社会网络角色识别方法综述

张树森 梁 循 齐金山
(中国人民大学信息学院 北京 100872)

摘 要 在社会网络研究中,角色识别是一个十分重要的研究问题,它对分析和理解社会网络、预测用户行为、研究用户之间的关系和交互过程具有重要意义.相对其他人或事物来说,社会网络中每个人都扮演着所在环境下的一个角色.社会角色可根据用户之间的交互水平来定义,这些角色可看作是对用户位置、行为或虚拟身份的刻画,并且随着时间的变化这些角色也在不断的改变和演化.当前,社会网络角色识别研究更多的是集中在新出现的社交网络平台上,如 Facebook、Twitter、微博等,也正是由于这些社交媒体网络的快速增长以及可被获得,使我们有了新的机会和条件来定义和识别社会角色.文中主要对近年来关于社会网络中角色识别的方法和研究现状进行了总结,并提出自己的想法和意见.文中首先阐述了社会网络中网络、角色等基本概念,提出了社会网络角色识别问题并给出社会网络角色识别研究中关键挑战问题;然后根据角色是否提前定义,将社会网络角色分为非明确角色和明确角色,并总体概括了当前这两种角色识别的主要方法和研究现状;最后对社会网络角色识别中的难点和未来研究方向进行了分析和展望.社会网络角色的识别是一个复杂的问题,不是单靠某一种方法能解决的,而是需要用"组合拳"方式来解决,这就要求我们综合考虑各种因素进行优化组合,识别出最终的社会角色.

关键词 社会网络;角色识别;网络分析;非明确角色;明确角色;方法综述;社会计算
中图法分类号 TP399 **DOI 号** 10.11897/SP.J.1016.2017.00649

A Review on Role Identification Methods in Social Networks

ZHANG Shu-Sen LIANG Xun QI Jin-Shan
(*School of Information, Renmin University of China, Beijing* 100872)

Abstract It is an important research issue in social networks to identify roles, which can help analyze and understand social networks, predict user's behavior, and study the relationship and interaction between users. In social networks, every person plays a role in one's environment relative to other people or things. The interaction level among users defines the appearance of several social roles which can be characterized as positions, behaviors, or virtual identities, and they keep changing and evolving over time. Currently, the research of role identification in social networks is more focused on the emerging social networking platform, such as Facebook, Twitter, micro blogging and so on. And, it is due to the rapid growth of these social media networks and these media data can be obtained, so that we have new opportunities and conditions to define and identify social roles. We summarize the methods and current research status of the role identification in social networks, and put forward some ideas and opinions. In this article, we firstly give the basic concept of networks and roles in social networks, and put forward the question of role identification in social networks and their main challenges. Then we introduce related research and state-of-the-art approaches from two different angles. The first one is the non-explicit roles

收稿日期:2016-01-26;在线出版日期:2016-08-11.本课题得到国家自然科学基金(71531012,71271211)、北京市自然科学基金(4172032)、中国人民大学自然科学基金(10XNI029)、北京高等学校青年英才计划(21147514040)资助.**张树森**,男,1988 年生,博士研究生,主要研究方向为数据挖掘、社会计算. E-mail: zss2446@163.com. **梁 循**(通信作者),男,1965 年生,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为互联网信息分析、数据挖掘、商务智能、社会计算. E-mail: xliang@ruc.edu.cn. **齐金山**,男,1977 年生,博士研究生,讲师,主要研究方向为数据挖掘、社会计算.

and the second is the explicit roles. At last some future research topics are given. Identifying roles in social networks is a complex problem, not rely on a particular method can be solved, but need to use the "combined fist" approach to resolve, which requires us to consider the optimal combination of various factors, identifying the ultimate social roles.

Keywords social networks; roles identification; network analysis; non-explicit roles; explicit roles; methods review; social computing

1 引 言

随着互联网技术的不断创新和发展,以及网络应用的多样化,网络用户已从简单的信息"消费者"逐步转变成为信息的"制造者"[1]. 而 Facebook、Twitter、微博以及其他在线社会网络应用的出现,使人类真实世界在网络虚拟世界中得以延伸. 现实社会中,人们相互之间的通信、交往、分享以及各种社交活动,可以在另一个虚拟的社会网络中进行,扩展了人们社会活动的空间. 尤其是国内的微博,这种基于 Web3.0 平台兴起的一类开放互联网社交服务[2],在技术与传播方式上对博客、BBS 等新媒体进行了"颠覆性创新",一出现就非常受欢迎[3]. 实际上,在线社会网络已成为信息传播的重要途径,不仅成为广大网络用户思想交流、感情沟通、信息获取的媒介,也成为政府、公司、团体等发布信息、商品营销、扩大宣传的载体.

当前,由于这种虚拟的社会网络变得越来越复杂,规模越来越大,对它们进行分析和想象也变得越来越困难. 为了分析这样的社会网络,目前常用的方法就是对网络中重要的特征进行提取,也就是将一个大的网络转换为一个较小的网络,而后者是对前者有效的总结且保持原有网络的重要特征[4]. 在社会网络中,这种方法可以通过两种方式来实现,一是识别用户群体并将它们之间的关系表示出来,也就是我们常说的社区发现. 另一个就是识别网络中具有相同结构、位置、作用或扮演相同角色的节点,构建一个小型网络,用角色间的关系代替节点间的关系,即角色识别. 由于近年来,像微博、论坛等社交网络媒体的快速增长以及可被获得,使我们有了新的机会和条件来定义和识别社会角色. 社会角色可根据用户之间的交互水平来定义,这些角色可看作是对用户位置、行为或虚拟身份的刻画. 社会网络角色的识别则主要根据社会网络的结构及其用户交互的内容来实现. 此外,随着时间的变化,这些角色也在不断的改变和演化.

对社会网络中的角色进行识别具有重要意义:(1)识别出社会网络中的具体角色具有重要的现实意义. 如识别出技术论坛中的专家角色,可以使我们获得最佳的问题答案;识别出社交网络中具有较大影响力的角色,对企业产品信息的扩散和良好口碑的建立具有重要的支持作用;识别出社交网络中的意见领袖角色,对于舆情的正确导向起着十分关键的引导作用等. 通过近几年社会网络尤其是在线社会网络的实际情况来看,一些如意见领袖、重要用户以及网络水军等特定类型网络用户的行为,对网络中信息的产生和传播有着非常重要的影响和作用,已成为能够左右网络舆论发展和走向的重要力量. 尤其是社交网络中的意见领袖角色,在社会舆论的形成过程中起着重要作用. 在其引导下,一些局部意见将演化为社会舆论,直接影响到现实社会[5]. 识别出社会网络角色,不仅对我们充分利用网络,而且对社会网络的监控和管理都具有重要的现实意义;(2)识别社会网络中的角色具有重要的研究意义. 识别和分析用户角色,为我们研究网络的动态特性提供了一个有价值的视角[6],对社会网络的深入研究具有重要的帮助. 如用户的行为在某种程度上代表了一种角色,当社会网络中一个新的社会关系出现,我们可根据对其中角色的识别,更好的理解他们之间的交互情况;通过分析社会网络中的角色还可以深入理解网络拓扑结构,了解网络时态演变对用户角色形成和变化产生的作用,也将使我们更容易了解信息在网络中的传播过程等. 此外,社区中角色的变化与社区的发展变化有直接关系,角色变化可引发社区变化[6]. 因此,社会网络角色的识别,对于我们分析和理解社会网络,预测用户行为,研究用户相互之间的关系和交互过程等都具有重要帮助和意义.

为了识别社会网络中的不同角色,研究人员提出了许多方法. 本文对近年来关于社会网络中角色

识别的方法和研究现状进行了总结，并提出自己的想法和意见. 本文组织结构为：首先对社会网络中角色识别的基本概念和社会网络角色识别问题的提出进行阐述，并提出当前所面临关键挑战. 然后根据角色是否提前定义，将社会网络角色分为非明确角色和明确角色两类，并对这两类角色的识别方法进行相关国内外研究现状的阐述、分析和总结. 最后得出结论并给出研究的难点和未来研究方向.

2　基本概念

2.1　网络和社会网络

2.1.1　网　络

简单的讲，网络是一组个体，我们称这些个体为顶点或节点，它们之间的连接称为边，如图 1 所示. 其实，在我们世界中，很多系统或组织都是以网络的形式存在的，像因特网，朋友或其他个体间的社会关系网络，机构网络以及自然网络，代谢网络，食物链，论文间的引用网络和通信网络等等[7]，如图 2 所示，存在的两种网络（互联网中获得）.

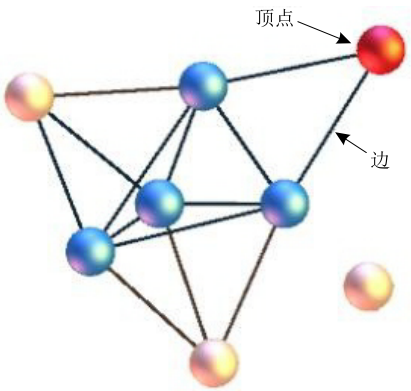
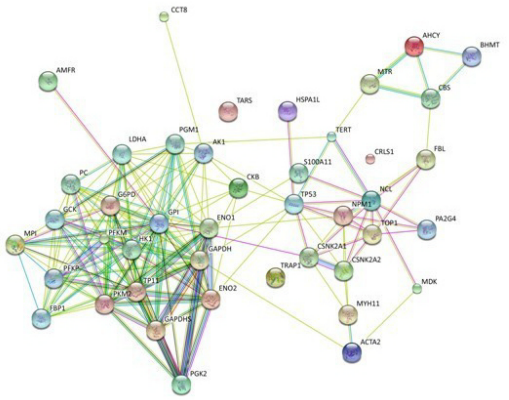


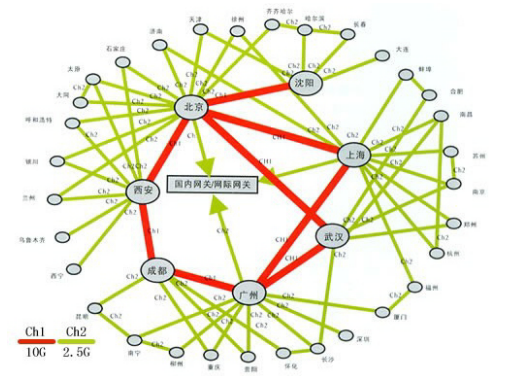
图 1　一个小型网络

网络的研究起源于数学中的图论. 1735 年欧拉提出著名的格尼斯堡桥问题解决方案，同时开创了数学新分支——图论. 到 20 世纪，图论已经成为一门学问. 在科学界，网络被广泛研究，如在社会学中，典型的网络研究是调查问卷，要求受访者描述与他人交往的情况，根据这些问卷，构建一个网络. 在这个网络中，用顶点来表示个体，用边来表示个体间的交互关系[7].

当前，对于网络的研究方法出现了许多新的变化，不再是通过研究和分析一些小型网络的属性来解释大规模网络的属性. 这些研究方法的改变，得益于我们能够通过计算机分析和计算大规模的网络数



（a）蛋白质代谢网络



（b）铁通网络

图 2　两种网络

据，而这在以前是无法做到的. 以往所要求研究的小型网络中的问题，在当前大规模网络中似乎不再有用或有意义. 例如，在一个小型网络中，移除一个节点后，剩下的节点中哪一个节点会成为该网络连接中最为关键的节点问题. 在数以百万计节点的网络中，这一问题就显得不再有意义了. 因为在这样的网络中，没有哪个节点移除后会对其他所有节点产生重大影响. 另一个使我们研究网络方法改变的原因是人们逐渐认识到网络的重要性. 以前几十、数百顶点的网络，是一个对真实世界的直接描述. 通过对其研究分析，可以回答许多关于网络结构的问题. 而当前，在面对数以百万计的节点时，一个人不可能直接将网络中数以百万计的节点直接简单的描述出来，直接用眼来分析的话根本就不现实. 由于网络的重要性，当前分析网络的方法就是在大型网络中试图最大限度地发现我们以前用眼能够发现的东西（如网络结构、特征等）[7].

2.1.2　社会网络

社会网络是描述人类社会的理论之一，而社会网络理论于 20 世纪 50、60 年代就已经开始出现. 社会网络这一概念的兴起，源于其对社会互动的恰当

描述,由德国社会学家齐美尔(Georg Simmelcon)从社会学角度开始提出[8].社会网络主要关注人们之间的互动和联系,并且这种互动会影响人们的社会行为.

社会网络是以各种连接或相互作用的模式而存在的一组人或群体[9].像人与人之间的朋友关系网络,家族、亲人之间的血缘关系网络,同学关系网络,同事关系网络,在线社交关系网络等,都属于社会网络.由于社会网络并不是一个关于个体的简单集合,也不是个体间相互连接关系的总和,而是包含了个体和个体间关系的网络.因此,社会网络一般会定义为:一个由一组代表社会成员(如人、组织)的节点和表示节点间关系的边或连线构成的社会结构,社会网络结构形式一般如图 3 所示.

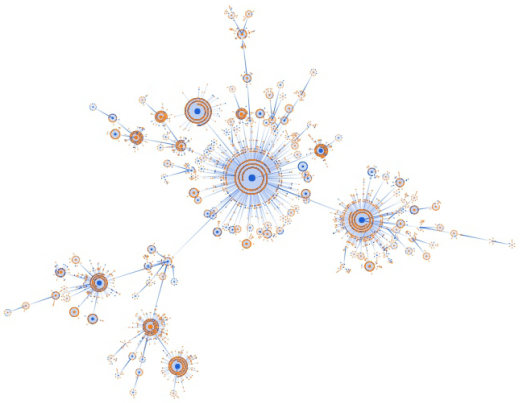


图 3 社会网络

事实上,人们很早就开始对社会网络进行研究,一直希望能够发现社会关系的内在机制、演化规律以及信息在社会网络中的传播模式等内容.在社会网络的研究中,阿纳托拉波波特(Anatol Rapoport)的数学模型[10]强调了社会网络中度分布的重要性,他也是最早提出这一理论的人之一.1969 年,米尔格兰姆(Milgram)做了一个比较著名的"小世界"(Small-world)实验,用以研究真实社会中,通过朋友关系网络传递信息的能力[11].虽然这个实验无法在现实网络中重现,但该实验使我们大体了解了社会网络的结构.在这个实验中,让随机选择的 296 名实验人员传递一封信件给位于波士顿郊区的一个陌生人,并且要求这些实验人员只能通过熟悉的朋友来传递.结果共收到了 64 封信件,且这些成功送达的路径平均长度约为 6.由此得出,我们的世界是"六度分离"的,任意两个人之间都存在较短的平均路径,社会网络中的个体能够根据局域信息高效地找到这些较短的路径.这也是"六度分离"概念的由来.虽然 Milgram 并没有书面提出,而几十年后由瓜雷[12](Guare)提出.1992 年,美国社会学家罗纳德·伯特(Ronald Burt)提出了著名的结构洞理论.社会网络中一些个体能够和其他个体之间发生直接的联系,而有些个体却不能产生直接的联系或者说是关系间断的情况,从整体来看,网络结构中出现了洞穴.如图 4 中,该网络中包含了 4 个节点,其中节点 2、3、4 之间没有直接的联系,都与节点 1 之间有联系.从该图中我们可以看出,节点 1 处于中心位置,其他三者需要通过该节点才能发生联系.所以,节点 1 具有 3 个结构洞:2～3、2～4 和 3～4.此外,社会网络中的关系可分为两种,即强联系和弱联系.强联系指的是二者在现实世界中自身就有关系,像同学、同事、亲友等.弱关系指的是通过应用社会网络使二者成为朋友,在现实生活中并不存在.在社会网路中这两种关系同时存在,共同增强社会网络的作用效果.美国斯坦福大学人文与科学学院社会学家马克·格兰诺维特(Mark Granovetter)对找工作的过程进行研究,得出这样的结论:能够提供更加有效的工作信息的人,往往不是和我们关系很熟或者很好的人(如 1～2 之间),相反,却往往是弱关系的人(如 2～4 之间).这主要是因为"弱连接"有着极快的以及可能具有低成本和高效能的传播效率,一个人的社会网络异质性越高,那么这个人通过弱关系获得各种社会资源的可能性就越大,也就是说,弱关系使我们与外界交流时得到跟多新的信息.

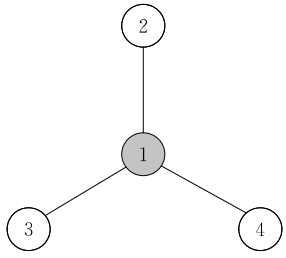


图 4 结构洞说明

由于早期社会网络研究条件的限制,社会网络数据的收集通常比较困难,且采集的样本规模也不够大,所得结论其实并不准确.因此,传统的社会网络研究往往会有准确性低,个人主观性强以及样本规模小三方面的问题[7].随着计算机及信息技术的发展,使大规模社会关系数据的获取和分析成为可能,社会网络研究成果也开始不断地涌现出来.

2.2 角色和社会角色

2.2.1 角 色

角色一词源于戏剧表演,是指演员通过道具、化

妆、场景等形式刻画个体在社会舞台上的身份、个性及其行为，在社会学和心理学中被广泛应用. 关于角色的定义有很多，许多社会学家和心理学家都对其进行过研究，但一直没有一个统一的定义.

20 世纪 20 年代，美国社会学芝加哥学派开始将角色这一概念应用到社会结构的研究中. 从此以后，角色的概念在社会学中被广泛应用[13-14]. 在社会学中，角色是被假定为真实的，客观的，在现实社会中有意义的特征[15]. 其中，最早定义社会角色概念的是人类学家拉尔夫·林顿(Ralph Linton)，他将社会角色定义为人们对于某种位置上人的行为的期望或要求[16]. 在社会网络中，有的将角色定义为用来帮助人们达到目标的一种资源，是用来建立社会结构的工具[17]. 例如，在有许多孩子的家庭中，父亲会扮演老师的角色，这样才能创造一个良好的家庭秩序以及惩罚那些犯了错误的孩子. 还有的将角色视为一组描述个体行为以及个体在社会环境中交互的特征[18]. 在社交媒体中，也有许多关于角色的研究，并试图去定义用户具体的角色. 例如，根据社会网络分析(SNA)中的特征度量，定义社交媒体(如博客)中的发起者和追随者角色，根据用户在社交网络中的影响值排名定义意见领袖角色等. 角色也可看作是简化行为模式、识别不同类型用户以及了解人们行为的工具. 识别用户角色，进而更有效的监督特定角色的用户. 此外，在未注解的民间故事文本中，如果识别出其中人物角色[19]，可以使我们更好的了解故事的情节；在网络搜索中，如果识别出内容作者所扮演的角色，则可能使我们获得更好的查询结果[20].

总之，角色的研究和分析一直处于社会学研究的中心[15]，在其他学科或领域中的应用也变得越来越有价值和意义. 例如，在社会网络中，从一个社会群体中寻找有影响力的领导者角色，在社交平台上寻找意见领袖角色，以及在各种论坛中发现专家角色. 在政治选举中，用户角色从支持者变为中性论者，再到具有较高的消极情感的持不同政见者的角色改变，可能是对一个政治家的警告标志[21]等.

2.2.2　社会角色

总体来讲，社会角色是研究者用来区分研究对象，以及人们以自己方式所表现出来的行为并对其进行的解释，是为我们研究当下社交网络结构所开启的一个重要窗口[15,19,22]. 通常情况下，社会角色可根据研究对象的行为特点和网络属性来定义.

简单讲的话，社会角色其实是用户行为和关系的划分. 在社会生活中，社会研究者主要关注两种基本的度量，即结构和文化. 社会结构是指在人群中关系和资源的分布，这包括像在线社区朋友关系的划分，一个城市中年龄的分布情况等. 社会结构即可以作为社会活动的机会，也可作为社会活动的制约因素. 其实，我们定义社会角色都来自一个简单的行为共性的结构基础. 如父亲的角色都来自于协调，保护和照顾后代的行为共性. 特殊的行为可能涉及多种文化，但父亲角色填补了某些基本的社会需求，是跨越不同文化的认识[19]. 同样，许多社会角色，尤其是那些新出现的角色，都有自己独特的社会结构基础，即使他们还没有发展到相同的认可水平和跨越不同的文化背景. 需要注意的是，在社会角色的许多概念中，文化因素是一个关键的部分. Callero[15]强调了文化重要性，特别是在有一定的社会背景或环境中识别既定的社会角色.

在社会网络中，社会位置是社会角色的基础，社会位置指的是在社会交往中个体所处的相应位置，不过社会位置和社会角色之间是有区别的. 社会位置往往侧重于社交关系网络中个体的结构位置以及与此位置相对应的各种行为、职责和权利. 而社会角色则更多是与所处环境有关. 也就是说，在不同环境中，处于相同位置的个体其角色可能是完全不一样的[14].

在实际社会生活中，一个个体在其所处的血缘关系网、地缘关系网、朋友关系网等众多社交关系网中的位置，决定了该个体所扮演的角色. 社会角色是以成员彼此间的交互形式存在的，也就是说社会角色是在用户互动中出现和演化的. 社会角色也可认为是在一个社会结构中用户所处位置的行为期望[16,22]. 用户可能扮演领导者、参与者、评论者等各种角色. 在一个社区中，当每个人都能扮演一个角色时，此时的社区是最强大的，而社区成员通过扮演各类角色不断参与也会增加成员自身价值[23]. 社会角色在价值网络中充当变化的资源，因为社会角色可以引导社会规范，建立社会地位. 由于用户可能改变他们的地位或立场，因此，他们在社会网络中的角色可能随着时间的变化而变化[24].

2.3　社会网络角色识别问题的提出

本文所说的社会网络识别问题指的是社会网络中社会角色的识别，尤其是当前蓬勃发展的在线社会网络中社会角色的识别问题. 社会网络角色其实

指的就是社会网络中的社会角色.在线社会网络简单的讲是 www 上人与人之间通过在线社会软件建立起来的人际关系网络.这些在线社会软件多运行在 Web 2.0 互联网模式下.当前,在线社会网络可以说十分常见,如 Face book、LinkedIn、Skype、Twitter、email、BBS、博客、在线社区、电子商务、医疗诊断、电子投票、在线社会媒体等等.

事实上,由于通信技术的快速发展和互联网的大量普及,越来越多的人们开始通过网络媒体平台进行交流和互动.与传统的社交活动形式不同,在线网络为我们搭建了一个虚拟的社交平台,相互之间无需见面就可直接进行交流,逐渐形成了一种新的社会网络,即在线社会网络.而这种新的社会网络的主要特点就是数据规模大,网络信息处于动态变化之中,而且十分复杂,没有多少规律可循.对这种社会网络进行分析和研究,不管我们借助数学方法还是网络拓扑分析的方法,都难以达到理想的效果.而以传统的方法(如图论、度的计算等)来分析这种社会网络也明显不能满足需要,不管是在运行时间方面还是结果准确性方面.但这也并不意味着我们对此就束手无策,通过角色替换以间接的方式对其进行研究分析(即构建一个小型网络,以角色间的关系代替节点间的关系),以及通过多种方法找出其中符合某些定义的社会角色加以应用,从目前来看还是一种比较有效可行的社会网络分析方法和实践应用.

由此,本文提出社会网络角色识别的问题,综合以往社会网络中关于社会角色识别的方法和技术,对其进行总结和概括,并提出自己的意见、想法和未来可能研究的方向.从而,使读者能够更好地了解当前社会网络中(主要是在线社会网络)角色识别的研究现状和技术概要,为社会网络中角色识别的进一步研究和发展提供便利和帮助.

2.4 面临挑战

根据以上所述,社会网络角色识别是一个复杂的研究问题,不是单靠某一种方法能解决的.当前,社会网络角色识别的主要方法为:(1)根据社会网络结构或用户所处社会位置的分析识别用户角色,即社会网络分析方法;(2)根据用户行为规律的分析来识别用户角色,即数学分析方法;(3)根据用户交互的内容的分析,识别用户角色,即内容分析方法;(4)机器学习方法.4 种方法以及这四种方法间的组合.当然,还有根据结构相似性、规则结构等价性、结构特征的分类、聚类、概率图模型等具体识别的方法,这些方法严格的讲也可归属到上述的方法中.这 4 种方法可以说是当前社会网络角色识别研究领域中,研究者一般采用的方法.但是,这些研究方法同时面临着一些挑战:社会网络复杂性分析问题、海量数据问题、评价问题等.

(1)社会网络复杂性分析问题

当前,由于网上人类行为的复杂性以及在线社交网络交流过程中人们各种反应及互动的复杂性,使得提取和识别社会角色的工作变得更加困难.由于社会网络的千差万别,很难找到一种通用的角色识别方法,只能是根据网络具体情况去寻找合适的角色识别方法.识别方法在不同社会网络中的准确性以及适用范围的广泛性上都存在一定的局限性.现在,各种识别方法只能根据一种或几种特征判断依据来识别,无法也不可能将所有的特征情况都考虑进去.因此,如何对复杂社会网络更好的分析进而识别出其中的角色,一直是研究者考虑的问题.

(2)海量数据问题

由于社会媒体的迅猛发展,社会网络尤其是在线社会网络中的数据量不断增加,给我们识别社会网络中角色带来巨大的挑战.例如,随着社交网络节点容量的不断增大,在利用图论对社会网络进行拓扑分析时会变得越来越复杂.当节点数以百万计时,更难以在短时间内得到有效结果.此外,在结构相似性角色识别中,假设在某个网络中,其中两个节点的邻居节点有很多是共有的,那么我们就可以说这两个节点是结构等价的.但在海量数据中,由于一些节点的度比较大,其本身的邻居数量太多,很容易造成影响,从而使角色识别问题变得更加难以解决.

(3)评价问题

在社会网络角色识别研究中,需要对识别的角色质量和相邻时间角色的变化情况进行评价.不同情况或环境下,角色定义不同,角色识别方法也不同,从而产生不同的角色质量评价方法.那么就会出现这样一种情况,即同一个社会网络,不同的角色识别方法得到的不同的角色,不同的评价体系得到不同的最优角色识别方法,而哪一种最好的,成为一个比较令人困惑的问题.

角色演化的评价问题就是社会网络角色演化情况的判别问题.一般情况下,根据角色相似性度量方法,对相邻时间片内角色的相似度进行评价,判断角色的演化情况.不同的相似性度量会产生不同的角

色演化，而寻找合适的角色评价成为难点问题. 其实，无论角色质量评价还是演化评价，他们的共同点都是事先设定了角色的特征、相似角色的特征，而忽视了不同社会网络间的差异性和角色自身特点，主观性较大. 因此，在社会网络角色识别中，很难找到一种有效的、通用的评价方法.

社会网络的复杂性，数据的海量性以及评价方法的主观性等具有挑战性问题的研究，推动了社会网络角色识别的进程，各种角色识别方法也在不断提出. 此外，用户数据采集问题，隐私和安全问题也是社会网络研究中面临的挑战，需要我们继续研究和解决.

根据对社会网络角色识别的理解，本文构建如图 5 所示的社会网络角色识别研究流程. 首先对采集到的社会网络数据进行预处理. 然后通过网络结构、内容分析，数学分析以及机器学习等方法来分析这些数据，进而得到不同用户群体. 再根据各种社会主题、环境和要求对这些用户群体命名，从而得到最终的用户角色. 此外，还可在研究过程中引入时间因子，进一步分析用户角色演化的情况.

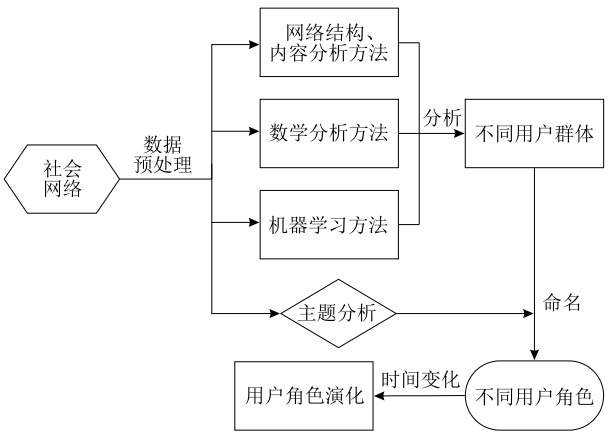


图 5　社会网络角色识别研究流程

本文将社会角色分为两类：非明确角色和明确角色. 其中，将没有预先定义的角色称为非明确角色，对该类角色的识别称为非明确角色的识别. 而将预先定义好的角色，如专家、意见领袖以及各种网络社区中的具体角色等，称为明确角色，对该类角色的识别称为明确角色的识别. 本文接下来将分别对这两类角色识别方法的国内外研究现状进行总结和叙述.

3　非明确角色的识别

本节将主要叙述和分析非明确角色识别方法的国内外研究现状. 当前，非明确角色的识别主要是通过机器学习和数学分析的方法来实现. 非明确角色意味着没有（或很少）背景知识预先对角色进行定义，将主要通过社会网络结构，特征描述，以及交互过程中产生的文本信息进行识别[4]. 其中，机器学习方法主要根据网络结构、交互信息或者同时将二者考虑进去，通过无监督的学习过程，自动将数据（或节点）分到不同角色类中[4]，最终识别出用户角色. 如 Laurent 等人[25]通过机器学习的方法，用小决策树取代决策树桩，提出一个改进算法，解决了广播新闻节目中说话者角色识别的问题. 通常机器学习角色识别过程如图 6 所示. 数学分析中，通常使用块模型（Blockmodels）、概率模型进行分析识别，这也是非明确角色识别的主要方法. 这里所识别出的角色可以理解为网络节点所处的位置[26]，就如同一个公司中“经理-秘书”在公司所处位置不同一样，“经理”位置可根据其使用的电子邮件中词汇的种类识别出来，再根据其邮件接收者的位置，判断这是一个“秘书”或是另一个“经理”[4].

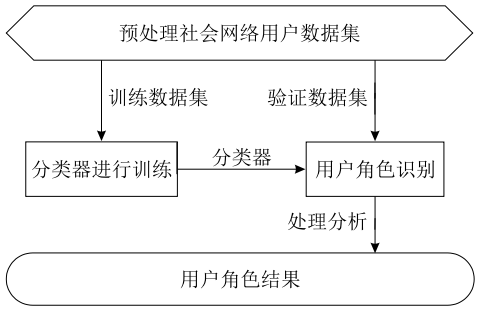


图 6　机器学习角色识别过程

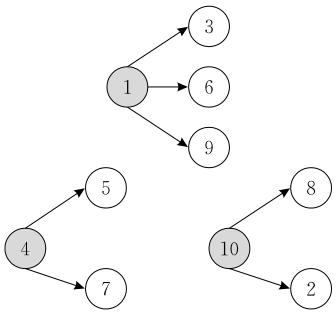
此外，还可通过图结构或者交互信息的聚类过程识别出非明确角色或节点位置. 如 Handcock 等人[27]提出了一个潜在位置聚类模型（Latent Position Cluster Model，LPCM）. 在该模型中，两个用户间连接的概率取决于一个隐含的欧几里德“社会空间”中二者间的距离，而用户的“位置”在该隐含社会空间中源于一个混合分布，他们分别对应一个聚类. 该模型能够轻易地模拟真实网络的聚类，从而识别其中潜在的位置.

本文接下来将主要叙述非明确角色识别方法中块模型（Blockmodels）识别方法和概率模型识别方法的国内外研究现状.

3.1　块模型识别方法

块模型（Blockmodels）是建立在一个社会关系网络的基础上使用预先定义的等价性，其目的是在

预先定义的等价 0/1 功能块中进行分类[26]. 块模型识别方法中,最重要的是根据所处理的问题,正确定义一个关系的等价. 当前,存在的几种关系等价取决于研究对象的结构等价、正则等价、强等价和自同构等价. 结构等价一般针对具有相似兴趣的成员分组,而正则和自同构等价则更多的用来表示社会角色的社会学概念:相同角色的人只能与共享同一角色的人联系. 通常,块模型是一个由二维 0/1 矩阵表示的简单图结构,该矩阵与某种类型的关系(如同事关系、同学关系、朋友关系等)相关联. 根据相关关系数据建立块模型,可用来解释复杂网络的活动. 通过块状建模,选择对应的等价性,将复杂的社会网络简化为一个和几个简单的图结构. 其处理过程如图 7 所示,由原始 10 人组成的社会网络(a),得到角色交互关系(e)的过程. 根据图(a),可以得到图(b)所示的二维关系矩阵,其中,二者之间的关系用 1 表示. 通过行列的置换得到重排的 0/1 矩阵(c),从该图中可看出该矩阵主要有 4 块:3 个 0 块和一个包含 1 的块. 使用预定义的标准,通过相关的等价概念,得到块模型(d). 块模型(d)可表示为位置 A 和 B 间的关系 A→B. 例如,可表示教师 A(T)和学生 B(S)的角色网络(e).



(a)

1	**0**	**0**	1	**0**	**0**	1	**0**	**0**	1	**0**
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	**1**	0	**1**	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	**1**	0	0	0	0	0	**1**	0	0

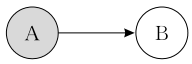
(b)

1	**0**	**0**	**0**	**0**	1	**0**	1	**0**	**0**	1
4	0	0	0	0	0	**1**	0	**1**	0	0
10	0	0	0	**1**	0	0	0	0	**1**	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0

(c)

A	0	1
B	0	0

(d)



(e)

图 7 块模型(Blockmodels)处理过程

块模型处理技术可以说是社会网络分析中应用最频繁的技术之一,在保证其数学理论基础上归纳块状建模(Blockmodeling),从而分析多种类型的网络结构[28]. 同时,块状建模也是获得社会网络结构模型的主要方法之一[29]. 虽然块状建模主要针对网络结构,但它也能够用来处理节点的属性和多重关系[27]. 在块模型角色识别过程中,节点位置可以由分析人员给出或者以无监督的方式估计出来. 如果位置的类型不是预先定义的,则需要同时计算出相同的集群及集群间的关系,这就是位置或角色,这意味着这里的角色是被估计的位置.

在以往研究中,一般会通过正则等价和结构等价对网络进行研究. 在 REGE 和 CATREGE[30] 算法中,将数据的关系(即网络)作为输入,集合分区作为输出的函数看作正则等价的定义,并计算网络节点对间的正则等价程度,建立块模型,识别出隐含位置. CONCOR[31] 算法则根据网络结构等价,通过迭代过程,不断将数据分割成两个块,构建一个分层的树(树状图). 最终,每个原始节点都会关联到图中的一个位置或角色. 此外,还有一些随机模型也都主要集中在结构等价,并且通过刻画网络节点的属性(例如,社会类别)得到划分结果[32-33],这可被看成是随

机块模型(Stochastic Blockmodels)后来又进一步将这些模型扩展到潜在类的分析[34-35]，只不过这些潜在类不假定聚类成员是已知的，而是从数据中估计它们. 与以往方法相比，Handcock 等人[27]集成了个体间距离概率模型，提出的 LPCM 模型考虑了图的传递性，而这种传递性可说明当人们有共同的属性(如年龄、性别、种族、地理位置等)时，往往更倾向于互相联系. Wolfe 等人[36]通过允许使用多角色又进一步地扩展了以往随机块模型. 此外，Airoldi 等人[37]提出了一个针对关系数据的混合成员随机块模型(Mixed Membership Stochastic Blockmodels，MMSB)和一个快速近似后推断的通用变分推理算法. 改进了先前潜在随机块模型，使每个对象允许同时属于几个不同的簇中. 也就是说一个人可同时扮演几种不同的潜在角色. 后来，Fu 等人[38]又将网络的自然演化考虑进去，即角色可随着时间变化而变化，提出了新的动态混合成员块模型(Dynamic Mixed Membership Blockmodel，DMMB).

总之，为了适应传统社会网络模型，提出了混合成员随机块模型(Mixed Membership Stochastic Blockmodel，MMSB). 与混合模型聚类方法类似，每个节点关联到一个隶属度向量，且该向量涉及到不同的簇，最终得到节点或个体的位置或角色.

3.2　概率模型识别方法

在我们分析如邮件、博客、学术论文等文本数据集时可能会发现，仅仅依靠关系结构进行分析或识别其中的角色是远远不够的. 在这种情况下，需要用到另一个主要分析、识别角色的方法，即概率模型方法. 概率模型方法主要处理文本数据集，通常会使用无监督的分级贝叶斯模型来实现. 在不考虑网络关系结构的情况下，将文本内容与图中的边关联起来. 其实现过程主要是依赖已有的主题模型，该模型首先假设一个概率生成模型，然后将每一个文本关联到多个主题，最终从文本中提取主题[4]. 当前，概率主题模型通常是依据这样的一种思路，即将文本看作是由若干主题随机混合组成. 在不同的模型中，通常会有不同的统计假设，并以不同的方法得到模型参数[39].

在主题模型中，一个主题通常被定义为一个给定词汇的多项式分布. 主题模型用一个数量较小词汇的分布对大量的文本进行总结，这些分布被称为“主题”[40]. 如作者-主题模型[41]中，每个作者都与一个多项分布的主题相关，而每个主题都与一个多项分布的词相关联. 在这个生成模型中，每一个文档表示为一个主题的混合. 如在潜在狄利克雷分布(Latent Dirichlet Allocation，LDA)[42]方法中，通过允许文档作者确定不同主题的混合权重，并将这些方法扩展到作者建模. 通过学习模型的参数，我们得到了一组出现在语料库和他们相关的不同文档的主题集，并识别出作者及使用的主题. 在 LDA 模型中，文档集合的生成可模拟为三个过程，首先，对每一个文档，从狄利克雷分布中抽取出该文档的主题分布. 然后，根据主题分布，对文档中的每一个词汇选择一个单一的主题. 最后，从具有特定采样主题的词汇多项式分布中，对每个词汇进行采样. 该生成过程对应的分层贝叶斯模型如图 8(a)所示. 其中 $\boldsymbol{\varnothing}$ 表示主题分布矩阵，通过 V 个词的多项分布，在已有的对称狄利克雷(β)分布中独立刻画 T 个主题. θ 是特定文档的 T 个主题混合权重的矩阵，这 T 个主题，在现有的对称狄利克雷(α)分布中，独立刻画每一个主题. 对每一个词，z 表示负责生成这个词的主题，从该文档的 θ 分布中得出，w 是词本身，由主题分布 $\boldsymbol{\varnothing}$ 对应到 z 得到. 估算 $\boldsymbol{\varnothing}$ 和 θ 参数的值，得到关于参与主体的主题信息和每个文档中这些主题的权重信息.

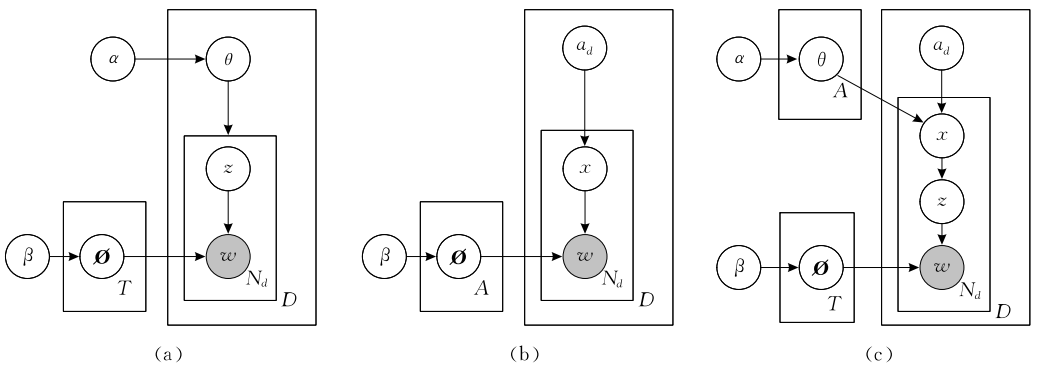


图 8　文档生成模型

在研究 LDA 基础上，Rosen-Zvi 等人[41]提出作者模型，该模型中每一位作者都与一个词的分布而不是一个主题的分布相关联，如图 8(b)所示. 其中，x 表示一个给定词的作者，从作者集合 a_d 中随机选择. 每一个作者对应一个词的概率分布 $\boldsymbol{\emptyset}$，这些概率分布也是从已有的对称狄利克雷(β)分布中得到. 通过估计 $\boldsymbol{\emptyset}$ 的值，得到有关作者的兴趣信息，同时可以得到相似作者的信息以及与该文档主题相似的作者. 根据这两个模型，最终提出了作者主题模型(Author-Topic Model，ATM)，如图 8(c)所示. 该图中，和作者模型相似，x 表示一个给定词的作者，同样从作者集合 a_d 中随机选取. 每一个作者对应一个主题的概率分布 θ，这些概率分布也是从已有的对称狄利克雷(α)分布中获得. 根据所选作者相应的混合权重，选择一个主题 z 并且得到与 $\boldsymbol{\emptyset}$ 分布对应主题的一个词 w，$\boldsymbol{\emptyset}$ 分布也是从已有的对称狄利克雷(β)分布中得到.

作者-主题模型可以说是 LDA 主题模型和作者模型的组合，和 LDA 主题模型中每个文档都有一个唯一的作者相对应一样，作者模型中每个作者都有一个唯一的主题相对应. 通过参数估计 $\boldsymbol{\emptyset}$ 和 θ 参数的值，我们就可以得到作者通常所写的主题信息，以及在这些主题所代表的每个文档的内容. 而对于这些参数，通常会采用吉布斯采样算法[43](Gibbs sampling algorithms)和变分近似推理(Approximate inference with variational methods)这两种方法来进行估计.

此外，McCallum 等人[44]为了识别邮件数据集中的角色，提出了 3 种贝叶斯分层模型，分别为 ART(The Author Recipient Topic model，ART)模型、RART1(Role Author Recipient Topicmodels)模型和 RART2 模型. ART 模型是一个词的有向图模型，而该词来源于给定的作者和一组收件人之间生成的消息. 由于要识别的角色存在于所关联到的每一个元组(作者、收件者)的主题组(或者说词汇种类)中，因此，ART 模型角色是未知的. 在 ART 模型的基础上，作者又进行了改进得到另外两个模型. 在这两种模式中，角色作为一个潜在的随机变量，在贝叶斯网络中被明确地建模. 因此，一个角色是一个刻画两人(作者和收件者)关系的主题混合. 此外，在会议挖掘[45](Conference Mining)中，ConMin 模型所识别的角色一般是具体的专家或内行角色. 此外，为了得到科学论文中主题的一个全局分析，往往会增加时间、来源等额外的维度进行分析. 如语义时态信息化专家(内行)搜索(Semantics and Temporal information-based Maven Search，STMS)[46]模型，该模型可同时得到作者潜在的主题，场地(会议或期刊)以及时间的信息，并识别内行角色.

总之，在非明确角色识别方法中，传统的块模型方法更多的是和图论有关，可能更加适用于社会网络中角色的识别. 但是，块模型主要是通过网络的关系结构来构建，忽视了用户间交互的信息内容. 而在社交网络中，用户的角色或多或少地表达在社交媒体网络生成的内容中，或者明确或隐含地表达在社交媒体网络生成的内容中. 此外，鉴于当前组合优化技术发展的现状，块模型方法在许多具体应用中，运行比较缓慢，效率比较低. 对于概率模型来说，这种模型可以说是使用更多自身信息的第一步. 基于主题的概率模型在识别角色过程中，虽然用到了节点间文本的关键信息，但却缺乏像块模型从全局的角度来识别. 因此，如何将这两种模式有效结合，是当前非确定角色识别方法中的一个挑战[4].

根据非明确角色的识别模型、主要识别依据和是否考虑时间因素，总结本文所述文献如表 1 所示.

表 1 非明确角色识别模型、依据及时间因素汇总

文献	作者(第一)	年份	模型	结构	内容	时间因素	说明
[31]	Breiger R L	1975	Blockmodels	√			块模型-结构等价
[32-33]	Fienberg S E, Holl P W	1981	Stochastic Blockmodels	√			随机块模型-结构等价
[34]	Wasserman S	1987	Stochastic Blockmodels	√			随机块模型-结构等价
[35]	Snijders T A B	1996	Stochastic Blockmodels	√			随机块模型-结构等价
[30]	Borgatti S P	1993	Blockmodels	√			块模型-正则等价
[36]	Wolfe A P	2004	Stochastic Blockmodels	√			随机块模型
[41]	Rosen-Zvi M	2004	ATM		√		作者-主题模型
[27]	Handcock M S	2007	LPCM	√			潜在位置模型
[44]	McCallum A	2007	ART，RART1，RART2	√	√		三种贝叶斯分层模型
[37]	Airoldi E M	2008	MMSB	√			混合成员随机块模型
[38]	Fu W	2009	DMMB	√		√	动态混合成员块模型
[45]	Daud A	2009	ConMin		√		会议挖掘模型
[46]	Daud A	2009	STMS		√		语义及时间的信息化专家(内行)搜索模型

4　明确角色的识别

在本节中，我们主要叙述明确角色识别方法的研究现状. 本文所说的明确角色，指的是预先定义好的角色. 在本节中，我们将分别对当前社会网络中比较常见的影响者角色、意见领袖角色、专家角色的识别方法的国内外研究现状进行叙述和总结. 最后又进一步叙述和总结了除这 3 种重要角色之外的其他角色识别方法的国内外研究现状.

4.1　影响者角色(Influencer Role)

当前，由于像 Facebook、Email、在线社区、微博、博客、微信等在线社会网络的快速发展，使人们有了更广阔的社会活动空间. 而这种在线社会网络却有可能影响我们现实社会中的人类活动. 比如，某个明星空间上发布的一个帖子，能够使他人(尤其是影迷或粉丝)在其生活中做出回应或疯狂的举动.

从在线社会网络的所有节点中，识别出能够影响他们邻居行为的节点是非常重要和有趣的. 在研究和实践中，识别在线社会网络中影响者角色以及对该角色在商业领域的应用也变得越来越流行[47]. 2012 年 10 月，Facebook 在线活跃用户突破 10 亿，超过 1400 亿的朋友连接关系，数据呈现出了爆炸式增长，引发社会的广泛关注，识别其中的影响者角色更是受到人们的重视[48]. 在社会网络中，如何有效的识别出这些能够动员他人并扩大影响力的人或节点，也被不断分析和研究[49]. 在本小节中，我们将讨论以往社会网络中，识别影响者角色的主要方法、模型以及衡量用户影响力大小的各种度量方法. 本文中影响者角色主要是指在社会网络中具有影响力的人或节点，即能够动员他人并扩大影响的人或节点.

总的来讲，在社会网络的相关的研究中，对于影响者角色的识别方法，主要是通过度量其影响力大小或对其进行排名来实现. 如果从技术角度来讲，则主要包括基于网络结构以及内容发掘的技术和方法.

在基于网络结构的识别方法中，Kayes 等人[50]通过使用网络中心性矩阵进行分析，得出博客通常是以“核心-边缘”的网络结构相连接的结论. 在该结论中，有影响力的博客会连接成为核心，而影响力小的博客会处于结构边缘. Wu 等人[51]在所研究的微博数据中，发现了一个近似幂律分布的跟随者分布和一个不服从幂律分布的朋友分布，提出了识别出影响者角色的 XinRank 算法. Wen 等人[52]通过同质现象来解释“对等”的存在，并扩展 PageRank 算法，提出能够识别有影响力的 Twitter 用户的 TwitterRank 算法. Zhang 等人[53]则将关系强度的有效性作为影响者角色的识别标准，并得出通过强连接用户数识别影响者角色要优于其他连接的度量. 而 Gliwa 等人[54]基于博客主题中评论间的关系以及对博客级别的定义，提出了一个更加合理、有效的影响者角色识别方法.

在一些社会网络中，仅仅依靠网络结构特性可能无法得到用户之间的真实影响，而用户的内容能使我们了解网络动态特性. 因此，结合用户内容发掘技术进行识别，也成为比较受欢迎的识别方法. 如 Tang 等人[55]从在线医疗保健论坛中提出一个检测和识别有影响力用户的框架. 该框架首先在医疗保健论坛中构建社会网络. 然后，通过结合了链接分析、内容分析的 UserRank 算法建立一个识别机制，识别有影响力的用户. Li 等人[56]则在博客数据中提出一个识别具有营销价值的市场影响力价值(Marketing Influential Value，MIV)模型. 该模型也是基于网络和内容两个因素度量博客特征维度，建立一种自适应的人工神经网络，识别出潜在的支持市场营销或广告商的影响力博客. 同样在博客中，Aziz 等人[57]则利用影响力衡量因子来识别有影响力的博客，提出了一个有效识别影响者角色的算法. 而这些因子主要是基于博客帖子的语义内容，定量分析帖子内容以及帖子中评论的跟帖者得到的.

在识别影响者角色的方法中，研究者还提出了多种衡量影响力大小的方式或指标. 如在衡量用户影响力大小的度量中，Moon 等人[58]提出基于博客间同质性和脆弱性的加权度量，建立量化影响模型(Quantifying Influence Model，QIM)来衡量博客的影响力得分，根据影响力得分的多少，识别出博客中的影响者角色. Bui 等人[59]提出了一种利用 h-指数衡量一个社区中博客产生的影响力大小的方法，并加强其中信息指数的计算，最终识别出其中有影响力的博客. h-指数是 2005 年由乔治·赫希(Jorge Hirsch)提出，是指一个人至多有 h 篇论文分别被引用了 h 次，论文按被引频次降序排列后，第 h 篇论文被引用 h 次以上如图 9 所示.

h-指数是一个衡量学术成就的指数，主要通过引用关系来衡量一个学者或科研人员的学术成果及影响力大小. 一个人的 h 指数越高，则说明这个人的学术成就越高，影响力也越大. Akritidis 等人[60]又将博客活动的时间因素考虑进去，提出了两种博客影响力排名方法，这两种方法分别基于具体的指标

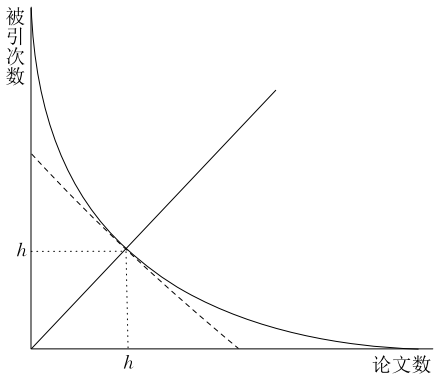


图 9 h-指数示意图

(如帖子的评论数量、发布日期、年龄等)计算博客帖子的得分.根据得分排名识别有影响力的博客.而 Moh 等人[61]通过研究现有的确定影响力博客的模型,基于独特性(Uniqueness)和 Facebook 数(Facebook Count)两种指标,构建一个改进的模型.该改进的模型也可通过链接和评论的数目,帖子发布和评论的时间,评论者的影响等因素,衡量出博客真正的影响力,并给博主相应的排名,有效识别出有影响力的博客.Khan 等人[62]则又根据博客的出产情况以及受欢迎程度等多种特点,提出了一种新的识别有影响力博客度量(Metric for Identification of Influential Bloggers,MIIB),有效的找到最有影响力的博客.

此外,Ding 等人[63]通过微博多关系数据的随机游走度量用户的影响力,提出了一个组合随机行走多关系影响网络的方法,有效识别出微博中的影响者角色.Akritidis 等人[64]则通过引入博客的产率指标(BP-index)、影响指标(BI-index),识别出具有高产率和高影响力的博客角色.而 Shalaby 等人[65]则通过找出在 Twitter 用户影响力的度量中起关键作用的因素,验证不同因素是如何对影响力的排名起作用以及怎样将这些因素表示成数学模型的过程,识别出 Twitter 用户中影响者角色.还有就是,Cai 等人[66]提出的一个有效挖掘前 k 个具有影响力博客的模型以及 Liu 等人[67]利用网络中异构连接信息和节点相关联的文本内容的生成图模型.其中,通过生成图模型来挖掘有关主题方面的直接影响,然后再得出网络节点间的间接影响.该作者虽然没有直接识别出影响者角色,但为我们提供一个通过两种影响力识别影响者角色的参考方法.同样,Agarwal 等人[68]在研究什么是有影响力的博客基础上,从一个博客网站上,评估收集到的决定博客帖子影响的各种度量统计,提出了一个量化有影响力博客的初步模型,为寻找各类型影响力角色铺平了道路.

总之,影响者角色的识别更多的是根据社会网络的拓扑结构和用户内容进行识别和衡量影响力的大小.社会网络中,在使用图论进行网络结构拓扑分析的过程中,有时会变得比较复杂,尤其是社会网络规模越来越大的今天.为了更好地识别在线社会网络影响者角色,未来应提出一个包含拓扑度量,内容权值以及链接极性值(Link Polarity Values)的新方法.除此之外,基于用户活动扩散历史的方法在识别过程中也是有效的,而基于时间和位置的方法上的扩展[69]也应是未来研究的重要方面.

4.2 意见领袖角色(Opinion Leader Role)

随着互联网技术的普及,越来越多的人们开始喜欢在网上获取信息,发表自己的观点.由于互联网用户素质的参差不齐和互联网自身的隐蔽性,一些用户可能会随意表达一些不负责任的观点、意见甚至是违反法律的言论而不被惩罚.其中的一些观点、言论可能会对人们的意识产生很大的影响,甚至一些言论会危害到社会治安.由此,适当引导网络舆情的发展就显得尤为重要.在社会网络中,意见领袖对社会网络舆情的产生和发展起着重要的作用.因此,在社会网络中,识别和分析意见领袖角色具有重要的实际意义.此外,意见领袖角色也是网络舆情分析领域一个重要的研究内容.意见领袖角色通常会将个人的观点、态度传递给其他网络用户,进而影响、改变这些用户的观点、态度及决策.本文中所说的意见领袖角色指的是社交网络中为他人提供信息(如发表微博、帖子及回复、评论其他用户发表的微博、帖子),同时对他人施加影响并起到加速信息扩散及中介作用的"活跃分子".

近年来,为了识别网络中的意见领袖角色,研究人员对此进行了广泛研究,提出了许多意见领袖角色的识别方法.总体来看,这些方法主要是根据用户特征、内容以及用户交互网络进行识别[70].

根据用户特征、内容的识别方法中,Duan 等人[71]提出了一种从在线股票留言板中识别出意见领袖角色的新方法.该方法首先根据发布的信息,计算出用户活动的特征,再通过聚类算法来处理用户数据,生成包含潜在意见领袖的簇.最后,通过情感分析的方法,分析用户和情绪与实际价格变动趋势之间的关联,进而识别出其中的意见领袖角色.而 Hudli[72]则在分析用户在线活动特征的基础上,为每个用户构造一个在线配置文件,并通过 k-means 聚类算法分析这些文件或意见,从而识别出论坛中

的意见领袖. Hung 等人[73]提出一个基于文本挖掘的意见领袖角色识别方法. 该方法则主要通过评估文本内容中的专业特征以及信息的新颖性、丰富性来识别出该网络中的意见领袖角色. 王君泽等人[74]根据得到的识别意见领袖角色的关键因素，如关注用户数量、是否验证身份、微博数量等，构建了一个多维识别模型，并为保证模型构建的科学性，提出用户重要性的评分公式. 最终，通过该模型识别出微博中的意见领袖角色. 而张伟哲等人[75]通过进一步分析基于内容的“影响力扩散模型”，进而提出一个能够有效识别网络论坛中意见领袖角色的基于语料阶梯评价算法. 此外，刘志明等人[76]分别从网络用户的影响力和用户活跃度两个角度入手，通过使用层次分析法和粗糙集决策分析理论分析意见领袖的特征，构建意见领袖指标体系，并对意见领袖的跨主题性进行研究，最终发现意见领袖是依赖于主题的.

根据用户内容和交互网络的意见领袖角色识别方法中，Cheng 等人[77]提出一种识别 BBS 中意见领袖的 IS_Rank 算法. 该算法主要是将用户影响值(包括内容影响值和情感影响值)作为用户之间链路的权值，有效提高了意见领袖角色识别的准确性. 宋昭君等人[78]针对传统链接分析方法忽略博文内容的问题，进一步提出基于链接分析和内容分析相结合的算法. 该算法通过计算博主的影响力得分(根据内链接数、外链接数、评论数和文章长度计算)，识别出了博客中的意见领袖角色. 同样，Bodendorf 等人[79]也结合文本挖掘和社会网络分析，提出一个识别在线社会网络中意见领袖角色的方法. Luo 等人[70]则提出一种基于交互网络和用户特征的混合数据挖掘方法，识别出微博中的意见领袖角色. 而 Zhang 等人[80]根据马尔可夫(Markov)逻辑网络，提出了基于关系数据的网络意见领袖角色识别方法. 此外，樊兴华等人[81]提出了一种新的影响力扩散概率模型(IDPM 模型)，在此基础上建立了具有开放性和包容性特点的网络意见领袖筛选模型. 王红瑞[82]运用中介中心性分析方法以一种定量的方法对意见领袖进行识别，构建了一个微博空间意见领袖识别模型. 尹衍腾等人[83]则基于用户关系，通过确定用户的中心性获得候选意见领袖. 其中，用户的中心性依据的理论是小世界网络理论(“六度分离”理论的拓广). 而 Song 等人[84]则根据网络中该节点重要性，并通过余弦距离衡量博客信息的新颖性，提出了一种影响等级算法(InfluenceRank algorithm)，从而识别出博客中的意见领袖角色.

其实，国内外研究人员在意见领袖角色的挖掘研究工作中，非常注重 PageRank 与 LeaderRank 算法. 如 Song 等人[85]借鉴熟悉的 PageRank 算法对用户进行排名，然后结合用户之间情感倾向分析、评论隐含关系分析以及评论的时间衰减等因素，挖掘出了复杂网络中的意见领袖. Zhou 等人[86]在介绍什么是意见网络的基础上，通过改进 PageRank 算法合并意见分数，提出了一种在意见网络中识别意见领袖角色的 OpinionRank 算法. 宁连举等人[87]通过构建有向加权评论网络并借鉴 PageRank 算法的思想，提出一种基于评论回复关系的意见领袖识别算法.

PageRank 算法根据网页间相互的超链接关系计算网页的排名. 我们一般会将 PageRank 算法看作是一种网页重要性排名的算法. 实际上，该算法是借鉴了传统引文分析的思想[88]. PageRank 算法基本思路是：假设网页 a 和 b 之间具有超链接且 a 指向 b，则网页 b 就得到网页 a 对它的贡献值，该值的大小由取决于网页 a 的重要性，重要性越大贡献值也就越高. 由于网页中的链接通常是相互指向的，因此贡献值的计算过程是一个迭代的过程，当迭代达到某个阶段时，就会根据网页所得分值进行检索排序，从而得到最终的网页排名结果. 一个网页的 PR 值(贡献值)大小，可由式(1)计算：

$$PR(a)=(1-d)+d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \tag{1}$$

通常假设每个网页的初始贡献值为 1，其中，$PR(a)$表示网页 a 的贡献值，d 表示用户访问某个网页的随机概率(通常为 0.85). T_i为指向页面 a 的其他网页，$C(T_i)$为网页 T_i指向外链接的个数，n 为网页总数，$\frac{PR(T_i)}{C(T_i)}$为网页 a 的链入网页 T_i的贡献值. PageRank 算法是一种比较成熟和常见的搜索引擎网页排序方法，也经常应用于文献统计学、社会和信息网络的分析，以及链接的预测和推荐等众多学科研究方面[89].

实际上，在使用 PageRank 算法时会有一个不足之处，PageRank 算法并不适合在结构快速变化的复杂网络中挖掘意见领袖. 为解决 PageRank 算法在此方面存在的不足，Lv 等人[90]提出了一种意见领袖挖掘的 LeaderRank 算法. LeaderRank 算法在用户排序过程中，会先为每个节点(基点除外，基点为通过双向链接与每个用户相连接的附加点)分配资源的一个单元，然后再平均分配给该节点的邻居节

点,直到达到平衡状态.该过程相当于直连网络中的自由行走,可通过随机矩阵 $\boldsymbol{P}$ 来表示,而随机矩阵 $\boldsymbol{P}$ 由公式 $p_{ij}=a_{ij}/k_i^{\text{out}}$ 得出,表示下步自由行走中节点 i 到节点 j 的概率.其中 $a_{ij}=1$(当节点 i 指向节点 j 时)或 $a_{ij}=0$(当节点 i 没有指向节点 j 时),k_i^{out} 为节点 i 的出度.在时间 t 时,节点 i 的得分为 $s_i(t)$,而

$$s_i(t+1)=\sum_{j=1}^{N+1}\frac{a_{ji}}{k_j^{\text{out}}}(t) \tag{2}$$

其中,所有节点初始值 $s_i(0)=1$(基点除外),基点初始值 $s_g(0)=0$.

定义关系值 S 为用户最终值:

$$S_i=s_i(t_c)+\frac{s_g(t_c)}{N} \tag{3}$$

其中 $s_g(t_c)$ 为基点状态稳定时的值.最终,根据每个用户该值排名,将位于首位的用户作为所得的意见领袖.与 PageRank 算法相比,LeaderRank 算法在意见领袖挖掘方面准确性更高,出现噪音和恶意攻击时的稳定性更强.

徐郡明等人[91]则又在 Lv 等人提出的 LeaderRank 算法上加以改进,加入用户间的情感倾向和用户活跃程度,使 LeaderRank 算法的准确性和抗干扰能力得到进一步提升.而 Xiao 等人[92]根据社区发现(本质上是一种文本分类过程)和情感挖掘(构建情感词汇词典)方法又提出一种 LeaderRank 算法,识别出了 BBS 中的意见领袖.同样依据文本分析和情感挖掘,肖宇等人[93]又结合聚类算法和分类算法的优势,提出一种基于话题内容分析的兴趣团体发现方法,然后通过分析用户回帖情感倾向来计算用户间链接的权重.在此基础上,提出一种新的 LeaderRank 意见领袖发现算法.

此外,徐会杰等人[94]针对现有意见领袖识别算法难以捕获网络的动态特性的现状,在网络论坛中根据时间变化图提出了一个意见领袖识别算法.该识别算法将论坛的演变描述为一连串静态图,每一幅图代表一个给定时间片内用户间的交互.依据构造的量化指标识别不同时间片内的潜在意见领袖,然后和其他时间片的意见领袖相匹配,从而识别出能够随时间推移的意见领袖.Volpentesta 等人[95]模拟了一个基于时间的商业社交网络,将其作为一个随时间变化的加权有向图.通过对商业社会网络中的行为和结构两个方面分析,提出了一种识别意见领袖角色的方法.蔡淑琴等人[96]提出结合 RFM 模型和情感词自动判别方法并考虑情感(S)这一个新的指标,提出了 RFMS 模型来衡量影响力,应用人工神经网络模型识别出意见领袖.人工神经网络模型通过大量的训练能够得到更加逼近于真实情况的结果,从而得到一个在线口碑传播中准确率更高的意见领袖识别方法.当前,在线评论中大多数意见领袖的识别方法更多的会去考虑评论中的正面意见.而 Chen 等人[97]则将反面意见也考虑进去,通过建立一个包括正反两方面意见链接的具有加权用户网络,设计了一种基于 PageTrust 的新模型 TrustRank 模型,进而识别出了在线评论中的意见领袖角色.

意见领袖角色是在线社会网络中比较常见的一种社会网络角色,同时也是社会网络研究领域中一个重要的研究课题.从目前来讲,对该角色的识别主要集中在网络分析,文本分析、情感分析等方法上.这些方法大多都是以单个节点为主要研究对象进行分析,未来可从群或团体的角度来进行分析和研究.此外,由于在社交媒体中,随着话题的进行或改变,随时会产生新的意见领袖主.因此,也可以从时空性角度对意见领袖进行进一步的分析研究.

4.3 专家角色(Experts Role)

近年来,为了能够充分利用专业知识,专家发现或识别在科研领域受到较大关注.在现实生活中,通过专家发现也可以帮助我们解决许多具有挑战性的实际问题.如期刊为提高发表论文的质量,论文评审过程中会强烈建议同行专家进行审核,而对于给定的论文主题,如何判断一个给定的审稿专家的专业知识以及匹配适当、合格的审稿人仍然是一个具有挑战性的问题[98].

在社会网络专家角色识别方面,研究人员先前已经做了许多研究工作,这些工作大多是使用各种度量来识别专家角色.根据其主要识别依据,也可将识别方法分为基于内容识别方法、基于链接结构识别方法以及二者相结合的方法.其中,专家的概念一般定义为:在社会网络中,具有关于所讨论话题的相关知识,并且他的意见和想法是可信的,我们就将这样一个人称为是专家[4].本文所叙述的专家角色识别,主要是指在社交网络中,尤其是具有专业知识的网络社区中对专家角色的识别.

社会网络专家角色的识别方法中,基于内容的识别方法可以说是最主要的识别方法,也是比较常见的方法.如 Pal 等人[99]在问答社区(CQA)服务中研究发现,专家更喜欢回答这样的问题,即能使他们获得更高有价值贡献机会的问题.我们称这种优先选择为问题选择性偏差,并提出一个数学模型来估

计它.该文研究结果表明,采用高斯分类模型,通过普通用户选择偏差,能够有效地从普通用户中识别出专家角色.在此基础上,Pal 等人[100]在问答社区服务(CQA)中又继续进一步的研究.他们根据用户选择回答的问题,得到用户优先选择偏好,根据优先选择偏好提出一种概率模型.最终,通过机器学习的方式来识别专家角色和潜在专家角色.文献[101-103]等则基于个人信息为中心的方法以及基于文档为中心的方法识别出相关的专家角色.在基于个人信息为中心的识别方法中,首先将一个潜在专家相关的所有文件、文本合并成一个单一的个人文件.然后,根据文件所对应的给定查询,估计出每一个潜在专家的排名分值.而基于文档为中心的方法则分别分析每个文档的内容,而不是建立一个单独的专业文件.文献[104-105]又将这两种方法进行结合,从而有效提高了专家角色识别的性能.

此外,基于内容专家角色识别方法中,还有一种主题建模的识别方法.早期 Hofmann[106]提出一个统计隐含语义标引(PLSI)算法,该算法是用来计算基于隐含主题层的文档生成一个词的概率.然而,PLSA 模型的参数过拟合现象严重,并且 PLSI 模型无法用一个直接的方法来推断文档.Blei 等人[42]通过提出一个称为狄利克雷分配(LDA)的三级分层贝叶斯模型解决了这些问题.在此基础上,文献[41,107]又提出作者主题模型和作者会议模型(Author Conference Topic,ACT),每个作者都与一个主题、所写词或出版会议的多项式分布相关.主题模型能够计算潜在专家之间的相关性以及推断相关主题,从而能够对它们进行排名.此外,Jung 等人[108]从获得的元数据和完整文本文档中,提出一种识别以主题为中心的专家角色的方法.该方法通过全文内容分析提取出主题信息构建局部分类,依此推断到个人和机构,从而识别出以主题为中心的专家角色.

链接结构是专家角色识别方法的另一个主要方面.Adamic 等人[109]通过研究 Yahoo Answers 问答论坛中人们提问和回答问题的模式和行为得出,相对于一个用户的专业,一个用户的链接结构能更好地表现该用户感兴趣的主题.在具体的识别方法中,算法 PageRank 和 HITS 可以用来分析学术网络中的关系,以发现其中权威专家角色.文献[110-111]提出对 PageRank 和 HITS 算法的扩展,通过度量合作关系权重以及用户的权威性,分别识别出数字图书馆研究社区或问题解答门户中的专家角色.文献[112-113]也是通过对 PageRank 算法的改进,得出了书目网络和作者共引网络中的专家角色.其实,这些识别专家角色的方法,都是在 PageRank 和 HITS 算法或其他基于链接方法上的改进,以解决一些指标上的限制(如引用数目),从而得到文献计量上的排名,但所得专家角色识别结果往往也不是最好的.文献[114]中,作者通过对一个 Java 论坛中帖子回复的分析,提出一种专家识别的技术.该作者用一个有向图表示帖子回复关系,边表示两个用户间的回复关系.专家角色认为是能够恰当地回答一个问题的人并用具体的指标(如入度、出度、问答的差异性等)来衡量这些专家.

最后一种主要的专家角色识别方法就是将基于内容和基于链接结构相结合,不过该方法不能同时以一个统一的方式对所有可能信息进行建模,而是通过在一个有限的潜在专家子集中排名的局部最优化来实现[115].如 Campbell 等人[116]通过搜集一个主题相关的所有邮件以及分析每对发件人、收件人之间的邮件,建立了一个对应的“专业图”.最后,通过改进 HITS 算法得到关于这一主题所有发件人和收件人等级.Zhang 等人[117]首先通过潜在专家的个人信息(如个人资料、联系信息和出版信息)估算出每个潜在专家的初始专家得分,并选择排名靠前的潜在专家构造一个子图.然后,提出了一个基于传播的方法来提高子图中专家发现的准确率.而 Lin 等人[115]在学术网中,通过研究引文网络中主题专家发现的问题,提出一个局部加权因子图模型(TWFG).该模型以统一的方式结合潜在专家的个人信息和学术网络信息,提高了主题级专家发现的有效性.

总之,当前大多数专家识别的方法往往注重其个人信息(如主题相关和引用数目)和网络信息(如引用关系),这些方法往往会导致一些潜在的专家被忽视.基于内容的专家角色识别方法,大多是计算专家和用户的查询主题或推断主题之间的相关值,却忽略专家间的社会关系.其他使用链接分析算法,如 PageRank、HITS,在识别专家角色过程中,都有一个共同的主题漂移问题,他们识别出的专家角色也不是最好的.因此,基于这两种方法的局限性,既考虑一个潜在专家的特殊主题的相关性,又考虑分析潜在专家网络结构的识别方法,相比较而言是比较理想的专家角色识别方法.

根据影响者角色、意见领袖角色和专家角色识别方法中的主要识别依据和是否考虑时间因素,总结本文所述文献如表 2 所示.

表 2 影响者、意见领袖、专家 3 种明确角色主要识别依据及时间因素汇总

文献	作者(第一)	年份	社会网络分析	内容分析	时间因素	社会角色
[56,66]	Li Yung-Ming, Caiv Y	2009	√	√		
[60]	Akritidis L	2009		√	√	
[52]	Weng J	2010	√			
[55,67]	Tang X, Liu L	2010	√	√		
[57-58]	Aziz M, Moon E	2010	√	√		
[64]	Akritidis L	2011	√	√	√	影响者角色
[50]	Kayes I	2012	√			
[53]	Zhang Y	2013	√			
[61,63]	Moh T S, Ding Z	2013	√	√	√	
[51]	Wu X	2014	√			
[54]	Gliwa B	2015	√			
[62]	Khan H U	2015		√		
[84]	Song X	2007	√	√		
[86]	Zhou H	2009	√			
[79]	Bodendorf F	2009	√	√		
[92]	Xiao X Y	2010		√		
[90]	Lv L	2011	√			
[74,76]	Wang Jun-Zhe, Liu Zhi-Ming	2011		√		
[85]	Song Kaisong	2011	√	√	√	
[72,75]	Zhang Wei-Zhe, Hudli S	2012		√		意见领袖角色
[77-78,93]	Cheng F, Song Zhao-Jun, Xiao Yu	2012	√	√		
[94-95]	Xu Hui-Jie, Volpentesta A P	2012	√	√	√	
[82,87]	Wang Hong-Rui, Ning Lian-Ju	2013	√			
[81,83,96]	Fan Xing-Hua, Yin Yan-Teng, Cai Shu-Qin	2013	√	√		
[71,73]	Hung C, Duan J	2014		√		
[80,97]	Zhang W, Chen Y	2014	√	√		
[91]	Xu Jun-Ming	2015	√	√		
[116]	Campbell C S	2003	√	√		
[110]	Liu X	2005	√			
[104]	Petkova D	2006		√		
[111,114]	Jurczyk P, Zhang J	2007	√			
[102,108]	Fu Y,Jung H	2007		√		
[117]	Zhang J	2007	√	√		专家角色
[112]	Fiala D	2008	√			
[101,103]	Balog K, Wu H	2009		√		
[113]	Ding Y	2010	√			
[99]	Pal A	2010		√		
[100]	Pal A	2012		√		
[115]	Lin L	2013	√	√		

4.4 其它角色

在社会网络中,除了上述 3 种最常见的社会网络角色之外,还有许多社会网络角色能够识别出来. 如在线论坛中,Forestier 等人[118]提出通过主成分分析的方法(PCA)得到三类社会群体(子群体),然后,最大化一种主题标准,从作为相关用户的子群体中,识别出参与一种主题的名人角色(与主题相最符合的人). 而为了能够自动识别在线论坛中的名人角色,Forestier 等人[119]又提出一些元标准(Meta-criteria). 通过三种元标准和一个基准(Baseline)构建新型数据框架. 最终得出,所提出的一个元标准能够成功识别出名人角色. 在社交网络中,Xu 等人[120]为帮助企业经济有效地进行市场营销和企业管理(如声誉、宣传等管理),通过分析用户的社会关系、等级以及相互间的交互关系等信息构建影响力网络,由此得到一个优化模型. 通过该模型,不仅使企业利润最大化,而且有效识别出最有价值用户(顾客)角色. 文献[121-122]分别在给定条件下分割社区,通过派系过滤(Clique Percolation Method,CPM)方法和快速模块化优化方法(Fast Modularity Optimization,FMO)以及基于兴趣多目标最优化方法,从博客以及给定网络中有效识别出关键人物(Key person/player)角色. 文献[123-124]分别根据微博质量和活跃度提出的 WeiboRank 方法以及通过贝叶斯分类器和反向传播神经网络的机器学习方法,识别出新浪微博中高价值用户和重要用户角色. 文献[125-127]分别通过频繁模式挖掘的方法,以及利用指数时间衰减函数、层级概念模拟分析用户影响,并分析

用户交互的信息以及对集群用户进行分类，最终识别出社会网络中的领导者、追随者角色. 而文献[128]则在定义发起者角色(Starters)和追随者角色(Followers)的基础上，首先提出一种基于竞争的随机游走访问磁盘的采样方法. 然后，根据对网络特征的度量，识别出社会媒体中发起者和追随者角色. Stadtfeld[129]分析了个体选择行为模式并根据个体行为的相似性将他们聚类，从而识别出动态社会网络中潜在的行为角色(Latent behavioral roles). 文献[130]提出了一种基于合作博弈理论的方法，基于社会网络普遍存在社区结构，利用 Owen 值得到每个节点的边际贡献. 由此，识别社会网络中的关键节点. 而 Lappas 等人[131]提出使用最佳动态规划算法来识别最优效应节点. 效应节点在这里指的是在给定的信息传播模型下，选择一组 k 个活动节点，而这组节点能够很好地解释所观察到的活动状态. Habiba 等人[132]通过对一些网络拓扑结构特征的度量，得出在网络中，简单的局部度量(如节点的度)是判断一个传播阻断角色好坏的最佳指标的结论，并分别在动态网络和静态网络中识别出最佳传播阻断角色(Spread Blockers). Pal 等人[133]为了识别给定主题下最权威的作者角色，首先刻画了一组社会媒体作者的特征，然后通过聚类、排序的过程，最终识别出给定主题下权威作者角色.

事实上，社会网络中的角色除上述这些之外还有许多. 如舆情扩散中的骨干节点[134](Backbone Nodes)、积极价值节点[135]、煽动者[136]角色、网络不实信息(谣言)专家角色[137]、以及水军角色[138]等等. 在当前社会网络角色识别研究中，这些角色虽然不是主要内容，但对我们研究和分析社会网络以及实际的应用都具有重要的意义.

根据其他角色的种类以及主要识别方法、依据，总结本文所述其他角色文献如表 3 所示.

表 3　其他角色种类及主要识别方法、依据汇总

文献	作者(第一)	角色名称	方法或模型	结构	内容	说明
[118-119]	Forestier M	名人	PCA，Meta-criteria		√	主成分分析法、元标准
[120]	Xu K	最有价值用户(顾客)	Optimization models	√		优化模型
[121-122]	Zygmunt A，Gunasekara R C	关键人物	FMO、CPM，Multi-objective	√	√	派系过滤、快速模块化优化、兴趣多目标最优化方法
[123-124]	Zhang G，Liu J	高价值/重要用户	WeiboRank，Machine Learning-based	√	√	WeiboRank、机器学习方法
[125-127]	Goyal A，Tsai M F，Shafiq M Z	领导者、追随者	Frequent pattern Discovery，APPM，LUCI		√	三种识别方法
[128]	Mathioudakis M	发起者、追随者	Random Sampling	√	√	随机抽样
[129]	Stadtfeld C	潜在行为角色	Individual Patterns，Clustering		√	个体模式、聚类
[130]	Wang Xue-Guang	关键节点	Cooperative Games	√		合作博弈论
[131]	Lappas T	最优效应节点	The Optimal DP algorithm	√		最佳动态规划算法
[132]	Yu Y	传播阻断	Topology Structure	√		拓扑结构
[133]	Pal A	最权威作者	Clustering、Sorting		√	聚类、排序
[134]	Sun W	骨干节点	FRT	√		转发关系树算法
[135]	Qiu De-Hong	积极价值节点	QTDG	√		定量时间有向图
[136]	Nakajima S	煽动者	Term Frequency，Cosine-correlation		√	两种分析方法
[137]	Liang C	网络不实信息专家	Tag-based		√	基于标签的方法
[138]	Han Zhong-Ming	水军	WGM		√	隐变量概率图模型

5　研究难点和未来方向

当前，社会网络角色识别研究更多的是集中在新出现的社会网络交互平台上，如微博、Facebook、Twitter、博客等. 由于这些平台中新出现的各种状况和其复杂特性，在实际的研究过程中，还存在一些难以解决的问题以及可进一步研究的方向. 在本文中，我们结合国内外研究现状概况了以下研究难点和未来研究的方向.

5.1　研究难点和方向

根据当前社会网络角色识别的国内外研究现状，总结其研究难点和方向如下：

(1) 如何处理海量和动态变化的社会网络数据的难点问题. 对于大规模在线社会网络的研究，如微博、Facebook，其节点数目往往会达到数以百万的级别，对各种角色识别算法的数据处理提出了更高的要求，不管是从处理的时间方面还是算法的准确率方面. 如果这一问题解决不好，将对社会网络的研究带来巨大的困难. 此外，现实中的这些社会网络数据

是在不断变化着的,我们也只能根据某一时间获得的数据作为识别的依旧,识别出来的角色也只是节点当时或某一时间段内的角色.现有的识别方法也难以解决数据的动态变化对角色影响的问题.因此,在将来的研究中,需要我们寻求解决海量数据问题的策略,研究更好的识别算法及计算方式,或者将现有识别算法改进并应用在分布式平台上,通过并行计算的方式来加以解决.对于数据的动态变化,则需要我们研究如何在识别方法中更好地引入时间因子,从而得到用户角色演化的过程.

(2)综合考虑宏观角度和微观角度的难点问题.在社会网络中,社会角色往往是通过内在的关系来定义,即一个角色仅仅依存于和其他人的关系中,这些人也是同样产生的社会角色.因此,有必要从结合个人行为和整个社会角色的宏观角度出发进行研究[19].在社会网络的非确定角色识别方法中,块模型通常依据社会网络中的关系结构,能够从宏观角度识别社会网络中的角色.而基于主题概率模型在角色识别过程中,则更多关注文本中的关键信息,是从微观角度识别社会网络中的角色,缺乏对全局的认识.因此,如何将两者有效的进行融合是目前研究的一个难点,因为这不仅涉及到识别效率问题,还有二者之间优化、组合的问题.而这两个问题从当前的研究现状来讲,其实都是不容易解决的.

(3)解决角色特征选择局限性的难点问题.当前大多数社会网络角色识别的方法,主要还是根据社会网络中的关系结构(或拓扑结构)、交互内容以及动态行为表现来确定用户的角色.未来还应考虑其他附加维度,如时间,具体的空间位置,用户所属社区的存在和影响等,并进一步强化这些角色识别方法[24].换句话说,用户角色识别的特征选择还具有一定的局限性.在识别过程中,对于更多网络特征并没有进行合理的选择和利用.因此,为了更好地识别社会网络角色,下一步我们可通过使用主成分分析方法或因子图模型(Factor Graph Model)的方法对更多的特征选择计算其贡献值,从而选择贡献较大的特征因子进行分析、研究和识别.总之,我们选择哪些特征会更加有效,及其所选特征如何进行优化组合,这些问题都是我们在具体的社会网络角色识别过程中需要认真考虑和分析的.

(4)处理用户角色环境(或主题)因素依赖性问题.用户角色与环境或主题通常具有强相关性.此处所说的环境或主题,主要指的是角色自身所在的环境或主题,本文中主要指识别的角色所处的社会网络.尤其是社会网络中用户所处环境的不同,因为环境不同,我们扮演的角色就有可能不同.如在现实生活中,当我们处在学校环境中时,对学校来说我们所扮演的是学生或老师的角色.而如果我们去一家商场购物,在这种环境中,对商场来说我们所扮演的就是顾客的角色.同样,在社会网络环境中,我们所识别出的角色其实是在所处网络环境下的一个角色,换一个网络、环境或主题,角色可能就会发生改变.就如同我们在某一社会网络中识别出一用户是影响者角色,在另一个网络或环境中,我们识别出该用户所扮演的可能就是领导者或专家的角色.

在上述非明确角色识别中,块模型识别方法主要是根据拓扑结构进行识别,对主题或环境的因素可以说是几乎没有考虑.虽然许多概率模型识别方法在识别过程中关联到了相关主题,但该角色识别方法则更多的是局限在特定领域(主要为学术领域)的文本数据集中.而在更为常见的明确角色识别方法中,角色的识别也主要是侧重于网络的结构或其交互信息内容的分析以及一些特定的衡量指标.因此,我们认为在社会网络角色识别的研究中,大多数的研究分析中其实很少考虑用户角色的环境(或主题)依赖性.在不同的环境下,用户所扮演的角色是不同的,需要在以后的工作中进一步研究,这也可以是我们未来研究社会网络角色识别中需要注意的重要方面.

(5)解决角色识别方法的评估难点问题.目前,角色识别方法的评估问题一直难以解决,虽然也有人提出,但也只针对某一类型或特定社会网络中的角色识别的方法,甚至很多时候需要通过人工来完成.如论坛中专家角色,评估者通过阅读论坛中发表的帖子,评价识别出的每一个用户的专长[114].因此,还没有一种评估方法,可以对所有角色识别方法进行有效评估,这也是我们未来的研究中一个非常重要的方面.

5.2 其它研究方向

此外,对社会网络角色识别的研究,未来还可在许多更加具体的方向进行研究和应用.如在现有人类行为动力学研究的基础上,进一步研究怎样通过应用群体行为来增加预测群体行为模式的准确度,以及如何将群体行为应用到群体中的角色识别.在文本和意见挖掘技术中,应当包含用户的内容分析,用户角色随时间交互的演化,以及社区对不同角色的影响方式等.在一些情况下,将基于用户活动的扩散历史的识别方法应用到在角色识别中(尤其是在

我们识别影响者角色时).此外,在主题模型中,考虑增加引文信息进一步合并文档,结合文体模型识别文档作者.社会网络中,分析网络演化一是为了规格化描述已发生的事件,二是为了预测分析未来的网络情况.因此,我们在下一步的研究中,可以考虑根据节点的角色,对网络的未来情况进行预测和分析等.

6　结　语

在社会网络中,每个人相对其他人或事物来说,都扮演着所在环境下的一个角色.识别出这些角色,不论是对社会、企业(如市场营销、产品宣传、口碑建立等)还是个人都具有重要实际意义和应用价值.如识别出在线网络中影响者角色,可通过该角色宣传企业的产品和理念.基于角色的相似性,在网络传播信息过程中识别出所扮演的角色,可根据这些角色对用户进行分类.根据用户角色的不同影响,对网络进行监控和舆情的引导等等.此外,识别社会角色对我们研究社会网络也有许多的帮助.如通过对社会网络中用户或节点的角色识别,能够使我们更好的理解该社会网络信息交互情况和网络拓扑结构,了解该网络的时态演变对用户角色形成和变化产生的影响、作用,以及使我们更加容易了解信息在网络中的传播过程等.因此,社会网络中用户角色的识别,不论是在现实社会中,还是在社会网络研究领域,都具有重要意义和价值.目前,大多社会网络角色识别的研究方法是试图去定义具体角色,用不同的角色来刻画如 Facebook、微博、论坛、博客等社会网络.在研究过程中,他们一般会基于这样一种假设,即在社会网络中,不同的社区会有显著不同的行为和意图,而角色能够代表和反映这些显著不同的行为和意图.此外,在具体识别中,即使是相同的社会网络,不同的研究人员会定义不同的角色,不同的命名和特征[63].

本文总体概括了当前社会网络中关于社会网络角色识别的主要方法.其中,大多数方法是根据社会网络的链接分析(网络结构)以及根据内容或用户动态行为表现来识别社会角色.识别方法中,块模型和概率模型方法,反映的是一个更加客观的真实情况,其识别出来角色,更加使人容易接受.而在线社会网络中,针对角色(如专家、影响者)识别的方法则主观性较强,因为这些方法并不总是简单的对具有相似特征的两个用户(或节点)进行排序[4],这些角色取决于他们的兴趣、行为活动、外界认可程度等,而这些特征并没有被所有人以相同的方式定义.社会网络角色识别未来主要的研究方向,其实就是如何继续强化这些识别方法,使其更加全面、有效,以及寻找更加有效的角色识别和评估的方法,并对这些角色识别方法进行更加有效和深入的评估等.如增加特征选择,引入新的特征度量,提高优化组合技术,迁移学习(Transfer Learning)方法的引入,跨领域知识的结合、应用等.

总之,社会网络的角色识别问题是一个值得研究的问题.在识别过程中,根据某一方面或者考虑几种因素识别出来的角色,虽然很多时候能够满足我们一些研究或应用的需要.但更多的是在特定条件或环境下才具有意义,没有哪一种方法或模型是通用或全面的.因此,我们认为社会网络角色的识别是一个复杂的问题,不是单靠某一种方法能解决的,而是需要用"组合拳"方式来解决,这就要求我们综合考虑各种因素进行优化组合,识别出最终的社会角色.

参　考　文　献

[1] Agarwal N, Liu H. Blogosphere: Research issues, tools, and applications. ACM Sigkdd Explorations Newsletter, 2008, 10(1): 18-31

[2] Yang Xiao-Ru. Micro-blog research in the perspective of communication. Contemporary Communications, 2010, (2): 73-74(in Chinese)
(杨晓茹. 传播学视域中的微博研究. 当代传播, 2010, (2): 73-74)

[3] Wang Jun-Chao, Zheng En. "Micro Communication" and the right to expression-on freedom of expression microblogging era. Modern Communications: Journal of Communication University of China, 2011, (4): 80-85(in Chinese)
(王君超, 郑恩. "微传播"与表达权——试论微博时代的表达自由. 现代传播: 中国传媒大学学报, 2011, (4): 80-85)

[4] Forestier M, Stavrianou A, Velcin J, et al. Roles in social networks: Methodologies and research issues. Web Intelligence & Agent Systems, 2012, 10(1): 117-133

[5] Liu Jian-Ming. Public Opinion Communication. Beijing: Tsinghua University Press, 2001(in Chinese)
(刘建明. 舆论传播. 北京: 清华大学出版社, 2001)

[6] Fagnan J, Rabbany R, Takaffoli M, et al. Community dynamics: Event and role analysis in social network analysis. Lecture Notes in Computer Science, 2014, 8933: 85-97

[7] Newman M E J. The structure and function of complex networks. SIAM Review, 2003, 45(2): 167-256

[8] Rainie L, Wellman B. Networked: The New Social Operating System. Massachusetts, American: MIT Press, 2012

[9] Scott J. Social Network Analysis: A Handbook. California, American: SAGE Publication Ltd., 2000

[10] Rapoport A. Contribution to the theory of random and biased nets. Bulletin of Mathematical Biology, 1957, 19(4): 257-277

[11] Travers J, Milgram S. An experimental study of the small world problem. Sociometry, 1969, 32(4): 425-443

[12] Guare J. Six Degrees of Separation: A play. Vintage Books. New York, USA: Vintage Books Press, 1990

[13] Qin Qi-Wen. Introduction to Role Theory. Beijing: China Social Sciences Publishing House, 2011(in Chinese)
(秦启文. 角色学导论. 北京: 中国社会科学出版社, 2011)

[14] Huang Ling-He, Zhu Qing-Hua. Identification of online community user types and relationships from the perspective of social roles. Information and Documentation Services, 2013, (2): 84-88(in Chinese)
(黄令贺, 朱庆华. 社会角色视角下网络社区用户类型及其关系的识别. 情报资料工作, 2013, (2): 84-88)

[15] Callero P L. From role-playing to role-using: Understanding role as resource. Social Psychology Quarterly, 1994, 57(3): 228-243

[16] Jahnke I. Dynamics of social roles in a knowledge management community. Computers in Human Behavior, 2010, 26(4): 533-546

[17] Baker W E, Faulkner R R. Role as resource in the Hollywood film industry. American Journal of Sociology, 1991, 97(2): 279-309

[18] Junquero-Trabado V, Dominguez-Sal D. Building a role search engine for social media//Proceedings of the 21st International Conference Companion on World Wide Web. Lyon, France, 2012: 1051-1060

[19] Valls-Vargas J, Zhu J, Ontanon S. Toward automatic role identification in unannotated folk tales//Proceedings of the 10th Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE). Raleigh, USA, 2014: 188-194

[20] Gleave E, Welser H T, Lento T M, et al. A conceptual and operational definition of social role' in online community//Proceedings of the 42nd Hawaii International Conference on Systems Science (HICSS'09). Los Alamitos, USA, 2009: 1-11

[21] Lee A J T, Yang Fu-Chen, et al. Discovering content-based behavioral roles in social networks. Decision Support Systems, 2014, 59(1): 250-261

[22] Himelboim I, Fisher D, Gleave E, et al. Reply magnets in online political discussions: Analysis of six months of discussion in 20 Usenet political newsgroups//Association for Education in Journalism and Mass Communication 58th Annual Convention. Washington, USA, 2007: 9-12

[23] Fournier S, Lee L. Getting brand communities right. Harvard Business Review, 2009, 87(4): 105-111

[24] Akaka M A, Chandler J D. Roles as resources: A social roles perspective of change in value networks. Marketing Theory, 2011, 11(3): 243-260

[25] Laurent A, Camelin N, Raymond C. Boosting bonsai trees for efficient features combination: Application to speaker role identification//Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Singapore, 2014: 76-80

[26] Borgatti S P, Everett M G. Notions of Position in Social Network Analysis. Sociological Methodology, 1992, 22(4): 1-35

[27] Handcock M S, Raftery A E, Tantrum J M. Model-based clustering for social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society), 2007, 170(2): 301-354

[28] Doreian P, Batagelj V, Ferligoj A. Generalized Blockmodeling. Cambridge, UK: Cambridge University Press, 2005

[29] White D R, Reitz K P. Graph and semigroup homomorphisms on networks of relations. Social Networks, 1983, 5(2): 193-234

[30] Borgatti S P, Everett M G. Two algorithms for computing regular equivalence. Social Networks, 1993, 15(4): 361-376

[31] Breiger R L, Boorman S A, Arabie P. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. Journal of Mathematical Psychology, 1975, 12(3): 328-383

[32] Fienberg S E, Wasserman S S. Categorical data analysis of single sociometric relations. Sociological Methodology, 1981, 49(12): 156-192

[33] Holl P W, Leinhardt S. An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association, 1981, 76(373): 33-50

[34] Wasserman S, Anderson C. Stochastic a posteriori blockmodels: Construction and assessment. Social Networks, 1987, 9(1): 1-36

[35] Snijders T A B, Nowicki K. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. Journal of Classification, 1996, 14(1): 75-100

[36] Wolfe A P, Jensen D. Playing multiple roles: Discovering overlapping roles in social networks//Proceedings of the ICML-04 Workshop on Statistical Relational Learning and Its Connections to Other Fields. Banff, Canada, 2004: 75

[37] Airoldi E M, Blei D M, Fienberg S E, et al. Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 2015, 9(5): 1981-2014

[38] Fu Wenjie, Song Le, Xing E P. Dynamic mixed membership blockmodel for evolving networks//Proceedings of the 26th International Conference on Machine Learning. Montreal, Canada, 2009: 329-336

[39] Shi Jing, Hu Ming, Shi Xin, et al. Text segmentation based on model LDA. Chinese Journal of Computers, 2008, 31(10): 1865-1873(in Chinese)

(石晶，胡明，石鑫等. 基于 LDA 模型的文本分割. 计算机学报，2008，31(10)：1865-1873)

[40] Steyvers M，Griffiths T. Probabilistic topic models. Handbook of Latent Semantic Analysis，2007，427(7)：424-440

[41] Rosen-Zvi M，Griffiths T，Steyvers M，et al. The author-topic model for authors and documents//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff，Canada，2004：487-494

[42] Blei D M，Ng A Y，Jordan M I. Latent Dirichlet allocation. Journal of Machine Learning Research，2003，3：993-1022

[43] Griffiths T L，Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America，2004，101(Supplement 1)：5228-5235

[44] McCallum A，Wang X，Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on enron and academic email. Journal of Artificial Intelligence Research，2007，30(1)：249-272

[45] Daud A，Li J，Zhou L，et al. Conference Mining via Generalized Topic Modeling//Proceedings of the European Conference on Machine Learning & Knowledge Discovery in Databases. Bled，Slovenia，2009：244-259

[46] Daud A，Li J，Zhou L，et al. A generalized topic modeling approach for maven search//Li Qing，Feng Ling，Pei Jian，et al，eds. Advances in Data and Web Management. Berlin Heidelberg，Germany：Springer，2009：138-149

[47] Nitzan I，Libai B. Social effects on customer retention. Journal of Marketing，2011，75(6)：24-38

[48] Hinz O，Schulze C，Takac C. New product adoption in social networks：Why direction matters. Journal of Business Research，2014，67(1)：2836-2844

[49] Probst F，Grosswiele D K L，Pfleger D K R. Who will lead and who will follow：Identifying influential users in online social networks. Business & Information Systems Engineering，2013，5(3)：179-193

[50] Kayes I，Qian X，Skvoretz J，et al. How influential are you：Detecting influential bloggers in a blogging community//Aberer K，Flache A，Jager W，et al，eds. Social Informatics. Berlin Heidelberg，Germany：Springer，2012：29-42

[51] Wu X，Wang J. Micro-blog in China：Identify influential users and automatically classify posts on Sina micro-blog. Journal of Ambient Intelligence & Humanized Computing，2014，5(1)：51-63

[52] Weng J，Lim E P，Jiang J，et al. TwitterRank：Finding topic-sensitive influential twitterers//Proceedings of the 3rd ACM International Conference on Web Search and Web Data Mining(WSDM 2010). New York，USA，2010：261-270

[53] Zhang Yifeng，Li Xiaoqing，Wang Tewei. Identifying influencers in online social networks：The Role of tie strength. International Journal of Intelligent Information Technologies，2013，9(1)：1-20

[54] Gliwa B，Zygmunt A. Finding influential bloggers. International Journal of Machine Learning and Computing，2015，5(2)：127

[55] Tang X，Yang C C. Identifying influential users in an online healthcare social network//Proceedings of the 2010 IEEE International Conference on Intelligence and Security Informatics (ISI). British Columbia，Canada，2010：43-48

[56] Li Yung-Ming，Lai Cheng-Yang，Chen Ching-Wen. Identifying bloggers with marketing influence in the blogosphere//Proceedings of the 11th International Conference on Electronic Commerce. Taipei，China，2009：335-340

[57] Aziz M，Rafi M. Identifying influential bloggers using blogs semantics//Proceedings of the 8th International Conference on Frontiers of Information Technology. Islamabad，Pakistan，2010：7

[58] Moon E，Han S. A qualitative method to find influencers using similarity-based approach in the blogosphere//Proceedings of the IEEE International Conference on Social Computing/IEEE International Conference on Privacy，Security，Risk and Trust. Minneapolis，USA，2010：225-232

[59] Bui D L，Nguyen T，Ha Q T. Measuring the influence of bloggers in their community based on the h-index family//van Do T，An Le Thi H，Nguyen N T eds. Advanced Computational Methods for Knowledge Engineering. New York，USA：Springer International Publishing，2014：313-324

[60] Akritidis L，Katsaros D，Bozanis P. Identifying influential bloggers：Time does matter//Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies(WI-IAT'09). Milano，Italy，2009，1：76-83

[61] Moh T S，Shola S P. New factors for identifying influential bloggers//Proceedings of the 2013 IEEE International Conference on Big Data. Santa Clara，USA，2013：18-27

[62] Khan H U，Daud A，Malik T A. MIIB：A metric to identify top influential bloggers in a community. Plos One，2015，10(9)：e0138359

[63] Ding Z，Jia Y，Zhou B，et al. Mining topical influencers based on the multi-relational network in micro-blogging sites. Wireless Communication Over Zigbee for Automotive Inclination Measurement China Communications，2013，10(1)：93-104

[64] Akritidis L，Katsaros D，Bozanis P. Identifying the productive and influential bloggers in a community. IEEE Transactions on Systems Man & Cybernetics-Part C：Applications & Reviews，2011，41(5)：759-764

[65] Shalaby M，Rafea A. Identifying the Topic-Specific Influential Users and Opinion Leaders in Twitter. Calgary，Canada：Acta Press，2013

[66] Caiv Y，Chen Y. Mining influential bloggers：From general to domain specific//Velásquez J D，Rios S A，Howlett R J，Jain L C eds. Knowledge-Based and Intelligent Information and Engineering Systems. Berlin Heidelberg，Germany：Springer，2009：447-454

[67] Liu L，Tang J，Han J，et al. Mining topic-level influence in heterogeneous networks//Proceedings of the ACM International Conference on Information & Knowledge Management.

Toronto, Canada, 2010: 199-208

[68] Agarwal N, Liu H, Tang L, et al. Identifying the influential bloggers in a community//Proceedings of the 2008 International Conference on Web Search and Data Mining. Stanford, USA, 2008: 207-218

[69] Rabade R, Mishra N, Sharma S. Survey of influential user identification techniques in online social networks//Thampi S M, Abraham A, Pal S K, Rodriguez J M C eds. Recent Advances in Intelligent Informatics. New York, USA: Springer International Publishing, 2014: 359-370

[70] Luo Jing, Xu Lizhen. Identification of microblog opinion leader based on user feature and interaction network//Proceedings of the 2014 11th Web Information System and Application Conference (WISA). Tianjin, China, 2014: 125-130

[71] Duan J, Zeng J, Luo B. Identification of opinion leaders based on user clustering and sentiment analysis//Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01. Warsaw, Poland, 2014: 377-383

[72] Hudli S, Hudli A, Hudli A V. Identifying online opinion leaders using *K*-means clustering//Proceedings of the 12th Intelligent Systems Design and Applications Conference. Kochi, India, 2012: 416-419

[73] Hung C, Yeh P W. Identification of opinion leaders using text mining technique in virtual community//Proceedings of the 1st Symposium on Information Management and Big Data. Cusco, Peru, 2014: 8-13

[74] Wang Jun-Ze, Wang Ya-Lei, Yu Hang, et al. Research of model on micro blog opinion leader identification. Journalism & Communication, 2011, (6): 81-88(in Chinese)
(王君泽, 王雅蕾, 禹航等. 微博客意见领袖识别模型研究. 新闻与传播研究, 2011, (6): 81-88)

[75] Zhang Wei-Zhe, Zhang Hong, Liu Xin-Ran, et al. Identifying public opinion leaders in network forum based on content ladder evaluation algorithm. Journal of Computer Research and Development, 2012, 49(S2): 145-152(in Chinese)
(张伟哲, 张鸿, 刘欣然等. 基于语料阶梯评价的互联网论坛舆论领袖筛选算法. 计算机研究与发展, 2012, 49(S2): 145-152)

[76] Liu Zhi-Ming, Liu Lu. Identification and analysis of opinion leaders in micro blog public opinion network. Systems Engineering, 2011, 29(6): 8-16(in Chinese)
(刘志明, 刘鲁. 微博网络舆情中的意见领袖识别及分析. 系统工程, 2011, 29(6): 8-16)

[77] Cheng F, Yan C, Huang Y, et al. Algorithm of identifying opinion leaders in BBS//Proceedings of the 2nd Cloud Computing and Intelligent Systems Conference. Hangzhou, China, 2012: 1149-1152

[78] Song Zhao-Jun, Dai Hang, Huang Dong-Xu. An algorithm for identifying and researching on opinion leaders in the blogosphere. Microprocessors, 2012, 33(6): 37-40(in Chinese)
(宋昭君, 戴航, 黄东旭. 一种鉴别博客空间意见领袖的算法研究. 微处理机, 2012, 33(6): 37-40)

[79] Bodendorf F, Kaiser C. Detecting opinion leaders and trends in online social networks//Proceedings of the 2nd ACM Workshop on Social Web Search and Mining. Hong Kong, China, 2009: 65-68

[80] Zhang W, Li X, He H, et al. Identifying network public opinion Leaders based on markov logic networks. Scientific World Journal, 2014, 2014(3): 435-444

[81] Fan Xing-Hua, Zhao Jing, Fang Bin-Xing, Li Yu-Xiao. Influence diffusion probability mode and utilizing it to identify network opinion leader. Chinese Journal of Computers, 2013, 36(2): 360-367(in Chinese)
(樊兴华, 赵静, 方滨兴, 李欲晓. 影响力扩散概率模型及其用于意见领袖发现研究. 计算机学报, 2013, 36(2): 360-367)

[82] Wang Hong-Rui. The model of micro blog space opinion leader identification based on social network analysis method. News World, 2013, (4): 219-221(in Chinese)
(王红瑞. 基于社会网络分析法的微博空间意见领袖识别模型. 新闻世界, 2013(4): 219-221)

[83] Yin Yan-Teng, Li Xue-Ming, Cai Meng-Song. Mining method of microblog opinion leader based on user relationship and attribute. Computer Engineering, 2013, 39(4): 184-189(in Chinese)
(尹衍腾, 李学明, 蔡孟松. 基于用户关系与属性的微博意见领袖挖掘方法. 计算机工程, 2013, 39(4): 184-189)

[84] Song X, Chi Y, Hino K, et al. Identifying opinion leaders in the blogosphere//Proceedings of the 16th ACM Conference on Information and Knowledge Management. Lisbon, Portugal, 2007: 971-974

[85] Song Kaisong, Wang Daling, Feng Shi, et al. Detecting opinion leader dynamically in Chinese news comments. Lecture Notes in Computer Science, 2011, 7142: 197-209

[86] Zhou H, Zeng D, Zhang C. Finding leaders from opinion networks//Proceedings of the 2009 Intelligence and Security Informatics conference. Texas, USA, 2009: 266-268

[87] Ning Lian-Ju, Wan Zhi-Chao. Identifying network opinion leaders based on group-buying product reviews. Journal of Intelligence, 2013, 32(8): 204-206(in Chinese)
(宁连举, 万志超. 基于团购商品评论的网络意见领袖识别. 情报杂志, 2013, 32(8): 204-206)

[88] Su Shu-Qing, Yang Kai, Zhang Ning. Comparative research on LeaderRank and PageRank algorithms. Information Technology, 2015, (4): 8-11(in Chinese)
(苏树清, 杨凯, 张宁. LeaderRank 与 PageRank 算法比较研究. 信息技术, 2015, (4): 8-11)

[89] Gleich D F. PageRank beyond the Web. Siam Review, 2014, 57(3): 321-363

[90] Lv L, Zhang Y C, Yeung C H, et al. Leaders in social networks, the delicious case. Plos One, 2011, 6(6): e21202

[91] Xu Jun-Ming, Zhu Fu-Xi, Liu Shi-Chao, et al. Identifying opinion leaders by improved algorithm based on LeaderRank. Computer Engineering and Applications, 2015, 51(1): 110-114(in Chinese)
(徐郡明，朱福喜，刘世超等. 改进 LeaderRank 算法的意见领袖挖掘. 计算机工程与应用，2015，51(1)：110-114)

[92] Xiao Yu, Xia Lin. Understanding opinion leaders in bulletin board systems: Structures and algorithms//Proceedings of the IEEE Local Computer Network Conference. Denver, Colorado, USA, 2010: 1062-1067

[93] Xiao Yu, Xu Wei, Xia Lin. Networking groups opinion leader identification algorithms based on sentiment analysis. Computer Science, 2012, 39(2): 34-37(in Chinese)
(肖宇，许炜，夏霖. 一种基于情感倾向分析的网络团体意见领袖识别算法. 计算机科学，2012，39(2)：34-37)

[94] Xu Hui-Jie, Cai Wan-Dong, Wang Jian-Ping, et al. Identifying algorithm for opinion leaders of forums based on time-varying graphs. Computer Science, 2012, 39(9): 51-54(in Chinese)
(徐会杰，蔡皖东，王剑平等. 基于时间变化图的网络论坛意见领袖识别算法. 计算机科学，2012，39(9)：51-54)

[95] Volpentesta A P, Felicetti A M. Identifying opinion leaders in time-dependent commercial social networks//Camarinha-Matos L M, Xu Lai, Afsarmanesh H eds. Collaborative Networks in the Internet of Services. Berlin Heidelberg, Germany: Springer, 2012: 571-581

[96] Cai Shu-Qin, Ma Yu-Tao, Wang Rui. Study on the method of identifying opinion leaders for online word-of-mouth communication. Chinese Journal of Management Science, 2013, 21(2): 185-192(in Chinese)
(蔡淑琴，马玉涛，王瑞. 在线口碑传播的意见领袖识别方法研究. 中国管理科学，2013，21(2)：185-192)

[97] Chen Y, Wang X, Tang B, et al. Identifying opinion leaders from online comments//Huang Heyan, Liu Ting, Zhang Hua-Ping, Tang Jie eds. Social Media Processing. Berlin Heidelberg, Germany: Springer, 2014: 231-239

[98] Mimno D, Mccallum A. Expertise modeling for matching papers with reviewers//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 500-509

[99] Pal A, Konstan J A. Expert identification in community question answering: Exploring question selection bias//Proceedings of the 19th ACM Conference on Information and Knowledge Management(CIKM 2010). Toronto, Canada, 2010: 1505-1508

[100] Pal A, Harper F M, Konstan J A. Exploring question selection bias to identify experts and potential experts in community question answering. ACM Transactions on Information Systems, 2012, 30(2): 10

[101] Balog K, Azzopardi L, Rijke M D. A language modeling framework for expert finding. Information Processing & Management, 2009, 45(1): 1-19

[102] Fu Y, Xiang R, Liu Y, et al. A CDD-based formal model for expert finding//Proceedings of the International Conference on Information & Knowledge Management. Lisbon, Portugal, 2007: 881-884

[103] Wu H, Pei Y, Yu J. Hidden topic analysis based formal framework for finding experts in metadata corpus//Proceedings of the 8th IEEE/ACIS International Conference on Computer and Information Science. Shanghai, China, 2009: 369-374

[104] Petkova D, Croft W B. Hierarchical language models for expert finding in enterprise corpora. IEEE Computer Society, 2006, 17(1): 599-608

[105] Serdyukov P, Rode H, Hiemstra D. Modeling multi-step relevance propagation for expert finding//Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008). Napa Valley, USA, 2008: 1133-1142

[106] Hofmann T. Probabilistic latent semantic indexing//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, USA, 1999: 50-57

[107] Tang J, Jin R, Zhang J. A topic modeling approach and its integration into the random walk framework for academic search//Proceedings of the IEEE International Conference on Data Mining. Pisa, Italian, 2008: 1055-1060

[108] Jung H, Lee M, Kang I S, et al. Finding topic-centric identified experts based on full text analysis//Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics. Busan, Korea, 2007: 56-63

[109] Adamic L A, Zhang J, Bakshy E, et al. Knowledge sharing and Yahoo answers: Everyone knows something//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 665-674

[110] Liu X, Bollen J, Nelson M L, et al. Co-authorship networks in the digital library research community. Information Processing & Management, 2005, 41(6): 1462-1480

[111] Jurczyk P, Agichtein E. Hits on question answer portals: Exploration of link analysis for author ranking//Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, Netherlands, 2007: 845-846

[112] Fiala D, Rousselot F, Ježek K. PageRank for bibliographic networks. Scientometrics, 2008, 76(1): 135-158

[113] Ding Y, Yan E, Frazho A, et al. PageRank for ranking authors in co-citation networks. Journal of the American Society for Information Science & Technology, 2010, 60(11): 2229-2243

[114] Zhang J, Ackerman M S, Adamic L. Expertise networks in online communities: Structure and algorithms. International Conference on World Wide Web. Banff, Canada, 2007: 221-230

[115] Lin L, Xu Z, Ding Y, et al. Finding topic-level experts in scholarly networks. Scientometrics, 2013, 97(3): 797-819

[116] Campbell C S, Maglio P P, Cozzi A, et al. Expertise identification using email communications//Proceedings of the 12th International Conference on Information and Knowledge Management. New Orleans, USA, 2003: 528-531

[117] Zhang Jing, Tang Jie, Li Juanzi. Expert finding in a social network//Kotagiri R, Krishna P R, Mohania M, Nantajeewarawat E eds. Advances in Databases: Concepts, Systems and Applications. Berlin Heidelberg, Germany: Springer, 2007: 1066-1069

[118] Forestier M, Velcin J, Zighed D A. Analyzing social roles using enriched social network on online sub-communities//Proceedings of the 6th International Conference on Digital Society. Valencia, Spain, 2012: 17-22

[119] Forestier M, Velcin J, Stavrianou A, et al. Extracting celebrities from online discussions//Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining. Istanbul, Turkey, 2012: 322-326

[120] Xu Kaiquan, Li Jiexun, Song Yuxia. Identifying valuable customers on social networking sites for profit maximization. Expert Systems with Applications, 2012, 39(17): 13009-13018

[121] Zygmunt A, Bródka P, Kazienko P, et al. Different approaches to groups and key person identification in blogosphere//Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Kaohsiung, China, 2011: 593-598

[122] Gunasekara R C, Mehrotra K, Mohan C K. Multi-objective optimization to identify key players in social networks//Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining(ASONAM). Beijing, China, 2014: 443-450

[123] Zhang G, Bie R. Discovering massive high-value users from Sina weibo based on quality and activity//Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. Beijing, China, 2013: 214-220

[124] Liu J, Cao Z, Cui K, et al. Identifying important users in Sina microblog//Proceedings of the International Conference on Multimedia Information Networking and Security (MINES). Nanjing, China, 2012: 839-842

[125] Goyal A, Bonchi F, Lakshmanan L V S. Discovering leaders from community actions//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Valley, USA 2008: 499-508

[126] Tsai M F, Tzeng C W, Lin Z L, et al. Discovering leaders from social network by action cascade. Social Network Analysis and Mining, 2014, 4(1): 1-10

[127] Shafiq M Z, Ilyas M U, Liu A X, et al. Identifying leaders and followers in online social networks. IEEE Journal on Selected Areas in Communications, 2013, 31(9): 618-628

[128] Mathioudakis M, Koudas N. Efficient identification of starters and followers in social media//Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. Saint Petersburg, Russia, 2009: 708-719

[129] Stadtfeld C. Discovering latent behavioral roles in dynamic social networks//Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust. Amsterdam, Netherlands, 2012: 629-635

[130] Wang Xue-Guang. Discovering Critical Nodes in Social Networks Based on Cooperative Games. Computer Science, 2013, 40(4): 155-159(in Chinese)
(王学光. 基于合作博弈论的社会网络关键节点发现研究. 计算机科学, 2013, 40(4): 155-159)

[131] Lappas T, Terzi E, Gunopulos D, et al. Finding effectors in social networks//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1059-1068

[132] Yu Y, Berger-Wolf T Y, Saia J. Finding spread blockers in dynamic networks//Giles L, Smith M, Yen J, Zhang H eds. Advances in Social Network Mining and Analysis. Berlin Heidelberg, Germany: Springer, 2010: 55-76

[133] Pal A, Counts S. Identifying topical authorities in microblogs//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 45-54

[134] Sun W, Zheng D, Hu X, et al. Microblog-oriented backbone nodes identification in public opinion diffusion//Proceedings of the 2014 International Conference on Audio, Language and Image Processing (ICALIP). Shanghai, China, 2014: 570-573

[135] Qiu De-Hong, Li Hao, Li Yuan. Identification of active valuable nodes in temporal online social network with attributes. International Journal of Information Technology & Decision Making, 2014, 13(4): 839-864

[136] Nakajima S, Tatemura J, Hara Y, et al. Identifying agitators as important blogger based on analyzing blog threads. Frontiers of WWW Research and Development-APWeb 2006, Berlin Heidelberg, Germany: Springer, 2006: 285-296

[137] Liang C, Liu Z, Sun M. Expert finding for microblog misinformation identification//Proceedings of the International Conference on Computational Linguistics (Posters). Mumbai, India, 2012: 703-712

[138] Han Zhong-Ming, Xu Feng-Min, Duan Da-Gao. Probabilistic graphical model for identifying water army in microblogging system. Journal of Computer Research and Development, 2013, 50(S2): 180-186(in Chinese)
(韩忠明, 许峰敏, 段大高. 面向微博的概率图水军识别模型. 计算机研究与发展, 2013, 50(S2): 180-186)

ZHANG Shu-Sen, born in 1988, Ph. D. candidate. His research interests include data mining, social computing.

LIANG Xun, born in 1965, professor, Ph. D. supervisor. His research interests include Internet information analysis, data mining, business intelligence, and social computing.

QI Jin-Shan, born in 1977, Ph. D. candidate, lecturer. His research interests include data mining, social computing.

Background

In social networks, we can say that every person plays a role in one's environment relative to other people or things. The interaction level among users defines the appearance of several social roles which can be characterized as positions, behaviors, or virtual identities, and they keep changing and evolving over time. It has practical significance and application value to identify roles in social networks for society, enterprise (E.g., marketing, publicity, building word of mouth, etc.), person, and social networks research.

Identifying roles is also an important research issue in social networks. Research the identification of the role in social networks offers a way for us to help understand the social network topology and its evolution. So we write this survey paper to show a full view for academicians about work of identifying the role in social networks. This paper introduces related research and state-of-the-art approaches from two different angles. The first one is the non-explicit roles and the second is the explicit roles. At present, most methods are based on the link analysis (network structure) of social networks and the content or the performance of user's dynamic behavior to identify the social roles. We think it is a complex problem to identify roles in social networks, not rely on a particular method can be solved, but need to use the "combined fist" approach to resolve.

This paper is supported by the National Natural Science Foundation of China under Grant Nos. 71531012, 71271211, the Beijing Natural Science Foundation under Grant No. 4172032, the Natural Science Foundation of Renmin University of China under Grant No. 10XNI029, the Beijing High School Youth Talent Plan under Grant No. 21147514040.