-anonymization of large sparse datasets[C]//Proceedings of the IEEE Symposium on Security and Privacy. [S.l.]: IEEE, 2008: 111-125.

[10] 王璐, 孟小峰. 位置大数据隐私保护研究综述[J]. 软件学报, 2014, 25(4): 693-712.
WANG Lu, MENG Xiao-feng. Location privacy preservation in big data era: A survey[J]. Journal of Software, 2014, 25(4): 693-712.

[11] 陈晨. 面向 Web 文本挖掘的主题网络爬虫研究[D]. 成都: 电子科技大学, 2017.
CHEN Chen. Research on web crawler for web text mining[D]. Chengdu: University of Electronic Science and Technology of China, 2017.

[12] 朱军芳, 陈端兵, 周涛, 等. 网络科学中相对重要节点挖掘方法综述[J]. 电子科技大学学报, 2019, 48(4): 595-603.
ZHU Jun-fang, CHEN Duan-bing, ZHOU Tao, et al. A survey on mining relatively important nodes in network science[J]. Journal of University of Electronic Science and Technology of China, 2019, 48(4): 595-603.

[13] LI Y J, LIU B. A normalized Levenshtein distance metric[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1091-1095.

[14] KONDRAK G, MARCU D, KNIGHT K. Cognates can improve statistical translation models[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. [S.l.]: ACM, 2003: 46-48.

[15] CALIANO D, FERSINI E, MANCHANDA P, et al. UniMiB: Entity linking in tweets using Jaro-Winkler distance, popularity and coherence[C]//Proceedings of the 6th International Workshop on Making Sense of Microposts. [S.l.]: Microposts, 2016: 70-72.

[16] LIU D, WU Q, HAN W, et al. User identification across multiple websites based on username features[J]. Chinese Journal Computers, 2015, 38(10): 2028-2040.

[17] VOSECKY J, HONG D, SHEN V Y. User identification across multiple social networks[C]//Proceedings of the 2009 First International Conference on Networked Digital Technologies. [S.l.]: IEEE, 2009: 360-365.

[18] 赵胜辉, 李吉月, 徐碧, 等. 基于 TFIDF 的社区问答系统问句相似度改进算法[J]. 北京理工大学学报, 2017, 37(9): 982-985.
ZHAO Sheng-hui, LI Ji-yue, XU Bi, et al. Improved TFIDF-based question similarity algorithm for community interlocution system[J]. Transactions of Beijing Institute of Technology, 2017, 37(9): 982-985.

[19] LI Y J, PENG Y, ZHANG Z, et al. A deep dive into user display names across social networks[J]. Information Sciences, 2018, 447: 186-204.

[20] FUGLEDE B, TOPSOE F. Jensen-Shannon divergence and Hilbert space embedding[C]//International Symposium on Information Theory. [S.l.]: IEEE, 2005, DOI: 10.1109/ISIT.2004.1365067.

[21] 吴铮. 跨社交网络用户多重身份识别算法研究[D]. 郑州: 解放军信息工程大学, 2017.
WU Zheng. Research on user identification algorithms across multiple online social networks[D]. Zhengzhou: The PLA Information Engineering University, 2017.

[22] REXFORD J, DOVROLIS C. Future internet architecture: clean-slate versus evolutionary research[J]. Communications of the ACM, 2010, 53(9): 36-40. DOI: 10.1145/1810891.1810906.

[23] ADAMIC L A, ADAR E. Friends and neighbors on the web[J]. Social Networks, 2003, 25(3): 211-230.

[24] GREENHALGH A, HUICI F, HOERDT M, et al. Flow processing and the rise of commodity network hardware[J]. ACM SIGCOMM Computer Communication Review, 2009, 39(2): 20-26.

[25] MACCHERANI E, FEMMINELLA M, LEE J W, et al. Extending the NetServ autonomic management capabilities using OpenFlow[C]//IEEE Network Operations and Management Symposium. [S.l.]: IEEE, 2012: 582-585.

[26] KIM H, FEAMSTER N. Improving network management with software defined networking[J]. IEEE Communications Magazine, 2013, 51(2): 114-119.

[27] QAZI Z A, TU C C, CHIANG L, et al. SIMPLY-fying middlebox policy enforcement using SDN[C]// Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM. [S.l.]: ACM, 2013: 27-38.

[28] MILOJEVIC S. Modes of collaboration in modern science: Beyond power laws and preferential attachment[J]. Journal of the Association for Information Science and Technology, 2010, 61(7): 1410-1423.

[29] ZHOU X P, LIANG X, ZHANG H Y, et al. Cross-platform identification of anonymous identical users in multiple social media networks[J]. IEEE Trans Knowl Data Eng, 2016, 28(2): 411-424.

[30] ZHOU X P, LIANG X, DU X Y, et al. Structure based user identification across social networks[J]. IEEE Trans Knowl Data Eng, 2018, 30(6): 1178-1191.

[31] NGUYEN H V, BAI L. Cosine similarity metric learning for face verification[C]//Asian Conference on Computer Vision. [S.l.]: Springer-Verlag, 2010: 709-720.

[32] LIU D H, CHEN X H, PENG D. Some cosine similarity measures and distance measures between q-rung orthopair fuzzy sets[J]. International Journal of Intelligent Systems, 2019, 34(7): 1-16.

[33] 黄丹阳, 王菲菲, 杨扬, 等. 基于网络结构与用户内容的动态兴趣识别方法[J]. 北京邮电大学学报, 2018, 41(2): 103-108.
HUANG Dan-yang, WANG Fei-fei, YANG Yang, et al. Dynamic interest identification based on social network structure and user generated contents[J]. Journal of Beijing University of Posts and Telecommunications, 2018, 41(2): 103-108.

[34] 毕娟, 秦志光. 基于概率主题模型的社交网络层次化社区发现算法[J]. 电子科技大学学报, 2014, 43(6): 898-903.
BI Juan, QIN Zhi-guang. Hierarchical community discovery for social networks based on probabilistic topic model[J]. Journal of University of Electronic Science and Technology of China, 2014, 43(6): 898-903.

[35] LI Y J, ZHANG Z, PENG Y, et al. Matching user accounts based on user generated content across social networks[J]. Future Generation Computer Systems, 2018, 83: 104-115.

[36] NIWATTANAKUL S, SINGTHONGCHAI J, NAENUDORM E, et al. Using of Jaccard coefficient for keywords similarity[C]//IAENG International Conference on Internet Computing. Hong Kong, China: IAENG, 2013: 380-384.

[37] MA J T, QIAO Y Q, HU G W, et al. Social account linking via weighted bipartite graph matching[J]. International Journal of Communication Systems, 2018, 31(7): e3471.

[38] MODI S, SHAGARI N M, WADATA B. Implementation of stable marriage algorithm in student project allocation[J]. Asian Journal of Research in Computer Science, 2018, 1(4): 1-9.

[39] ZAFARANI R, LIU H. Connecting corresponding identities across communities[C]//International Conference on Weblogs and Social Media. [S.l.]: AAAI, 2009, 9: 354-357.

[40] PERITO D, CASTELLUCCIA C, KAAFAR M A, et al. How unique and traceable are usernames?[J]. International

Symposium on Privacy Enhancing Technologies Symposium, 2011, 6794: 1-17.

[41] LIU J, ZHANG F, SONG X Y, et al. What's in a name?: An unsupervised approach to link users across communities[C]//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. [S.l.]: ACM, 2013: 495-504.

[42] ZAFARANI R, LIU H. Connecting users across social media sites: A behavioral-modeling approach[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13). [S.l.]: ACM, 2013: 41-49.

[43] WANG Y B, LIU T W, TAN Q F, et al. Identifying users across different sites using usernames[J]. Procedia Computer Science, 2016, 80: 376-385.

[44] LI Y J, PENG Y, JI W L, et al. User Identification based on display names across online social networks[J]. IEEE Access, 2017, 5: 17342-17353.

[45] MOTOYAMA M, VARGHESE G. I seek you: Searching and matching individuals in social networks[C]// Proceedings of the 11th International Workshop on Web Information and Data Management. HongKong, China: ACM, 2009: 67-75.

[46] RAAD E, CHBEIR R, DIPANDA A. User profile matching in social networks[C]//Proceedings of the 13th International Conference on Network-Based Information Systems. [S.l.]: IEEE, 2010: 297-304.

[47] IOFCIU T, FANKHAUSER P, ABEL F, et al. Identifying users across social tagging systems[C]//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. [S.l.]: AAAI, 2011: 522-525.

[48] YE N, ZHAO Y L, DONG L L, et al. User identification based on multiple attribute decision making in social networks[J]. China Communications, 2013, 10(12): 37-49.

[49] 吴铮, 于洪涛, 刘树新, 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.
WU Zheng, YU Hong-tao, LIU Shu-xin, et al. User identification across multiple social networks based on information entropy[J]. Journal of Computer Applications, 2017, 37(8): 2374-2380.

[50] DENG K K, XING L, ZHENG L S, et al. A user identification algorithm based on user behavior analysis in social networks[J]. IEEE Access, 2019, 9: 47114-47123.

[51] GOGA O, PERITO D, LEI H, et al. Large-scale correlation of accounts across social networks[EB/OL]. [2019-05-06]. http://www.icsi.berkeley.edu/pubs/techreports/ICSI_TR-13-002.pdf.

[52] LI H X, ZHU H J, DU S G, et al. Privacy leakage of location sharing in mobile social networks: Attacks and defense[J]. IEEE Transactions on Dependable and Secure Computing, 2018, 15(4): 646-660.

[53] VENI R H, REDDY A H, KESAVULU C. Identifying malicious web links and their attack types in social networks[J]. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2018, 3(4): 1060-1066.

[54] ZAMANI K, PALIOURAS G, VOGIATZIS D. Similarity-based user identification across social networks[C]//International Workshop on Similarity-based Pattern Recognition. [S.l.]: Springer, 2015: 171-185.

[55] ESFANDYARI A, ZIGNANI M, GAITO S, et al. User identification across online social networks in practice: Pitfalls and solutions[J]. Journal of Information Science, 2016, 44(3): 377-391.

[56] LI Y J, PENG Y, ZHANG Z, et al. Matching user accounts across social networks based on username and display name[J]. World Wide Web, 2018, 22(7): 1-23.

[57] LI Y J, PENG Y, ZHANG Z, et al. Understanding the user display names across social networks[C]//Proceedings of the 26th International World Wide Web Conference Committee (IW3C2). [S.l.]: ACM, 2017: 1319-1326.

[58] MISHRA R. Entity resolution in online multiple social networks[J]. Emerging Technologies in Data Mining and Information Security, 2019, 813: 221-237.

[59] NARAYANAN A, SHMATIKOV V. De-anonymizing social networks[C]//The 30th IEEE Symposium on Security and Privacy. [S.l.]: IEEE, 2009: 173-187.

[60] CUI Y, PEI J, TANG G T, et al. Finding email correspondents in online social networks[J]. World Wide Web, 2013, 16(2): 195-218.

[61] KONG X N, ZHANG J W, YU P S. Inferring anchor links across multiple heterogeneous social networks[C]//ACM International Conference on Information & Knowledge Management. [S.l.]: ACM, 2013: 179-188.

[62] KORULA N, LATTANZI S. An efficient reconciliation algorithm for social networks[J]. Proceedings of the VLDB Endowment, 2014, 7(5): 377-388.

[63] TAN S L, GUAN Z Y, CAI D, et al. Mapping users across networks by manifold alignment on hypergraph[C]// Proceedings of the 28th AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2014, 14: 159-165.

[64] JIN T S, YU Z T, GAO Y, et al. Robust $\ell 2-$ Hypergraph and its applications[J]. Information Sciences, 2019, 501: 708-723.

[65] ZHANG Y T, TANG J, YANG Z L, et al. COSNET: Connecting heterogeneous social networks with local and global consistency[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2015: 1485-1494.

[66] LEE J Y, HUSSAIN R, RIVERA V, et al. Second-level degree-based entity resolution in online social networks[J]. Social Network Analysis and Mining, 2018, 8: 19.

[67] ZHANG W, SHU K, LIU H, et al. Graph neural networks for user identity linkage[EB/OL]. [2019-11-03]. https://arxiv.org/pdf/1903.02174.pdf.

[68] WANG N, ZHOU Y D, SUN Q D, et al. A study on influential user identification in online social networks[J]. Chinese Journal of Electronics, 2016, 25(3): 467-473.

[69] SHI C, LI Y T, ZHANG J W, et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 17-37.

[70] DENG K K, XING L, ZHANG M C, et al. A multiuser identification algorithm based on internet of things[J]. Wireless Communications and Mobile Computing, 2019, DOI: 10.1155/2019/6974809.

[71] ALMISHARI M, TSUDIK G. Exploring linkability of user reviews[J]. Computer Security-ESORICS, 2012, 7459: 307-324.

[72] LIU S Y, WANG S H, ZHU F D, et al. HYDRA: Large-scale social identity linkage via heterogeneous behavior modeling[C]//ACM SIGMOD International Conference on Management of Data. [S.l.]: ACM, 2014: 51-62.

[73] NIE Y P, JIA Y, LI S D, et al. Identifying users across social networks based on dynamic core interests[J]. Neurocomputing, 2016, 210: 107-115.

[74] SHA Y, LIANG Q, ZHENG K J. Matching user accounts across social networks based on users message[J]. Procedia Computer Science, 2016, 80: 2423-2427.

[75] ROEDLER R, KERGL D, RODOSEK G D. Profile matching across online social networks based on geo-tags[J]. Advances in Nature and Biologically Inspired Computing, 2016, 419: 417-428.

[76] GOAG O, LEI H, PARTHASARATHI S H K, et al. Exploiting innocuous activity for correlating users across sites[C]//Proceedings of the 22nd International Conference on World Wide Web. [S.l.]: ACM, 2013: 447-458.

[77] CAO W, WU Z W, WANG D, et al. Automatic user identification method across heterogeneous mobility data sources[C]//IEEE 32nd International Conference on Data Engineering. [S.l.]: IEEE, 2016: 978-989.

[78] HAO T Y, ZHOU J B, CHENG Y S, et al. User identification in cyber-physical space: A case study on mobile query logs and trajectories[C]//Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. [S.l.]: ACM, 2016, 71: 1-4.

[79] HAN X H, WANG L H, XU S J, et al. Linking social network accounts by modeling user spatiotemporal habits [C]//IEEE International Conference on Intelligence and Security Informatics. [S.l.]: IEEE, 2017: 19-24.

[80] RIEDERER C, KIM Y, CHAINTREAU A, et al. Linking users across domains with location data: Theory and validation[C]//Proceedings of the 25th International Conference on World Wide Web. [S.l.]: ACM, 2016: 707-719.

[81] HAN X H, WANG L H, XU L J, et al. Social media account linkage using user-generated geo-location data[C]//IEEE Conference on Intelligence and Security Informatics. [S.l.]: IEEE, 2016: 157-162.

[82] QI M J, WANG Z Y, HE Z, et al. User identification across asynchronous mobility traijectories[J]. Sensors, 2019, 19(9): 2020.

[83] JAIN P, KUMARAGURU P, JOSHI A. @ i seek 'fb. me': Identifying users across multiple online social networks[C]//Proceedings of the 22nd international Conference on World Wide Web Companion. [S.l.]: ACM, 2013: 1259-1268.

编　辑　叶　芳