社交关系挖掘研究综述

赵 姝[1),2)] 刘晓曼[1),2)] 段 震[1),2)] 张燕平[1),2)] 唐 杰[2),3)]

[1)](安徽大学计算智能与信号处理教育部重点实验室 合肥 230601)

[2)](安徽大学信息保障技术协同创新中心 合肥 230601)

[3)](清华大学计算机科学与技术系 北京 100084)

摘 要 随着在线社交应用和媒体的迅速扩散,在线社交网络(Online Social Network,OSN)已将我们的日常生活与网络信息空间连接起来.这些连接产生了大量的数据,不仅包括传播信息,还包括用户行为.社交关系挖掘的研究是社交网络挖掘中的一个重要领域,为我们对网络的形成机理、用户的交互模式和动态机制的理解提供了一个机会.社交关系(Social Ties)是社交网络中人与人连接和交互的纽带,也是社交网络中信息传播的基础.从计算学的观点来看,社交关系挖掘的研究包括社交关系的形成机理、社交关系的语义化以及基于社交关系人与人之间的交互.该文综述性地分析了这 3 个方面的研究现状,具体来说,在社交关系形成机理方面介绍关系链接预测,在基于社交关系的交互方面介绍关系交互预测,在社交关系语义化方面介绍关系类型预测.首先给出社交网络分析问题的形式化描述和相关概念、常用数据,然后分别介绍关系链接预测、关系类型预测和关系交互预测 3 个方面的方法、理论和模型,并给出重要的应用实例及其效果.最后,该文给出了未来工作的展望.

关键词 社交关系;关系链接预测;关系类型预测;关系交互预测;在线社交网络

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2017.00535

A Survey on Social Ties Mining

ZHAO Shu[1),2)] LIU Xiao-Man[1),2)] DUAN Zhen[1),2)] ZHANG Yan-Ping[1),2)] TANG Jie[2),3)]

[1)](*Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei* 230601)

[2)](*Center of Information Support and Assurance Technology, Anhui University, Hefei* 230601)

[3)](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract With the emergence and rapid proliferation of online social applications and media, online social networks already connect our physical daily life with the web-based information space. The connection produces huge volume of data including not only the spreading information, but also user behavior. Social network mining aims to provide a comprehensive understanding of global and local patterns, mechanism of the network formation, and dynamics of user behaviors. Research on social ties mining is one of the most important researches in social network mining. Social ties are a bridge of connection and interaction between people, and also the foundation of information diffusion in social networks. From the computing viewpoint, researches on social ties mining include the mechanism of social ties formation, the semantization of social ties, and the interaction of social ties between people. This paper reviews the current stage of these fields. Specifically, it discusses the social ties link prediction, social ties type prediction, and social ties

收稿日期:2015-10-30;在线出版日期:2016-06-24.本课题得到国家"八六三"高技术研究发展计划项目基金(2015AA124102)、国家社会科学重大基金(13&ZD190)、国家自然科学基金(61402006,61175046)、安徽省高等学校省级自然科学研究基金(KJ2013A016)、安徽省自然科学基金(1508085MF113)、教育部留学回国人员科研启动基金(第 49 批)、安徽大学高层次人才需求计划项目资助.**赵 姝**,女,1979 年生,博士,教授,中国计算机学会(CCF)会员,主要研究方向为机器学习、智能计算.E-mail:zhaoshuzs2002@hotmail.com.**刘晓曼**,女,1989 年生,硕士,助理实验师,中国计算机学会(CCF)会员,主要研究方向为机器学习、社交网络.**段 震**,男,1976 年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为机器学习、社交网络.**张燕平**,女,1962 年生,博士,教授,中国计算机学会(CCF)会员,研究领域为机器学习、粒度计算.**唐 杰**(通信作者),男,1977 年生,博士,特别研究员,中国计算机学会(CCF)杰出会员,主要研究领域为数据挖掘、机器学习、社交网络.E-mail:jietang@tsinghua.edu.cn.

interaction prediction. Firstly, the formalized descriptions of these three problems in social network analysis are given and related concepts, some used data sets are introduced. Then, the methods, theories and models of social ties link prediction, social ties type prediction, and social ties interaction prediction are discussed. Also for each of these three problems, some real applications and experimental results are presented. Finally, the future research directions are discussed.

Keywords social ties; link prediction; type prediction; interaction prediction; online social network

1 引 言

在线社交网络(Online Social Network,OSN)是指由成千上万的互联网用户通过自组织方式构建成关系连接而组成的集合[1],也是真实物理世界的社交网络在虚拟网络世界的一种映射,其本质是人与人之间的关系网络.

近年来,在线社交网络取得了飞速发展,众多网站如 Facebook、Twitter 以及国内的新浪微博、人人网、腾讯网等迅速崛起. 2004 年成立的 Facebook 公司到目前已经有超过 22 亿的注册用户,2006 年发布的 Twitter 有超过 5 亿的注册用户;国内的腾讯公司也拥有超过 8 亿的活跃用户,新浪最新公布的数据表明新浪微博的注册用户数已经超过 5 亿[2]. 图 1 给出国内外流行在线社交网站基本统计信息. 据报道,在美国,人们 16% 的上网停留在 Facebook 上,这一数字超过了人们使用传统搜索引擎(如 Google)的 10%. 毫无疑问,在线社交网络已经成为连接物理社交世界和虚拟网络空间的桥梁. 网络用户和信息的交互以及用户之间的交互在社交网络上留下了各种"足迹",直接促成了网络大数据时代的到来. 在线社交网络存储了大量用户资料,用户之间的社交关系以及用户之间的交互,这些海量社交数据有着巨大的研究价值,同时也在广告、推荐系统等方面具有广阔的应用前景.

在线社交网络已经成为连接我们现实日常生活与基于网络的信息空间的桥梁. 这种连接产生了庞大的数据,不仅包括传播信息,还包括用户行为. 无所不在的社交网络和巨大的社交数据为我们学习用户之间的交互模式以理解不同网络下的动态机制提供了前所未有的机遇,这在缺乏可用数据的过去是很难进行的.

社交关系(Social Ties)是社交网络中人与人连接和交互的纽带,也是社交网络中信息传播的基础. 从计算学的观点来看,社交关系挖掘的研究包括:社交关系的形成机理(关系链接预测)、社交关系的语义化(关系类型预测)以及基于社交关系人与人之间的交互(关系交互预测).

图 1 国内外流行在线社交网站基本统计信息

关系链接预测,即预测和推荐未知的链接. 对给定的社交网络,Liben-Nowell 和 Kleinberg[3] 系统地研究了推断用户之间新链接的问题. 他们引入几种无监督学习的方法来处理这类基于网络中节点的"接近度"或同质性原则[4](物以类聚[5])的问题,这类问题主要是基于用户的内容或者结构的相似性. Backstrom 等人[6] 提出了监督学习的随机游走算法来估算社交关系的强度.

关系类型预测,即自动地识别与每一个社交关系相关联的语义. Leskovec 等人[7] 使用 Logistic 回归模型预测在线社交网络中的正/负关系. Diehl 等人[8] 通过学习排序函数识别"经理-下属"关系. Menon 等人[9] 针对二元预测提出了对数线性矩阵模型. Wang 等人[10] 提出一种概率模型用于从出版物网络中挖掘"指导者-受指导者"的关系. Pentland 等人[11] 提出了几种模型用于挖掘 Mobile 数据,并用这些模型推断朋友关系. Tang 等人[12] 进一步提出一种通用的框架用于异构网络的社交关系类型的分类.

除了关系链接预测和关系类型预测之外,社交

关系研究中第 3 个重要分支是关系交互预测.

关系交互预测，即研究单向的社交关系怎样发展成双向的社交关系，及其产生的原因. 近年来，Hopcroft 等人[13]探索了关系交互预测问题，Lou 等人[14]研究了社交关系如何发展成三元闭包. 他们提出一种学习框架，将关系交互预测问题形式化成一种图模型，并在 Twitter 数据集上评估这个方法，该模型可以准确地预测出动态网络中 90%的交互关系.

鉴于社交关系分析的重要研究意义和实用价值，本文对目前社交关系的分析进行总结，主要从关系链接预测、关系交互预测和关系类型预测 3 个方面介绍社交关系分析相关的方法、理论和模型，最后给出一些应用实例.

图 2 给出社交关系分析中主要的研究课题. 它们的纵向关系由浅到深依次是关系链接预测、关系交互预测、关系类型预测. 其中，关系链接预测和关系类型预测都是单向预测，关系交互预测是双向预测，关系类型预测是对预测边进行语义研究.

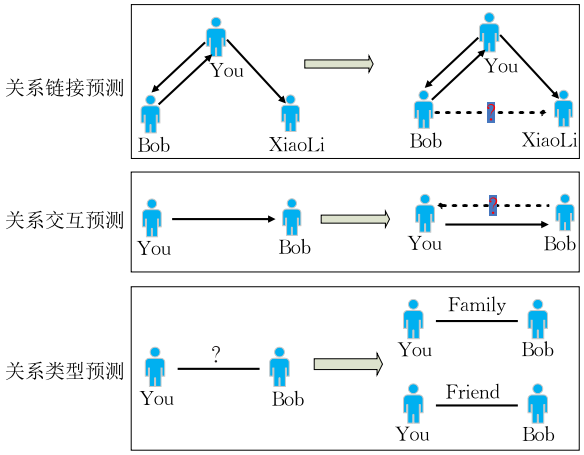


图 2　社交关系分析研究

本文第 2 节给出社交关系分析问题的形式化描述，详细介绍问题的定义和相关数据集；第 3，4，5 节分别介绍关系链接预测、关系类型预测和关系交互预测 3 个方面的方法、理论和模型，并给出重要的应用实例及其效果；最后对关系链接预测、关系类型预测和关系交互预测 3 个方面的内容、主要的进展和存在的问题进行整体的梳理，并对其未来的研究热点和新方向进行展望.

2　问题的形式化描述

2.1　问题的输入

首先，给出本文所讨论的问题的输入.

给定一个社交网络 $G=(V,E)$，其中 V 是 $|V|=N$ 个用户的集合，$E\subset V\times V$ 是 $|E|=M$ 条用户关系的集合. $e_{ij}\in E$ 表示从用户 v_i 到用户 v_j 的一条有向关系. 在无向网络中，$e_{ij}=e_{ji}$，在这种情况下，我们用 e_{ij} 或 e_{ji} 来表示用户 v_i 与 v_j 之间的关系. 下面我们将以无向图为例进行介绍相关定义，这些定义也可以很容易扩展到有向图上.

2.2　问题形式化描述

2.2.1　关系链接预测问题的形式化描述

问题 1.　关系链接预测. 给定一个已知社交网络 $G=(V,E)$ 和预测算法 $Pred()$，已知的社交关系集合 E，未知的社交关系集合为 B. 设用户数目为 $|V|=N$ 的网络的全关系集合为 U，且 $|U|=N\times(N-1)/2$，可知 $B=U-E$. 预测目标：$b_{rt}\in B$，表示网络中节点 v_r 与 v_t 之间没有连接. 对于任意 b_{rt}，由 $Pred(E,b_{rt})$ 输出一个数值 p_{rt}（该数值用以描述节点 v_r 与 v_t 产生链接关系的可能性）.

例 1.　关系链接预测的例子.

图 3(a)和(b)给出了一个以科学家合作网络为例的关系链接预测实例. 问题的输入：科学家之间的合作关系，$E=\{e_{12},e_{13},e_{23},e_{24},e_{34}\}$ 和一个给出关系链接预测算法 A. 预测目标：未有过合作的科学家的集合 $B=\{b_{14},b_{15},b_{24},b_{35},b_{45}\}$. 输出结果：虚线上的数值是预测算法 A 输出的预测分值，预测数值越大说明科学家未来合作的概率越大.

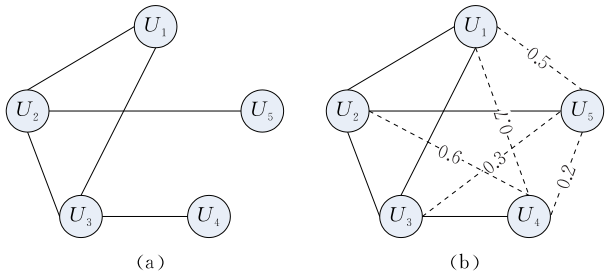


图 3　关系链接预测的例子

2.2.2　关系类型预测问题的形式化描述

问题的输入仍是一个社交网络 $G=(V,E)$，输出是社交关系的语义表示.

定义 1.　语义关系. 语义关系可以用一个三元组 (e_{ij},r_{ij},p_{ij}) 来描述，其中 $e_{ij}\in E$ 表示社交关系，$r_{ij}\in y$ 表示关系的语义标签，y 是所有标签的集合，p_{ij} 是算法所推断出的关系类型正确的概率（信度）.

社交关系在某些网络中可能是无向的（如从移动通信（Mobile）网络中发现的朋友关系），在另一些网络中是有向的（如 Publication 网络中的“指导者-受指导者”关系）. 此外，随着时间的推移，关系可能

是静态的(如家庭成员关系)或动态的(如同事关系).这里重点介绍静态关系.

为了推断出关系语义,需要考虑不同的因素,如特定用户的信息、特定链接的信息,以及全局约束.例如,要从 Publication 网络发现"指导者-受指导者"关系,可以考虑两位作者合作了多少篇论文,每位作者发表了多少篇论文,什么时候发表第 1 篇论文等.除此之外,部分关系可能存在标记.因此,在形式上,可以定义问题的输入为一个部分标签网络.

定义 2. 部分标签网络.部分标签网络表示为 $G=(V,E^L,E^U,R^L,W)$,其中 E^L 表示有标签的关系集合,E^U 表示无标签的关系集合,$E^L \cup E^U = E$. R^L 是与 E^L 中的关系对应的标签集;W 是 V 中与用户相关的属性矩阵,其中每行对应一个用户,w_{ij} 表示用户 v_i 的第 j 个属性.

基于上述概念,可以定义关系类型预测的问题,即给定一个部分标签网络,目标是预测网络中所有未知关系的类型(标签).以下给出问题形式化描述.

问题 2. 关系类型预测.给定一个部分标签网 $G=(V,E^L,E^U,R^L,W)$,社交关系推荐问题的目标是学习预测函数

$$f: G=(V,E^L,E^U,R^L,W) \rightarrow R.$$

针对这类问题,在许多情况下,标签数据是有限的,获取标签的代价也非常昂贵,是否可以设计一个策略,用最小的标记成本来主动地学习这个模型?给出主动学习的关系类型预测问题的描述.

问题 3. 基于主动学习的关系类型预测.给定一个部分标签网络 $G=(V,E^L,E^U,R^L,W)$ 和一个标记预算 b(用户交互的数量).目标是在 b 的约束范围下选择未知关系 $A' \subset E^U$ 的子集进行标记,从而尽可能的提高预测函数 f 的性能.

因此,关键问题是如何找到函数 f,可以利用有标签的关系和无标签的关系对未知关系进行推断.

问题 4. 基于迁移学习的关系类型预测.给定源网络 G_S 和目标网络 G_T,其中 G_S 中有大量带有标签的关系,而 G_T 中只有部分关系有标签.目标是学习预测函数 $f:(G_T|G_S) \rightarrow Y_T$,该函数利用源网络中的监督信息(标记的关系)推断目标网络中的关系类型.

该问题的输入由两个部分标记网络 G_S(源网络)和 G_T(目标网络)组成,$|E_S^L| \gg |E_T^L|$(极端的情况是 $|E_T^L|=0$).注意,这两个网络可能是完全不同的(具有不同的顶点集,即 $V_S \cap V_T = \varnothing$,且边定义的属性不同).

例 2. 关系类型预测的例子.

图 4 给出了在 Mobile 网络中关系挖掘的例子.左边的图是问题的输入:一个由用户、用户之间的呼叫和短信,以及用户的位置记录等组成的 Mobile 社交网络,目标是推断这个网络中关系的类型.在右边的图中是家庭成员的用户用实线相连,朋友用破折线相连,同事用虚线相连.与每对关系相关的概率表示在检测关系类型上的置信.

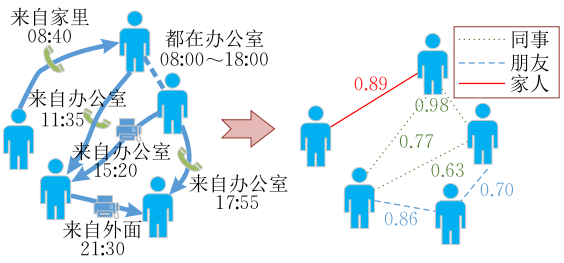


图 4 在移动通信网络中推断社交关系的例子[49]

例 3. 迁移学习的例子.

图 5 给出一个在跨产品评价网络和移动通信网络上推断社交关系的例子.在图 5 中,左边的子图是问题的输入:评论家网络,它由评论家和他们之间的关系组成;移动网络,它由移动用户和他们之间的通讯关系(通过呼叫或者文本短信建立)组成.右边的子图是问题的输出:在这两种网络中推断社交关系.在评论家网络中,推断信任/不信任关系,在通信网络中,推断朋友关系、同事关系和家人关系.图 5 的中间部分是用于在不同网络中推断社交关系的知识转移的组件.

2.2.3 关系交互预测问题的形式化描述

假设在 t 时刻,用户 v_i 创建了到 v_j 的链接,但之前 v_j 不存在到 v_i 的链接,则称 v_i 新关注 v_j.当用户 v_i 创建到 v_j 的链接时,v_j 已经存在到 v_i 的链接,则称 v_i 回粉 v_j.

问题 5. 回粉预测.令 $\langle 1,\cdots,t\rangle$ 表示某个时间粒度(日、星期等)上的时间序列.给定 t 时刻的社交网络 $G^t=(V^t,E^t,Y^t)$,其中 Y^t 是在 t 时刻的回粉标记集合.回粉预测的任务是得到一个预测函数 $f:(\{G^1,\cdots,G^t\}) \rightarrow Y^{t+1}$,从而可以预测在 $t+1$ 时刻的回粉行为.

在社交网络中,用户之间通过关注行为产生有向边,称这种方式下形成的网络为显式网络.除此之外,用户之间还会进行回复和转发等交互行为.如果用户 v 转发了用户 w 的帖子,则形成一条从 v 到 w 的有向边,称这种方式下形成的网络为隐式网络.

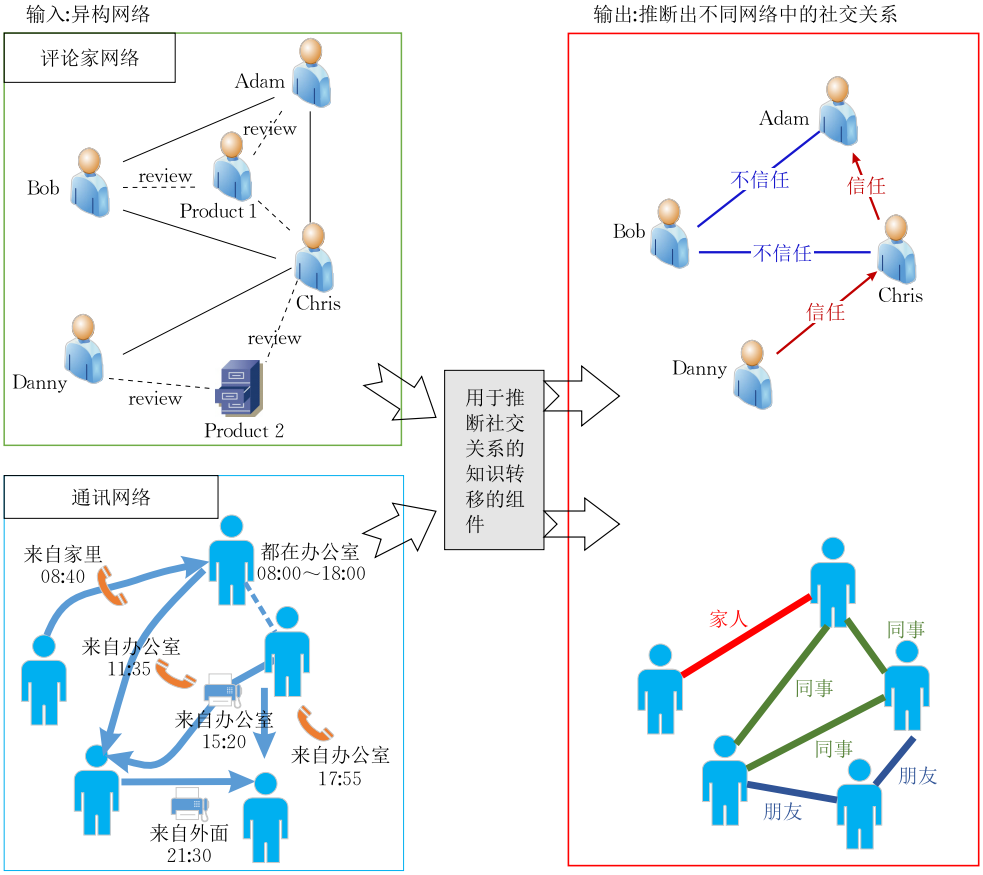


图 5　迁移学习的例子[12]

问题 6.　隐式关系交互预测. 给定图 G 和节点对$\{v,w\}$,且已知有向边(v,w)和(w,v)中至少存在一条边,但具体方向未知. 隐式关系交互预测的任务是判断连接 v 和 w 的边的方向. 具体可以再细分为两种问题:一是判断这条边是单向还是双向;二是在已知存在(v,w)的条件下,判断(w,v)是否存在.

问题 7.　三元闭包预测. 给定 t 时刻的社交网络$G^t=(V^t,E^t,X^t,Y^t)$,其中 X^t是在 t 时刻的回粉行为,如 v_i回粉 v_j. 三元闭包预测的任务是给定$\{G^1,G^2,\cdots,G^t\}$求预测函数 f,判断 v_i是否会在 $t+1$时刻新关注 v_j的某个被关注者 v_k,从而使(v_i,v_j,v_k)形成三元闭包.

给定 3 名用户 A,B,C,如果 A 和 B 以及 B 和 C 之间均存在链接,而 A 和 C 之间没有链接,则称三元组(A,B,C)是一个开放式三元组. 如果 A,B,C中任意两名用户之间均存在链接,则称三元组(A,B,C)是一个封闭式三元组. 给定 t 时刻的开放式三元组,需要预测该三元组是否会在 $t+1$ 时刻变为封闭式三元组,即 A 和 C 之间是否会产生链接,由此给出如下描述.

问题 8.　封闭式三元组预测. 记 t 时刻的网络$G^t=(V,E)$,用 Tr^t表示某个候选的开放式三元组,并为每个开放式三元组设置隐变量 y^t. 如果在 $t+1$时刻,三元组的状态变为封闭的,则 $y^{t+1}=1$,否则$y^{t+1}=0$. 给定所有历史信息,需要对 y 进行预测,即求:$f:(\{G^t,Y^t\}_{t=1,\cdots,T})\rightarrow Y^{t+1}$,其中 Y^{t+1}是 $t+1$ 时刻所有隐变量的值.

2.3　数据集

表 1 给出了本文介绍的实验中使用的主要社交网络信息.

表 1　实验中使用的网络数据集概括

名称	顶点数	边数	用途描述
Epinions[15-16]	131 828	841 372	信任/不信任网络
Slashdot[16]	77 357	516 575	朋友/敌人网络
Mobile[11]	107	5463	朋友网络
Coauthor[10,15]	815 946	2 792 833	合作者网络
Enron/Email[8]	151	3572	邮件通信网络
DBLP[16]	12 563	49 779	引文图
Hep-th[16]	27 766	352 807	引文图
Advogato[16]	7385	57 627	信任网络
WWW[16]	325 729	1 497 135	超链接图
WT10G[16]	1 601 787	8 063 026	超链接图
Publication[10]	1 036 990	1 990 260	合作者网络

Epinions 是产品评论网络. 每一个用户可以对任意产品发表评论,其他用户也可以对这个评论标

记信任或者不信任.通过评论者之间的信任和不信任关系构建网络.这个数据集包含 131 828 个节点(用户)和 841 372 条边,其中大约 85%是信任链接.该数据集用于推断用户之间的信任关系.

Slashdot 是朋友网络.Slashdot 是一个用来分享技术相关消息的网站.2002 年 Slashdot 引进 Slashdot Zoo 允许用户用"朋友"(喜欢)或"敌人"(不喜欢)标记其他人.这个数据集由 77 357 名用户和 516 575 条边组成,其中 76.7%的边是"朋友"关系.该数据集用于推断用户之间的"朋友"关系.

Mobile 是移动用户网络.这个数据集来源于文献[11],其中包含 10 个月中 107 名用户的呼叫记录、蓝牙扫描数据和发射塔 ID.Mobile 数据集总共包含 5463 条边.数据集用于推断两个用户之间是否存在朋友关系.

Coauthor 是合作者网络.这个数据集来自于 ArnetMiner.org[15],包含 815 946 名作者和 2 792 833 条合作关系.文献[10]中也用到了该数据.

Enron 是 Enron 公司的邮件通信网络,在有些文献中被称为 Email.它包含 Enron 公司的 151 名员工和员工之间的 136 329 封邮件.员工之间存在两种关系,即经理-下属关系和同事关系.该数据集由文献[8]提供,目标是推断用户之间的经理-下属关系.数据集中共有 3572 条边,其中 133 条是经理-下属关系.

DBLP 是学术引文图,数据来源于计算机科学电子书目网站.DBLP 中包含 12 563 名作者,49 779 条合著关系.文献[16]中也使用了这个数据集.

Hep-th 是 Arxiv 的高能物理学/理论方面的学术引文图,数据来源于文献[16].Arxiv 是一个收集物理学、数学、计算机科学与生物学论文预印本的网站(http://arxiv.org/).Hep-th 中包含 27 766 名作者,352 807 条合著关系.

Advogato 是具有 3 层信任的信任网络.包含 7385 个节点,57 627 条边.该数据集用于推断节点之间的信任/不信任关系.

WWW 和 WT10G 是从万维网的一个子集中提取的超链接网络数据集,来源于文献[16].其中,WWW 中包含 325 729 个节点,1 497 135 条边;WT10G 中包含 1 601 787 个节点,8 063 026 条边.

Publication 是 Coauthor 数据集的子集,由文献[10]提供.该数据集包含 ArnetMiner[15] 中 1936 年到 2010 年间的 1 632 442 本出版物,其中包含 1 036 990 个作者,1 984 164 条无标记关系和 6096 条有标记关系.

3 关系链接预测

3.1 关系链接预测的主要方法

目前社交网络关系链接预测方法可以分成如表 2 所示的几组.

表 2 关系链接预测的主要方法

种类	方法	描述
相似性度量	LPP[3]	多种相似性指标
	PPM[17]	拓扑特征
	FIF[18]	基于社会学领域知识的特征
	GFT[19]	图形特征追踪
	HSLP[20]	基于相似度的分层递阶结构
	SVC[21]	基于社交属性的向量模型
	VCP[22]	基于节点的属性
	DLP[23]	局部和全局属性
	LP-MPSN[24]	移动电话社交网络的特征
	LP-LSN[25]	基于社交网络的位置特征
	CTLP[26]	隐特征的降维
	NLFM[27]	隐特征模型
	LP-PS[28]	声望 vs. 相似度
	LP-SSN[29]	稀疏网络中基于相似度方法
矩阵因子分解	SGT[16]	谱图变换
	NMD[30]	非对角线矩阵分解
图模型	LPGM[31]	局部概率图模型
	SBM[32]	随机块模型
	TPFG[10]	有时间约束的概率因子图模型
	CSLP[33]	Bootstrap 概率图模型
	LINKREC[34]	随机游走
	SRW[6]	监督学习的随机游走
	DBN[35]	动态贝叶斯网络
	MTLM[36]	混合主题链接模型
	LFPM[37]	潜在朋友关系传播模型
	PFGM[38]	因子图模型
	DLFPM[39]	动态潜在特征传播模型

吕和周等人在 2011 年发表的一篇关于关系链接预测的综述文章,并在之后出版《链接预测》专著,将在这之前的国内外学者关于关系链接预测的研究成果做了很好的综述,详细见文献[40-41],与吕的综述相同的部分我们在这里仅作简要介绍,以保持完整性.

3.1.1 基于相似性度量的方法

基于相似性度量的方法一般通过计算节点(用户)对之间的相似值,再基于获得的相似值进行关系链接预测——相似值越高,产生链接的可能性就越高.

在文献[3]中,Liben-Nowell 和 Kleinberg 提出基于相似性度量的方法解决关系链接预测问题.给定一个社交网络,他们首先通过各种基于图的相似

性度量方法计算出节点对之间的相似度，然后利用相似度值预测两个节点之间的链接. Lee 等人[21]提出基于社交向量时钟特征求解关系链接预测问题，基于这种特征的计算代价较小. 在文献[22]中，Lichtenwalter 等人介绍了顶点排列轮廓(VCP)的概念进行拓扑链接分析和预测. VCP 提供了关于嵌入式节点对的局部结构周围几乎全部的信息. De 等人[23]把全局属性、局部属性，以及社团的中间层连接密度结合在一起，用在有识别能力的关系链接预测器中. 文献[24-25]分别为 Mobile 网络和基于位置的社交网络提取了不同的特征进行关系链接预测. Oyama 等人[26]提出一种降维的方法，从训练数据中学习一组特征投影矩阵，进行跨时间的关系链接预测. Zhu[27]提出一种最大间隔非参隐特征模型用于发现有识别力的隐含特征，并且自动地推断未知的隐含的社会维度. Papadopoulos等人[28]指出声望只是吸引力的其中一维，另一维是相似性，他们开发了一种框架能够精确地预测技术、生物和社交网络中的新链接，在这个框架中，新的链接优化了声望和相似度之间的平衡. 在文献[29]中，Lichtenwalter 等人发现了许多在影响和指导分类中有重大意义的因素，提出了一种有效的基于流的预测算法，这个算法给出了稀疏网络关系链接预测的不均衡性的边界. Zhu 等人[42]考虑路径的异质性，在相似度计算中利用中间节点的度，提出重要路径指标解决链接预测问题.

国内外众多的关系链接预测方法大多都适用于无权网络，只有极少部分可扩展到加权网络. Zhao 等人[43]提出了一种有效的基于“可靠路径”的适用于加权网络的预测方法.

3.1.2　基于矩阵因子分解的方法

基于矩阵因子分解方法将社交网络建模为一个矩阵，然后用线性代数或图论的矩阵分解法解决关系链接预测问题.

在文献[16]中，Kunegis 等人基于图的代数变换，提出一个统一框架用于学习大规模网络中的关系链接预测和边权预测的函数. 他们的方法概括了几种图内核和降维的方法，并且提供一种有效方法进行参数估计. 在文献[30]中，他们又提出一种基于有向网络的非对称邻接矩阵的非对角线分解的方法用于信任关系预测. 他们将非对角线的分解用于有向部件(DEDICOM)以获得网络的邻接矩阵的矩阵多项式的系数. 该方法可以用于计算网络的邻接矩阵的多项式的低阶近似，比奇异值分解法更好.

3.1.3　基于概率图模型的方法

基于概率图模型的方法是采用贝叶斯图模型方法对节点之间的联合概率进行建模，用贝叶斯图模型直接挖掘网络的隐关系通常是复杂的.

在文献[31]中，Wang 等人介绍了一种局部概率图模型的方法可以估计大规模网络中节点对的联合同现概率. Guimera 等人[32]提出一种基于随机分块模型的通用数学与计算框架处理复杂网络中数据可靠性问题，通过这个框架，可以从带有噪声的网络观测值中可靠地识别出丢失的和虚假的交互.

Wang 等人提出一种有时间约束的概率因子图模型(TPFG)，它将 Publication 网络作为输入，用联合似然目标函数对“指导者-受指导者”之间关系的挖掘问题进行建模[10]. 在文献[33]中，Leroy 等人提出一种基于引导概率图的两阶段方法. 在文献[34]中，Yin 等人提出使用随机游走算法在具有属性和结构信息的增广社交网络上评估链接的相关性. Backstrom 等人提出一种有监督的随机游走算法，这个算法自然地将节点的网络结构信息和边层次的属性结合在一起[6].

Marthur 等人[35]利用动态贝叶斯网络对带手动标注数据的合作团队检测其交互链接. Zhu 等人[36]将主题模型中的经典思想与混合成员分块模型的一种变型结合起来，提出混合主题链接模型用于无监督学习的主题分类和关系链接预测. Zhang 等人[37]将链接信息作为个人交友行为与个人兴趣相结合的结果. 他们提出潜在友谊传播网络(LFPN)描述个人的自我中心网络的演化发展，并使用潜在友谊传播模型(LFPM)对个人的社会行为进行建模. Wu 等人[38]提出一种交互式学习框架，这个框架将推荐专利合作伙伴的问题形式化描述为因子图模型. Heaukulani 等人[39]提出一种潜在属性传播模型，该模型通过捕获网络中已经观察到的社交关系如何影响未来没有被观察的网络结构进行关系链接预测.

3.2　相似性度量

首先介绍用于关系链接预测的各种相似性指标，包括基于节点邻居的指标、基于路径的指标、基于节点和边属性的指标以及基于隐特征的指标. 其中，基于节点邻居的指标 Lü 等人[40]已做详细介绍，这里我们将不再赘述，其他大部分的指标在文献[44-45]中也都有相关介绍，这里我们仅对一些经典指标加以阐述.

3.2.1 基于路径的度量

(1)最短路径距离

该类方法的基本思想是如果在一个社交网络中两个节点之间的距离(通过最短路径)比较短,那么这两个节点之间就很有可能产生一个链接.然而,根据六度分离理论,每个人和事物之间只有六步或者更短的距离.因此,这个属性有时候就不太起作用.

(2) Hitting Time

Hitting time 的概念来自于图的随机游走.给定图中两个节点 u 和 v,击中时间 $H_{u,v}$ 表示从 u 开始随机游走到 v 所需的期望步数.击中时间越短表明节点间越相似,因此它们在将来会有更高的机会进行链接.由于这个指标是不对称的,对无向图,可以使用往返时间 $C_{u,v}=H_{u,v}+H_{v,u}$ 表示.

(3) Rooted PageRank

PageRank 值可以作为关系链接预测的一个指标.然而,由于 PageRank 本身是量化单个节点重要性的指标,因此需要对它进行修正以估算节点对 u 和 v 之间的相似性.PageRank 最初的定义是:对某个固定的概率 α,一个网民以概率 α 从一个网页跳转到一个随机网页,以概率 $1-\alpha$ 关注一个已经链接的超链接.在这种随机游走下,网页 v 的重要性是链接到 v 的所有网页 u 的重要性的期望总和.在随机游走术语中,我们可以用"平稳分布"来取代"重要性".为了进行关系链接预测,原始 PageRank 随机游走的假设可以更改为:每一步随机游走中节点 v 以概率 $1-\alpha$ 返回到节点 u,以概率 α 移到一个随机的邻居节点的平稳概率来度量 u 和 v 之间的相似性.这个指标是不对称的,可以通过计算节点 u 和 v 互换后的对应项之和使它变得对称,这个指标也称为 Rooted PageRank.

3.2.2 基于节点属性的度量

顶点和边的属性在关系链接预测中扮演着非常重要的角色.注意,在一个社交网络中,代表节点的个体效用是链接产生的直接动机,这种效用是节点和边的属性函数.许多研究表明,节点或边的属性作为近似特征可以明显地提高关系链接预测的性能.

(1)择优连接值

一般来说,一个节点与网络中其他节点相连接是基于它们度的概率.因此,如果把邻居节点集的大小作为特征值,那么将它们相乘就是一个聚合函数,称为"择优连接值":

$$\text{Preferential-Attachment-Score}(u,v)=\Gamma(u)\cdot\Gamma(v) \quad (1)$$

(2)聚类系数值

节点 u 的聚类系数表示如下:

$$\text{Clustering-Coefficient}(u)=\frac{2\times n}{k\times(k-1)} \quad (2)$$

其中 n 是节点 u 的所有 k 个邻居间的边数.

为计算顶点 u 和 v 之间用于关系链接预测的得分,我们可以把 u 和 v 的聚类系数值求和或相乘.

3.2.3 基于隐特征的度量

在许多关系链接预测问题中,尽管表示数据对象的特征向量是高维的,但对预测真正起作用的隐特征的数量却相对较少.因此,可以通过识别和使用低维隐特征空间来提高关系链接预测的准确性.在监督学习的线性降维方法中,从原始的 D 维特征空间到 $d(<D)$ 维隐特征空间的线性投影 W 是通过训练数据学习到的,这些训练数据包括的数据对象是已知在它们之间有或没有链接的.在学习过程寻求线性投影 W 使得映射空间的距离 $\|W_x-W_y\|$ 尽可能小,其中 x 和 y 是两个已知它们之间存在链接的节点.学习过程完成后,两个具有未知链接状态的数据对象通过使用 W 被映射到隐空间.如果这两个数据对象映射后彼此足够近,就认为他们之间存在一个链接.

假设有 N 个训练数据对象,$x_1,\cdots,x_N$,每一个数据对象 x_i 用 D 维特征向量表示[26].可以使用保留局部性投影,通过求解下面的优化问题发现优化的线性投影矩阵 $\boldsymbol{W}^*$:

$$\boldsymbol{W}^*=\arg\min_{\boldsymbol{W}}\sum_{i,j}A_{ij}\|W_{x_i}-W_{x_j}\|_2^2 \quad (3)$$

其中:$\|\cdot\|_2$ 是欧几里德范数(2-范数);$\boldsymbol{A}'=\{A_{ij}\}$ 是邻接矩阵,定义为

$$A_{ij}=\begin{cases}1, & \text{如果 } x_i \text{ 和 } x_j \text{ 之间存在链接}\\ 0, & \text{其他}\end{cases} \quad (4)$$

3.3 矩阵分解

本节以预测无向链接为例,介绍谱变换的基本方法.

如果用代数的方法处理关系链接预测问题,可以考虑图的邻接矩阵 $\boldsymbol{A}'$,寻找一个函数 $F(\boldsymbol{A}')$ 返回一个相同大小的矩阵,使得它的元素可以用来预测.Kunegis 等人提出的方法包括两个过程:计算矩阵分解 $\boldsymbol{A}'=\boldsymbol{U}'\boldsymbol{D}'\boldsymbol{V}'^{\mathrm{T}}$;求解函数 $F(\boldsymbol{A}')=\boldsymbol{U}'F(\boldsymbol{D}')\boldsymbol{V}'^{\mathrm{T}}$,其中 $F(\boldsymbol{D}')$ 将实数集上的函数分别运用到图谱 $\boldsymbol{D}'$ 的每一个元素上[16].Kunegis 等人证明了有为数不少的共有链接和边权预测算法可以映射成这种形式.该方法为这类的关系链接预测算法提供了一种评估任意参数的机制.类似地,他们也考虑了将网络

的拉普拉斯矩阵作为关系链接预测的基础.

3.4　概率图模型

众所周知,朋友关系的传递性是社交网络中流行的社会学原理.然而,人们的交友行为在何种程度上遵循这个原理以及这个原理在多大程度上有益于关系链接预测工作仍然是未知的.在文献[37]中,Zhang 等人尝试采用这个社会学原理来解释网络的发展,并学习潜在朋友关系传播的问题.

3.4.1　问题描述

Zhang等人提出了潜在朋友关系传播网络(LFPN)捕获朋友关系传递性的趋势.LFPN g'由 3 层构成:ego 层,local 层和 global 层.接下来,首先给出 local 层和 global 层的定义,再给出 LFPN 的定义.

定义 3.　潜在朋友关系传播三元组(LFPTriple).给定 G 中的 3 个顶点 u,z 和 v,如果 u 通过 z 和 v 交朋友,那么(u,z,v)就称为 LFP 三元组.u 是这个元组的发起人,z 是中间人.

定义 4.　Local 层.在 LFPN g' 上 ego u 的 local层表示为 $g'^{L(u)}$,由 u 的所有朋友组成.对每一个由 u 发起的 LFP 三元组(u,z,v),在 $g'^{L(u)}$ 中有一个对应的 local LFP 边$\langle z,v\rangle$,它的权重是 $w_{z,v}^{L(u)}=p(z|\langle u,v\rangle)$.

定义 5.　对每一个朋友 $z\in g'^{L(u)}$,在 local 层中有一个中间倾向边$\langle u,z\rangle$从 ego u 指向 z,它的权重是 $w_{u,z}^{E}=p(u\rightarrow z)$.

定义 6.　每一个 LFP 三元组(u,z,v)源自一个 LFP 模式 $z\rightarrow v$,表明 z 的朋友关系可能传播给 v.

定义 7.　Global 层.LFPN g'的 global 层,记为 g'^{G},由给定社交网络的所有个体组成.对每个 LFP 模式 $z\rightarrow v$,存在一个相应的 global LFP 边$\langle z,v\rangle$,其权重是 $w_{z,v}^{G}=p(z\rightarrow v)$.

定义 8.　潜在朋友关系传播网络(LFPN).给定社交网络的潜在朋友关系传播网络(LFPN)g'是一个加权三层网络,由 ego 层$\{u\}$,每个 ego 的 local 层$\{g'^{L(u)}\}$和 global 层 g'^{G}组成.

LFPN 推理问题.给定一个演化中的网络 $G^{\#}=\{G_1,G_2,\cdots,G_t\}$,其相关交互集合为 $X^{\#}=\{X_1,X_2,\cdots,X_t\}$,LFPN 推断问题的目的是推断 LFPNs $g'_1,g'_2,\cdots,g'_t,g'_t$其中是 LFPN 的 G_t.

3.4.2　潜在朋友关系传播模型(LFPM 模型)

每个 LFP 三元组(u,z,v)隐含这样一种假设,即交友行为有两步过程:首先 u 选择 z 作为中间人,然后通过 z 和 v 交朋友.同时这个过程必然受到 u 的兴趣的影响.LFP 模式是 LFP 三元组的聚合,它们都是 LFPN 的基础元素.因此 LFPN 推断的初步问题是发现所有的 LFP 三元组.换句话说,给定 u 和 v 之间的朋友关系,我们需要推断中间人 z.为了模拟这个过程,并推断每种朋友关系的中间人,Zhang 等人[37]提出了潜在朋友关系传播模型(LFPM),它是社交网络中的一种社会行为的概率生成模型.

3.5　异质信息网络

目前,社交网络链接预测的研究成果绝大多数都是在单一网络或者同质网络上实现的,但是也有很多学者逐渐将研究重点转向异质信息网络,使得异质信息网络的链接预测成为链接预测研究的一大重点方向.相对单一网络或者同质网络来说,异质网络中节点和链接的类型不再是单一的,而是多样的,从而增加了链接预测的复杂度.

Zhang 和 Philip[46]的关于异质信息网络的链接预测方面的研究综述详细介绍了国内外该领域学者的研究成果.此外,Liu 等人[47]提出了均衡的因子图(AFG)模型解决均衡的异质网络链接预测问题,并提出均衡结构算法降低因子图的规模,从而提高预测性能.Dong 等人[48]提出用广义的耦合张量分解框架进行异质信息网络的链接预测.

4　关系类型预测

关系类型预测的目标是有效地推断两个用户之间社交关系的类型.具体说来,给定用户行为历史和用户间的交互,能否估计他们是家庭成员或同事的可能性?面临的一个挑战就是如何设计一个统一的模型,使其能够很容易应用到不同领域(或不同网络).目前已有一些相关研究.如,Diehl 等人[8]尝试通过学习排序函数识别这些关系.Wang 等人[10]提出一种无监督算法用于从出版物网络中挖掘“指导者-受指导者”关系.然而,Diehl 等人只考虑通信记录,Wang 等人提出的是一种特定领域的无监督算法.两种算法都不容易扩展到其他领域.

另一个挑战是当今的社交网络变得越来越复杂多变,甚至最先进的算法获得的最优性能也低于 90%,这一结果无法令人满意.比较好的解决方案是:设计交互式界面允许用户提供对推断结果的反馈.但交互过程可能是乏味的、容易出错,而且很耗费时间.比较理想的情况是,这个算法应该能够主动

选择少数可能存在错误的结果让用户检查,而不是被动地等待用户反馈.这个问题被称为通过主动学习来推断社交关系.

因此,基本的问题是,如何针对不同的网络设计一个灵活的模型从而有效地推断社交关系.这个问题面临一系列的挑战.首先,哪些潜在因素决定特定类型的社交关系的形成?其次,作为输入的社交网络是部分标记的.有些关系是有标记的,但大多数关系是未知的.要得到一个高质量的预测模型,不仅要考虑有标记的关系提供的信息,也要考虑如何利用无标记的网络信息.第三,如何最好地利用用户交互.在选择用于进行用户交互的内容时,应该考虑不确定性和网络结构信息.最后,实际的社交网络的规模越来越大,拥有数百万甚至更多的节点,设计一种适用于大规模实际网络的方法非常重要.

本节主要介绍无监督学习、监督学习和主动学习、迁移学习四类框架下的关系类型预测模型.

4.1 无监督学习的关系类型预测

获得数量充足的标签关系的代价是昂贵的,Wang 等人[10]提出一种无监督学习的方法来推断不带标签数据的社交关系,用来推断合作网络中的"指导者-受指导者"关系,称为"有时间约束的概率因子图模型(TPFG)".其主要思想是利用有时间限制的概率因子图模型对每个作者的指导者的联合概率进行分解.通过最大化因子图的联合概率,在候选图上推断关系,计算每个关系的排列名次.

框架按如下两个阶段进行处理.

阶段 1:预处理.预处理的目的是生成候选图 H',使得大多数情况下保证指导者不会被排除在候选池之外,减少搜索空间.

首先根据合作者信息,通过逐一处理网络中的论文,产生一个均匀的作者网络 G'.对每篇论文 p_i,可以对其中每一对作者构造一条边.

然后通过执行过滤过程,排除那些不可能的"指导者-受指导者"关系.对图 G' 上的每条边 e_{ij},表明 a_i 和 a_j 有合作.为了确定 a_j 是否可能是 a_i 的指导者,需要检查以下条件.首先,只有当 a_j 开始发表文章的时间比 a_i 早时,才考虑这种可能.其次可以基于"指导者-受指导者"关系的直观知识来启用一些启发式规则.详细的规则定义可参考文献[10].

阶段 2:因子图模型.根据候选图 H' 中的局部信息,可以知道每个作者可能的指导者以及这种可能性.对整个网络进行建模,结合结构信息和时序约束,从而更好地分析个体链接间的关系.

使用因子图模型,为候选人图 H' 的每个节点定义一个隐变量,通过最大化目标函数,得到隐变量的取值.学习该模型,可以考虑使用和积算法(sum-product)与联合树(junction tree)算法[13].

为了对该方法进行评价,使用 Publication 数据库,从一些在线资源中收集了有标签的"指导者-受指导者"关系.在该数据库上,TPFG 模型的 $F1$ 测度达到了 81%~85%.

4.2 监督学习的关系类型预测

4.2.1 PLP-FGM 模型

进行关系类型预测时,Tang 等人[49]结合下面描述的一些直观知识,提出一种部分标记的成对因子图模型(PLP-FGM).首先,特定用户或特定链接的属性中可能包含关系的隐含信息.例如,两个在工作时间经常打电话的用户可能是同事;而两个经常在晚上互相联系的用户更可能是家庭成员或好友.第二,不同用户之间的关系可能有相关性.例如,在移动网络中,如果用户 v_i 打电话给 v_k 后立即打电话给 v_j,那么 v_i 可能与 v_j 和 v_k 有类似的关系(家庭成员或同事).第三,还需要考虑一些全局约束,例如常识或对特定用户的约束.

图 6 是 PLP-FGM 的图形化描述.部分标签网络 G 中的每一个关系 (v_{i_1},v_{i_2}) 或 $e_{i_1i_2}$ 在 PLP-FGM 上都被映射为关系节点 $r_{i_1i_2}$.一组关系节点可表示为 $Y=\{y_1,y_2,\cdots,y_M\}$.G 中的关系是部分标记的,因此 PLP-FGM 中的所有节点可以分为两个子集 Y^L 和 Y^U,分别对应有标签和无标签的关系.对每个关系节点 $y_i=(v_{i_1},v_{i_2},r_{i_1i_2})$,可以把属性 $\{W_{i_1},W_{i_2}\}$ 组合成关系属性向量 $\boldsymbol{x}_i$.

在 PLP-FGM 中,输入的关系表示为模型中的关系节点.与之前所提到的 3 个直观知识相对应,可定义以下 3 个因子.

(1)属性因子.$f(y_i,x_i)$ 代表给定属性向量 $\boldsymbol{x}_i$ 时关系 y_i 的后验概率;

(2)相关因子.$g^*(y_i,G(y_i))$ 表示关系之间的相关性,其中 $G(y_i)$ 表示与 y_i 有关的一组关系.

(3)约束因子.$h^*(y_i,H(y_i))$ 反映了关系之间的约束,其中 $H(y_i)$ 表示对 y_i 有约束的一系列关系.

给定一个部分标记网络 $\boldsymbol{G}=(V,E^L,E^U,R^L,W)$,可以定义 Y 上的联合概率分布

$$p(Y|G)=\prod_i f(y_i,x_i)g^*(y_i,G(y_i))h^*(y_i,H(y_i)) \quad (5)$$

3 个因素可以用不同的方法实现.这里使用指数-线性函数.具体说来,定义属性因子为

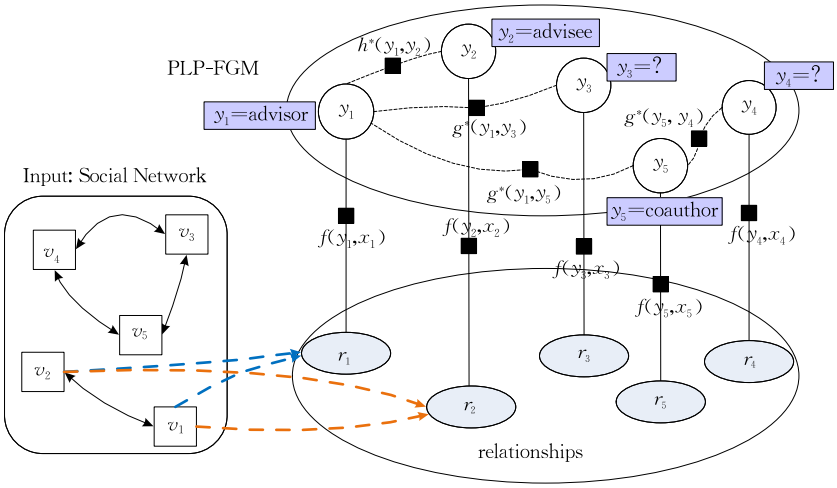


图 6　PLP-FGM 模型的图形化表示[49]

$$f(y_i,x_i)=\frac{1}{Z_\lambda}\exp\{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Phi}(y_i,x_i)\} \tag{6}$$

其中：$\boldsymbol{\lambda}$ 为权重向量；$\boldsymbol{\Phi}$ 是特征函数向量. 类似的，相关因子和约束因子可定义为

$$\boldsymbol{g}^*(y_i,G(y_i))=\frac{1}{Z_\alpha}\exp\Big\{\sum_{y_j\in G(y_i)}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{g}(y_i,y_j)\Big\} \tag{7}$$

$$\boldsymbol{h}^*(y_i,H(y_i))=\frac{1}{Z_\beta}\exp\Big\{\sum_{y_j\in H(y_i)}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{h}(y_i,y_j)\Big\} \tag{8}$$

其中 $\boldsymbol{g}$ 和 $\boldsymbol{h}$ 函数可以定义为指示函数向量，这种特征定义在马尔科夫随机场①和条件随机场[50]等图模型中经常使用.

(1) 模型学习

PLP-FGM 的学习是估计参数 $\theta=(\lambda,\alpha,\beta)$，使观测信息(标记关系)的对数似然函数最大化. 为简单起见，将关系节点 y_i 的所有因子函数合并为 $s(y_i)=(\boldsymbol{\Phi}(y_i,x_i)^{\mathrm{T}},\sum_{y_i}\boldsymbol{g}(y_i,y_i)^{\mathrm{T}},\sum_{y_i}\boldsymbol{h}(y_i,y_i)^{\mathrm{T}})^{\mathrm{T}}$.

式(5)中定义的联合概率可以重写为

$$\begin{aligned}\boldsymbol{p}(Y\mid G)&=\frac{1}{Z}\prod\exp\{\boldsymbol{\theta}^{\mathrm{T}}s(y_i)\}\\&=\frac{1}{Z}\exp\{\boldsymbol{\theta}^{\mathrm{T}}\sum s(y_i)\}\\&=\frac{1}{Z}\exp\{\boldsymbol{\theta}^{\mathrm{T}}S\}\end{aligned} \tag{9}$$

其中，$Z=Z_\lambda Z_\alpha Z_\beta$ 为归一化因子(也称为配分函数)；S 是所有关系节点的因子函数的聚合(aggregation)，即 $S=\sum_i s(y_i)$.

学习 PLP-FGM 模型时的一个挑战是：输入数据是部分标记的. 在计算配分函数 Z 时，需要对所有节点(包括无标记节点)可能状态的似然性求和. 为了解决这一问题，可使用有标签的数据来推断未知标签. 用 $Y\mid Y^L$ 表示根据已知标签推断出来的标签 Y，定义如下的对数似然目标函数 $O'(\boldsymbol{\theta})$

$$\begin{aligned}O'(\boldsymbol{\theta})&=\log p(Y^L\mid G)\\&=\log\sum_{Y\mid Y^L}\frac{1}{Z}\exp\{\boldsymbol{\theta}^{\mathrm{T}}S\}\\&=\log\sum_{Y\mid Y^L}\exp\{\boldsymbol{\theta}^{\mathrm{T}}S\}-\log Z\\&=\log\sum_{Y\mid Y^L}\exp\{\boldsymbol{\theta}^{\mathrm{T}}S\}-\log\sum_{Y}\exp\{\boldsymbol{\theta}^{\mathrm{T}}S\}\end{aligned} \tag{10}$$

为了求解目标函数，考虑梯度下降法，首先计算参数 $\boldsymbol{\theta}$ 的梯度

$$\begin{aligned}\frac{\partial O'(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}&=\frac{\partial\big(\log\sum_{Y\mid Y^L}\exp\boldsymbol{\theta}^{\mathrm{T}}S-\log\sum_{Y}\exp\boldsymbol{\theta}^{\mathrm{T}}S\big)}{\partial\boldsymbol{\theta}}\\&=\frac{\sum_{Y\mid Y^L}\exp\boldsymbol{\theta}^{\mathrm{T}}S\cdot S}{\sum_{Y\mid Y^L}\exp\boldsymbol{\theta}^{\mathrm{T}}S}-\frac{\sum_{Y}\exp\boldsymbol{\theta}^{\mathrm{T}}S\cdot S}{\sum_{Y}\exp\boldsymbol{\theta}^{\mathrm{T}}S}\\&=E^{\#}_{p\boldsymbol{\theta}(Y\mid Y^L,G)}S-E^{\#}_{p\boldsymbol{\theta}(Y\mid G)}S\end{aligned} \tag{11}$$

另一个挑战是，PLP-FGM 中的图形结构可以是任意的，可能包含循环，这时难以直接计算期望 $E^{\#}_{p\boldsymbol{\theta}(Y\mid G)}S$. 研究者提出了一些近似算法，这里采用环路信度传播(LBP)[51]算法对边缘概率 $p(y_i\mid\boldsymbol{\theta})$ 和 $p(y_i,y_i\mid\boldsymbol{\theta})$ 进行近似估计，结合边缘概率，对所有关系节点求和得到梯度. 值得注意的是，在每次迭代中需要执行两次 LBP，一次用于估计边缘概率 $p(y\mid G)$，另一次估计 $p(y\mid Y^L,G)$. 根据梯度，用学习率 η 更新每个参数. 算法 1 对 PLP-FGM 的学习过程进行了总结.

① Hammersley J M, Clifford P. Markov field on finite graphs and lattices. http://www.citeulike.org/group/14833/article/8970271, 1971

算法 1. PLP-FGM 学习过程.

输入:学习速率 η

输出:学习参数 θ

初始化 θ;

重复

使用 LBP 计算 $E^{\sharp}_{p_{\boldsymbol{\theta}}(Y|Y^L,G)}S$;

使用 LBP 计算 $E^{\sharp}_{p_{\boldsymbol{\theta}}(Y|G)}S$;

根据方程(18)计算 $\boldsymbol{\theta}$ 的梯度

$$\nabla_\theta = E^{\sharp}_{p_{\boldsymbol{\theta}}(Y|Y^L,G)}S - E^{\sharp}_{p_{\boldsymbol{\theta}}(Y|G)}S$$

使用学习速率 η 更新参数 $\boldsymbol{\theta}$

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \nabla_\theta$$

直到收敛;

(2) 推断未知社交关系

根据学习到的参数,通过最大化式(5)的联合概率,得到对应的标签,从而预测关系的标签.

$$Y^* = \arg\max_{Y|Y^L} p(Y \mid G) \tag{12}$$

因此,再次使用 LBP 计算每个关系节点的边缘概率 $p(y_i|Y^L,G)$,然后具有最大边缘概率的标签作为待预测关系的类型,边缘概率是预测结果的信度.

(3) 时间复杂度分析

用 m_1,m_2,m_3 分别表示 PLP-FGM 中的属性因子、相关因子和约束因子的数量. 在 LBP 的每一次循环中,传播的时间代价是 $O(m_1 \cdot \dim(\boldsymbol{\Phi}) + m_2 \cdot \dim(\boldsymbol{g}) + m_3 \dim(\boldsymbol{h}))$,其中 $\dim(\cdot)$ 是向量维度. 对算法执行 n 次迭代,在每迭代中执行 n_{LBP} 次 LBP. 因此时间复杂度为 $O(m_1 \cdot \dim(\boldsymbol{\Phi}) + m_2 \cdot \dim(\boldsymbol{g}) + m_3 \dim(\boldsymbol{h})) \times n \times n_{\text{LBP}}$.

4.2.2 评 估

Tang 等人[49] 在 Publication、Email 和 Mobile 这 3 种不同的数据集上对方法的性能进行了评估. 在 3 种数据集上使用不同的方法进行关系挖掘并比较性能,结果如表 3 所示.

表 3 在 3 种数据集上使用不同方法进行关系挖掘的性能[49] (单位:%)

数据集	方法	*Accuracy*	*Precision*	*Recall*	*F1*
Publication	SVM	76.6	72.5	54.9	62.1
	TPFG	81.2	82.8	**89.4**	86.0
	PLP-FGM-S	84.1	77.1	78.4	77.7
	PLP-FGM	**92.7**	**91.4**	87.7	**89.5**
Email	SVM	82.6	79.1	**88.6**	83.6
	PLP-FGM-S	85.6	85.8	85.6	85.7
	PLP-FGM	**88.0**	**88.6**	87.2	**87.9**
Mobile	SVM	80.0	**92.7**	64.9	76.4
	PLP-FGM-S	80.9	88.1	71.3	78.8
	PLP-FGM	**83.1**	89.4	**75.2**	**81.6**

(1) SVM. 使用关系属性向量 $\boldsymbol{x}_i$ 训练分类模型,然后利用该模型预测关系.

(2) TPFG. 即上一节中给出的无监督学习方法. 由于该方法与领域有关,因此仅在 Publication 数据集上进行比较.

(3) PLP-FGM-S. 与基于部分标记网络的 PLP-FGM 模型相比,PLP-FGM-S 只使用有标签的节点进行训练.

4.3 主动学习推断社交关系

如 2.2.2 节中问题 4 所述,针对主动学习的关系类型预测,Zhuang 和 Tang 等人[52] 定义质量函数 $\boldsymbol{Q}(A)$ 用于衡量标记集合 A 中的关系后所带来的预测性能的改善,此时问题可以定义为 $\boldsymbol{Q}(A)$ 的优化问题,即

$$A^* = \arg\max_{A \subset Y^U} \boldsymbol{Q}(A), \ |A| = b, \ b > 0 \tag{13}$$

为了量化 $\boldsymbol{Q}(A)$,可以考虑一个被选中的节点是如何影响其他节点的. 例如,纠正中心位置关系可能会把这种纠正传播出去,从而有利于相关关系的推断.

基于上面的直观想法,设计了影响力最大化选择(IMS)和信任最大化选择(BMS)两种模型用于主动地推断社交关系类型. IMS 模型利用从 PLP-FGM 中获得的网络结构和不确定性选择最具影响力的节点. BMS 模型进一步将主动选择过程与 PLP-FGM 的学习过程相结合.

4.3.1 影响力最大化选择(IMS)

IMS 模型为每个节点设置分数以反映它在模型中影响传播的强度. 传播过程可描述为:

(1) 图形化表示同图 6 所示. 当一个节点的标记被确定后,称这个节点是激活的. 初始化的激活节点集合是 Y^L. 为每个节点指定一个阈值 $\varepsilon_i = \sum_{y \in y'} \left| p(y_i = y|G,Y^L) - \frac{1}{|y'|} \right|$,其中 y' 表示 y 的取值空间,这样不确定性较高的节点将更容易被激活.

当一个节点 i 被激活时,它将所获得的分数增益 $(g'_i - \varepsilon_i)$ 以权重 $b_{i,j}$ 传播给相邻节点 $j \in NB(i)$,即 $g'_i \leftarrow g'_i + b_{i,j}(g'_i - \varepsilon_i)$,其中 g'_i 表示节点 i 的分数,$NB(i)$ 表示节点 i 的邻居节点集. 获得的分数增益反映了用户标记带来的信度提升,因此标记不确定关系将比标记较确定关系带来更大的影响力. 为了简化问题,设置权重 $b_{i,j} = \frac{1}{|NB(j)|}$.

若用户标记了一个节点,则将该节点设置为激活状态,并将它获得的分数设置为 1. 开始时,其他

节点获得的分数均设置为 0. 一旦某个非激活节点 k 所得的分数超过了阈值，即 $g'_k > \varepsilon_i$，就变成激活状态，并传播它所获得的分数. 一个激活节点只能传播一次分数，并保持其状态.

将质量函数 $\boldsymbol{Q}_{\mathrm{IMS}}(A)$ 定义为传播完成之后激活节点的总数. 要找到一个能最大化质量函数 $\boldsymbol{Q}_{\mathrm{IMS}}(A)$ 的集合 A 是 NP-难问题. 与文献[53]类似，可以采用贪心策略进行近似求解.

4.3.2　信任最大化选择(BMS)

使用 LBP 过程中所获得的每个节点的信度来量化一个节点对其他节点的影响力，并从信度的分数中移除属性所带来的影响，由此获得一个启发式规则，表示为 $B'(y_i|G,Y^L)$.

$$B'(y_i|G,Y^L)=\exp\{\boldsymbol{\theta}^{\mathrm{T}}s(y_i)-\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Phi}(y_i,x_i)\} \quad (14)$$

通过对关系节点的信度进行归一化，可获得其信度的边缘概率.

$$pB'(y_i|G,Y^L)=\frac{1}{Z_{B'}}B'(y_i|G,Y^L) \quad (15)$$

其中 $Z_{B'}$ 是归一化因子. 式(15)用于估计关系节点的边缘概率分布，其中不存在属性相关的信息.

一个直观的想法是，关系节点的信度随着相同类型关系节点数量的增加而单调增加，即 $B'(y_i=y|G,Y^L)$ 随着带有标签 y 的关系的数量而单调增加. 不失一般性，首先考虑二值关系挖掘问题，即只有两种可能的关系标签($y'=\{0,1\}$). 在二值情况下，进一步针对每种类型分别考虑主动选择，因为不同类型的关系混合在一起的时候，无法保证得到一个解析解. 因此，当用户仅提供正反馈时，目标是发现正的节点集. 对应地，定义正向 BMS 策略的质量函数为

$$\boldsymbol{Q}_{\mathrm{BMS}^+}(A)=\sum_{y_i\in Y^U_{(1)}}pB'(y_i=1|G,Y^L\cup A) \quad (16)$$

其中 $Y^U_{(1)}=\{y_i\,|\,y_i\in Y^U\wedge B'(y_i=1(G,Y^L)\geq B'(y_i=0|G,Y^L))\}$.

质量函数 $\boldsymbol{Q}_{\mathrm{BMS}^+}(A)$ 和 $\boldsymbol{Q}_{\mathrm{BMS}^-}(A)$ 的最优化都是 NP-难问题. 但这两个质量函数都是子模函数，可以通过贪心算法获得最优解($1-1/\mathrm{e}$)的近似解. 具体策略为每次选择那些能使质量函数获得最大边缘增长的关系. 这里注意一点，由于主动学习算法在选择阶段是不知道标签的，在最优化 $\boldsymbol{Q}_{\mathrm{BMS}^+}(A)$ 时，将关系节点 y_i 作为正类，在最优化 $\boldsymbol{Q}_{\mathrm{BMS}^-}(A)$ 时，将 y_i 作为负类.

为了防止进行不平衡选择，使用 $\boldsymbol{Q}_{\mathrm{BMS}^+}$ 选择 $b/2$ 个节点(其中 b 是每次需要询问用户的关系数量)，然后使用 $\boldsymbol{Q}_{\mathrm{BMS}^-}$ 选择剩下的 $b/2$. 这种选择策略被称为 BMS，它结合了 BMS^+ 和 BMS^-，但无法保证近似解得到更低的误差界.

4.3.3　评　估

这里仍然使用 4.2.3 节中使用的数据集评估不同的主动学习算法的性能. 算法执行了多次，表 4 列出了不同数据集上不同选择策略的平均 $F1$ 值.

表 4　所有选择策略的$F1$值[52]　(单位：%)

数据集	随机	MU[12]	ID[54]	BMS	IMS
Publication	60.6	63.7	64.8	66.4	**66.8**
Email	85.6	86.2	87.3	**87.6**	86.3
Mobile	79.2	80.0	74.3	**80.4**	79.9

在每个数据集中，首先随机选择 10 条关系作为初始标签集 Y^L，然后迭代地执行主动选择算法，每次选择 $b=10$ 条关系进行查询.

4.4　通过迁移学习进行跨网络的关系类型预测

传统的方法中，通常需要数量充足的有标签关系来获得好的预测模型以推断社交关系. 但是，在不同网络中获得关系标签的难度是不同的. 有些网络中，比如 Slashdot，可能更容易收集有标签的关系(比如用户之间的信任/不信任关系)，而在大部分其他网络中，获得标签信息可能是困难的(甚至是不可行的). 一个挑战问题是：是否可以利用一个网络中有标签关系的信息来推断另一个完全不同网络中的关系类型？

Tang 等人[12]提出基于迁移的因子图模型 G 用于跨网络进行社交关系类型的学习和预测. 首先在单个网络中进行学习，然后将网络提供的监督信息迁移到另一个网络.

4.4.1　模型框架

(1)跨异构网络学习

在基于迁移的因子图模型 TranFG 中结合社会学理论(如社会平衡、结构洞、社会状态、意见领袖等[55])，在源网络和目标网络上定义如下的对数似然目标函数

$$\begin{aligned}O'(\alpha,\beta,\mu)&=O'_S(\alpha,\mu)+O'_T(\beta,\mu)\\&=\sum_{i=1}^{|V_S|}\sum_{j=1}^{d}\alpha_j g_j(x_{ij}^S,y_i^S)+\\&\quad\sum_{i=1}^{|V_T|}\sum_{j=1}^{d'}\beta_j g''_j(x_{ij}^T,y_i^T)+\\&\quad\sum_k\mu_k\Big(\sum_{c\in G_S}h_k(Y_c^S)+\sum_{c\in G_T}h_k(Y_c^T)\Big)-\\&\quad\log Z\end{aligned} \quad (17)$$

其中：d 和 d' 分别是源网络和目标网络中属性的数量；$|V_S|$ 和 $|V_T|$ 分别是源网络和目标网络中节点的数量；μ_k 表示第 k 个相关功能函数的权重；h_k 表示第 k 个相关功能函数；Y_c^S 和 Y_c^T 分别表示源网络和目标网络中 c 领域的标签集合. 在这个目标函数中，前两项分别定义源网络和目标网络上的似然函数，第 3 项定义这两个网络上共有特征的似然函数，共有特征函数根据社会学理论进行定义.

(3) 模型学习和推断

TranFG 模型的学习是估计参数 $\theta=(\{\alpha\},\{\beta\},\{\mu\})$，从而最大化对数似然目标函数 $O'(\alpha,\beta,\mu)$. Tang 等人使用梯度下降方法来求解目标函数. 以 μ 为例来介绍如何学习参数. 目标函数关于每个 μ_k 的梯度可以写为如下形式：

$$\frac{O'(\theta)}{\mu_k}=E^{\sharp}[h_k(Y_c^S)+h_k(Y_c^T)]-E^{\sharp}_{P_{uk}(Y_c|X_S,X_T,G_S,G_T)}[h_k(Y_c^S)+h_k(Y_c^T)] \quad (18)$$

其中：$E^{\sharp}[h_k(Y_c^S)+h_k(Y_c^T)]$ 是给定数据分布的因子函数 $h_k(Y_c^S)+h_k(Y_c^T)$ 的期望；$E^{\sharp}_{P_{uk}(Y_c|X_S,X_T,G_S,G_T)}[\cdot]$ 是在估计模型给出的分布 $P_{\mu_k}(Y_c|x_S,x_T,G_S,G_T)$ 下的期望. 类似的，可以求出参数 α_j 和 β_j 的梯度.

4.4.2 评 估

为了证明 TranFG 模型的通用性，Tang 等人[12]在 Epinions，Slashdot，Mobile，Coauthor 和 Enron 这 5 种不同类型的网络上模型进行了验证，并与其他的关系类型预测方法(SVM，CRF，PFG)进行对比. 表 5 和表 6 概括了不同方法在推断朋友或信任关系下性能对比. 可以看出，TranFG 明显地提高了关系分类的性能.

表 5 不同方法推断朋友关系(或信任关系)的性能对比[12]

数据集	方法	*Prec.*	*Rec.*	*F*1-score
Epinions(*S*)到 Slashdot(*T*)(40%)	SVM	0.7157	0.9733	0.8249
	CRF	0.8919	0.6710	0.7658
	PFG	0.9300	0.6436	0.7607
	TranFG	**0.9414**	0.9446	**0.9430**
Slashdot(*S*)到 Epinions(*T*)(40%)	SVM	0.9132	**0.9925**	0.9512
	CRF	0.8923	0.9911	0.9393
	PFG	**0.9954**	0.9787	**0.9870**
	TranFG	**0.9954**	0.9787	**0.9870**
Epinions(*S*)到 Mobile(*T*)(40%)	SVM	0.8983	0.5955	0.7162
	CRF	0.9455	0.5417	0.6887
	PFG	**1.0000**	0.5924	0.7440
	TranFG	0.8239	**0.8344**	**0.8291**
Slashdot(*S*)到 Mobile(*T*)(40%)	SVM	0.8983	0.5955	0.7162
	CRF	0.9455	0.5417	0.6887
	PFG	**1.0000**	0.5924	0.7440
	TranFG	0.7258	**0.8599**	**0.7872**

注：(*S*)表示源网络，(*T*)表示目标网络. 对于目标网络，使用标签数据的 40%进行训练，其余的用于测试.

表 6 不同方法推断有向关系的性能对比[12]

数据集	方法	*Prec.*	*Rec.*	*F*1-score
Coauthor(*S*)到 Enron(*T*)(40%)	SVM	0.9524	0.5556	0.7018
	CRF	0.9565	0.5366	0.6875
	PFG	**0.9730**	0.6545	0.7826
	TranFG	0.9556	**0.7818**	**0.8600**
Enron(*S*)到 Coauthor(*T*)(40%)	SVM	0.6910	0.3727	0.4842
	CRF	**1.0000**	0.3043	0.4666
	PFG	0.9916	0.4591	0.6277
	TPFG	0.5936	**0.7611**	0.6669
	TranFG	0.5936	0.5525	**0.7065**

注：(*S*)表示源网络，(*T*)表示目标网络. 对于目标网络，使用标签数据的 40%进行训练，其余的用于测试.

5 关系交互预测

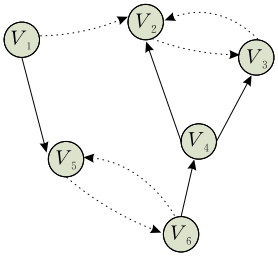
关系交互预测的目标是分析两个用户之间交互关系的形成原因及对社交网络演化的进一步影响.

在社会学中，个体之间的关系划分成两类：单向关系和双向关系[56]. 最常见的单向关系形式是明星和他们的粉丝之间的关系，而好友之间则是双向关系. 反映到社交网络中，用户之间也存在这样的单向和双向关系. 例如，当用户 A 关注用户 B 之后，两者之间就产生了单向联系. 用户 B 可以选择也关注 A(称为回粉)，从而形成一个双向关系. 有些社交网络中用户之间的联系是无向图，但也存在类似的问题. 那么，双向关系是如何从单向关系演化而来的？它对社交网络的进一步演化有什么影响？本节主要介绍关系交互预测的几个不同模型.

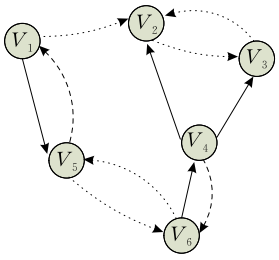
5.1 回粉预测

回粉预测与第 3 节中讨论的关系链接预测看起来相似，但有着明显的区别. 文献[57]表明，关系链接预测中所使用的特征在回粉预测中并非是最重要的. 从问题求解的定义域来说，关系链接预测的定义域是网络中所有尚不存在的边，而回粉预测的定义域是网络中所有的单向边[58].

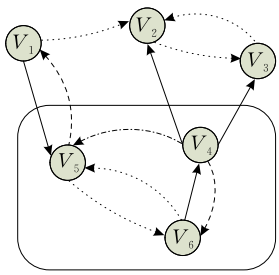
文献[13-14]探索了关系的相互性预测问题，他们提出一个学习框架 TriFG，使用图模型来解决关系的相互性预测，并在 Twitter 数据上对该模型的性能进行了评价. 对关系的相互性预测，给定从 1 到 t 时刻所有用户关注行为的历史日志，我们希望得到一个预测模型，从而判断当用户 B 在 t 时刻关注 A 之后，A 是否会在 $t+1$ 时刻关注 B. 图 7 给出了该问题的一个形象描述. 图 7(a)是一个关注关系形成的网络，其中实线箭头表示的边表示在 t 时刻新生成的关注关系，图 7(b)中的破折线箭头表示的边表示在 $t+1$ 时刻形成的回粉关系.



(a) t时刻形成的关注关系网络



(b) $(t+1)$时刻形成的回粉关系



(c) $(t+2)$时刻形成的闭合三元组

图 7　回粉问题示意图[14]

新关注和回粉这两种行为分别代表了之前所提到的单向关系和双向关系. 在关系的相互性预测中，重点研究回粉行为，如问题 5 所述.

Lou等人设计了一个三元组因子图模型(TriFG)，用于对关系的相互性进行预测. 根据问题 6 的描述，如果用户 v_i^s在 t 时刻关注了 v_i^u，则存在边 $e_i \in E$. 对回粉预测来说，任务是预测在 $t+1$ 时刻 v_i^u是否会回粉 v_i^s，即 y_i的值是 0 还是 1. 该模型根据观察结果为每条边定义了若干属性，记为 X_i. 则整个数据集可以描述为一个大小为 $|E| \times d$ 的属性矩阵，其中 d 是属性的维数. 例如，可以定义一个属性用来描述两个用户是否来自同一时区.

图 8 给出了 TriFG 模型的结构. 左边是输入的关注网络，其中包含 6 个用户在 t 时刻的关注状态. 实线箭头表示在 t 时刻之前就已经存在的关注，虚线箭头表示在 t 时刻新增的关注. 图 8 的右侧是根据输入网络所得到的因子图模型. 每个椭圆表示用户之间的关系，每个圆形是对应的隐变量 y_i. $y_i=1$ 时，表明 v_i^s回粉了 v_i^u，$y_i=0$ 时，则没有产生回粉. 如果 $y_i=?$，则表示未知，这正是模型所要预测的. $h(\cdot)$表示定义在三元组上的平衡因子，用于反映边之间的约束关系. $f(v_i^u, v_i^s, y_i)$表示与边 e_i相对应的属性因子.

TriFG 模型的推理和预测过程与 4.1 节相似，此处不再赘述，有兴趣的读者可以参考文献[14].

Lou 等人选取了 Twitter 网络上的 13 442 659 名用户，观察他们从 10/12/2010 到 12/23/2010 期间的关注变化情况，最终获得 56 893 234 条关注关系. 为了理解有哪些因素会影响双向关系的形成，他们从多个角度对数据进行分析，得到一些有趣的结果，这里给出部分结论，有助于我们了解目前社交网络的一些特性.

(1) 地理距离. 数据分析表明，在线社交网络正在变得越来越全球化，即使用户间隔了几个时区的距离，产生回粉行为的概率仍然是相近的. 而从另一个角度来看，相同时区内用户的双向关系数量是那些相距 3 个时区用户之间数量的 50 倍，社交网络仍然表现出很强的局部特性，即绝大多数的朋友仍然身处同一地域.

(2) 同质性. 同质性准则表明具有相似特征的用户更愿意彼此之间产生联系. 从链接同质性来看，随着两名用户之间共同邻居数量的增多，他们之间产生关注的可能性也迅速增大. 而对于社会地位同质性，如果将网络中的用户划分为知名用户和普通

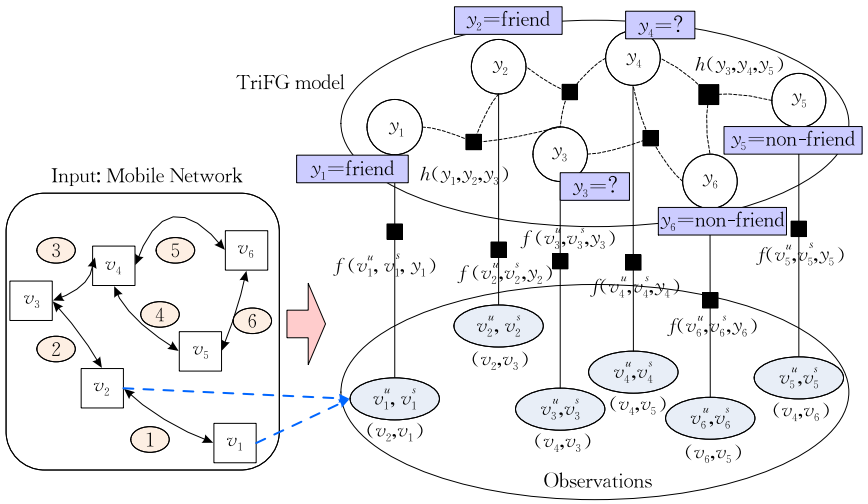


图 8　TriFG 模型的结构[14]

用户两类,则知名用户之间更乐于互相关注.

(3)结构平衡. 结构平衡理论是社会心理学中的一个基本理论. 对 3 个用户形成的一个三元组,结构平衡理论表明,或者这 3 个人彼此都是朋友,或者仅有其中的某两人之间是朋友关系. Lou 等人将双向关系和单向关系分别映射为朋友关系后分析网络是否满足该理论. 结果表明,在 Twitter 网络上,对于具有双向关系的用户,大部分用户所形成的三元组均满足结构平衡理论;而对仅具有单向关系的用户,网络结构是非常不平衡的. 比较常见的情况是两名普通用户都关注了某个明星,但这两名用户互不相识,这也非常符合日常生活中的情况.

为了评估 TriFG 模型的性能,Lou 等人将 TriFG 模型与支持向量机(SVM)、Logistic 回归(LRC)、条件随机场(CRF)等模型进行了对比. 同时,为了考察社会学理论对预测性能的影响,还与不考虑结构平衡时的条件随机场(CRF-balance)和不考虑社会地位同质性和结构平衡的 TriFG 模型(Weak TriFG)进行了对比. 结果如表 7 所示. 可以看出,TriFG 模型取得了比较好的预测结果,而且社会学理论也对提升预测性能有很大的帮助.

表 7 不同模型在回粉预测问题下的性能对比[14]

算法	*Prec.*	*Rec.*	*F*1
SVM	0.6908	0.6129	0.6495
LRC	0.6957	0.2581	0.3765
CRF-balance	0.9968	0.5161	0.6801
CRF	**1.0000**	0.6290	0.7723
wTriFG	0.9691	0.5483	0.7004
TriFG	**1.0000**	**0.8548**	**0.9217**

Gong 等人[58]将回粉预测看作一个离群值检测问题,并在 Google$^+$ 和 Flickr 数据集上对问题进行了研究. Google$^+$ 数据集抓取了 2011 年 Google$^+$ 刚刚推出时用户关系在 98 天内的连续变化信息,能够展现在线社交网络的早期演化过程. Flickr 数据集抓取于 2007 年,由于 Flickr 诞生于 2002 年,因此该数据集描述了在线社交网络的稳定发展阶段,两个数据集可以互补地反映社交网络在不同时期的一些性质.

他们首先比较了双向边和单向边的结构及演化过程,使用同配系数[59]分析了度同质性,结论是网络在双向关系下表现为高同配的,而在单向关系下表现为不同配的. 这一现象表明,双向关系一般发生在团体内部,而单向关系产生于团体之间,即:度相近的用户之间更容易产生双向关系,而在普通用户和明星用户之间则偏向于产生单向关系.

通过观察还发现,用户行为、节点属性和边的属性,这些都对双向边的形成有着重要的影响. 以用户的行为为例,一个直观的想法是,如果一个用户在过去一段时间比较乐于接受其他人的朋友请求,那么他今后会接受请求的可能性也很高;类似的,如果一个人发出去的朋友请求比较容易被其他人所通过,那么以后也会有这种趋势,数据分析的结果对上述想法给出了肯定的结论. 与此类似,结点属性也对双向边的形成有着显著的影响,且不同属性的影响力也存在差异. 如果给单向边一个关于年龄的描述,即单向边形成之后所经历的天数,统计结果显示,随着单向边年龄的增大,会产生回粉的可能性也在降低. 基于这些观察结果,可以定义一组特征用于回粉预测.

Gong 等人认为,之前的相关研究中,将关系的相互性预测看作是监督学习或半监督学习,将产生了回粉行为的边看作正样本,反之则看作负样本. 但由于网络是动态变化的,当前网络快照下的负样本有可能在未来变为正样本. 因此,他们将回粉预测看作是一个已知正样本的离群值检测问题,使用 SVM 方法进行检测,在 Google$^+$ 和 Flickr 数据集上均取得了较好的实验结果.

5.2 隐式关系交互预测

Cheng 等人[57]在 Twitter 数据集上对隐式关系交互预测问题进行了研究. 他们将这种隐式网络称为@-信息网络,其中@-信息可能是由于回复或转发而产生的. 由于 Twitter 用户所发布的信息数量符合长尾分布,因此他们选择 Twitter 上发布@-信息数量的活跃用户作为研究对象.

一个非常直观的想法是,能够反应用户 v 和 w 地位相似性和社交圈相似性的特征应当有助于这种关系的交互预测. 他们定义了一组特征,用于描述用户各自的属性以及结点对(v,w)的属性. 对每个单一特征,通过为其选择一个最佳的阈值,使得分类的准确性最高. 同时,他们在多种属性上使用决策树和 Logistic 回归模型进行预测. Cheng 等人发现,“出度/入度比”以及“两步路径比”这两个属性在隐关系的交互预测问题上发挥了重要的作用. 比如,只使用“出度入度比”这一个属性进行预测时,其准确性(82.0%)仅比使用所有属性的决策树模型的准确性(86.2%)略低.

Steurer 等人[60]使用 Logistic 回归模型在虚拟社会网络 My Second Life 上对用户之间的交互进行了类似研究. 他们将网络分解成社会网络和位置

网络两部分，其中社会网络根据用户之间的交互行为生成，位置网络依据用户是否曾在不同时间出现在相同的位置生成.在两个网络中，均分别根据拓扑结构和同质性得到两组特征.使用不同的特征组合，他们讨论了如何判断用户之间是否会进行交互，以及这种交互行为是单向还是双向的.

5.3　三元闭包预测

在回粉预测问题的基础上更进一步，如果 A 在关注 B 之后，又继续关注 B 所关注的某个用户 C，则(A,B,C)形成一个有向的三元闭包.仍然以图 7 为例，在图 7(c)中，当 v_4 回粉 v_6 之后又关注了 v_5，在 v_4，v_5 和 v_6 之间就形成了一个有向三元闭包.此时产生了一个新的问题，这种两两之间的关系最终是如何形成三元闭包的?

Lou 等人[14]在回粉预测的基础上进一步讨论了三元闭包的形成过程.与回粉预测中的分析类似，Lou 等人对影响三元闭包形成的因素进行分析.首先，仍然把用户分为明星用户和普通用户两类，Lou 等人对比了不同组合情况下形成三元闭包的可能性，得到了一组有趣的结论.比如，如果明星用户会回粉普通用户，那么他很有可能会去了解一下这名普通用户所关注的对象.而从链接同质性来看，观察结果表明，随着共同邻居数的增加，形成三元闭包的可能性也在逐步提高.基于这些观察结果，可以通过定义新的特征函数和平衡因子，将 TriFG 模型直接用于三元闭包的预测.

Huang 等人[61-62]针对问题 9 对三元闭包预测进行了更为系统的研究，综合考虑网络结构、人口学特征和社会角色等因素，提出了一个概率图模型 TriadFG 用来预测 3 名用户之间是否会形成闭包.

Huang 等人使用新浪 Weibo 数据集，在有向图上对三元闭包预测问题进行了研究.数据集中包括 1776950 名用户之间的 308489739 条关注关系，同时，获取了所有用户的描述信息，包括他们的昵称、性别、地理位置以及发出的微博内容.

同样地，他们首先对数据集进行观察，得到如下一些结论.(1)地址位置对形成三元闭包的影响不大；(2)女性用户比男性用户更乐于形成三元闭包；(3)明星用户对形成三元闭包没有什么影响，但是明星用户之间非常乐于形成三元闭包；(4)对结构洞用户来说，如果他原先是开放式三元组中间位置的节点，那么对形成三元闭包没什么影响，但他如果是两边的节点，则会非常愿意形成三元闭包，因为这样可以获取到更多的资源.基于这些观察结果，Huang 等人提出了一个三元组因子图模型 TriadFG，用于预测三元闭包的形成.有兴趣的读者可以参考文献[61-62]了解更多的细节.

为了评估 TriadFG 模型的性能，Huang 等人使用支持向量机(SVM)和 Logistic 回归进行了对比.结果如表 8 所示.可以看出，TriadFG 模型的预测性能与其他两种模型相比均有明显提高.

表 8　不同方法在三元闭包预测问题上的性能对比[61]

算法	*Prec.*	*Rec.*	*F*1	*Accu.*
SVM	0.7683	0.7420	0.7344	0.7422
Logistic	0.7657	0.7393	0.7316	0.7394
TriadFG	**0.8360**	**0.9084**	**0.8564**	**0.8444**

6　未来工作展望

(1)关系链接预测中的类别偏斜

在机器学习中，存在的一个难点是监督学习中的类别偏斜问题.由于模型估计的方差和类分布的不平衡，即使只有很少一部分负例的预测值与正例类似，模型最终还是会产生很多假的正例.而在基于相似度指标的监督式关系链接预测中也存在同样的问题.在社交网络中，可能的链接数是顶点数的二次方倍，然而网络中实际的链接(即图中的边)仅仅是数量很小的一部分.这就会导致大量的类别偏斜，使训练和推理都变得困难起来[39].如何从不均衡的数据集中进行学习是一项非常重要的研究，而文献[63]就解决此问题的各种技术进行了深入讨论.

(2)大数据与动态数据的处理

大数据是目前比较热门的研究方向，随着在线社交网络数据不断增长，从技术上，我们面临挑战，同时也拥有机遇.首先，社交网络的数据是动态的并且以流数据的形式产生的，因此，针对大规模动态网络，研究高效的模型和算法进行社交关系挖掘是非常有必要的.其次，人们认为并不是所有的大数据都会产生价值，在许多实际应用中，一部分数据很有可能以较高的性能解决了问题，若用大数据则会引发较大的计算代价问题.那么一个挑战就是，什么时候用大数据解决问题，什么时候只需用采样的数据就可以解决问题.再次，在线社交网络数据越来越大，有必要研究有效的算法以牺牲一定的精度保证计算的速度，那么挑战是，如何设计有效的算法并能从理论上保证求解的近似值.

(3)将社会理论与复杂网络理论融合到计算模型中，以提高预测精度

如何无缝地将社会理论、复杂网络理论以及社会心理学理论等融合到挖掘算法中,指导模型建立,以提高预测精度将是一个有意义的研究方向.目前已有一些算法和模型融入了社会理论,但通常都是针对某个特定问题进行求解的,难点在于如何建立一个通用的模型,使得社会理论、社会心理学等理论更容易融入.同时,社交网络是动态的,关系链接和关系类型是不断变化的,更重要的是如何根据社会理论等进行动态模式挖掘和关系预测.

(4)异质网络的社交关系挖掘

目前,在单一网络中研究关系链接预测、关系交互预测和关系类型预测的较多,但在真实社会中,没有人是只存在于单一的社交网络中的,人们可能有自己的工作圈也有自己的朋友圈,通常,同一个人在不同的社交网络中所在位置和所起作用是不同的.如何研究异质网络的社交关系挖掘面临挑战,主要有如下 3 个难点:① 无共同特征.两个网络之间可能没有共同特征,甚至没有交集.已有的迁移学习无法直接使用 ,需要建立两个或者多个网络之间的桥梁;② 网络规模不均衡.如 Facebook 和企业邮件网络,两个网络的规模可能相差千万倍,很难用一个网络做训练集另一个网络做测试集,直接进行预测;③ 统一的模型.目前的模型大部分都是针对具体问题设计的,如何设计一个统一的模型求解这类问题将是难点.

(5)应用

有许多实际应用是基于上述问题求解的结果,如可以根据关系链接预测进行信息推荐,根据关系类型预测研究人们在不同网络中的不同影响力.下一步可以进一步将理论研究成果应用于更多实际领域.

7 结束语

近年来,社交关系挖掘的研究已经取得了飞速的发展,而无所不在的社交网络和巨大的社交数据为我们学习用户之间的交互模式以理解不同网络下的动态机制提供了前所未有的机遇.社交关系(Social Ties),也称为人际关系,被定义为人与人之间信息传输的连接.从计算的观点看,社交关系挖掘的相关研究主要包括:关系链接预测、关系类型预测、关系交互预测等.

本文首先介绍了社交关系挖掘研究相关的一些基础知识,即相关问题的形式化描述,然后以关系链接预测、关系类型预测和关系交互预测三个社交关系挖掘的研究方向为重点,结合国内外相关的研究成果,对社交关系挖掘研究进行了详细的阐述,并对同方向不同方法进行对比说明.

社交关系挖掘研究是目前社交网络领域的一个热点,我们将社交关系挖掘目前的研究现状归纳总结并介绍给读者,希望更多感兴趣的研究同行更多地了解社交关系挖掘的研究工作,促进这一研究方向及其相关研究的发展.

参 考 文 献

[1] Fang Ping. Research of Community Detection's Algorithm Based on Friends Similarity from Online Social Networks [Ph. D. dissertation]. Huazhong University of Science & Technology, Wuhan, 2013(in Chinese)
(方平. 基于好友相似度的在线社交网络社区发现算法研究[博士学位论文]. 华中科技大学, 武汉, 2013)

[2] Xu Ke, Zhang Sai, Chen Hao, Li Hai-Tao. Measurement and analysis of online social networks. Chinese Journal of Computer, 2014, 37(1): 165-188(in Chinese)
(徐恪, 张赛, 陈昊, 李海涛. 在线社会网络的测量与分析. 计算机学报, 2014, 37(1): 165-188)

[3] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031

[4] Lazarsfeld P F, Merton R K. Friendship as a social process: A substantive and methodological analysis. Freedom and Control in Modern Society, 1954, 18(1): 18-66

[5] McPherson M, Smith-Lovin L, Cook J M. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 2001, 27(1): 415-444

[6] Backstrom L, Leskovec J. Supervised random walks: Predicting and recommending links in social networks//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 635-644

[7] Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks//Proceedings of the 19th International Conference on World Wide Web. North Carolina, USA, 2010: 641-650

[8] Diehl C P, Namata G, Getoor L. Relationship identification for social network discovery//Proceedings of the 22nd Conference on Artificial Intelligence. Vancouver, Canada, 2007: 546-552

[9] Menon A K, Elkan C. A log-linear model with latent features for dyadic prediction//Proceedings of the 10th International Conference on Data Mining (ICDM). Sydney, Australia, 2010: 364-373

[10] Wang C, Han J, Jia Y, et al. Mining advisor-advisee relationships from research publication networks//Proceedings

of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 203-212

[11] Pentland A, Eagle N, Lazer D. Inferring social network structure using mobile phone data. Proceedings of the National Academy of Sciences, 2009, 106(36): 15274-15278

[12] Tang J, Lou T, Kleinberg J. Inferring social ties across heterogenous networks//Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, USA, 2012: 743-752

[13] Hopcroft J, Lou T, Tang J. Who will follow you back?: Reciprocal relationship prediction//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Glasgow, UK, 2011: 1137-1146

[14] Lou T, Tang J, Hopcroft J, et al. Learning to predict reciprocity and triadic closure in social networks. ACM Transactions on Knowledge Discovery from Data, 2013, 7(2): 5

[15] Tang J, Zhang J, Yao L, et al. Arnetminer: Extraction and mining of academic social networks//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 990-998

[16] Kunegis J, Lommatzsch A. Learning spectral graph transformations for link prediction//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 561-568

[17] Kashima H, Abe N. A parameterized probabilistic model of network evolution for supervised link prediction//Proceedings of the 6th International Conference on Data Mining (ICDM'06). Hong Kong, China, 2006: 340-349

[18] Schifanella R, Barrat A, Cattuto C, et al. Folks in folksonomies: Social link prediction from shared metadata//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 271-280

[19] Richard E, Baskiotis N, Evgeniou T, et al. Link discovery using graph feature tracking//Proceedings of the Conference on Neural Information Processing Systems 2010. Vancouver, Canada, 2010: 1966-1974

[20] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks. Nature, 2008, 453(7191): 98-101

[21] Lee C, Nick B, Brandes U, et al. Link prediction with social vector clocks//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 784-792

[22] Lichtenwalter R N, Chawla N V. Vertex collocation profiles: Subgraph counting for link analysis and prediction//Proceedings of the 21st International Conference on World Wide Web. Lyon, France, 2012: 1019-1028

[23] De A, Ganguly N, Chakrabarti S. Discriminative link prediction using local links, node features and community structure//Proceedings of the IEEE 13th International Conference on Data Mining (ICDM). Dallas, USA, 2013: 1009-1018

[24] Wang D, Pedreschi D, Song C, et al. Human mobility, social ties, and link prediction//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 1100-1108

[25] Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 1046-1054

[26] Oyama S, Hayashi K, Kashima H. Cross-temporal link prediction//Proceedings of the 11th International Conference on Data Mining (ICDM). Vancouver, Canada, 2011: 1188-1193

[27] Zhu J. Max-margin nonparametric latent feature models for link prediction. arXiv preprint arXiv 2012: 1206.4659

[28] Papadopoulos F, Kitsak M, Serrano M Á, et al. Popularity versus similarity in growing networks. Nature, 2012, 489(7417): 537-540

[29] Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 243-252

[30] Kunegis J, Fliege J. Predicting directed links using nondiagonal matrix decompositions//Proceedings of the 12th International Conference on Data Mining (ICDM). Brussels, Belgium, 2012: 948-953

[31] Wang C, Satuluri V, Parthasarathy S. Local probabilistic models for link prediction//Proceedings of the 7th IEEE International Conference on Data Mining (ICDM). Omaha, USA, 2007: 322-331

[32] Guimerà R, Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences, 2009, 106(52): 22073-22078

[33] Leroy V, Cambazoglu B B, Bonchi F. Cold start link prediction//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 393-402

[34] Yin Z, Gupta M, Weninger T, et al. LINKREC: A unified framework for link recommendation with user attributes and graph structure//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 1211-1212

[35] Mathur S, Poole M S, Pena-Mora F, et al. Detecting interaction links in a collaborating group using manually annotated data. Social Networks, 2012, 34(4): 515-526

[36] Zhu Y, Yan X, Getoor L, et al. Scalable text and link analysis with mixed-topic link models//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 473-481

[37] Zhang J, Wang C, Yu P S, et al. Learning latent friendship propagation networks with interest awareness for link prediction//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dubin, Ireland, 2013: 63-72

[38] Wu S, Sun J, Tang J. Patent partner recommendation in enterprise social networks//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome, Italy, 2013: 43-52

[39] Heaukulani C, Ghahramani Z. Dynamic probabilistic models for latent feature propagation in social networks//Proceedings of the 30th International Conference on Machine Learning (ICML-13). Atlanta, USA, 2013: 275-283

[40] Lü Lin-Yuan, Tao Zhou. Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170

[41] Lü Lin-Yuan, Zhou Tao. Link Prediction . Beijing: Higher Education Press, 2013(in Chinese)
(吕琳媛, 周涛. 链接预测. 北京: 高等教育出版社, 2013)

[42] Zhu X, Tian H, Cai S, et al. Predicting missing links via significant paths. EPL(Europhysics Letters), 2014, 106(1): 18008

[43] Zhao J, Miao L, Yang J, et al. Prediction of links and weights in networks by reliable routes. Scientific Reports 5: 12261, 2015

[44] Al Hasan M, Zaki M J. A survey of link prediction in social networks. Social Network Data Analytics. USA: Springer, 2011: 243-275

[45] Wang Yu, Gao Lin. Social circle-based algorithm for friend recommendation in online social networks. Chinese Journal of Computer, 2014, 37(4): 801-808(in Chinese)
(王玙, 高琳. 基于社交圈的在线社交网络朋友推荐算法. 计算机学报, 2014, 37(4): 801-808)

[46] Zhang J, Philip S Y. Link Prediction Across Heterogeneous Social Networks: A Survey [Ph. D. dissertation]. University of Illinois, Chicago, USA, 2014

[47] Liu F, Xia S T. Link prediction in aligned heterogeneous networks//Cao T, Lim E-P, Zhou Z-H, et al, eds. Advances in Knowledge Discovery and Data Mining. Berlin, Germany: Springer International Publishing, 2015: 33-44

[48] Dong Y, Tang J, Wu S, et al. Link prediction and recommendation across heterogeneous social networks//Proceedings of the 12th International Conference on Data Mining (ICDM). Brussels, Belgium, 2012: 181-190

[49] Tang W, Zhuang H, Tang J. Learning to infer social ties in large networks//Gunopulos D, Hofmann T, Malerba D, et al, eds. Machine Learning and Knowledge Discovery in Databases. Berlin Heidelberg, Germany: Springer, 2011: 381-397

[50] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data//Proceedings of the 18th International Conference on Machine Learning. Williamstown, USA, 2001: 282-289

[51] Murphy K P, Weiss Y, Jordan M I. Loopy belief propagation for approximate inference: An empirical study//Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Stockholm, Sweden, 1999: 467-475

[52] Zhuang H, Tang J, Tang W, et al. Actively learning to infer social ties. Data Mining and Knowledge Discovery, 2012, 25(2): 270-297

[53] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 137-146

[54] Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. Honolulu, USA, 2008: 1070-1079

[55] Huang Li-Wei, Li Yi-De, Ma Yu-Tao, et al. A meta path-based link prediction model for heterogeneous information networks. Chinese Journal of Computers, 2014, 37(4): 848-858(in Chinese)
(黄立威, 李德毅, 马于涛等. 一种基于元路径的异质信息网络链接预测模型. 计算机学报, 2014, 37(4): 848-858)

[56] Horton D, Richard Wohl R. Mass communication and para-social interaction: Observations on intimacy at a distance. Psychiatry, 1956, 19(3): 215-229

[57] Cheng J, Romero D M, Meeder B, et al. Predicting reciprocity in social networks//Proceedings of the 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE 3rd International Conference on Social Computing (SocialCom). Boston, USA, 2011: 49-56

[58] Gong N Z, Xu W. Reciprocal versus parasocial relationships in online social networks. Social Network Analysis and Mining, 2014, 4(1): 1-14

[59] Newman M E J. Assortative mixing in networks. Physical Review Letters, 2002, 89(20): 208701

[60] Steurer M, Trattner C. Predicting interactions in online social networks: An experiment in second life//Proceedings of the 4th International Workshop on Modeling Social Media. Paris, France, 2013: 5

[61] Huang H, Tang J, Wu S, et al. Mining triadic closure patterns in social networks//Proceedings of the 23rd International World Wide Web Conference. Seoul, Korea, 2014: 499-504

[62] Huang H, Tang J, Liu L, et al. Triadic closure pattern analysis and prediction in social networks. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(12): 3374-3389

[63] Weiss G M. Mining with rarity: A unifying framework. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 7-19

ZHAO Shu, born in 1979, Ph. D., professor. Her research interests include machine learning, intelligent computing.

LIU Xiao-Man, born in 1989, M. S., assistant experimentalist. Her research interests include machine learning, social network.

DUAN Zhen, born in 1976, Ph. D., lecturer. His research interests include machine learning, social network.

ZHANG Yan-Ping, born in 1962, Ph. D., professor. Her research interests include machine learning, granular computation.

TANG Jie, born in 1977, Ph. D., researcher. His research interests include data mining, machine learning, social network.

Background

Social network mining aims to provide a comprehensive understanding of global and local patterns, mechanism of the network formation, and dynamics of user behaviors. Social network analysis and mining is an inherently interdisciplinary academic field which emerged from sociology, psychology, statistics, and graph theory. However, due to the lack of efficiently computational models and the nonavailability of large-scale social networking data, traditional research on social network has mainly focused on qualitative study in small-scale network. More recently, with the emergence and rapid proliferation of online social applications and media, online social networks already become a bridge to connect our physical daily life with the web-based information space. The connection produces huge volume of data including not only the spreading information, but also user behavior.

Social Ties are a bridge of connection and interaction between people, and also the foundation of information diffusion in social networks. Mining social ties is an important problem in social network analysis.

This paper provides a summary of previous works and researches. From the computational perspective, the authors summary three aspects on social ties analysis: predicting missing links, inferring social ties, and predicting reciprocity. Related concepts are introduced first and then the important algorithms are presented. Finally, the future research directions are discussed.

This work is supported by the National High Technology Research and Development Program (863 Program) of China under Grant (2015AA124102), the National Social Science Foundation of China (No. 13&ZD190), the National Natural Science Foundation of China under Grants (61402006, 61175046), the Provincial Natural Science Research Program of Higher Education Institutions of Anhui Province under Grant (KJ2013A016), the Provincial Natural Science Foundation of Anhui Province under Grant (1508085MF113), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (Forty-ninth batch), and the Recruitment Project of Anhui University for Academic and Technology Leader.