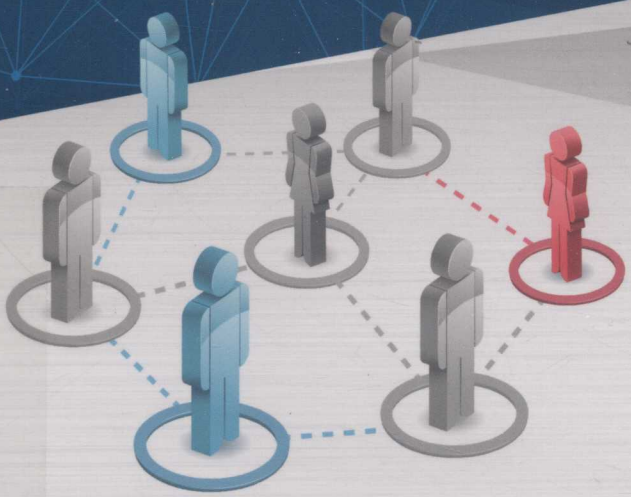
社交网络大数据融合
——关联用户挖掘

周小平　梁　循　著



科学出版社

社交网络大数据融合
——关联用户挖掘

周小平　梁　循　著

科学出版社
北京

内 容 简 介

社交网络融合为社会计算等各项研究提供更充分的用户行为数据和更完整的网络结构，从而更有利于人们通过社交网络认识和理解人类社会，具有重要的理论价值和实践意义。社交网络中的关联用户挖掘旨在通过挖掘不同社交网络中属于同一自然人的不同账号，从而实现社交网络的深度融合。因此，关联用户挖掘是大型社交网络融合的基础问题，近年来已引起人们的广泛关注。考虑真实世界的朋友圈极具个性化，即现实中没有两个人具有完全一致的朋友圈，同时，相同的用户在不同的社交网络中往往具有部分相同的好友关系。本书基于社交网络的好友关系，充分利用好友关系的唯一性、稳定性和一致性，探索关联用户挖掘的方法。

本书适用于从事社交网络、大数据挖掘等领域的研究人员。

图书在版编目（CIP）数据

社交网络大数据融合：关联用户挖掘 / 周小平，梁循著. —北京：科学出版社，2019.6

ISBN 978-7-03-060417-0

Ⅰ. ①社… Ⅱ. ①周… ②梁… Ⅲ. ①互联网络-数据处理 Ⅳ. ①TP393.4

中国版本图书馆 CIP 数据核字（2019）第 012966 号

责任编辑：阚 瑞 / 责任校对：张凤琴

责任印制：吴兆东 / 封面设计：迷底书装

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2019 年 6 月第 一 版 开本：720×1000 B5

2019 年 6 月第一次印刷 印张：7 3/4

字数：150 000

定价：58.00 元

(如有印装质量问题，我社负责调换)

前　言

社交网络是当前学术界和产业界的研究热点。然而，现阶段大多数研究都集中于单一的社交网络内部。社交网络融合为社会计算等各项研究提供更充分的用户行为数据和更完整的网络结构，从而更有利于人们通过社交网络认识和理解人类社会，具有重要的理论价值和实践意义。准确、全面、快速的关联用户挖掘是大型社交网络融合的根本问题。社交网络中的关联用户挖掘旨在通过挖掘不同社交网络中同属于同一自然人的不同账号，从而实现社交网络的深度融合，近年来已引起人们的广泛关注。然而，社交网络的自身数据量大，用户属性相似、稀疏且存在虚假和不一致等特点给关联用户挖掘带来了极大的挑战。

用户关系，尤其是好友关系，是社交网络中较稳定、不易受攻击且可获取的信息。目前，基于用户关系的最相关研究大都针对匿名化的社交网络在线发布数据的还原(又称“去匿名化”)。然而，“去匿名化”方法大多适用于部分子网高度重叠的两个网络，不能直接应用于节点和关系都部分重叠的社交网络融合。考虑真实世界的朋友圈极具个性化，也即现实中没有两个人具有完全一致的朋友圈，同时，相同的用户在不同的社交网络中往往具有部分相同的好友关系，为此，本书提出基于社交网络的好友关系探索关联用户挖掘的方法。

第 1 章介绍了社交网络大数据融合及其核心问题和面临的主要挑战。

第 2 章系统给出了关联用户挖掘所涉及的相关术语和关联用户挖掘定义，并从社交网络重叠阐述了基于好友关系关联用户挖掘的可行性。

第 3 章总结了关联用户挖掘总体研究框架，从用户属性、用户关系及其综合等三个方面梳理并总结了当前关联用户挖掘的研究现状，给出了关联用户挖掘的性能评价方法。

第 4 章介绍了基于好友关系的半监督关联用户挖掘方法，从半监督的角度解决了基于用户属性的关联用户挖掘所存在的易受攻击、健壮性差等问题。最后，讨论了所提出的基于好友关系的半监督关联用户挖掘方法对知识管理的应用。

第 5 章介绍了基于好友关系的无监督关联用户挖掘方法，解决了基于好友关

系的半监督关联用户挖掘方法受先验关联用户限制的问题。最后，讨论了所提出的基于好友关系的无监督关联用户挖掘方法对知识管理的应用。

第 6 章给出了一种综合用户属性和用户关系的关联用户挖掘模型及其近似求解和并行计算方法。

第 7 章为总结与展望，对本书的研究工作进行了总结，并指出未来有价值的研究问题。

本书受国家社会科学基金重大项目(基金编号：18ZDA309)的资助，特此感谢。

目　录

第 1 章　社交网络大数据融合

1.1　社交网络与社交网络大数据融合

社交网络(social network)是指人们用于创建、分享、交流信息和观点的虚拟社区和网络[1]。近年来，随着 Facebook、Twitter 的影响力不断提高，微博、微信在人们日常生活中的深入渗透，社交网络已使得当前社会经济文化问题日益呈现出了动态性、快速性、开放性、交互性和数据海量化等特点[2]。得益于社交网络所产生的海量用户行为数据，研究人员使用社交网络进行社区发现[3]、影响力分析[4]、链接分析[5]、情感分析[6]、观点挖掘[6]、商务智能[7]、企业决策支持[8]等[2, 9, 10]。

由于社交网络功能和需求的差异性，越来越多的用户同时使用多个社会网络。例如，人们可以在人人网上发一些动态文章，同时也会在微博上分享他们的旅游照片。据研究报告显示，截至 2013 年，约有 42%的用户同时使用多个社交网络，其中，93%的 Instagram 用户同时使用 Facebook，53%的 Twitter 用户同时使用 Instagram。用户由于不同的目的使用多个社会网络，因此，分析用户在单一网络里的行为是无法全面了解用户的性格及兴趣特征。然而，大多数的社交网络研究都仅局限于单一的社交网络内部。

定义 1-1　社交网络大数据融合。社交网络大数据融合，又称社交网络融合，是指通过匹配不同社交媒体中的相同节点，将多个社交媒体融合形成一个规模更大、信息更完备的社交媒体。

社交网络大数据融合为社交网络的相关研究提供了更完备的用户行为数据，是现阶段社交网络研究的一个重要和热点问题，已成为社交网络领域研究的新趋势和新方向。

首先，社交网络大数据融合为社交网络各项研究提供更为完善的用户行为数据，将使社交网络的研究更全面、更准确，也更有利于人们认识社交网络，进而通过社交网络认识人类社会。以图 1.1 为例，当某个自然人计划去青海湖骑行时，他很可能会先在豆瓣上注册用户，并在上面寻找骑友。当他到青海湖骑行的过程

中，他会通过微博或者微信等发布他实时的动态和美好风景等。旅程结束时，他很可能会撰写完整的骑行游记或攻略，并发布在人人网或者马蜂窝等社交平台上。因此，如果我们能够融合上述提及的社交网络，那么，将可能得到该用户在多个社交网络的完整用户关系，更详细的用户个人资料和全流程的用户行为数据。这些数据将更有利于社交网络的各项分析。

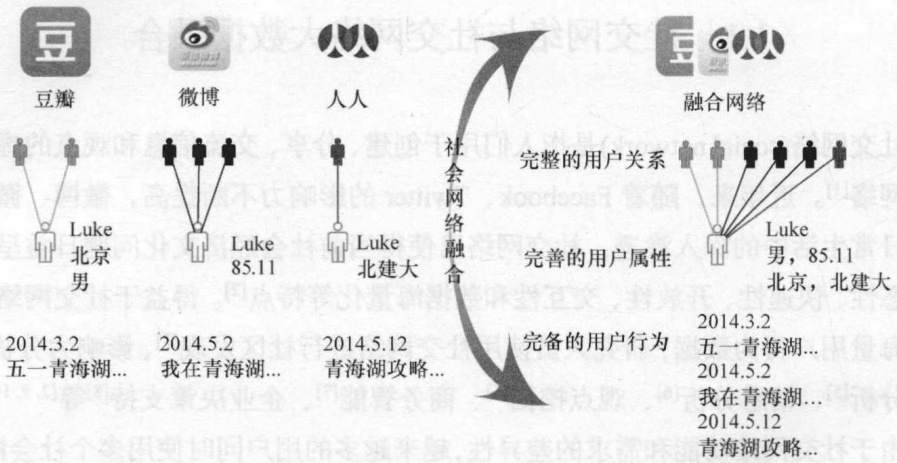


图 1.1　社交网络融合示意图

其次，社交网络大数据融合能够帮助解决只利用单一网络数据无法解决的问题，如冷启动问题[11]和数据稀疏性问题等。例如，一个新成立的社交网络没有充足的历史数据来给用户进行推荐，如果我们能够在其他已建立的社交网络上识别这些用户，那么就可以从成熟的社交网站上转移数据到新的社交网站上，从而解决数据稀疏性和冷启动问题。

然后，社交网络大数据融合能帮助我们分析用户的迁移模式，并给网站的发展提供指导。通常，用户从一个网络迁移到另一个网络反映了用户所经历的网站的发展。跨平台的用户身份关联能够很好地研究用户的迁移行为。

此外，社会网络融合还有以下功能。

(1) 增强好友推荐机制。

在线用户参与可以提升好友推荐机制[12,13]。目前大多数的好友推荐算法基本是：推荐不相关的好友；推荐共同好友。比如在社交网络 SN_1 上，两个用户 U_1 和 U_2 不是好友关系，但是他们均和用户 U_3 是好友，那么 U_1 很有可能会被推荐给 U_2。如果用户 U_1 和 U_2 同时也是社交网络 SN_2 的用户，他们在 SN_2 上也不是好友关系，

也没有共同好友，根据从 SN_1 上获取的信息，推荐系统在社交网络 SN_2 上可以把 U_1 推荐给 U_2。这种推荐方式可通过跨平台来实现。

(2) 信息扩散。

信息扩散的研究基本集中在单一网络。实际上，信息和谣言可在社交网络内部以及多个社交网络间传播。因此，研究跨平台的信息传播是更有意义的。此外，不同类型的信息在网络内和网络间传播的速度差异性也是未来研究的一个方向。

(3) 动态网络分析。

单一网络的动态性分析已有很多文献涉及。这些网络具有幂率分布、平均路径较短、高聚类等特征。然而，用户活跃于多个网络，这些网络特征也应该推广到多网络，尤其是单一网络与多网络动态性的异同。目前很多研究在寻找用户参与的网络类型，他们的度特征分布(如好友数)以及用户在不同网络上的好友差异。

1.2　社交网络大数据融合的核心问题

用户是社交网络的主体。由于不同的使用需求，人们在不同的社交网络上注册用户。因此，用户是社交网络融合的天然桥梁。

定义 1-2　关联用户。假定 U_i^A 和 U_i^B 分别是大型社交网络 SN^A 和 SN^B 中的用户。若 U_i^A 和 U_i^B 是现实世界中同一自然人分别在 SN^A 和 SN^B 中的账户(用户)，则 U_i^A 和 U_i^B 是关联用户，记为 $U_i^A = U_i^B$。

定义 1-3　关联用户挖掘。关联用户挖掘是指根据已知信息 Υ，获取 SN^A 和 SN^B 中所有关联用户的方法。通常，关联用户挖掘将转化为关联用户识别问题，即判定两个来自 SN^A 和 SN^B 的用户 U_i^A 和 U_i^B 在已知信息 Υ 下是否同属于一个自然人 Γ，即

$$f(a,\hat{a}|\Upsilon)=\begin{cases}1, & a=\hat{a}\\ 0, & a\neq\hat{a}\end{cases} \tag{1-1}$$

社交网络间的关联用户挖掘旨在发现准确、全面的关联用户以实现社交网络的深度融合(图 1.2)。显然，关联用户挖掘将直接从社交网络节点上融合社交网络。因此，构建准确、全面、快速的关联用户挖掘模型和方法是社交网络融合的核心问题。

图 1.2　关联用户挖掘示意图

1.3　社交网络大数据融合的主要挑战

早期，研究人员通过 Email 构建“Find Friend”机制构建关联用户挖掘方法[14]。绝大多数的社交网络都通过 Email 注册账号(近年来兴起的移动社交网络中，有部分使用手机号注册账号)。由于 Email 的唯一性，“Find Friend”使用社交网络所提供的“Email 查找用户”功能挖掘不同社交网络间的关联用户。近年来，随着人们对自身网络隐私的重视以及社交网络对用户数据的保护，可获取的用户属性信息越来越少。据统计，用户平均在一个社交网络中公开 4 项属性信息[15]，这给关联用户挖掘及社交网络融合带来了极大的挑战。大型社交网络是指用户数达到千万级以上的社交网络，如新浪微博、人人网、Twitter、Facebook 等，它们所提供的海量社会行为数据更有利于各领域的研究。因此，大型社交网络关联用户挖掘更具有研究价值，且其理论和方法也能应用于小型社交网络。目前，大型社交网络关联用户挖掘所面临的挑战包括以下几点。

(1) 相似性。随着用户数量的增加，大型社交网络出现了大量的具有相似或相同属性信息但不关联的用户。如图 1.2 所示，新浪微博和人人网都有上千用户名包含 luke 的用户。

(2) 稀疏性。因许多用户未填写某项(些)属性而导致该项(些)属性信息较为稀疏。例如，头像是社交网络中的一项重要属性，而只有 66%的用户会上传头像[16]。

(3) 虚假性。社交网络用户属性的虚假性主要源于：①用户因不愿公开某项(些)属性而填写虚假的属性值；②恶意用户因其需要设定用户属性与某(些)其他用户相同；③用户填写属性信息时的随意性也容易造成虚假信息。

(4) 不一致性。同一用户在不同的社交网络中对同一属性填写不同的值。

(5) 大数据。社交网络往往包含千万级以上的用户，其给社交网络融合带来了极大挑战。

用户属性是挖掘关联用户的最直接方法。现阶段，大多数的关联用户发现方法都基于用户属性(如昵称、头像)相似度的计算。然而，大型社交网络中用户属性的相似性、稀疏性、虚假性和不一致性使得单纯使用用户属性挖掘关联用户方法易受恶意用户的攻击，健壮性较差。

用户关系，尤其是好友关系，是社交网络中较稳定、不易受攻击且可获取的信息。目前，基于用户关系挖掘关联用户的研究大都针对匿名化的社交网络在线发布数据的还原(又称去匿名化)[17]。然而，“去匿名化”方法大多适用于部分子网高度重叠的两个网络，不能直接应用于节点和关系都部分重叠的社交网络融合。基于好友关系建立关联用户挖掘方法，将从网络结构角度为建立准确、全面、健壮的关联用户挖掘模型提供重要的理论和方法补充。其相关理论和方法可为“去匿名化”和大数据融合等领域提供借鉴，有利于解决协同过滤中的“冷启动”问题，具有重要的理论价值和应用意义。

1.4　本书主要内容

本书主要研究面向社交网络大数据融合的关联用户挖掘方法。当前，基于用户属性的关联用户挖掘方法已经取得了较多的研究成果，而仅有少数研究利用了社交网络的网络结构(好友关系)。针对现有基于用户属性的关联用户挖掘方法所存在的健壮性较差、易受恶意用户攻击等问题，本书充分利用好友关系的稳定性和一致性，建立半监督和无监督的关联用户挖掘方法，其主要内容可以概括为以下几点。

(1) 系统总结关联用户挖掘的研究现状。总结关联用户挖掘的总体研究框架，从用户属性、用户关系及其综合使用三个方面综述关联用户挖掘的研究现状，给出关联用户挖掘的性能评价指标和数据集。

(2) 提出一种基于好友关系的半监督关联用户挖掘方法。分析不同社交网络

好友关系特征，建立不同社交网络用户和好友关系部分重叠的随机抽样模型，建立在给定部分关联用户情况下的好友相似度计算模型，最终形成基于好友关系的半监督关联用户挖掘方法。

(3) 提出一种基于好友关系的无监督关联用户挖掘方法。在研究对好友关系特征的基础上，借鉴现有深度学习的前沿理论和方法，将不同空间的高维、稀疏、离散的好友关系映射到统一空间中低维、连续的向量，而后，建立不同用户的用户关系相似度模型，形成基于好友关系的无监督关联用户挖掘算法。

(4) 分析关联用户挖掘方法对企业知识管理的应用。在社交网络环境下，基于半监督的关联用户挖掘方法可以帮助企业快速应对企业内部人员变动对知识管理的影响，基于无监督的关联用户挖掘方法可以帮助企业迅速反映企业外部社交网络变迁对知识管理的影响。

第 2 章　关联用户挖掘定义

社交网络是指人们创建、分享和/或交换信息和想法的虚拟社区和网络[1]。在社交网络中，人们可以在有界系统内构建公开或半公开的个人资料，以及与其他用户建立连接关系并进行信息交流[18]。因此，社交网络通常包含三个关键要素：公共或半公开的用户个人资料，用户发布因社交需要而产生的内容及其时间和位置等信息以及用户之间的连接关系(或网络)。本章系统介绍涉及社交网络关联用户挖掘的相关术语和社交网络关联用户挖掘的数学定义。

2.1　基本术语定义

定义 2-1　社交网络。社交网络定义为 $\mathrm{SN}=\{U, C, I\}$，其中，U，C 和 I 分别表示用户及其公开或半公开的个人资料信息，用户连接和用户因社交需要而发布的内容和交互信息以及这些行为的时间和位置等信息。

通常，社交网络中 U 主要包含用户名、用户头像、用户签名、用户出生日期、用户教育背景和用户的工作职业等。I 包括用户之间的关注关系、好友关系、评论关系、转发关系、@关系和私信关系等。C 包括用户发布的内容(user generated content，UGC)、用户相互发送的私信信息以及这些行为的发生时间和位置等信息。

深入研究 SN 的主要组成部分，不难发现 C 和 I 都是由 U 生成的。也即，C 和 I 也可作为 U 的特殊属性。因此，U 是 SN 中的核心元素。从该意义上说，关联用户挖掘是跨社交网络研究中最重要的问题之一。

由于跨社交网络研究往往涉及多个社交网络，为此，在本书中，$\mathrm{SN}^A=\{U^A, C^A, I^A\}$用于表示社交网络 A，其中，U^A，C^A 和 I^A 分别表示社交网络 A 中的用户集合、用户连接集合和用户交互内容集合。例如，$\mathrm{SN}^{\mathrm{Twitter}}=\{U^{\mathrm{Twitter}}, C^{\mathrm{Twitter}}, I^{\mathrm{Twitter}}\}$表示社交网络 Twitter。

不失一般性，本书所用符号的上标为社交网络的标识，下标为社交网络中用户的标识。例如，U_i 表示未知社交网络中的用户 i，U_i^A 表示社交网络 SN^A 中的用户 i。

本书重点研究基于网络结构或用户连接的关联用户挖掘方法，因此，在本书中社交网络简化为 $SN=\{U, C\}$。

定义 2-2　好友关系。社交网络中的用户连接分为单向连接和双向连接。单向连接又称关注关系，双向连接又称好友关系。在微博类社交网络中，如果用户 a 关注了用户 b，而用户 b 没有关注用户 a，则称 a 和 b 之间建立单向连接。若 a 和 b 同时彼此关注了对方，则称 a 和 b 建立了双向连接或好友关系。在 Facebook 或人人网社交网络中，好友关系的建立需要双方的确认，也即一方发起好友请求，另一方对请求进行确认。

在微博等社交网络中，一个用户可以随意的关注另一用户，而好友关系是连接双方共同承认的关系，也更能反映真实世界的人际关系。因此，在关联用户挖掘中，单向连接由于随意性而容易被伪造，而好友关系由于需要双方的确认而更健壮，也更适用于关联用户挖掘。为此，本书所讨论的网络结构或用户连接又指好友关系。此时，社交网络可以表示为 $SN=\{U, F\}$，其中，F 为社交网络 SN 的好友关系集合。

定义 2-3　度/好友数。社交网络 SN 中用户 U_i 的好友集合为 F_i，用户 U_i 的度为其好友数，表示为 $d_i=|F_i|$，其中 $|\bullet|$ 表示集合中的元素数目。

定义 2-4　关联用户/关联用户对。若社交网络 SN^A 中的用户 U_i^A 和社交网络 SN^B 中的用户 U_i^B 同属于同一自然人的账户，则称 U_i^A 和 U_i^B 为关联用户，记为 $U_i^A=U_i^B$。此时，(U_i^A, U_i^B) 组成一对关联用户，记为 $I(U_i^A, U_i^B)$。

定义 2-5　先验关联用户。在关联用户算法执行前，作为已知先验知识给定的部分关联用户集合，称为先验关联用户集合 $\mathcal{P}$。先验关联用户集合中的关联用户则为先验关联用户。

先验关联用户集合是有监督和半监督关联用户挖掘的必要条件。先验关联用户的获取方式通常有以下几种。

(1) 从用户个人网站中获取。例如，用户在 Google+和 About.me 等网站中会关联其 Facebook、Twitter 账号信息等[19]。

(2) 通过比较个人资料、内容和网络特征等获取。例如，有些用户会在其 Twitter 账号中关联其 Facebook 账号等[13]。

(3) 当上述两种方法都较难获取先验关联用户时，需要进行人工标注。此时，人工标注的工作量将相当繁复。例如，在新浪微博和人人网中进行先验关联用对

标注。

定义 2-6　候选关联用户/候选关联用户对。任何社交网络 SN^A 中的待关联用户 U_i^A 和社交网络 SN^B 中的待关联用户 U_j^B 组成一对候选关联用户，表示为 $\ddot{I}(U_i^A, U_j^B)$ 或 $\ddot{I}_{A\sim B}(i,j)$。一对候选关联用户又称候选关联用户对。本书将所有的候选关联用户集合标识为 $\ddot{I}(\cdot,\cdot)$，将包含社交网络 SN^A 中的待关联用户 U_i^A 的所有候选关联用户集合标识为 $\ddot{I}(U_i^A,\cdot)$，将包含社交网络 SN^B 中的待关联用户 U_j^B 的所有候选关联用户集合标识为 $\ddot{I}(\cdot,U_j^B)$。初始情况下，在无监督关联用户挖掘中，$\ddot{I}(\cdot,\cdot)$ 包含 $|U^A| \times |U^B|$ 个候选关联用户，$\ddot{I}(U_i^A,\cdot)$ 中包含 $|U^B|$ 个候选关联用户。图 2.1 为候选关联用户的示例。

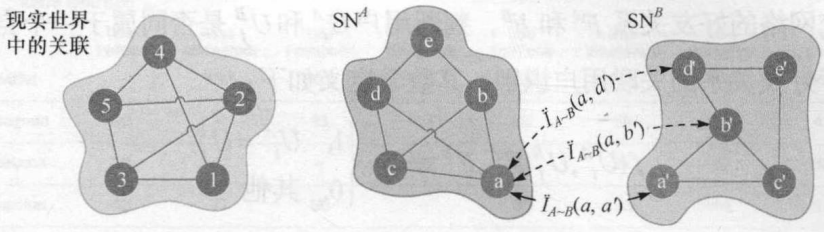


图 2.1　术语定义示例

社交网络 SN^A 中的用户 U_4^A 与社交网络 SN^B 中的任一用户形成一对候选关联用户

定义 2-7　相似度/匹配度。相似度定量度量在给定一定已知条件的情况下候选关联用户 $\ddot{I}(U_i^A, U_j^B)$ 中两个用户 U_i^A 和 U_j^B 的相似性。在关联用户挖掘中，有些文献又将相似度称为匹配度。

在给定先验关联用户集合的情况下，已知共同好友(好友共献)、Dice 系数等都可以用于计算候选关联用户 $\ddot{I}(U_i^A, U_j^B)$ 中两个用户 U_i^A 和 U_j^B 的相似度。在无先验关联用户集合的情况下，U_i^A 和 U_j^B 的相似度计算是一件极有挑战的任务。本书所提出的基于好友关系的无监督关联用户挖掘算法采用深度学习提取 U_i^A 和 U_j^B 的好友特征，形成好友特征向量，而后通过计算好友特征向量的欧式距离在度量 U_i^A 和 U_j^B 的相似度。

2.2　关联用户挖掘问题定义

通常，关联用户挖掘将转化为判定两个来自 SN^A 和 SN^B 的用户 U_i^A 和 U_j^B 在