跨社交网络用户身份链接回顾与展望

薛 晖[1,2], 孙 波[1,3], 司成祥[3], 张 伟[3], 房 婧[3]

[1] 中国科学院信息工程研究所 北京 中国 100093
[2] 中国科学院大学网络空间安全学院 北京 中国 100049
[3] 国家互联网应急中心 北京 中国 100029

摘要 随着互联网的飞速发展, 社交网络平台(又称在线社交网络)也日益普及和多样化, 为了更好地利用每个社交网络平台提供的服务, 用户往往会加入多个社交网络平台。链接同一个自然人在不同社交网络平台中的账户, 称为用户身份链接。通过用户身份链接可以充分了解用户的兴趣, 极大地丰富用户画像, 进而用于数字营销和推荐系统。本文首先通过回顾用户身份链接方法在发展过程中所使用的不同特征类型, 提出了一种用户身份链接问题的通用形式化定义, 适用于属性、网络、内容、行为等各种特征类型及其任意组合。然后根据用户身份链接的特征提取和模型构建两个阶段对现有用户身份链接方法进行了分类分析, 并分别从性能、计算开销、鲁棒性维度对各类方法进行了比较和评价。而后分析了现有方法使用的不同数据集和评价指标, 说明了数据集的主要获取方式, 并给出了目前用户身份链接领域无公开公认的基准数据集的原因。最后讨论了用户身份链接存在的问题与挑战, 展望了用户身份链接的未来研究趋势。本文通过提出一种用户身份链接问题的通用定义、比较分析已有用户身份链接方法、讨论存在的问题和展望未来研究趋势, 将用户身份链接问题的现状和未来以清晰的结构化的方式进行分析展示, 有助于研究人员对该领域的相关研究形成系统性的理解和把握, 进而做出更加深入的研究工作。

关键词 社交网络; 用户身份链接; 账号关联; 锚链接预测; 用户画像
中图法分类号 TP391.4 DOI 号 10.19363/J.cnki.cn10-1380/tn.2026.03.19

Advance in user identity linkage across online social networks

XUE Hui[1,2], SUN Bo[1,3], SI Chengxiang[3], ZHANG Wei[3], FANG Jing[3]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
[3] National Internet Emergency Center, Beijing 100029, China

Abstract With the rapid development of the Internet, social network platforms (also known as online social networks) have become increasingly popular and diversified. In order to make better use of the services provided by each social network platform, users often join multiple social network platforms. Linking the accounts of the same natural person in different social network platforms is called user identity linkage. Through user identity linkage, we can fully understand the user's interests, and greatly enrich the user portrait, which is used in digital marketing and recommendation system. In this paper, by reviewing the different feature types used in the development of the user identity linkage method, a general formal definition of the user identity linkage problem is proposed, which can be applied to various feature types such as attribute, network, content, behavior and any combination of them. Then, according to the two stages of feature extraction and model construction of user identity linkage, the existing user identity linkage methods are classified and analyzed, and different methods are compared and evaluated in terms of performance, computing cost and robustness. Then, different datasets and evaluation indicators used by existing methods are analyzed, the main methods of obtaining datasets are explained, and the reason why there is no publicly recognized benchmark datasets in the field of user identity linkage is given. Finally, the problems and challenges of user identity linkage are discussed, and the future research trend of user identity linkage is forecasted. By proposing a general definition of user identity linkage problem, comparing and analyzing existing user identity linkage methods, discussing existing problems and looking forward to future research trends, this paper analyzes and presents the current situation and future of user identity linkage problem in a clear and structured way, which helps researchers to form a systematic understanding and grasp of related research in this field, and then make more in-depth research work.

Key words online social network; user identity linkage; account association; anchor link prediction; user profile

通讯作者: 孙波, 博士, 正高级工程师, 主要研究领域为网络攻防、大数据, Email:sunbo@cert.org.cn。
收稿日期: 2020-12-30; 修改日期: 2021-03-15; 定稿日期: 2023-08-11

1 简介

据中国互联网络信息中心统计, 截至 2020 年 3 月, 我国网民规模为 9.04 亿, 较 2018 年底增长 7508 万, 互联网普及率达 64.5%, 较 2018 年底提升 4.9 个百分点[1]。随着互联网的飞速发展, 社交网络平台(又称在线社交网络)也日益普及和多样化, 国内有微信朋友圈、QQ 空间、新浪微博、豆瓣、人人网等, 国外则有 Facebook、X、Instagram、LinkedIn 等。由于功能的多样性, 不同的社交网络平台具有不同的功能特点, 如社会关系维护、信息共享等。例如, 用户可以在微信朋友圈分享生活近况, 而使用新浪微博评论新闻事件, 或者使用 X 发布对政治事件的看法, 而在 Instagram 中分享他们的休闲活动。为了更好地利用每个社交网络平台提供的服务, 用户往往会加入多个社交网络平台, 因而用户在多个社交网络平台上拥有账户(也称为用户身份)变得越来越流行。研究显示, 69%的在线用户同时使用多个社交网络平台, 93%的 Instagram 用户同时使用 Facebook, 53%的 X 用户也在使用 Instagram, 82%的 LinkedIn 用户和 91%的 X 用户也使用 Facebook 来享受不同社交网络平台的服务[2-4]。

链接同一个自然人在不同社交网络平台中的账户, 称为用户身份链接(User Identity Linkage, UIL, 也被称为 Social Identity Linkage、User Identity Resolution、Social Network Reconciliation、User Account Linkage Inference、Profile Linkage、Anchor Link Prediction、User Alignment、Detecting Me Edges), 是一个具有重大研究挑战和实用价值的重要问题。由于人们使用各种社交网络平台的目的不同, 各个平台上的信息通常是不完整且具有互补性的, 因而通过用户身份链接将用户在多个社交网络平台上的信息整合起来可以充分了解用户的兴趣, 极大地丰富用户画像, 进而用于数字营销和推荐系统[5], 以获得更好的商业智能。此外, 用户身份链接还有助于促进各种应用, 包括用户行为理解、朋友推荐、分析用户迁移模式、影响评估和社交网络中的专家发现等[2-3,6-9]。

在用户身份链接问题相关研究中, 研究人员尝试使用各种类型的数据、各种技术路线, 对该问题进行研究和实验, 采用不同的评价指标对实验结果进行分析。研究当中对用户身份链接问题的定义往往由于所使用的数据特点和技术路线不同而形式各异。

此外, 对已有用户身份链接方法的系统的分类比较分析, 也是一个需要解决的重要问题。

本文的贡献如下:

1) 我们提出了一种用户身份链接问题的通用定义, 可覆盖现有的基于属性、网络、内容、行为的定义方法。

2) 从特征提取和模型构建两个方面对现有用户身份链接方法进行了分类分析, 并分别从性能、计算开销、鲁棒性维度对各类方法进行了比较和评价。

3) 讨论了现有方法使用的不同数据集和评价指标, 说明了数据集的主要获取方式, 并给出了目前用户身份链接领域无公开公认的基准数据集的原因。

4) 讨论了用户身份链接问题目前存在的挑战, 以及未来研究趋势。

本文的其余部分组织如下: 第 2 节回顾了用户身份链接方法的发展历史; 第 3 节根据用户身份链接的特征提取和模型构建两个阶段对现有用户身份链接方法进行了分类分析; 第 4 节总结了现有方法所采用的数据集和评估指标; 第 5 节讨论了存在的问题与挑战; 第 6 节展望了用户身份链接的未来研究趋势; 第 7 节对全文进行了总结。

2 发展历史

Zafarani 等人[10]于 2009 年首次提出用户身份链接问题, 即将现实世界中属于同一个自然人的多个社交网络平台的用户连接起来, 并利用用户名提出了一种基于 Web 搜索的方法来解决该问题。后来, 文献[9,11-13]通过对包含更丰富信息的用户网络档案进行特征描述来实现用户身份链接。Motoyama 等人[13]首先使用用户网络档案信息(如性别、位置、职业和大学)来链接账户并帮助新用户找到他们朋友的账户。为了合并来自不同社交服务的用户联系人或为许多社交应用程序构建更完整的社交图, 文献[11]设计了一种基于条件随机场的用户网络档案匹配方法, 特别适合网络档案数据不完整或由于隐私设置而被隐藏的情况。Liu 等人[9]首先提出了关于网络用户名使用的人类行为来解决用户身份链接问题, 并提出了别名消除歧义的步骤来区分具有相同用户名的用户。此外, Zafarani 等人[12]利用人们选择用户名时所表现出的特定行为模式构建行为模型 MOBIUS, 以发现多个社交网络平台上个人身份之间的映射。

与上述工作不同, Liu 等人[14]提出了一个通过异构行为建模实现用户身份链接的框架 HYDRA, 主要利用沿着时间维度的用户行为轨迹和用户的核心社交网络结构两类社交数据, 框架由三个部分组成, 即用户行为建模、结构一致性图构建、映射函数学习。在后续的研究中, Zhang 等人[15]提出了一种新的

基于能量的模型 COSNET 来解决多网络用户身份链接潜在的不一致性问题, 该模型考虑了多网络之间的局部一致性和全局一致性。文献[16]研究利用位置数据进行用户身份链接, 其中位置和时间被划分为 bin, 并根据这些 bin 测量用户对的得分。

另一些研究则使用用户生成内容(User Generated Content, UGC)来进行用户身份链接。Jain 等人[17]利用用户在X上的社交网络结构和发布的内容(tweet)来找到他在 Facebook 上的身份。Goga 等人[18]利用了 Yelp、Flickr 和 X 上的三个特定特征: 附加在用户帖子上的地理位置、帖子的时间戳以及由语言模型捕获的用户的写作风格, 作者证明该方法超过了仅基于用户名的用户身份链接方法。Sha 等人[19]使用词嵌入技术将所有用户历史消息转化为向量, 并通过计算这些向量之间的相似性来实现用户身份链接。

总之, 用户身份链接从最初仅基于用户名逐渐扩展到利用用户网络档案的多种属性, 之后又出现了使用用户社交网络结构、用户生成内容以及用户行为模式的方法, 如图 1 所示。

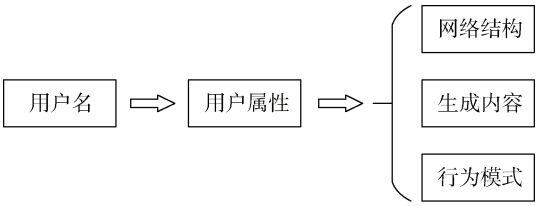

图 1 用户身份链接的发展历史

Figure 1 History of user identity linkage

用户身份链接问题可形式化定义如下:

用户身份链接问题定义

符号表示: 令 $G=\{V, E, A, C, B\}$表示一个社交网络平台, 其中 $V=\{v_1, v_2, \ldots, v_n\}$表示该社交网络平台中的所有 n 个用户; $E \subseteq V \times V$ 表示用户间的社交关系; A、C、B 都是 n 维向量, 向量元素分别为用户的属性、内容和行为。

输入: 一个源社交网络平台 $G^s=\{V^s, E^s, A^s, C^s, B^s\}$, 和一个目标社交网络平台 $G^t=\{V^t, E^t, A^t, C^t, B^t\}$。

输出: 一个用户对集合 $P \subseteq V^s \times V^t$, 表示从源社交网络平台到目标社交网络平台已链接的用户身份。

3 用户身份链接方法的分类

用户身份链接通常包括两个主要阶段: 特征提取和模型构建。在特征提取阶段, 选择可有效区分用户的数据特征并进行特定表示; 在模型构建阶段, 利用提取的特征构建模型以对用户表示进行分类或聚类, 再进一步进行用户身份链接。

3.1 根据特征提取阶段分类

根据用户身份链接所使用的特征类型, 可以将其分为基于属性、基于网络、基于内容、基于行为四种类型, 如图 2 所示。部分研究同时采用了多种类型的特征。

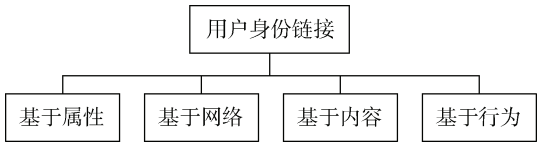

图 2 基于特征提取方法的用户身份链接分类

Figure 2 User identity linkage methods classification by feature extraction stage

1) 基于属性的 UIL

最初, 研究人员试图获取用户的个人资料特征, 即用户名、性别、出生日期、兴趣、职业、地理位置等, 以识别同一个自然人在不同社交网络平台上的身份。

Zafarani 等人[10]的研究表明, 66%的用户在不同的社交网络平台上使用相似的用户名, 他们通过添加或删除用户名的常见前缀或后缀来利用用户名进行用户身份链接。此外, Perito 等人[20]通过建立马尔可夫链模型来估计用户名的唯一性, 并使用编辑距离和词频反文档频率(TF-IDF)两种相似度方法对用户名进行比较。类似地, Liu 等人[9]提出了一种无监督方法, 该方法采用 n-gram 模型来估计用户名的唯一性。Iofciu 等人[21]通过结合用户名相似性和标签相似性来进行跨平台的用户身份链接。此外, Zhang 等人[15]使用 Jaro-Winkler 距离来度量用户名相似性, 并通过训练马尔可夫链模型来计算用户名唯一性。他们将与用户个人资料相关的所有信息合并到一个文档中, 并将文档转换为词袋向量, 其中的词由 TF-IDF 加权。然后用内积和余弦距离来测量两个个人资料的相似度。类似地, Malhotra 等人[22]使用 Jaro-Winkler 相似性和基于 WordNet 的本体技术来测量不同社交网络上用户个人资料的相似性。Vosecky 等人[23]为每个用户属性分配权重, 为了确定属性相似度, 区分了精确匹配、部分匹配和模糊匹配三类匹配。Raad 等人[24]通过提供一个合适的匹配框架来解决全局匹配用户网络档案的问题, 该框架能够考虑所有网络档案属性, 允许用户给予某些属性更多的重要性, 并为每个属性分配不同的相似性度量。

用户在多个社交网络中的属性大多是相似的,

如用户名、性别、出生日期等, 基于属性的 UIL 利用这一特点, 通过文本相似性比较、知识图谱、加权等技术对来自不同社交网络的用户属性进行比较, 根据用户属性的相似程度进行用户身份链接。如图 3 所示, u 和 v 是来自不同社交网络平台的两个用户身份, 基于属性的 UIL 通过综合计算它们的各项个人资料属性信息的相似度来进行身份链接。

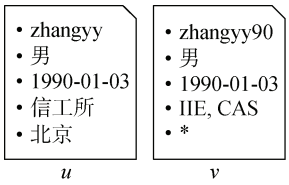

图 3 基于属性的 UIL
Figure 3 UIL based on attribute

虽然这类方法可以取得良好的性能, 但最大的挑战是真实性和完整性, 因为由于隐私问题用户通常不会提供真实和完整的个人资料。因此, 在无法保证个人资料准确性的情况下, 基于属性的用户身份链接无法获得良好的结果。

2) 基于网络的 UIL

用户在社交网络平台上维护的与该平台其他用户的社会关系, 称为社交网络结构。研究显示, 用户在多个社交网络平台的网络结构之间具有一定的重叠度, 因此, 许多研究利用这一特点进行用户身份链接。

Tan 等人[25]通过超图上的流形对齐, 将两个社交网络平台中的用户映射到一个公共的嵌入空间, 然后采用基于超图的子空间学习算法学习最优空间, 最后通过余弦相似度计算用户相似性。类似地, Cui 等人[26]使用图形匹配方法来识别跨社交网络平台的电子邮件联系人, 该方法将用户属性信息和社交网络结构信息组合在一起, 以发现电子邮件网络和 Facebook 之间的用户关系。此外, Man 等人[27]提出了一种新的监督模型 PALE, 该模型将已知锚链接作为监督信息, 采用网络嵌入方法来捕获主要和特定的结构规律, 并进一步学习一个稳定的跨网络映射用于预测锚链接。Zhang 等人[15]提出了一种新的基于能量的模型 COSNET 来解决多网络用户身份链接潜在的不一致性问题, 该模型考虑了多网络之间的局部一致性和全局一致性。Zhou 等人[28]提出了基于友谊结构的用户识别算法 FRUI, 该算法将已知匹配用户作为种子集, 每次迭代仅考虑种子集的邻居节点, 性能好且可扩展性强。Bartunov 等人[11]提出了一种基于条件随机场的用户身份链接新方法, 该方法结合使用了用户属性和用户社交网络结构, 适合用户属性不完整或被隐藏的情况。Koutra 等人[29]将用户身份链接问题转化为二部图对齐问题, 提出了一种使用随机梯度下降的优化算法, 并给出了一种快速有效的求解方法。Liang 等人[30]定义了一个用户亲密度指标, 除源用户的好友外, 将亲密度足够高的用户也作为源用户社交网络结构的一部分, 并结合用户属性和用户社交网络结构进行用户身份链接。

用户在不同社交网络中的社会关系往往具有一定的相似性, 基于网络的 UIL 利用这一特点, 通过图匹配、二部图对齐、加权等技术对来自不同社交网络的用户社会关系进行比较, 根据用户社会关系的相似程度进行用户身份链接。如图 4 所示, 左边的图(1)和右边的图(2)是来自不同社交网络平台的两个用户身份的社会关系网络, 基于网络的 UIL 综合考虑这两个网络的相似度和节点间的对应关系, 以此来进行身份链接。

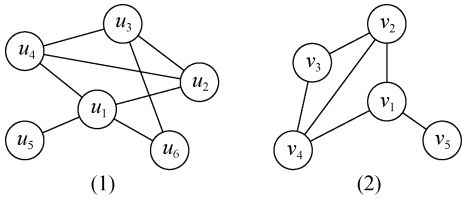

图 4 基于网络的 UIL
Figure 4 UIL based on network

当用户在多个社交网络平台的网络结构之间的重叠度较低时, 基于网络的 UIL 方法效果不佳。此外, 该方法很难推广到大规模的社交网络平台中, 因为该方法对于稀疏社交网络平台的计算开销较大。

3) 基于内容的 UIL

用户生成内容(User Generated Content, UGC)通常指用户在社交网络平台创作发布的文本、图片、视频等多媒体内容信息。用户生成内容中体现出的写作风格、兴趣、语义等可以用于识别用户, 从而进行用户身份链接。

许多用户喜欢在不同的社交网络平台上同步他们的活动, 比如在 X 和 Flickr 上发布了同样的照片。Goga 等人[18]利用了用户生成内容的一些特性来进行用户身份链接, 主要包括帖子的地理位置、帖子的时间戳和用户的写作风格。作者证明其性能超过了仅基于用户名的方法。Nie 等人[31]提出了一种动态核心兴趣映射(dynamic core interests mapping, DCIM)算法, 该算法结合用户的社交网络结构和用户的文章内容来识别平台上的用户。该算法首先对用户的核心兴趣进行建模, 然后利用 DCIM 计算两个目标用户的相似度。Mishari 等人[32]发现, 用户的写作风格可以作为

链接用户账户的有用特征。他们使用用户的语言模型和写作风格来建立一个特征模型和链接用户账户。由于许多社交网络平台中的用户个人资料页面是不可见的, Liu 等人[9]从用户生成内容中提取头像、位置、签名、与其他用户的交互关系、写作风格等作为特征进行用户身份链接。Sha 等人[19]使用词嵌入技术将所有用户历史消息转化为向量, 并通过计算这些向量之间的相似性来实现 UIL。Van Le 等人[33]利用潜在狄利克雷分布(LDA)发现用户生成内容中的隐藏主题, 进而根据主题分布进行用户身份链接。Jain 等人[17]利用用户在 X 上的社交网络结构和发布的内容(tweet)来找到他在 Facebook 上的身份。

用户生成内容可以体现出用户的兴趣爱好、写作风格等, 而用户的这些特点在一段时期内往往比较稳定。基于内容的 UIL 通过文本相似度比较、自然语言处理、图片相似度比较等技术, 发现来自不同社交网络的用户生成内容之间的相似性, 从而实现用户身份链接。如图 5 所示, 左图是豆瓣某用户发表的评论内容, 右图是微博某用户发表的评论内容, 基于内容的 UIL 计算这些内容间的文本相似度、图片相似度, 并可能结合知识图谱技术计算内容中体现出的写作风格、兴趣爱好等的相似度, 最终通过对这些因素进行综合加权来实现身份链接。

豆瓣　　微博

图 5　基于内容的 UIL

Figure 5　UIL based on content

各个社交网络平台的内容存在文本、图片、视频等多种形式, 而且很多用户发布内容的频率较低, 异构性与稀疏性使基于内容的 UIL 存在一定的局限性。

4) 基于行为的 UIL

实证研究和社会行为研究都表明, 在足够长的一段时间内, 用户的社会行为在不同社交网络平台上表现出惊人的高度一致性。与用户属性相比, 用户行为是独特的、不容易被模仿的, 因此, 一些研究利用用户在互联网上的各种活动进行用户身份链接, 如时空轨迹、Web 日志、在线兴趣等, 这种方法称为基于行为的 UIL。

由于环境、个性或人的局限而产生的独特行为, 会在社交网络平台上产生冗余信息。基于社会学和心理学中的行为理论, MOBIUS[12]利用人们在选择用户名时的行为模式导致的信息冗余来识别社交网络平台上的用户。HYDRA[14,34]利用多维相似向量对社交网络用户之间的行为相似度进行建模, 模型包含以下信息: (a) 用户属性的相对重要性, 衡量两个用户在其中一个属性相同时属于同一个人的可能性; (b) 主题分布的统计差异, 描述用户长期的潜在倾向; (c) 行为轨迹的整体匹配程度, 捕捉一定时间内用户账号之间相同的动作。HYDRA 通过结合行为相似度建模和社交网络结构建模来进行用户身份链接。此外, Kong 等人[35]提出了一种基于用户的社交网络结构、地理位置、时间和文本信息的多网络锚链接方法来识别不同社交网络平台中的相关账户。Zhang 等人[36]使用用户生成内容中涉及的位置信息来链接用户账户。Roedler 等人[37]使用由社交网络生成的时间戳和由设备生成的地理位置标签进行用户身份链接。Vosoughi 等人[38]使用语言模型和时间模式来匹配来自 Facebook 和 X 的用户。Hazimeh 等人[39]提出的 SocialMatching 由两个阶段组成, 分别基于生活大事和个人描述进行用户身份链接。

在一定时期内, 用户在社交网络中的行为是比较稳定的, 如在线兴趣、时空轨迹等。基于行为的 UIL 利用该特点, 通过文本分析、时空轨迹匹配等技术对来自不同社交网络的用户行为信息进行建模分析, 进一步发现这些行为信息之间的相似性, 从而实现用户身份链接。如图 6 所示, 左边的图(1)和右边的图(2)是来自不同社交网络平台的两个用户身份的 GPS 轨迹, 通过对这些轨迹进行预处理、降噪、建模、匹配, 可以得到轨迹相似度, 然后从时间维度进行扩展, 考察一定时期内两个用户身份的若干轨迹相似度, 最终可以达到身份链接的目的。

(1)　　(2)

图 6　基于行为的 UIL

Figure 6　UIL based on behavior

然而, 出于隐私考虑, 用户行为数据的可用性比以前更低了, 而且某些行为数据仅局限于特定的社交网络平台, 这使得基于用户行为的链接方法无法扩展到一般的社交网络平台。

最后从多个维度分析比较了根据特征提取方法分类的四种用户身份链接方法, 如表 1 所示。

表 1 根据特征提取方法分类的四种用户身份链接方法分析

Table 1 Analysis of four user identity linkage methods classified by feature extraction stage

方法	性能	计算开销	鲁棒性
基于属性	好	低	低, 个人资料的真实性和完整性难以保证
基于网络	较好, 网络结构之间的重叠度较低时效果不佳	高	高, 不易被伪造
基于内容	较好, 受限于社交网络平台的内容异构性与稀疏性	中	中, 时效性好
基于行为	较好, 用户行为数据可用性低, 局限于特定社交网络平台	中	高, 用户行为具有独特性, 不易被模仿

3.2 根据模型构建阶段分类

根据模型构建方法的不同, 现有的用户身份链接方法可以分为有监督的、半监督的与无监督的三种类型, 如图 7 所示。模型的训练数据可以分为两类: 有标签的(即已知是否属于同一个自然人的多个社交网络平台账号)和无标签的(即未知是否属于同一个自然人的多个社交网络平台账号)。将有标签数据作为训练数据, 通过提取的特征训练模型, 然后进行预测的方法称为有监督的 UIL; 同时利用有标签数据和无标签数据作为训练数据的方法称为半监督的 UIL; 仅利用无标签数据作为训练数据的方法称为无监督的 UIL。

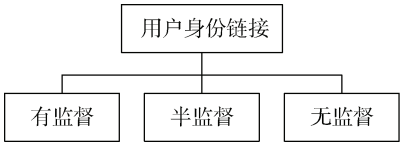

图 7 基于模型构建方法的用户身份链接分类

Figure 7 User identity linkage methods classification by model construction stage

1) 有监督的 UIL

现有的大多数工作都是有监督的, 将 UIL 问题视为一个典型的二元分类问题, 将用户身份对分为匹配的和不匹配的两类。Motoyama 等人[13]提取用户网络档案信息(生日、年龄、位置、家乡、教育程度等)作为特征, 利用已知的用户匹配关系来训练分类器, 以预测新的用户匹配关系, 并帮助新用户找到他们朋友的账户。与之类似, Vosecky 等人[23]利用 8 个用户属性特征以及人工标记的用户匹配关系训练分类器, 然后进行用户身份链接。Man 等人[27]提出了一种基于监督嵌入的 UIL 框架 PALE, 该框架将社交网络结构映射到一个低维空间中, 然后引入一种投影方法在潜在空间中进行用户身份链接。Mu 等人[40]提出的基于潜在用户空间建模的 ULink 算法首先将多个网络中的用户身份映射到一个潜在的用户空间中, 然后最小化同一人的用户身份之间的距离, 最大化不同人的用户身份之间的距离。Peled 等人[41]使用监督学习技术和提取的基于用户属性和社交网络结构的多种特征, 构建了不同的分类器, 然后训练这些分类器并用于对来自两个不同社交网络的两个用户网络档案属于同一个人的概率进行排序。Perito 等人[20]利用用户名字符串和用户名唯一性作为特征, 以监督方式训练分类器, 并将分类器用于用户身份链接。Zhang 等人[42]利用用户名、头像、语言、个人简介等用户属性特征和已知的用户匹配关系, 以监督方式训练概率分类器 OPL, 用于解决现实世界中的大型用户身份链接问题。

在已知一定数量真实用户身份关联关系的情况下, 有监督的 UIL 可以在训练集上有较好的表现, 但获取真实用户身份关联关系的代价很高, 有时甚至是难以实现的, 而且获取的训练数据是否具有代表性会直接影响模型的性能表现。

2) 半监督的 UIL

由于获取跨多个社交网络平台的有标签数据非常困难, 而且非常昂贵, 多项研究提出了几种半监督方法来综合利用有标签数据的规律性和无标签数据之间的相似性来链接用户身份。Korula 等人[43]引入了标签传播这一流行的半监督模型, 根据基于邻居的网络特征来执行 UIL 任务, 其核心思想是利用高度节点发现用户之间所有可能的映射。COSNET[15]是一个基于能量的模型, 通过考虑多网络之间的局部和全局一致性来链接用户身份。COSNET 首先提取基于距离的属性特征和基于邻居的拓扑特征, 然后使用聚合算法获得局部一致性。HYDRA[14,34]是一个多目标半监督框架, 利用传播算法对异构行为和结构一致性进行联合建模。Liu 等人[44]通过将每个用户的关注与被关注关系显示建模为输入与输出上下文向量表示来学习一个网络嵌入, 以保证拥有相似的关注与被关注关系的用户在嵌入空间中更加接近。为了促进上下文信息在网络上的传输, 他们在网络上添加了已知的和潜在的锚用户, 同时利用了有标签数据和无标签数据。Tan 等人[25]首先为每个社交

网络构建一个社交超图, 然后进行社交超图上的半监督流形对齐。通过学习一个所有用户的低维公共空间, 再比较用户相关性来完成用户映射任务。Zhong 等人[8]提出了一个通用的半监督 UIL 问题框架 CoLink, 其中使用了两个独立的模型, 一个基于属性的模型和一个基于关系的模型, 然后采用一个协同训练算法来迭代地增强它们。

半监督的 UIL 同时利用了有标签数据的规律性和无标签数据的相似性, 能够在有监督的 UIL 和无监督的 UIL 之间取得一个折中的效果, 但依然存在有标签数据获取困难等问题。

3) 无监督的 UIL

与使用有标签数据的 UIL 模型相比, 无监督用户身份链接问题的研究较少。Liu 等人[9]提出了一种无监督方法, 该方法采用 *n*-gram 模型来估计用户名的稀有性。他们先根据两个社交网络中用户名的稀有性自动生成一组训练样本, 然后利用这些样本训练二分类器。Lacoste-Julien 等人[45]提出了一种基于启发式字符串相似性的贪婪方法来对齐用户属性。POIS[16]使用基于时空轨迹的属性特征来链接用户身份。POIS 首先根据带有时间戳的位置数据计算用户身份之间的亲缘性得分, 然后利用最大加权匹配方案寻找最可能的匹配对。

无监督的 UIL 不需要有标签的训练数据, 但现有的无监督 UIL 模型往往严重依赖于强鉴别特征或跨网络公共属性, 极大地限制了其通用性。

从多个维度分析比较了根据模型构建方法分类的三种用户身份链接方法, 如表 2 所示。

表 2 根据模型构建方法分类的三种用户身份链接方法分析

Table 2 Analysis of three user identity linkage methods classified by model construction stage

方法	性能	计算开销	鲁棒性
有监督	有标签数据的数量和质量理想时性能好	高	低, 受有标签数据的质量影响大
半监督	性能较好, 有标签数据数量少或质量差时优于有监督方法	中	中, 受有标签数据的质量影响中
无监督	性能较好	低	高, 不需要有标签的训练数据

最后总结了部分参考文献的特征类型及建模方法, 如表 3 所示。

表 3 部分参考文献的特征类型及建模方法分析

Table 3 Analysis of feature types and modeling methods of some references

文献	特征类型	建模方法	技术路线
Buccafurri 等[46]	用户名+网络	半监督	基于公共邻居方法
Zafarani 等[12]	用户名	有监督	利用行为模式导致的用户名信息冗余性, 朴素贝叶斯分类
Liu 等[34]	属性+内容+网络+行为	有监督	提出一个包含异构行为模型和核心社交网络结构的多目标学习框架
Gao 等[47]	属性+网络	无监督	以集合方式集成属性和网络特征
Zhou 等[28]	网络结构	半监督	基于好友关系的 FRUI
Zhang 等[15]	属性+网络	半监督	能量模型, 考虑多个网络的全局一致性
Mo 等[48]	用户名+内容	有监督	软件开发社区 UIL, 决策树+启发式贪婪算法
Han 等[49]	时空数据	无监督	在时间和空间维度上同步进行聚类, 互相增强结果, 提高了聚类结果的一致性
Almishari 等[50]	内容	有监督	分析写作风格, 在多个级别上运行可链接性实验, 每个级别使用不同的特征类别
Liu 等[44]	网络	有监督	社交有向图, 网络对齐
Mu 等[40]	属性	无监督	提出“用户潜在空间”概念
Han 等[51]	时空数据	无监督	通过主题模型来捕获用户在空间和时间维度上的习惯模式
Ma 等[6]	属性+网络	有监督	对个人资料和网络结构两个因素加权
Li 等[7]	属性+内容+网络	无监督	从用户空间分布的层次来执行 UIL
Zhong 等[8]	属性+网络	半监督	协同训练
Zhou 等[52]	网络	半监督	深度学习+强化学习
Chen 等[53]	地理位置	无监督	核密度估计(KDE), 划分网格后计算单元格熵
Qiao 等[54]	网络流量	有监督	提取用户的数字身份、在线指纹和用户的时空行为等特征用于分类
Ma 等[4]	属性+时间分布+兴趣	无监督	加权二部图最大匹配
Wang 等[55]	时空数据	无监督	通过研究用户活动的时空局部性来跨多个服务链接用户身份
Xie 等[5]	属性+网络	无监督	Factoid 嵌入
Feng 等[56]	时空数据	无监督	深度学习, 基于地理位置
Zhang 等[57]	网络	半监督	图神经网络

4 数据集和评估指标

4.1 数据集

用户身份链接相关文献中使用的数据集多种多样, 包括属性、网络、内容、时空数据等多种类型, 如表 4 所示。文献[2]使用了两个数据集, TF 和 TI。TF 包含 5223 个 X 用户和 5392 个 Foursquare 用户, 它们共享 3388 个锚用户; TI 是自行爬取的 X 和 Instagram 数据。文献[56]使用了两个移动数据集, ISP-Weibo 和 Foursquare-X。ISP-Weibo 是中国最大的互联网服务提供商(ISP)之一提供的移动网络记录数据集, 以及从微博获取的社交网络位置服务记录。Foursquare-X 是来自 Foursquare 和 X 的相同用户的位置签到记录。文献[52]使用了三个数据集, Foursquare-X、Lastfm-MySpace 和 Livejournal-MySpace, 都是网络结构类型数据。文献[8]使用了两个真实数据集, 其中一个社交网络是 LinkedIn, 而另一个网络是一个内部企业用户网络。LinkedIn 包含 240 万份 LinkedIn 资料, 信息包括姓名、组织机构、职位名称、位置等属性, 以及查看列表中的人。企业内网包含 22 万以上用户的姓名、职称和办公地点, 以及参与共同活动的信息。文献[7]使用了两对社会网络数据集(X-Flickr 和 Weibo-Douban)和三对学术合著数据集(DBLP15、16、17), 包括属性和网络类型的数据。文献[53]使用了两个时空数据类型的真实数据集, Foursquare-X 和 Instagram-X, 其数据记录的形式是(*userid*、纬度、经度、时间戳)。文献[5]爬取了 X、Facebook 和 Foursquare 的部分数据作为数据集, 可用信息包括用户名、昵称、头像和网络结构。文献[55]使用了两个时空数据类型的真实数据集, ISP Dataset 和 X-Foursquare。ISP Dataset 包括 412455 个真实用户在 QQ、淘宝、微博、大众点评上的 815117 个用户 ID。记录了用户通过宽带网络(与地理位置相关)的访问记录。X-Foursquare 包含 385 个用户在两个网站上的 24556 条签到记录, 涉及 770 个账号。文献[58]使用了三个时空数据的真实数据集。其中 Beijing Walk Trajectories-Beijing Car Trajectories 是由微软亚洲研究院 GeoLife 项目在北京收集的移动轨迹数据, 具有不同的交通模式, 如步行、汽车、公共汽车、自行车等。Foursquare-X 和 Instagram-X 是社交网络签到数据。文献[44]使用了两个网络结构类型的真实数据集, 收集自 Foursquare 和 X。文献[16]使用了三个时空数据类型的真实数据集, Foursquare-X、Instagram-X 和 Cell Phone-Credit Card Record, 其中前两个数据集爬取自社交网络, Cell Phone-Credit Card Record 是由电信公司和信用卡公司关联的用户地理位置记录。文献[27]使用了两个网络结构类型的真实数据集, Facebook 数据集和合著网络数据集。Facebook 数据集是 2009 年 Facebook 新奥尔良地区的用户好友关系和评论关系, 作者通过一定策略将该网络分为两个子网, 然后对其中的用户进行关联。合著网络数据集从 Microsoft Academic Graph (MAG)中获取的 AI 和 DM 两个领域的各 10 个会议的论文组成。文献[28]分别在合成网络和真实网络中做了实验。他们合成了三个各 10000 个节点的网络, 分别是 ER(正态分布全随机)、WS(正态分布)、BA(幂律分布)类型。另外从新浪微博和人人网爬取了数百万用户的好友关系。文献[50]使用的数据集包含了来自 Yelp、X 和 Flickr 的用户, 它非常庞大, 包含了超过 3.5 亿条 tweet、2900 万条 Flickr 帖子和 100 万条 Yelp 评论, 包括文本、时间戳和每个帖子的地理位置。文献[40]使用了 Weibo、Renren、36.cn、Zhaopin 四个数据集, 数据类型是各种属性数据, 如性别、生日、教育背景、地理位置等。

数据集的获取方式主要包括从社交网站爬取、组织自有数据和从第三方获取。目前, 用户身份链接领域无公开公认的基准数据集, 原因可能包括:

1) 由于现有各种方法所使用的特征类型和组合不同, 很难建立一个包含各种特征的综合数据集。如表 4 所示, 文献当中使用的数据集类型多样, 如属性、网络、内容、时空数据等, 建立一个包括所有数据类型的公开数据集是相当困难的。

表 4 近年部分参考文献所使用的数据集类型

Table 4 Types of datasets used in some references in recent years

文献	时间	数据集类型
TransLink[2]	2019	属性+网络+内容+行为
DPLink[56]	2019	时空数据
DeepLink[52]	2018	网络
CoLink[8]	2018	属性+网络
EMD-based[7]	2018	属性+网络
KDE-based[53]	2018	时空数据
Factoid Embedding[5]	2018	属性+网络
Mobility Traces based[55]	2018	时空数据
STUL[58]	2017	时空数据
IONE[44]	2016	网络
Location-based[16]	2016	时空数据
PALE[27]	2016	网络
FRUI[28]	2016	网络
stylometric-based[50]	2016	内容+时空数据
ULink[40]	2016	属性

2) 获取用户身份链接任务的 ground truth 是非常困难的。文献当中获取 ground truth(锚链接)的方式主要包括人工标注和利用用户公开信息两种。人工标注方式费时费力易出错, 而且有时是难以实现的; 利用用户公开信息的方式则存在可用信息数量十分有限且代表性弱的问题。

3) 隐私原因。由于用户和社交网站对隐私问题的关注, 获取用户信息变得越来越困难, 增加了建立公开数据集的难度。

4.2 评估指标

用户身份链接结果的优劣需要使用相关指标来进行评估, 文献中使用的评估指标主要包括 precision@k、hit-precision@k、precision、recall、F1 等, 如表 5 所示。

表 5 近年部分参考文献所使用的评估指标

Table 5 Evaluation indicators used in some references in recent years

文献	时间	评估指标
TransLink[2]	2019	precision@k, recall@k, mean-rank
DPLink[56]	2019	F1, AUC, hit-precision@ k
DeepLink[52]	2018	precision@k, MAP, AUC, hit-precision
CoLink[8]	2018	precision, recall, F1
EMD-based[7]	2018	hit-precision@20
KDE-based[53]	2018	precision, recall, F1
Factoid Embedding[5]	2018	HitRate@K(precision@k), MRR
Mobility Traces based[55]	2018	precision, recall, AUC
STUL[58]	2017	precision@k
IONE[44]	2016	precision@k
Location-based[16]	2016	precision, recall
PALE[27]	2016	F1, MAP@30
FRUI[28]	2016	precision, recall
Stylometric analysis[50]	2016	precision@k
ULink[40]	2016	hit-precision@5

precision@k(前 k 准确率)表示在返回的锚链接中真正(true positive)的锚链接所占的比例, 其中@k 表示只要锚链接 r 属于 top-k 列表, 即认为锚链接 r 是真正的锚链接。其中 n 表示算法返回的锚链接数(链接对数)。

$$\text{precision@}k=\frac{\sum_1^n 1@k}{n} \tag{1}$$

hit-precision@k(前 k 击中率)表示在返回的锚链接中真正(true positive)的锚链接的平均排名得分, 其中@k 表示前 k 个排名的得分依次降低, k 之后的排名得分为 0, 该指标比 precision@k 更能体现算法的排名性能。其中 n 表示算法返回的锚链接数(链接对数)。

$$\text{hit-precision}@k=\frac{\sum_1^n \frac{k+1-\text{rank}}{k}}{n} \tag{2}$$

precision 表示在返回的锚链接中真正(true positive)的锚链接所占的比例。该指标只关注预测结果中的第一个, 比 precision@k 和 hit-precision@k 更加严格。recall 表示返回的锚链接中的真正(true positive)的锚链接占实际锚链接的比例。F1 表示 precision 和 recall 的综合平衡值。其中 TP、FP、FN 分别表示真正类、假正类、假负类。

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{3}$$

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{4}$$

$$\text{F1}=\frac{2*\text{precision*recall}}{\text{precision}+\text{recall}} \tag{5}$$

此外, 也有少量文献使用 recall@k(前 k 召回率)、mean-rank(所有真正锚链接的平均排名)、AUC(ROC 曲线下方的面积)、MRR(Mean reciprocal rank)、MAP(mean Average Precision)作为评估指标。

5 问题与挑战

在线社交网络上的用户身份链接是一个非常重要和具有挑战性的问题, 因此它已成为一个研究热点并受到越来越多的研究关注。用户身份链接是一项复杂的任务, 主要存在以下挑战:

1) 数据质量较低

用户身份链接主要使用用户的属性、网络、内容、行为四类数据, 而在实际中这些数据往往具有相应的质量问题。在进行大规模用户身份链接时, 用户个人资料属性的区分度降低, 而且真实性和完整性难以保证。由于社交网络差异和用户隐私安全, 用户个人资料属性的可用性也是一个需要面对的重要问题。基于网络的 UIL 方法需要处理社交网络结构多样性的问题, 在网络结构重叠度较低时效果不佳, 而且由于计算开销较大难以推广到大规模的社交网络平台中。基于内容的 UIL 方法依赖于大量同构内容, 需要面对自然语言的不规范性, 而且实际中用户生成内容往往还存在稀疏性和异构性的问题。如果对异构内容进行模态转换或者直接计算相关性, 则会引入新的误差, 进一步降低基于内容的 UIL 方法的准确性。用户在不同社交网络中的活跃程度往往是不同的, 由此导致的行为不对称和不一致的现象, 给基于行为

的 UIL 带来了很大困难。总之, 社交网络中的用户数据的真实性、完整性、可用性难以保证, 总体数据量大而每个用户的数据又具有稀疏性, 往往是带有噪声的、无结构的, 不同社交网络之间的用户数据还存在异构和不一致的问题。

2) 无公开公认数据集

用户身份链接可以使用的数据类型众多, 已有文献大多使用其中的一种或多种类型的数据, 因此他们所使用的数据集往往只包含他们所选择类型的数据(如属性、内容等)。建立一个包括所有类型数据的公开公认的数据集是相当困难的。一方面, 由于用户和社交网站对隐私问题的关注, 获取用户信息变得越来越困难; 另一方面, 获取用户身份链接任务的 ground truth 是非常困难的。人工标注方式费时费力易出错, 而且有时是难以实现的; 利用用户公开信息的方式则存在可用信息数量十分有限且代表性弱的问题。由于以上诸多原因, 目前用户身份链接领域没有公开公认的基准数据集。

3) 模型构建困难

根据模型构建方法的不同, 现有的用户身份链接方法可以分为有监督的、半监督的与无监督的三种类型。在已知一定数量真实用户身份关联关系的情况下, 有监督的 UIL 可以在训练集上有较好的表现, 但获取真实用户身份关联关系的代价很高, 有时甚至是难以实现的, 而且获取的训练数据是否具有代表性会直接影响模型的性能表现。半监督的 UIL 同时利用了有标签数据的规律性和无标签数据的相似性, 能够在有监督的 UIL 和无监督的 UIL 之间取得一个折中的效果, 但依然存在有标签数据获取困难等问题。无监督的 UIL 不需要有标签的训练数据, 但现有的无监督 UIL 模型往往严重依赖于强鉴别特征或跨网络公共属性, 极大地限制了其通用性。

6 未来研究趋势

在最近的相关研究中, 我们发现了一些用户身份链接的未来研究趋势, 可能会引起越来越多的关注。

1) 新的用户数据类型

一些研究利用了多种新的用户数据类型, 如用户属性改变历史、写作风格、时空数据等。Jain 等人[59]利用用户属性改变历史记录进行跨社交网络的用户身份链接。Almishari 等人[50]利用写作风格在三个异构社交网络上进行了大规模的用户身份链接研究。Chen 等人[53]提出了一种基于核密度估计(KDE)的方法, 提高了使用地理位置数据进行 UIL 的测量精度。Chen 等人[58]提出了 STUL 模型, 利用时空数据进行用户身份链接。他们使用基于密度的聚类方法提取空间特征, 使用高斯混合模型提取时间特征。Chen 等人[60]使用自编码器对属性、网络、内容类型的用户数据进行融合嵌入, 并进一步使用图神经网络对用户身份链接结果进行增强。

2) 新的分析技术

很多新的技术被融入了用户身份链接问题中, 如用户空间分布、指数分布族、深度学习等。Li 等人[7]使用搬土距离(EMD)作为衡量用户间亲密度的指标, 并从用户空间分布的层次来执行无监督 UIL。他们使用了属性、UGC、网络、兴趣等多种特征, 并将无监督的 UIL 问题转化为一个投影函数的学习, 以最小化两个社交网络中用户身份分布之间的距离。Mu 等人[40]提出“潜在用户空间”的概念, 认为每个社交网络账号都是真实用户在相应社交网络平台上的投影, 并提出了 ULink 算法, 用于解决多平台 UIL 问题。Veiga 等人[61]通过分析 UGC 中可能包含的其他社交网络平台链接, 提出了一种跨社交网络平台关联用户数据集的收集方法。Liu 等人[44]明确地将每个用户的被关注关系和关注关系建模为输入和输出上下文, 通过学习一个对齐的网络嵌入来进行 UIL。Gao 等人[47]利用指数分布族对异构用户行为进行建模, 提出了一种无监督的方法, 集体网络链接(CNL), 来连接跨异构社交网络的用户。Zhou 等人[2]提出了一种翻译模型方法 TransLink 来解决 UIL 问题, 该方法基于翻译概念来解释用户之间的交互关系。Feng 等人[56]提出了一种基于端到端深度学习(LSTM)的 DPLink 框架, 用于完成从不同服务中收集的异构移动数据的用户身份链接任务。

3) 半监督和无监督的方法的探索

对于尚未被广泛研究的半监督和无监督的方法的探索越来越多。Zhong 等人[8]提出了一个通用的半监督 UIL 问题框架 CoLink, 其采用了一种协同训练算法, 该算法将基于属性的模型和基于关系的模型两种独立的模型进行迭代, 使它们相互增强。Zhou 等人[52]利用深度学习在自动特征提取和表示方面的成功, 提出了一种基于深度神经网络的半监督 UIL 算法 DeepLink。Xie 等人[5]提出了一种采用无监督方法的新框架 Factoid 嵌入, 它将用户属性、社交关系等事实形式化描述为 Factoid, 然后将涉及对象的 Factoid 中的对象(如用户昵称)嵌入到一个空间中, 最后进行 Factoid 嵌入, 得到 UIL 结果。

除以上最近相关研究中体现出的研究趋势外, 我们分析认为以下几点也将成为未来用户身份链接领域的重要研究方向。

1) 人工智能技术的应用

随着近几年人工智能技术在众多领域的成功应用和长足发展, 机器学习、深度学习、强化学习等代表性的人工智能技术越来越受到研究人员的关注和重视。与各领域的传统技术方法相比, 这些技术能够更加高效、智能、深入地解决一些研究问题。在用户身份链接领域的研究中, 人工智能技术的应用相对较少, 是一个值得探索的研究方向。

2) 大规模用户身份链接

目前, 用户身份链接相关研究使用的数据集规模相对较小, 用户数在数千左右, 记录数在数十万左右。而在现实中, 一个社交网络中的用户数可能多达数亿。这样的数据规模, 可能引起现有用户身份链接方法在准确率和效率方面的下降, 将给用户身份链接带来新的挑战, 需要进一步的深入研究。

3) 隐私问题

随着用户身份链接技术的发展, 隐私问题越来越受到研究人员、社交网站和用户的关注。进行用户身份链接能够更好地了解用户, 给推荐系统、商业智能、用户体验提升等带来帮助, 但同时也带来了用户隐私泄露的风险。如何在进行用户身份链接的同时最大限度地保护用户隐私, 是一个需要解决的重要问题。

7 总结

如今, 人们倾向于为了不同的目的加入多个在线社交网络。通过在线社交网络连接用户身份在许多应用领域(如推荐、链接预测等)具有很大的价值。本文首先介绍了用户身份链接及其发展历史, 然后根据用户身份链接的两个阶段——特征提取阶段和模型构建阶段——对用户身份链接方法进行了分类分析, 根据特征提取阶段可将用户身份链接方法分为基于属性、基于网络、基于内容与基于行为四类, 根据模型构建阶段可将用户身份链接方法分为有监督的、半监督的与无监督的三类。接下来我们对相关文献中的数据集和评估指标使用情况进行了分析总结, 并阐述了目前用户身份链接问题存在的挑战。最后对用户身份链接的未来研究趋势进行了分析和展望。

参考文献

[1] 中国互联网络信息中心. 第 45 次中国互联网络发展状况统计报告[EB/OL]. (2020)[2020-09-07]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202004/P020200428596599037028.pdf.

[2] Zhou J Y, Fan J X. TransLink: User Identity Linkage across Heterogeneous Social Networks via Translating Embeddings[C]. *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019: 2116-2124.

[3] Shu K, Wang S H, Tang J L, et al. User Identity Linkage across Online Social Networks[J]. *ACM SIGKDD Explorations Newsletter*, 2017, 18(2): 5-17.

[4] 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.

[5] Xie W, Mu X, Lee R K W, et al. Unsupervised User Identity Linkage via Factoid Embedding[C]. *2018 IEEE International Conference on Data Mining*, 2018: 1338-1343.

[6] Ma J T, Qiao Y Q, Hu G W, et al. Balancing User Profile and Social Network Structure for Anchor Link Inferring across Multiple Online Social Networks[J]. *IEEE Access*, 2017, 5: 12031-12040.

[7] Li C Z, Wang S Z, Yu P S, et al. Distribution Distance Minimization for Unsupervised User Identity Linkage[C]. *The 27th ACM International Conference on Information and Knowledge Management*, 2018: 447-456.

[8] Zhong Z X, Cao Y, Cao Y, et al. CoLink: An Unsupervised Framework for User Identity Linkage[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 32(1): 5714–5721.

[9] Liu J, Zhang F, Song X Y, et al. What's in a Name? : An Unsupervised Approach to Link Users across Communities[C]. *The sixth ACM international conference on Web search and data mining*, 2013: 495-504.

[10] Zafarani R, Liu H A. Connecting Corresponding Identities across Communities[J]. *Proceedings of the International AAAI Conference on Web and Social Media*, 2009, 3(1): 354-357.

[11] Bartunov S, Korshunov A, Park S, et al. Joint link-attribute user identity resolution in online social networks[C]. *The 6th SNA-KDD Workshop '12*, 2012.

[12] Zafarani R, Liu H. Connecting Users across Social Media Sites: A Behavioral-Modeling Approach[C]. *The 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013: 41-49.

[13] Motoyama M, Varghese G. I Seek You: Searching and Matching Individuals in Social Networks[C]. *The eleventh international workshop on Web information and data management*, 2009: 67-75.

[14] Liu S Y, Wang S H, Zhu F D, et al. HYDRA: Large-Scale Social Identity Linkage via Heterogeneous Behavior Modeling[C]. *The 2014 ACM SIGMOD International Conference on Management of Data*, 2014: 51-62.

[15] Zhang Y T, Tang J, Yang Z L, et al. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency[C]. *The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015: 1485-1494.

[16] Riederer C, Kim Y, Chaintreau A, et al. Linking Users across Domains with Location Data: Theory and Validation[C]. *The 25th International Conference on World Wide Web*, 2016: 707-719.

[17] Jain P, Kumaraguru P, Joshi A. @i Seek 'Fb.me': Identifying Users across Multiple Online Social Networks[C]. *The 22nd International Conference on World Wide Web*, 2013: 1259-1268.

[18] Goga O, Lei H, Parthasarathi S H K, et al. Exploiting Innocuous Activity for Correlating Users across Sites[C]. *The 22nd international conference on World Wide Web*, 2013: 447-458.

[19] 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.

[20] Perito D, Castelluccia C, Kaafar M A, et al. How Unique and Traceable Are Usernames?[G]. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2011, 6794 LNCS: 1-17.

[21] Iofciu T, Fankhauser P, Abel F, et al. Identifying Users across Social Tagging Systems[J]. *Proceedings of the International AAAI Conference on Web and Social Media*, 2021, 5(1): 522-525.

[22] Malhotra A, Totti L, Meira Jr W, et al. Studying User Footprints in Different Online Social Networks[C]. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2013: 1065-1070.

[23] Vosecky J, Hong D, Shen V Y. User Identification across Multiple Social Networks[C]. *2009 First International Conference on Networked Digital Technologies*, 2009: 360-365.

[24] Raad E, Chbeir R, Dipanda A. User Profile Matching in Social Networks[C]. *2010 13th International Conference on Network-Based Information Systems*, 2010: 297-304.

[25] Tan S L, Guan Z Y, Cai D, et al. Mapping Users across Networks by Manifold Alignment on Hypergraph[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014, 28(1): 159–165.

[26] 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.

[27] Man T, Shen H W, Liu S H, et al. Predict Anchor Links across Social Networks via an Embedding Approach[C]. *The Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016: 1823-1829.

[28] Zhou X P, Liang X, Zhang H Y, 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.

[29] Koutra D, Tong H H, Lubensky D. BIG-ALIGN: Fast Bipartite Graph Alignment[C]. *2013 IEEE 13th International Conference on Data Mining*, 2014: 389-398.

[30] Liang W X, Meng B, He X S, et al. GCM: A Greedy-Based Cross-Matching Algorithm for Identifying Users across Multiple Online Social Networks[C]. *Pacific-Asia Workshop on Intelligence and Security Informatics*, 2015: 51-70.

[31] 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.

[32] Almishari M, Tsudik G. Exploring Linkability of User Reviews[G]. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2012, 7459 LNCS(1): 307–324.

[33] Van Le T, Nghia Truong T, Vu Pham T. A Content-Based Approach for User Profile Modeling and Matching on Social Networks[C]. *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, 2014: 232-243.

[34] Liu S Y, Wang S H, Zhu F D. Structured Learning from Heterogeneous Behavior for Social Identity Linkage[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2015, 27(7): 2005-2019.

[35] Kong X N, Zhang J W, Yu P S. Inferring Anchor Links across Multiple Heterogeneous Social Networks[C]. *The 22nd ACM international conference on Information & Knowledge Management*, 2013: 179-188.

[36] Zhang J W, Kong X N, Yu P S. Transferring Heterogeneous Links across Location-Based Social Networks[C]. *The 7th ACM international conference on Web search and data mining*, 2014: 303-312.

[37] Roedler R, Kergl D, Rodosek G D. Profile Matching across Online Social Networks Based on Geo-Tags[C]. *Advances in Nature and Biologically Inspired Computing*, 2016: 417-428.

[38] Vosoughi S, Zhou H, Roy D. Digital Stylometry: Linking Profiles across Social Networks[C]. *International Conference on Social Informatics*, 2015: 164-177.

[39] Cudré Mauroux P, Abou Khaled O, Hazimeh H, et al. Linking User Profiles in Social Networks: A Comparative Review[J]. *International Journal of Social Network Mining*, 2017, 2(4): 333.

[40] Mu X, Zhu F D, Lim E P, et al. User Identity Linkage by Latent User Space Modelling[C]. *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016: 1775-1784.

[41] Peled O, Fire M, Rokach L, et al. Entity Matching in Online Social Networks[C]. *2013 International Conference on Social Computing*, 2014: 339-344.

[42] Zhang H, Kan M-Y, Liu Y, et al. Online Social Network Profile Linkage[G]. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014, 8870: 197–208.

[43] Nitish K, Silvio L. An Efficient Reconciliation Algorithm for Social Networks[J]. *Proceedings of the VLDB Endowment*, 2014, 7(5): 377-388.

[44] Liu L, Cheung W K, Li X, et al. Aligning Users across Social Networks Using Network Embedding[C]. *KAMBHAMPATI S. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*. New York, NY, USA: IJCAI/AAAI Press, 2016: 1774–1780.

[45] Lacoste-Julien S, Palla K, Davies A, et al. SIGMa: Simple Greedy Matching for Aligning Large Knowledge Bases[C]. *The 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013: 572-580.

[46] Buccafurri F, Lax G, Nocera A, et al. Discovering Links among Social Networks[M]. *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 467-482.

[47] Gao M, Lim E P, Lo D, et al. CNL: Collective Network Linkage across Heterogeneous Social Platforms[C]. *2015 IEEE International Conference on Data Mining*, 2016: 757-762.

[48] Mo W K, Shen B J, Chen Y T, et al. TBIL: A Tagging-Based Approach to Identity Linkage across Software Communities[C]. *2015 Asia-Pacific Software Engineering Conference*, 2016: 56-63.

[49] Han X H, Wang L H, Xu L J, et al. Social Media Account Linkage Using User-Generated Geo-Location Data[C]. *2016 IEEE Conference on Intelligence and Security Informatics*, 2016: 157-162.

[50] Almishari M, Oguz E, Tsudik G. Trilateral Large-Scale OSN Ac-

count Linkability Study[C]. *2016 AAAI Fall Symposia*, Arlington, Virginia, USA, November 17-19, 2016. AAAI Press, 2016.

[51] Han X H, Wang L H, Xu S J, et al. Linking Social Network Accounts by Modeling User Spatiotemporal Habits[C]. *2017 IEEE International Conference on Intelligence and Security Informatics*, 2017: 19-24.

[52] Zhou F, Liu L, Zhang K P, et al. DeepLink: A Deep Learning Approach for User Identity Linkage[C]. *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018: 1313-1321.

[53] Chen W, Yin H Z, Wang W Q, et al. Effective and Efficient User Account Linkage across Location Based Social Networks[C]. *2018 IEEE 34th International Conference on Data Engineering*, 2018: 1085-1096.

[54] Qiao Y, Wu Y, He Y, et al. Linking User Online Behavior across Domains with Internet Traffic[J]. *Journal of Universal Computer Science*, 2018, 24(3): 277–301.

[55] Wang H D, Li Y, Wang G, et al. You are how You Move: Linking Multiple User Identities from Massive Mobility Traces[M]. *Proceedings of the 2018 SIAM International Conference on Data Mining*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2018: 189-197.

[56] Feng J, Zhang M Y, Wang H D, et al. DPLink: User Identity Linkage via Deep Neural Network from Heterogeneous Mobility Data[C]. *WWW'19*: *The World Wide Web Conference*, 2019: 459-469.

[57] Zhang W, Shu K, Liu H, et al. Graph Neural Networks for User Identity Linkage[EB/OL]. 2019: arXiv: 1903.02174. https://arxiv.org/abs/1903.02174

[58] Chen W, Yin H Z, Wang W Q, et al. Exploiting Spatio-Temporal User Behaviors for User Linkage[C]. *The 2017 ACM on Conference on Information and Knowledge Management*, 2017: 517-526.

[59] Jain P, Kumaraguru P, Joshi A. Other Times, other Values: Leveraging Attribute History to Link User Profiles across Online Social Networks[J]. *Social Network Analysis and Mining*, 2016, 6(1): 85.

[60] Chen S Y, Wang J H, Du X, et al. A Novel Framework with Information Fusion and Neighborhood Enhancement for User Identity Linkage[EB/OL]. 2020: arXiv: 2003.07122. https://arxiv.org/abs/2003.07122

[61] Veiga M H, Eickhoff C. A Cross-Platform Collection of Social Network Profiles[C]. *The 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016: 665-668.

薛晖 于 2016 年在北京航空航天大学计算机技术专业获得硕士学位。现在中国科学院信息工程研究所网络空间安全专业攻读博士学位。主要研究兴趣包括: 网络信息安全, 用户画像, 机器学习和深度学习。Email: xuehui@iie.ac.cn

司成祥 现任国家互联网应急中心高级工程师。主要研究兴趣为网络安全。Email: sichengxiang@cert.org.cn

房婧 现任国家互联网应急中心高级工程师。主要研究兴趣为网络安全。Email: fj@cert.org.cn

孙波 现任国家互联网应急中心正高级工程师。主要研究兴趣包括: 网络攻防, 大数据。Email: sunbo@cert.org.cn

张伟 现任国家互联网应急中心高级工程师。主要研究兴趣为网络安全。Email: zhangwei@cert.org.cn