社交网络研究综述

李立耀[1,2] 孙鲁敬[2,3] 杨家海[2,3]
(福建师范大学福清分校电子与信息工程学院 福清 350300)[1]
(清华大学网络科学与网络空间研究院 北京 100084)[2]
(清华大学清华信息科学与技术国家实验室 北京 100084)[3]

摘 要 社交网络已成为 Web2.0 时代最流行的应用,其服务范围已逐步从社交关系管理扩展到媒体信息、应用集成、电子商务等领域。社交网络中大量的活跃用户为研究网络行为、数据安全、信息传播以及其他跨学科问题提供了宝贵的数据和场景。自 Facebook 出现以来,研究者先后从不同的角度对社交网络进行了大量的研究,这些研究对人们认识社交网络内部规律、促进 ICP 服务改进具有重大意义。首先对社交网络的发展进行了简单的回顾;然后从社交网络的数据采集技术、社交网络用户行为分析、社交网络中的信息传播及社交网络中的用户隐私 4 个方面对已有的研究工作进行总结评价;最后,总结了当前研究中出现的问题并对未来研究发展趋势进行了展望。希望能为该领域的研究者提供一些有益的启示。

关键词 社交网络,数据测量,用户行为分析,信息传播,用户隐私
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.002

Research on Online Social Network

LI Li-yao[1,2] SUN Lu-jing[2,3] YANG Jia-hai[2,3]
(School of Electronic and Information Engineering, Fuqing Branch of Fujian Normal University, Fuqing 350300, China)[1]
(The Institute of Network Science and Cyberspace, Tsinghua University, Beijing 100084, China)[2]
(Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China)[3]

Abstract The online social network (OSN) has become the most popular application in the Web2.0 age. Its services have extended from social relationship management to media information, application integration and e-commerce. With its huge amounts of active users, OSN has played a critical role in the academic research area such as online behavior, data security, information diffusion and other interdisciplinary studies. Since the appearance of Facebook, various research works have been emerging, which help us look inside the OSN. We firstly summarized the development of OSN and its challenges. Then we categorized and evaluated the research work in four aspects including measurement technology, user behaviors analysis, information dissemination and user privacy. Finally, we summarized the problems of current work and the prospects of future research topics. The purpose of this paper is to provide some enlightenment for researches of this field.

Keywords OSN, Data measurement, User behavior analysis, Information dissemination, User privacy

社交网络是在线社交网络(Online Social Network,OSN)的简称。社交网络服务是基于六度分隔理论,以互动交友、用户之间共同的兴趣、爱好、活动或者用户间真实的人际关系为基础,以实名或者非实名的方式在网络平台上构建的一种社会关系网络服务。目前社交网络主要以综合性的 Web 站点作为实现形式,向用户提供在线个人信息管理服务、人际关系管理服务,以及多模式的信息交流服务。2004 年 Facebook 上线,它基于真实的人际关系网向用户提供综合性的社交服务,被认为是第一个真正意义上的社交网站。自此以后,社交网络快速发展,综合性社交社区、专注特定领域的垂直社区、以信息流为主的社区相继出现。当今热门的 Facebook、Twitter、LinkedIn、微博客、人人网、开心网都属于社交网络,他们为社交参与者为建立、拓展、维系各类人际关系而进行的个人展示、互动交流、休闲娱乐等活动提供交流平台。截止到 2012 年 8 月,世界上最大的社交网站 Facebook 拥有注册用户量约 10 亿,其网络流量曾一度超过网络巨头 Google;Sina 微博的最新注册用户量已达到了 3 亿;人人网用户量在 2 亿左右。

到稿日期:2014-11-26 返修日期:2015-03-22 本文受国家重点基础研究发展计划(2012CB315806),国家自然科学基金(61170211,61202356,61161140454),博士点基金(20110002110056,20130002110058),福建省教育厅科技项目(JA12352)资助。
李立耀(1970-),男,硕士,副教授,CCF 会员,主要研究领域为云计算、计算机网络;孙鲁敬(1988-),男,硕士,主要研究领域为网络测量、社交网络测量及用户行为分析;杨家海(1966-),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络、网络管理与测量、协议工程学,E-mail:yang@cernet.edu.cn(通信作者)。

社交网络的深入发展让其本身变得异常复杂。当前的社交网络常常拥有上亿级别的用户，社交图谱变得异常庞大，大量用户持续交互，各种信息在网络中快速流转，这些特点给相关研究者在深入研究社交网络规律等工作上带来了巨大的挑战。

(1)从数据角度来讲，社交网络每天产生大量的结构化或非结构化的文本数据、多媒体数据。数据的获取是学术界分析用户行为等一切研究工作的基础。目前多数 ICP 向第三方开发商提供 API 接口，但由于缺乏有效的数据共享机制或是出于商业机密的考虑，ICP 对于学术界共享数据的工作还远远不够，这给社交网络的学术研究工作带来了巨大的困难。研究者通常的做法是使用基于 API 或基于网页的传统爬虫来获取数据，但由于 ICP 在访问策略上的控制，学术界在数据集上面临着数据不完整、数据有偏差、爬取周期长、消耗资源过多等挑战。

(2)从用户角度看，活跃用户是社交网络的核心，它们主导整个网络的交互行为。用户行为分析是认识社交网络的起点。然而社交网络用户组成结构复杂，具有多样文化背景，这造成了不同的社交网络用户在社交关系建立、社交性行为、内容分享等方面都有各自鲜明的特点。这一领域面临的挑战在于研究者既需要对社交网络用户进行宏观性分析，又需要找到有效的方法对不同群体在显式行为、隐式行为、行为动机进行深入的研究。

(3)从信息传播角度看，社交网络特别是微博这样具有强媒体性质的社交网络已经成为个人、机构等进行舆论表达、信息发布、营销等的主要场所。不同的社交网络因侧重点以及用户结构的差异而导致其在信息传播速度、时效性、传播范围等方面都有很大的区别。学者们如何对大量异构社交网络的信息传播规律进行分析，进而总结归纳出有效的信息传播和预测模型是这方面面临的挑战。

(4)从隐私保护角度看，用户个人信息通过互联网传播。用户信息泄露、钓鱼攻击、僵尸用户操作舆论、垃圾信息泛滥等问题已严重影响了用户的社交体验，并成为用户流失的重要原因。研究者如何定义现有的安全现状，提出更加精确的僵尸用户检测算法，乃至提出新的隐私保护模型是这一领域存在的挑战。

针对社交网络领域面临的挑战，本文主要总结了学术界在社交网络测量、用户行为分析、信息传播、用户隐私 4 个领域的研究成果。

(1)社交网络测量的研究工作。社交网络数据采集是一切研究的基础。早期的研究工作包括使用网页爬虫来对社交网络进行全网数据抓取[1,2]，随着数据规模的扩大，研究者把重心转向了数据采样算法[3-6]，并在对隐式用户分析的需求的驱动下提出了基于点击流的数据获取模型[7,8]。

(2)用户行为分析的研究工作。研究初期，学者们主要关注社交图谱分析[9,10]和显性用户行为分析[11]。社交图谱分析从图论的角度来研究整个社交网络的宏观性质，而显性用户行为分析则基于留言或者评论等信息来分析用户进行社交活动的规律。点击流数据模型出现以后，研究者开始对页面浏览等隐式用户行为进行分析，至此用户行为分析才算完整。

(3)信息传播领域的研究工作。学术界主要关注信息传播规律发现、信息传播和预测模型构建等问题 。一些研究者对 Blog、Twitter、Flicker 这些不同性质的社交网络信息传播规律进行了对比研究[12-14]，并提出类似于病毒感染、新事物传播等具体模型来预测信息走向[15,16]，这些工作为最大化信息影响力提供了有力的理论依据。

(4)用户隐私的研究工作。为了定义隐私方面存在的问题，部分学者关注社交网络中的隐私现状，并对一些社交网络权限模型的不足进行分析[17-20]。针对这些不足或问题，一些学者提出了具有针对性的解决方案，如构建 Sybil 用户的检测算法[21,22]、提出基于公私钥体系的新权限模型[23]等。

本文接下来将对上述研究领域进行详细的阐述。第 1 节将介绍社交网络数据测量领域主要的采样算法和点击流数据模型；第 2 节介绍用户行为分析的最新研究成果；第 3 节主要关注信息传播的研究；第 4 节总结用户隐私相关的研究成果；第 5 节对已有的研究工作进行总结展望。

1 社交网络测量

社交网络的数据获取是一切研究的基础。由于缺乏有效的数据共享机制或是出于商业机密的考虑，ICP 对数据的开放很谨慎。学术界在社交网络领域研究面临的主要困难之一就是缺乏可用的数据集。社交网络的飞速发展使得社交网络中的图谱数据、内容数据迅速地膨胀，不同的测量技术相继出现。

全网测量技术出现在社交网络初期，爬虫系统使用广度优先(BFS)算法对社交图谱和节点信息进行采集。区域测量则针对社交网络特定划分的子图进行数据采集，如对社交网络中某学校的用户进行采集。这种测量方法避免了全网采集的巨大开销，采集的数据也具有一定的代表性。采样算法是针对全图的均匀采样，获得的数据可以估算某些全局的信息，是当前大规模社交网络研究中最常用的测量方法。

上述测量方法总的特点是利用 Web 爬虫或者 API 接口向社交网络平台进行数据请求，它们常常受到 ICP 在访问策略上的限制，就算是采样测量，获取可观的数据量也需要巨大资源开销，同时上述算法获取的数据都是可见的交互数据(如留言、评论等)，不足以支持社交网络中诸如页面访问之类的隐式行为分析。为此，学者们提出了一种全新的社交网络测量方法即点击流模型，该方法利用 ISP 提供的 Http traces 还原社交网络的 session，并以此作为用户隐式行为研究的基础。

在社交网络的测量研究中，爬虫系统的设计是基础，采样算法是研究的重点，而点击流模型则是 ISP 角度的一种全新的测量方法。本节余下部分对上述方法进行详细介绍。

1.1 爬虫系统的设计

社交网络具有封闭、数据展示复杂、数据量大等特点，在设计社交网络爬虫系统时需要考虑系统对不同社交网络、不同数据采集任务的通用性，并对大规模数据采集进行系统架构考量。针对不同的社交网络以及不同的采集任务，现有采集手段通常是开发专用爬虫，每个爬虫都需要重复开发权限控制、URL 分配、数据抽取存储、数据更新等模块，缺乏对统一采集框架的研究。对于社交网络大规模的数据采集，通常存在单节点速率瓶颈问题，因此需要研究分布式的数据采集系统，能够使用多 IP、多账号来提高抓取的效率。

1.1.1 专用爬虫的设计

不同的社交网络由于其权限、API 开放策略、数据展示特

点的不同，需要对爬虫系统进行特定的研究。胡亚楠以Twitter、Facebook、RenRen 3个典型的社交网络为案例进行不同的爬虫设计[2]，并分别就不同的权限问题设计了基于API和基于模拟登录的爬虫系统。考虑到当前社交网络的Web2.0特性（Ajax技术的大量使用），作者还分析了使用HttpUnit解析Ajax页面以深度获取数据的问题，并就爬取频率、单次请求数据量等参数进行讨论并给出实践建议。冯典则研究了新浪微博的数据获取技术[24]。针对不同的数据采集任务，作者使用新浪微博的API分别开发了针对好友关系、微博数据内容、微博标签的定向爬虫，并使用多线程和多APPkey复用技术加快了系统的数据采集速度。

由于新浪微博在API开放策略上的限制越来越严，学术界出现了以众包的方式来采集新浪微博数据的联盟——中国爬盟[25]。爬盟中的用户使用专用的爬虫以模拟登录的方式访问微博并抓取数据。爬盟运营者使用中心服务器向众包节点发送需要下载的任务，众包节点提供自己的账号、IP、主机资源在后台完成抓取任务并上传数据、赢取积分，用户根据爬盟的积分政策就可以下载到相应的数据集，目前在爬盟中发布的数据集包括用户数据、微博数据、关键词数据。爬盟这种众包的方式能够集合数据需求者的资源以完成一定量的数据抓取任务，为广大研究者提供了一定场景下有用的数据集。然而对众包用户而言，其获得的数据集也是相对碎片化和分散的，在有些研究中并不适用。比如，在基于社区的研究中，通常需要定义子图并获取局部数据，众包的方式无法提供这样的数据集，研究者仍然需要一些专有的爬虫系统来完成这样的数据采集任务。

1.1.2 分布式爬虫系统

由于社交网络数据采集通常需要对一个网站进行多次访问，单个节点的访问吞吐量会受到网站的严格控制，因此对于较大规模的数据采集，需要通过扩展节点来提高抓取速度，这需要爬虫有支持分布式扩展的能力，并能处理分布式环境下社交网络数据采集在权限控制、任务划分、数据抽取等方面面临的问题。DuenHorng Chau等人于2007年提出了一个并行爬取的爬虫框架[26]，该框架主要分为3层：Application Server层为爬虫系统提供UI交互功能，用户可以设置相应的采集参数；Coordinator&DataMaster层为协调服务器，采用队列记录BFS搜索节点；Crawler Agent层则为分布式的爬取节点，每个Crawler从不同的社交节点进行数据采集，并和中心服务器进行数据交互。作者给出了一个通用的架构设计，但并没有就爬虫系统面临的挑战如权限、爬取速率、节点选择等问题进行深入的分析。

随着云平台的流行，基于Hadoop的Nutch分布式爬虫系统被越来越多地用于大规模数据采集工作。程锦佳基于分布式文件系统HDFS和MapReduce计算模型提出了分布式爬虫的设计方案，确定了其系统布局以及流程控制，并对传统爬虫的URL分配、数据下载、URL更新等模块进行了基于MapReduce的模式改造，使得传统爬虫能够很好地移植到云平台的分布式环境下[27]。李娜娜基于Storm和Zookeeper技术研究了海量社交网络数据获取的技术，设计了实时云计算平台下的数据获取任务调度策略及社交网络协议解析方法，采用基于数据流的技术来获得大规模的数据集[28]。然而现有的基于Hadoop的分布式爬虫主要关注于对传统爬虫的分布式改造，对其应用在社交网络场景中所面临的问题的研究较少。与传统的Web爬虫相比，社交网络爬虫需要考虑如下的问题：

（1）对于权限控制，传统的Web爬虫通常不需要考虑权限问题，而社交网络是封闭的社区，访问数据需要进行模拟登录来获取Cookie或者使用API获取基于Authtoken的授权。

（2）对于抓取策略，传统的Web爬虫通常基于外链链接关系构图，在图上使用BFS、DFS、PageRank等策略产生下一轮抓取的URLs，而社交网络爬虫基于社交关系构图，在图上使用BFS或是采样算法产生下一轮抓取的用户IDs。

（3）对于数据下载，传统的爬虫使用HTTP协议下载页面，使用正则表达式和DOM模型抽取正文和外链，而社交网络通常需要解析Ajax事件流程，进行URL扩展，对下载的页面使用JSON对象模型和DOM模型进行解析。

（4）对于数据更新，传统的Web爬虫通常根据预定的时间和策略来重新抓取整个页面，而社交网络的更新粒度更细，需要对图谱进行定期更新，对用户产生的内容数据按照时间戳进行增量式更新，对内容下的交互数据进行增量更新。

使用Nutch爬虫的插件机制来解决上述的问题是一个值得尝试的研究方向。Nutch自身提供了传统爬虫所需的URL分配、数据下载、数据解析等核心模块，并预留了扩展点来满足定制化需求。开发者可以在其协议插件上实现社交网络登录和API授权等工作来获取相应的Cookie或Token；对于不同的社交网络下不同的采集任务，开发者可以定制特定的URL分配器、数据下载、数据解析插件，然后通过配置文件进行集成。使用Nutch的插件机制可以避免专有爬虫重复开发权限、URL分配器、抽取等模块的工作，同时又能使用其底层的分布式环境进行大规模的数据采集。

1.1.3 社交网络数据抽取

与传统Web页面不同的是，社交网络在数据展示上大量使用Ajax技术，很多数据需要一系列的JS事件驱动才能完整呈现。以获取个人用户新鲜事或微博信息为例，爬虫系统从URL分配器获得的URL通常是指向用户主页的一个链接，单次请求下的数据只是最近某个时间段的快照，要想获得用户整个生命周期下的数据，需要对下载链接进行扩展，即需要弄清楚整个事件请求的过程，包含多少个子请求，最后组合参数，形成扩展的URL来获取完整的数据。这种数据展示策略在社交网络是普遍存在的，爬虫要有能力去触发“更多新鲜事”、“下一页”这样的事件来下载完整的信息，这就要求爬虫的信息抽取模块具有完整的Ajax事件处理技术。

Ajax技术大量出现在Web2.0的页面中，为了让爬虫系统能够获得这些异步传输的深层内容，学者们主要研究了Web页面中Ajax事件的识别和触发。袁小节提出了事件驱动模型，通过识别网页中的异步JS函数来识别Ajax事件，并使用JS引擎执行事件来获取完整的异步传输网页内容[29]；曾伟辉设计出支持Ajax站点的网络爬虫系统，在理论和技术方法上总结并提出Ajax框架网站中URL关联信息提取、切片代码的有序执行以及程序切片模块、爬虫模块、脚本执行模块之间的互操作，为Ajax框架网站网络爬虫提供了新的解决方案[30]；夏天认为每个页面触发Ajax请求时对应一个状态，

爬虫需要不重复地抓取一个页面不同的状态，针对 Ajax 元素的识别、页面状态的标识、页面状态的程序可控性转换、页面状态的内容动态获取和状态重复检测 5 个方面的问题，其进行了详细的论述[31]。

总的来说，对于社交网络中的数据采集任务，信息抽取模块通常需要识别特定采集任务下 Ajax 请求链接，并对其进行扩展，最后使用脚本执行引擎来触发事件，直到所有的 Ajax 事件都执行完毕，爬虫就可以获取完整的交互数据。在获得数据之后，数据抽取任务基本上就解决了一大半，剩下的工作就是基于 DOM 或者正则表达式或 HTML 标签等一系列文本处理技术进行内容的抽取，其与传统的 Web 数据抽取技术相似，在这里就不再详述。

1.2 数据采样算法

社交网络可以使用基于点和边的有向或无向图建模，节点代表用户，边代表社交关系（如好友）。采样算法的目的是从整个社交图上获取不完整的数据样本使得该样本能够大致反映社交网络的性质。采样算法主要关注如下问题：采样效率问题，即发现新节点和边的速率；采样敏感性问题，即网络中权限不可见的用户对采样结果的影响；采样偏差问题，即采样结果是否偏向某种属性的用户（比如偏向度数高的用户）；种子节点的选择问题；采样过程中下一个节点的选择问题；采样过程的收敛问题。其中下一节点的选择问题是研究的重点，以采样结果是否有偏差为分界线，笔者把采样算法分为有偏采样算法和无偏采样算法。考虑到当前的研究工作主要是面向无向图，如无特殊说明本节所说的采样算法都是针对无向图建模的社交网络。

1.2.1 有偏采样算法

总的来说，有偏采样算法主要依据节点度数的大小来选择下一节点，它大致有如下几种变形：

（1）BFS 算法：对于爬取到的节点队列，选择队列中第一个节点作为下一个节点。

（2）Greedy 算法：对于爬取到的节点队列，选择节点度数最大的节点作为下一个节点。

（3）Lottery 算法：根据节点度数的不同，以不同概率选择下一个节点，节点度数越大，被选择的概率越大。

（4）Hypothetical Greedy 算法：选择节点度数最大的节点作为下一节点，但该最大度数为对全局的一个估计。

算法的目标是以尽量小的采样比达到尽量大的节点覆盖率（NC）和边覆盖率（LC），其中 $NC=\frac{V_{seen}}{V}$（V_{seen} 为发现的节点），$LC=\frac{E_{seen}}{E}$（E_{seen} 为发现的边）。Shaozhi Ye 等人[5]使用上述算法在 Flicker、LiveJournal、Orkut、Youtube 数据集[24]上进行采样测试，就采样效率、节点选择等主要问题进行对比实验，并为上述算法的实际应用提供了有益的经验。实验表明，Greedy 算法在相同的采样比下能达到更高的覆盖率；爬取小部分节点能够发现整个社交图谱的一大部分，如采样 10% 的节点能发现 50% 左右的节点和边；开始节点的选择不影响节点覆盖率和边覆盖率；爬取过程中会大量爬取重复节点，爬虫系统应设计协调服务器来避免因重复而导致的资源浪费；社交图谱中的一部分设置了隐私权限的节点不影响覆盖率；采样样本趋向于节点度高的节点，如图 1 所示。

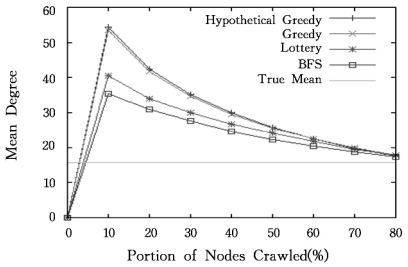

图 1 有偏采样算法节点度的分布[5]

从图 1 可以看出，上述算法能使用较低的采样比达到较高的覆盖率，但其数据样本偏向于节点度数较高的节点，这给图谱性质估算带来了很大的误差，因此我们也称其为有偏采样算法。

1.2.2 无偏采样算法

无偏采样算法大都基于随机游走理论，随机游走的采样过程可以使用马尔可夫随机过程建模。马尔可夫链的稳定分布就是每个节点被访问的概率，算法的目标是使得图谱上每个节点被访问的概率相等，以达到均匀采样的目的[25]。目前学术领域主要研究的无偏采样算法如下。

（1）UNI[3]：根据 OSN 用户 ID 的分配方案，使用拒绝式采样方法获得均匀样本，该策略常常用作算法分析的基准。对于特定的 OSN 网站，程序使用均匀随机算法生成 0～$MaxUserID$ 的一个 ID，并使用该 ID 向 OSN 查询其主页信息，如果有数据则进行采样，否则被拒绝。理论上讲，采用 UNI 方法获得的数据样本是绝对均匀的，其不足之处在于需要预先知道用户 ID 空间，OSN 需要对每一次查询进行响应，并且 OSN 的用户空间不能太大，例如 Facebook 用户空间从 32 位变成 64 位后 UNI 方法因其巨大的开销就不适用了。同时对于使用非数字 ID 的 OSN，该方法也不适用，如 YouTube。

（2）RW（Random Walk）[3,6]：随机游走算法，根据概率选择下一个节点，偏向节点度大的点。该算法首先随机选择一个 *seed* 节点，采用如下概率策略选择下一个节点 W：

$$p(v,w)=\begin{cases}\frac{1}{K_v}, & w\text{ 为 }v\text{ 的邻节点}\\ 0, & \text{else}\end{cases}$$

使用马尔科夫过程建模可以得到其平稳分布的概率亦即每个节点被访问的概率 $\pi_v=\frac{K_v}{2|E|}$（K_v 为节点 v 的度数）。由该平稳概率可以看出，RW 算法以更大的概率采集度数大的节点。

（3）Re-Weighted RW[3,6]：首先使用 RW 算法获得一个有偏节点集合，然后依据需要估算的性质（如节点度分布、Network 大小）把节点分散到特定的集合包，最后在该集合包上使用特定的估算因子来消除偏差。以估算节点度数的分布为例，算法使用 RW 获得的一个有偏点集合 $V=V_1V_2V_3\cdots V_n$，设集合 A 为节点度的集合包：$A=(A_1A_2A_3\cdots A_n)$，A_i 为度为 i 的采样点的集合，则全网中度数为 i 的节点的比例 $p(A_i)$ 可估算为：

$$\hat{P}(A_i)=\frac{\sum_{u\in A_i}\frac{1}{k_u}}{\sum_{u\in V}\frac{1}{k_u}}$$

$\hat{P}(A_i)$ 被称为估算因子，对于 OSN 中不同的性质如图聚合度等，需要开发不同的估算因子。

(4) MHRW[3]：在选择下一个节点时会适当降低高度数节点的访问概率，从而修正偏向高度数节点的偏差。但与 Re-Weighted RW 不同的是，MHRW 算法是在采样的过程中就对转移概率进行修正。采样过程的转移概率为：

$$P_{v,w}^{MH}=\begin{cases}\min(\frac{1}{K_v},\frac{1}{K_w}), & w \text{ 为 } v \text{ 的邻节点}\\ 1-\min(\frac{1}{K_v},\frac{1}{K_w}), & w=v\\ 0, & \text{otherwise}\end{cases}$$

从转移概率可以看出，MHRW 算法使用算式 $\min(\frac{1}{K_v},\frac{1}{K_w})$ 减少了大度数节点的访问概率。MHRW 的不足之处是需要知道当前节点 v 和邻居节点 w 的度数，这在某些社交网络上常常成为限制。另外 MHRW 算法会有一定的概率停留在当前节点，对于连通性不强的图谱，自循环的问题会降低采样的效率。

(5) USRS[33]：修正 MHRW 在低度数节点时自循环概率过大的缺陷。MHRW 算法要求图谱的连通性较好，因为对于低度数的节点，其自循环的概率可能达到 0.5 以上。USRS 算法的核心思想就是使用一个 ReduceLoop 函数在采样过程中降低每一个节点的自循环概率。USRS 在算法收敛时间、点覆盖率和边覆盖率上都优于 MHRW 算法，其不足之处是需要计算当前节点的邻居节点的自循环概率，这需要知道邻居节点的邻居的度数信息，所以其可行性常常受到 OSN 权限策略的限制。

(6) AS[4]：修正 MHRW 在低连通性子图中陷入 trap 的缺陷，加入随机跳跃过程。转移概率为：

$$P_{v,w}^{MH}=\begin{cases}\min(\frac{1-p}{K_v},\frac{1-p}{K_w})+\frac{p}{|V|}, & \text{if } w \text{ 为 } v \text{ 的邻节点}\\ 1-\sum\min(\frac{1-p}{K_v},\frac{1-p}{K_w})+\frac{p}{|V|}, & \text{if } w=v\\ \frac{p}{|V|}, & \text{otherwise}\end{cases}$$

从转移概率可以看出，该算法一方面降低了 $v=w$ 即自循环的概率，另外一方面增加了随机跳跃的概率。而 AS 算法的一个问题就是随机跳跃时选择一个随机节点的开销。一般的做法是根据用户空间 USERID 的构成，随机选择一个有效的节点，但是如果用户空间很大并稀疏分布（如 Facebook 的 64 位的 ID），那么会造成巨大的开销。

(7) FS[6]：随机边采样算法，对于连通性较弱的图谱有较高的效率。与上面谈到的算法不同的是，FS 是一个针对随机游走而提出的边抽样方法。算法首先选择一组随机节点构成种子节点 S，并以一定的概率在集合 S 选择顶点 v，从顶点 v 的出边中随机选取一条边 (v,w) 并加入采样到的边队列，同时用顶点 w 替换种子集合 S 中的顶点 v，直到算法收敛到一个稳定分布。FS 算法需要特殊的估计函数来修正偏差，与 ReWeighted RW 算法一样，对于不同性质需要不同的估算因子。

(8) USDSG[34]：前面涉及的采样算法都是针对无向图而设计的，对于 Twitter 等基于弱关系（有向图）的社交网络并不适用。有向图中存在出度为 0 的节点，MHRW 之类的算法将在该节点出现死循环。USDSG 算法将有向图中的有向边看作无向的，把图 G_d 转换成一个无向对称图 G，节点的度数为出入度之和，在对称图 G 上运行 MHRW 算法，同样可以实现无偏采样。

总的来说，上述算法基本上解决了数据采样的偏差问题，但却增加了实现的复杂度。在实际的数据采集中要仔细分析社交网络的图谱特点来选择最合适的采样算法。表 1 对上述采样算法的优缺点、适用范围、采样效率进行了对比。

表 1　数据采样算法比较

采样算法	优缺点	适用范围	采集效率
BFS 系列	实现简单，采样时偏向于度数大的节点	适合对特定区域的子图爬取全部的数据，不宜作为采样算法	可避免重复爬取，效率高
UNI	实现简单，数据无偏差，但拒绝式采样资源消耗大	适合采用数字空间作为 ID 策略的 OSN，其 ID 空间不宜过大	根据 ID 空间大小而定，一般情况下效率不高
Re-Weighted RW FS	实现简单，可在事后消除样本偏差，但可用的估算因子有限	适用于对 OSN 度数分布和聚合系数的估计	存在重复爬取，效率不高
MHRW	能无偏采样，收敛过程易判断，但实现相对复杂	适用于连通性较好的 OSN	每一步需要获得邻居节点的度数，且容易陷入自循环，效率不高
USRS	无偏采样，实现复杂	对低连通和高连通性的 OSN 都适应	需计算邻居节点的自循环概率，但降低了自循环概率，效率高于 MHRW
AS	无偏采样，实现相对简单	对低连通和高连通性的 OSN 都适应	降低了自循环概率，如果 ID 相对较小，效率会高于 MHRW 和 USRS
USDSG	同 MHRW	适用于出入度都可见的 OSN	同 MHRW

1.3　点击流数据模型

爬虫系统主要通过浏览节点用户的页面来获取其好友列表，并构建图谱结构或者抓取交互信息进行用户行为分析。但实际上，在社交网络中除了评论、留言、内容发布这类会显性地产生数据的用户行为外，还存在大量诸如页面访问的隐式交互行为。隐式交互行为因为在网站上不留下痕迹（Ren-Ren 好友访问列表是个特例）而无法被普通爬虫抓取。然而任何的交互行为都会产生一个 Http 请求亦即点击流，如果能够获得某个网络入口的 Http trace，就可以抓取到去往该入口的所有社交网络交互行为。

点击流的数据模型就是从 Http 头信息出发，通过还原用户访问 OSN 的 session 进行用户行为定义，理论上其可以采集所有的显式和隐式交互行为。通常情况下获取点击流数据的途径有：

(1) 通过站点服务器获取点击流的数据，如 Web 服务器的日志文件、内容服务器日志文件、网络监视器日志文件。这些服务器通常都由 ICP 控制，因此从 ICP 的角度来获取点击流信息是最快捷和高效的。

(2) 与 ISP 合作在特定的网络出口截取 head traces 并进行社交网络流量的提取，以此来获得相应的点击流数据[8]。

(3) 从第三方代理如 social network aggregator 获得点击流数据[7]。

在获得点击流数据之后，需要进行社交网络流量的抽取和用户行为识别。首先，需要对 Http 头进行过滤以便抽取去往 OSN 的 request-response pair(rr-pair)。接着需要根据 IP 和 Cookie 的不同，把这些 rr-pair 分组到不同的会话，如区分特定 pair 是属于 Facebook 还是 LinkedIn。在区分会话之后需要根据请求的 URI 来区分用户的点击行为，并且完成 URI 与用户行为的映射。至此就可以把不同的 rr-pair 与特定的社交网络以及特定的功能模块进行关联(如对 Facebook 主页进行浏览的 rr-pair)。

点击流模型虽然能够获取所有的用户交互行为，包括页面浏览等隐式行为，但是它也具有如下不足：

(1)通过 ISP 获取 Http trace 对采集存储资源的消耗巨大，且研究者通常只能拿到部分 ISP 接入用户的数据，即数据是不完整的。

(2)从抽取 rr-pair 到 URI 的映射，整个分析过程相当繁琐和耗时。

(3)URI 的 pattern 很多且不规范，如果 OSN 更新服务，相应的 URI 可能会更新，这给采集数据的更新带来很大的困难。

(4)如果要从多个不同 ISP 获得数据，需要处理数据一致性的问题。

2 用户行为分析

社交网络作为开放性的社交平台聚集了大量的用户，这些用户在文化背景、社交目的上都大不相同，这也决定了其用户行为的差异性。对用户行为的分析一方面可以帮助 ICP 了解用户使用社交网络服务的习惯，让 ICP 针对不同的用户群体提供定制化的服务。同时，对用户行为的研究也是定义活跃用户、发现核心圈子以及检测僵尸用户的有效手段。维持社交网络的活跃度、避免核心用户的流失关系到社交网络的战略目的，新应用的推出、内容推荐、好友推荐等刺激用户活跃的策略也都需要有深入的用户行为分析作为理论支持，用户行为分析也因此成为 ICP 和学术界重点关注的领域。

用户行为的产生依托于社交关系图谱，对社交图性质的研究是该领域研究的基础和起点。社交图谱节点度分布、路径长度、聚合系数等性质能够宏观地反映整个社交网络的用户行为特征，我们称之为广义用户行为。考虑到社交网络中存在大量的不活跃用户和僵尸用户，社交关系的建立不一定会导致用户行为，一些学者由此提出了通过构建显式活动交互图和隐式活动交互图来代替社交图谱进行行为分析。

除了宏观性的行为分析，学者们关注比较多的是社交网络中的功能性行为分析，即用户与社交网络本身的交互规律，包括社交功能的流行度及各功能之间的访问关系。更细致的研究工作则还包括对用户行为的社交性分析，如用户之间的互访规律。值得一提的是点击流数据模型让基于 session 的用户行为分析成为可能，这可以让学者从使用网络的角度分析社交网络用户的行为。

本节余下部分将就前面提到的宏观和微观社交网络用户行为分析做详细的介绍。

2.1 社交图谱分析

社交网络可以看做一张巨大的图谱，节点代表用户，边代表社交关系，各节点的用户依托社交关系进行社交活动。图的整体性质能够反映社交网络用户的群体特点，如分析社交图的节点度数分布和稀疏度可以获知整个社交网络连接关系的强弱，对关键点的分析有助于分析用户影响力，对图聚合性的分析有助于发现不同的社区，基于图论对社交图谱进行分析可以让研究者从宏观上把握社交网络用户的社交行为。目前学术界对图谱性质研究的主要关注点如下。

(1)节点度的分布。社交网络中节点的度数反映了用户平均拥有好朋友的个数。研究表明，大多数社交网络的节点度数满足幂律分布，大部分的节点度数较低，拥有较少的朋友。对于使用有向图建模的社交网络，需要分开对出入度分布进行研究，如 Alan Mislove 等人对 Youtube、LiveJounary、Flickr 的出入度研究发现，高入度的用户和高出度的用户存在很大比例的重合[32]，这说明了在上述社交网络中活跃的用户也更容易受到别人的关注。

(2)聚合系数(Cluster Coefficient)。聚合系数反映的是用户朋友间成为好朋友的比例。在社交网络中聚合系数一般为正数，如 Facebook 中用户聚合系数为[0.05，0.35]，且随着度数的增加其聚合系数减小[35]，这说明小型朋友圈之间的用户更容易趋向于建立好友关系。

(3)对称系数(Assortativity Coefficient)。对称系数量化了社交网络中一个节点是否趋向于与具有相近度数的其他节点建立联系。该系数的取值为[−1，1]，正的系数说明相同度数节点具有连接趋势，而负数则相反。如 Youtube 的对称系数为−0.033，而 Flickr 为 0.202，LiveJournal 为 0.179，Orkut 为 0.072[32]。这主要是因为 Youtube 的社交性质较弱，以内容为主，用户节点的度数与社交关系的建立相关性不强。

(4)路径长度(Path Length)。路径长度通过对社交图谱进行广度优先搜索取平均值计算得到。路径长度反映了一个社交网络的连接程度，而社交网络的路径平均长度基本上在 6 上下，如 RenRen 为 5.38[9]，这验证了六度分隔理论的正确性。

(5)强连通分支分析(SCC)。像 RenRen 或 Facebook 这些基于 Network 概念的社交网络，其 SCC 子图通常由某些特定区域的用户组成。强连通分析可以发现社交网络中的核心圈子以及孤立的用户[9]。

不同的社交网络由于其关注的重点和组成的用户群体不同，通常在图谱性质上存在一定的差异，分析和解释这种差异能够让我们更加深刻地认识到社交网络的内在规律。如从图 2 可以看到，Twitter 由于媒体性质多于社交性质，其平均度数最低，但其对称系数却最高。一种解释是 Twitter 上的明星用户、媒体用户趋向于与其他明星媒体用户建立关注关系。从路径长度上看，Facebook 和 Cyworld 比 RenRen 具有更好的连通性。

Network	Users Crawled (k)	Links Crawled (k)	Avg. Degree	Cluster Coef.	Assor-tativity	Avg. Path Len.
RenRen	42115	1657273	78.70	0.063	0.15	5.38
Facebook	10697	408265	76.33	0.164	0.17	4.8
Cyworld	12048	190589	31.64	0.16	−0.13	3.2
Orkut	3072	223534	145.53	0.171	0.072	4.25
Twitter	88	829	18.84	0.106	0.59	N/A

图 2 常见的社交网络图谱的性质[9]

2.2 活动图分析

社交图谱的点边模型反映了真实生活中的社交关系，利用社交图谱可以进行信息传播、信任机制等研究。然而通过好友关系建立起的图谱因为非活跃用户、僵尸用户等的存在大大地降低了其有效性，因为图谱中很大一部分的连接并没有激发实际的交互行为。因此一些学者提出基于交互行为的活动图，图中的点仍代表用户，如果用户间具有交互行为则产生一条边，交互的次数可以用来定义边上的权值。考虑到社交活动是对社交关系的真实反映，学者们认为活动图更能反映社交网络的本质。

给定一个特定的社交网络，根据其好友关系可以构建社交关系图，根据留言等行为可以构建显式交互图，根据用户页面访问可以构建隐式交互图。通常 3 种图谱会在聚合系数、对称系数、路径长度等性质上存在较大的差异。图 3 是对 RenRen 分析的结果。从中可以看出活动图在边、聚合系数、对称系数上都明显减少，这验证了 RenRen 中很大一部分社交关系并不能带来用户行为。同时该研究工作中基于隐式交互图的信息传播试验显示，隐式交互图具有更快的信息传播速度[9]。

Network	Edges	Power-Law Fit Alpha	Cluster Coef.	Assortivity	Avg. Path Len.
Social Graph	753297	3.5	0.18	0.23	3.64
Visible Interaction Graph	27347	3.5	0.05	0.05	5.43
Latent Interaction Graph	240408	3.5(in) 3.39(out)	0.03	−0.06	4.02

图 3 RenRen 图谱研究[9]

Hyunwoo Chun 等人通过对 Cyworld 活动图的研究发现这种差异并不明显。作者根据 Cyworld 的留言板构建用户活动图并就聚合系数、节点度数分布、k-core 等指标与社交关系图进行详细的对比，发现 Cyworld 的留言交互图和朋友关系图结构相近，交互行为具有很高的相互性。一种可能的解释是早期(2007 年)Cyworld 中的留言活动很流行，所有的好友几乎都使用过留言进行交互活动[36]。

活动图代表真实的交互关系，一些学者用它代替社交图谱来对用户行为的演变以及信任机制进行研究。Bimal Viswanath 等人使用 postwall 为 Facebook 构建不同时段的活动图快照，并刻画不同快照之间共同边的覆盖率。作者使用边覆盖率来定义活动图的 Core 核心，并通过分析不同的快照中新出现的边来研究活动图的变化趋势。研究发现，在测量周期中尽管活动图中活跃边更新很快，但其在度数分布、聚合系数、路径长度等度量指标上都趋于稳态[10]。而 Christo Wilson 等人则关注信任机制，作者使用评论内容为 Facebook 构建活动图，并对比社交图和活动图在防范 Spam 邮件、Sybil 攻击时的性能表现。活动图由于更确切地反映了交互行为，其信任机制更加可靠[37]。

2.3 Session 分析

社交网络的一个 session 是指一个用户在某一次 login 和 logout 的时间内与社交网络的会话。社交图谱研究是对社交网络群体性行为的宏观分析，而 session 分析则更能反映社交网络服务中用户的网络行为。基于 session 的用户行为分析产生于点击流数据模型出现之后，该方法依赖于 Http 请求响应对来定义 session，是对用户行为分析领域的补充，而且还有助于从流量特性的角度区分社交网络服务与其他 Web 应用间的不同。

Fabian Schneider 等学者利用从 ISP 获得的点击流数据分析了 Fackbook 等 4 个社交网络的 session 特点[8]。从流量角度上讲，社交网络中单个 session 的 bytes 大小符合长尾分布，但长尾并没有像其他 Http 服务表现得那么明显。session 的持续时间同样也符合长尾分布，由于 session 的持续时间不同，单个 session 内的用户活跃程度也不相同，session 中用户从不活跃到活跃所使用的社交功能也不一样，如在一个 5min 的 session 内，用户更多的是使用 message 功能来恢复活跃状态，而在 20min 的 session 中，用户则选择浏览主页。

Fabricio 等人则利用社交网络聚合器的点击流数据对基于 session 的用户行为做了更为细致的分析[7]。就 session 的流量特征而言，他们在 Orkut 等 4 个网站上得出了与文献[8]相同的结论，即 session 大小和长度符合长尾分布。从 session 的频率看，不同社交网络的访问频率也都满足长尾分布，使用社交网络频率高的用户，其 session 时间也不一定长，即 session 频率和时长没有明显的相关性。为了了解 session 的动态特点，作者对 Orkut 的 session 到来时间、session 活跃度、session 中 request 的间隔进行了建模，并得出了一些有用的结论。session 的到来时间整体上服从对数正态分布；以单个 session 中 request 的个数来定义的 session 活跃程度符合长尾分布；session 中 request 的时序也满足对数正态分布，这说明用户倾向于在短时间的 session 会话中进行持续的交互。另一个有些意外的结论是 session 的持续长度和活跃程度与 session 的开始时间无关，这在一定程度上反映了社交网络用户的特质，他们在社交网络上的活跃度与上网时间不具有强相关性。

2.4 功能活动分析

每一个社交网络都对用户提供相应的功能集合(如照片、评论、留言、分享、主页)，这些功能的使用构成了用户行为的主体。研究这些功能在用户间的流行度或变化趋势能为 ICP 提供重要的运营支持，如升级服务、开发新应用等。

在点击流数据模型出现之前，通常通过爬虫来抓取显性的交互数据(如留言、评论)以分析用户行为。如 Hyunwoo Chun 等人根据 ICP 提供的两年数据分析了用户在评论、留言、交友等行为上的增长趋势[32]。Christo Wilson 等人则通过爬取 Facebook 的 postwall 和照片评论来分析用户行为并构建活动交互图[10,37]。Jing Jiang 等人利用 RenRen 记录的用户访问主页信息的特性对隐式交互行为(主页访问)进行了研究[9]，并得出主页访问占据用户行为 90%以上的结论。

然而上述研究工作通常只是抓住功能活动中的一部分，对于社交功能更全面的研究是在点击流数据模型被提出之后。Fabian 等人把社交功能分为主页、相册、朋友关系、留言、应用、视频、搜索等，并通过 URI 与社交功能的映射以及 rr-pair 数据的统计分析给出了每个功能的流行程度[8]。Fabricio 等学者基于同样的方法分析了 Orkut、Hi5、MySpace、LinkedIn 上各个活动的流行度[7]。两篇论文都得出了主页访问是最流行应用的结论，这也说明隐式交互的重要性。而其

他功能由于各社交网络的侧重不同而具有不同的流行度，如 Orkut 与 Hi5 的相册功能较流行，Myspace 和 LinkedIn 的留言板更具吸引力。

为更精细地研究社交网络中的用户行为特征，文献[11，12]对一个具体的 session 之内用户都使用哪些功能以及各功能间的状态如何转移做了深入的分析。作者创造性地使用单一 session 中功能被点击的顺序进行一阶马尔科夫过程建模，并计算各功能直接的转移概率。图 4 为 Orkut 各功能的条件转移概率图。基于这样的转移图，我们可以清晰地得出关于 OSN 的一些有趣的结论，如用户趋向从主页和朋友关系开始使用社交网络，在一些社区活动上持续地交互，除了主页应用外，相册、搜索等应用的流行度很高。

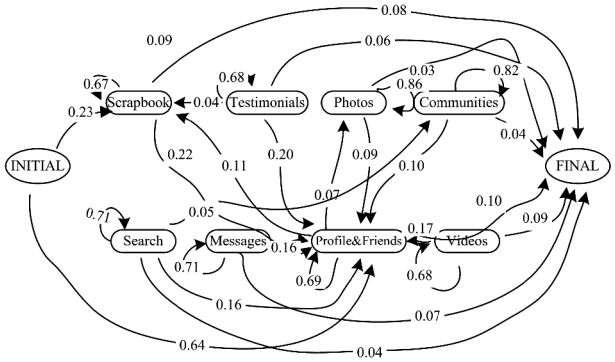

图 4　Orkut 活动概率转移图[7]

2.5　社交性分析

对社交网络中各功能的流行性分析主要立足于用户与社交网络本身的交互，为了更精确地研究用户与用户之间的交互，一些学者就用户之间社交活动的内容、相互访问的对称性、用户间访问的行为动机、用户访问朋友的频率和范围等有关社交性的问题进行了深入分析。

Jing Jiang 等人以用户页面被访问次数作为度量，分析了 RenRen 中人气用户在社交网络的比例、被访问次数等社交信息。文章就朋友、陌生用户、特定距离的用户在访问中的比重、互访率等问题进行了探讨。总的来说，用户主页的互访率在 10％左右，而陌生人的互访率更低[9]。

Fabricio 等人分析了 Orkut 中朋友间的交互行为与朋友间的距离之间的关系，以及社交网络的哪些功能通常把用户导向朋友主页[7]。作者发现在测量的时间内用户访问朋友的平均数目相对较少（仅为 3.2），80％的用户行为仅发生在一跳的范围之内，用户在访问朋友的主页之前通常是在浏览自己的主页。

Christo Wilson 等人通过爬取 Facebook 的 postwall 和照片评论的内容来分析社交性行为[37]。作者发现 90％左右的用户通常只与 20％的朋友进行交互，社交图谱的连接关系并没有广泛地转化为真实的用户行为。交互行为在用户间的分布大致符合幂律分布，即小部分的活跃用户贡献了大量的交互行为（1％的用户贡献了 20％的 postwall 的交互事件），大部分用户处于不活跃状态，活跃的用户与其节点度数正相关。该研究还选择数据集中注册时间最久的 10％用户与最新的 10％用户研究交互行为与生命周期的关系。老用户见证了 OSN 的发展过程，整体活跃程度保持稳定且小幅增长，而新用户的整体活跃度大都随时间呈下降趋势。

3　信息传播机制研究

社交网络的一个重要功能就是信息的交流和共享，而随着社交网络用户的迅猛增长以及类似微博的强媒体性质的社交网络的出现，社交网络在信息传播方面正扮演着举足轻重的角色，也吸引着越来越多的学者对其信息传播机制进行研究，从而为社交网络上信息传播的引导和控制提供服务。

信息传播机制研究的起点是信息传播规律的发现。不同的学者对早期的博客、以照片共享为主要功能的 Flickr 和具有强媒体性质的 Twitter 等社交网络上信息传播的热度变化、拓扑特性等性质进行了研究。而为了深入解释信息传播规律，研究者又基于病毒传播模型以及解释创新事物传播的独立级联模型和线性级联模型，提出了 AsIC 和 AsLT 等模型。为充分利用社交网络中信息传播的规律，给社交网络中的信息监管、商业营销提供指导，部分学者对社交网络中信息传播预测模型进行了研究，相继提出了回归分析、概率预测、机器学习、LIM 等方法来对社交网络信息传播进行预测和分析。

本节余下部分将从信息传播规律、信息传播模型、传播预测模型、意见领袖挖掘 4 个方面对已有的研究工作进行总结。

3.1　信息传播规律的分析

为了对社交网络的信息传播机制进行研究，首先必须对社交网络中信息传播的规律进行观察和分析。而由于社交网络的种类繁多，包括早期的博客、以照片共享为主要功能的 Flickr 以及具有强媒体性质的 Twitter 等，不同类型的社交网络有着不同的功能和结构，其信息传播规律也往往表现出不同的特征。从已有的研究成果来看，这些差异主要集中在信息传播的周期性、信息传播的速度、信息到达的范围以及传播拓扑上。

Jure Leskovec 等人对通过博文中的 URL 链接构成的博客网进行研究，发现在博客网中博文数呈现出以周为单位的周期性；博文的热度变化呈现出指数为－1.5 的幂律分布，与 Barabasi 关于人类行为学的长尾理论良好吻合；博文的链入数、传播拓扑的节点数等均满足幂律分布；传播拓扑可以划分为星形和链状两种基本组成部分，其中以星形拓扑最为常见[12]。

Meeyoung Cha 等人以 Flickr 中 250 万用户对 1100 万照片的收藏行为为数据集，研究了 Flickr 中的信息传播规律。在 Flickr 中，即使是热门图片，其传播的范围并不广，速度也比较慢；好友间的信息交互占据了全部交互的 50％以上，但是好友间的消息传递时延仍然比较大[13]。

Wojciech Galuba 等人基于在 270 万用户间传播的 1500 万 URL，研究了 Twitter 中 URL 的传播规律。Twitter 中的传播拓扑的特点是浅而广，即绝大多数节点与源节点的距离只有一跳，而拓扑中节点距源节点的最大跳数满足幂律分布；传播拓扑上相邻的两个节点的传播时延满足中值为 50 分钟的对数正态分布[14]。

3.2　信息传播模型的研究

针对社交网络中信息传播呈现出的规律，研究人员一直致力于提出一个合理的模型，以对社交网络中的信息传播过程进行建模。社交网络可以被建模为一个图，图中的节点往往受到参与消息传播的邻居节点的影响而参与到传播过程，并进一步地去影响其他节点，这和病毒的扩散和创新事物的

传播有诸多相似之处。实际上，关于病毒扩散和创新事物的传播规律的研究，已经成为了社交网络中信息传播规律的基础。

病毒的扩散规律更多关注的是时序上的变化，它通常将节点分为易感、已感和已恢复3类节点，不同的节点之间存在一定的转换概率。Jure Leskovec等人基于病毒传播的SIS模型将博客网节点分为易感(Suspectible)和已感(Infected)节点，已感节点以概率β去感染其未感染的邻居节点，模拟产生的结果从拓扑的大小分布、入度分布等指标上都能与实际情况良好吻合[15]。Dechun等学者基于该模型对Twitter中谣言的传播进行了研究，作者将Twitter中的节点分为3种：未知者、传播者和知而不传者[16]。若谣言的接收者是未知者，他会以概率α转换成传播者；若接受者是传播者，他会以概率β转换成知而不传者；若传播者未从的邻居再次听到谣言，他会继续充当传播者直至一定时间。在如Facebook的好友关系为双向的网络中模拟结果显示，网络中最终的未知者的比例为20%左右，而在互粉率不高的Twitter类网络中，该比例仅仅为10%，说明微博网络中的信息传播的效率更高。同时，作者还提出了一个信任模型，即用户对一个谣言的信任程度取决于他的所有传播该谣言的关注者的关注粉丝之和，并认为该和超过一定阈值时，用户就会相信该谣言，但是用户对谣言的相信与否与用户是否参与谣言的传播是独立的。

对创新事物的传播规律的解释有两个比较经典的模型：独立级联模型(Independent Cascade，IC)和线性阈值模型(Linear Threshold，LT)。独立级联模型认为每一条社交网络的边都有一定的传播概率，消息从源节点开始沿着这些边依次向外扩散。线性阈值模型则定义了好友之间两两的影响值，以及每一个用户的影响力阈值，并认为当用户与参与传播的好友节点的边影响力之和超过用户的阈值时，用户就会参与到传播中来。由于这两个模型是同步的，不能很好地反映信息传播的时序上的特征，因此Saito等人在传播拓扑的每条边上加入了传播时延参数，提出了这两个模型的异步版本AsIC和AsLT，并将其应用到社交网络中的信息传播和用户行为规律的研究上[40]。通过在独立级联模型和线性阈值模型中引入扩散时延因素，AsIC和AsLT能够更精确地反映和预测社交网络中用户行为和信息传播随时间变化的趋势，与实际网络情况更吻合。

3.3 信息传播的预测

对社交网络信息传播模型的研究，并不仅仅是为了社交网络中信息传播规律的解释，更重要的是提出一个合理的模型，并利用该模型对社交网络中的信息传播进行预测。这也成为社会网络信息传播研究中最活跃的方向之一，研究者相继提出了回归分析、概率预测、机器学习、LIM等方法对社交网络中的信息传播进行预测和分析。

J Yang等人利用回归分析的方法对Twitter中的信息传播的速度、规模和范围进行了建模和预测[41]。文章将信息传播的速度定义为某消息从一个节点出发的最小的传播时延；规模定义为一个节点带来的直接转发量；范围定义为从一个节点出发最多经过多少跳消息停止传播。以这些变量为因变量，以用户的微博数、被@次数、转发(创建)消息的时间等为自变量，利用线性回归分析的方法求出其函数表达式，以此对Twitter中信息传播的速度、规模和范围进行预测。结果表明，在所有因素中，被@的次数所起到的决定作用最大。

Wojciech Galuba等人则提出一个概率模型来对Twitter中的信息传播进行预测[14]。文章提出了两个模型来模拟Twitter中信息传播的概率。在ALO(At Least One)模型中，用户会因受他的某一个参与消息传播的关注者的影响而参与消息的传播，用户参与一条消息传播的概率至少等于受他的一个参与消息传播的关注者的影响而参与传播的概率；而在LT(Linear Threshold)模型中，用户参与一条消息传播的概率与他的所有参与消息传播的关注者的综合影响相关。这两个概率模型都综合考虑了时效性和用户的模仿行为等特性。实验结果表明，预测结果能达到85%的精确率(Precision)和55%的召回率(Recall)。

Adrien Guille等人则采用了机器学习的方法，综合考虑了网络、语义和时间3方面的因素，对社交网络中信息传播的时序特征进行预测[42]。作者以用户的活跃程度、@的频率、用户间的社交网络的相似性等指标量化网络因素，以用户历史博文的关键词和消息的关键词是否存在交集为语义因素，以用户在一天中不同时间段的活跃程度代表时间因素。将这些指标作为输入，两个用户之间消息传播行为的发生与否为输出，采用C4.5决策树、线性感知、多层感知、贝叶斯逻辑回归等算法，利用机器学习的方式进行训练，获得相应参数之后即可对后续消息传播的概率进行预测。实验表明，该方法能准确地预测消息热度变化的大体趋势，但无法在量上精确地与实际情况吻合。

Jaewon Yang等人则提出了一个与拓扑无关的模型LIM(Liner Influence Model)来预测网络中的信息传播的热度变化趋势[42]。作者认为每一个用户都有相应的影响函数，即用户转发(创建)一条消息以后在各个时间段带来的转发量。而信息转发量的变化即是参与消息传播的用户的影响函数在相应时延作用下的线性组合。利用同一网络中的部分消息的传播数据，可以求出网络中各个用户的影响函数的近似值，进而可以利用这些影响函数对该网络中其他消息的传播情况进行预测。文章在没有明显网络结构的在线媒体和Twitter上均进行了实验，结果表明在线媒体中的信息热度变化往往由少数几个大媒体的影响所决定；而Twitter中的信息传播则由更多用户的共同影响所决定。

表2就适用问题、所需数据、预测效果等指标对上述4种算法进行了分析比较。

表2 信息传播预测算法比较

算法	适用问题	所需数据	预测效果
回归分析	预测用户带来的首次转发的时间、直接转发量和带来转发的最大跳数	用户微博数、@和被@的次数、转发消息的时间等	相关系数较低(小于0.5)，预测精度不高
概率预测	预测微博中的用户是否会转发一条URL	(用户、URL、转发时延、转发与否)四元组	预测模型能达到85%以上的精确率和55%以上的召回率
机器学习	预测微博中的用户是否会转发一条微博	用户的微博数、被@频率、用户间共同好友比例、博文关键词、用户历史博文等	预测模型能达到85%以上的精确率，预测得到的转发量变化趋势与实际良好吻合
LIM	预测微博中#标签的转发量变化，或者获取各个在线媒体在不同话题上的影响力函数	用户(在线媒体)转发消息的时间与各个时间段消息的转发量	比较适合在线媒体间消息传播的预测，微博中#标签传播由更多用户的影响决定

3.4 意见领袖的挖掘

除了信息传播的预测以外，意见领袖的发掘、用户影响力的量度也是社交网络信息传播引导与控制过程中重要的一环。

在目前各主要的社交网络中，使用最广泛的用户影响力衡量指标是用户的入度，即认为用户的粉丝越多，影响力越大。考虑到不同的粉丝对用户影响力的贡献不同，有学者借鉴了搜索引擎中对网页进行排序的 PageRank[43] 算法对社交网络中的用户影响力进行排序[44]。作者定义用户 v 的影响力为 $Influence(v)$，出度为 N，则用户 v 对每一个他指向的节点的影响力贡献值为 $Influence(v)/N$，且认为用户的影响力等于他的所有粉丝对他的影响力贡献值之和，由此可得到用户影响力的函数关系式。从一定的初值开始对该函数式进行迭代直至收敛，最终可得不同用户的影响力值，依此对用户进行排序，可发掘其中的意见领袖。

考虑到用户在不同话题上的影响力明显是不同的，Jianshu Weng 等提出了 TwitterRank 算法[44]，在 PageRank 算法的基础上，加入话题因素来对用户的影响力进行排序。文章认为，在话题 t 上，用户 i 对关注者 j 的影响力贡献值与用户 i 的影响力、j 的博文数占 i 的所有关注的博文数的比例、i 与 j 在话题 t 上的感兴趣程度这 3 个因素成正比，由此可得到基于话题的用户影响力函数关系式。与 PageRank 类似，对该函数式进行若干次迭代，可得用户在不同话题上的影响力，从而获知不同领域的意见领袖。通过对不同话题赋予不同的权值，对用户在不同话题上的影响力进行加权求和，可得到用户的综合影响力。

实际上，由于未考虑时间因素，InDegree、PageRank、TwitterRank 等都未能很好地发掘那些入度不高但对话题传播起关键作用的首批引领者。为此，Diego 等人在PageRank 算法中引入了时间因素[45]。文章首先将 Twitter 中的＃标签划分为不同的话题，并认为若用户 u 常常在用户 v 之后转发话题 t 的标签，且时间差越小，则用户 v 在话题 t 上对用户 u 的影响 $I_t(u,v)$ 越大。利用 $I_t(u,v)$ 对用户 v 在话题 t 上影响的所有用户 u 的影响力进行加权，即可得到用户 v 在话题 t 上的影响力。

4 用户隐私研究

社交网络是真实社交关系的虚拟表达，社交网络最初的目的是给真实社交圈子中的同学、朋友等在网络上提供交流平台，所以多数网站要求用户使用真实资料注册，并在网站的个人主页上提供了包括身份资料、联系方式、生活动态在内的大量隐私数据。针对用户的数据，社交网络大都提供不同粒度的权限管理来帮助用户保护隐私，然而其用户隐私保护的力度仍然广受诟病。

对于社交网络中隐私保护上存在的问题，研究者首先需要解决的问题就是定义当前社交网络中的隐私现状，即现有的安全策略存在哪些不足之处。在已发现的问题中，Sybil 攻击即伪造用户现象较为严重，一些学者就相应的解决方案进行了探索。除了对特定隐私问题的解决方案进行研究外，许多学者也在考虑从根本上解决隐私安全的问题，即提出全新的权限理论模型并进行原型试验，如基于加密体系的模型、co-author 权限模型等。

本节余下部分将从隐私现状的研究、Sybil 攻击检测、新权限模型 3 个方面对用户隐私领域的研究工作进行总结。

4.1 隐私现状研究

社交网络中的个人信息数据、行为习惯数据可以说是用户的一部数据卷宗。通过强大的数据挖掘，数据拥有者可能会比用户更了解其自身。虽然社交网络 ICP 承诺会尽量保护用户的隐私数据，但是隐私问题还是无处不在。账号被盗、铺天盖地的垃圾信息推送、僵尸用户扎堆这些事件都在说明隐私问题的严重性，因此隐私现状的研究十分有必要。

隐私问题与社交网络用户密切相关，但对于哪些信息涉及到隐私问题，用户并没有明确的认识。基于此，孙剑等人定义了当前社交网络中的安全性需求，即社交网络用户的个人信息不被泄露，社交网络中的数据与现实中保持一致。文章以人人网、qq 空间、新浪微博为案例，深入分析了 3 大社交网络在信息保密性、数据完整性、数据可用性上的表现，并就存在的问题建议 ICP 加强用户默认权限的安全级别，开发内容审查技术，提供基于对象的安全策略[20]。

在用户隐私保护上，社交网络面临的最主要挑战在于用户的隐私保护意识淡薄，社交网络无法很好地引导用户进行权限设置，由此造成用户信息容易泄露。Balachander 等人分析了社交网络权限设置功能的使用情况，如 Facebook 可以针对一些信息项（如照片）设置对自己可见、朋友可见、所有人可见等，但多数的用户趋向使用默认权限即所有人可见，这说明社交网络用户的权限意识不强[17]。Yabing Liu 等人用 AMT 亚马逊土耳其机器人对 Facebook 的权限设置进行人工调研，该调研使用随机图片和敏感图片的权限数据作为对比，发现用户的隐私设置与他们期望的权限设置相差很大。作者量化了这种差异，发现大概只有 39％左右的权限设置与预期相符[46]，这说明社交网络的权限设置功能远远没有达到用户的期望。Markus Huber 等人则总结了当前社交网络造成用户信息泄露的场景[19]，主要包括对图谱数据进行提取，定位人的信息、地理信息；爬虫对用户数据进行二次聚集；主页抢注，钓鱼攻击等身份窃取；利用社会工程学的社交攻击；第三方应用乱用用户数据等。

除了由于用户或社交网络本身引起的隐私问题外，一些学者还分析了第三方带来的隐私问题。一些组织会在互联网上公开一些社交网络中的数据集以促进科研工作。Prateek Joshi 等人考虑了相关组织发布社交网络的运营数据所带来的问题，即匿名化的数据也可能会被破解从而造成隐私泄露，如使用图节点和图像结构进行异构映射的图匿名算法容易受到还原攻击和邻居识别攻击。作为可能的解决方案，文章分析了 q-Anon 概率模型和去中心化的 OSN 体系结构如何能够在保留足够的图谱研究信息的同时满足隐私保护的需求[47]。Mainack Mondal 则阐述了大规模爬虫系统在数据重聚集、数据非法使用等方面带来的隐私问题。传统的反爬虫技术包括 IP 和频率限制，然而这无法限制大规模 Sybil 用户对数据的爬取。文章假设 Sybil 用户无法与正常用户建立大规模的边，并提出了一个信用模型，当 Sybil 节点与其他节点路径上某一通路的信任值为 0 时，则禁止访问。该方法能限制 Sybil 节点的大量访问，但使用不当可能对正常用户造成 DDoS 攻击，并且 Credit 模型的建立尚未有好的方法[18]。

4.2 Sybil 攻击检测

Sybil 攻击起源于传统的 P2P 网络，攻击者使用大量伪造

的用户节点来达到某种目的。在社交网络中 Sybil 用户通常指机器账号，一些恶意用户或机构操作大量 Sybil 用户来传递广告信息，构建舆论趋势。Sybil 用户的存在给正常用户带来了极差的用户体验，并一定程度造成了社交网络繁荣的假象。因此，Sybil 攻击检测是隐私保护中一类很重要的问题，从已有的研究工作看，Sybil 节点的检测主要采用基于社交图谱和基于 Sybil 行为分析这两大类方法。

在基于社交图谱的 Sybil 检测领域里，学者们相继提出 SybilGuard[48]、SybilLimit[49]、SybilInfer[50]、SumUp[51] 4 个 Sybil 用户检测算法，这些算法共同的核心假设是所有 Sybil 节点之间趋向于高密度的互联，以便使自己看起来更像正常的用户；Sybil 节点和正常用户节点间难以建立大量的边。前 3 个算法通过随机游走理论在社交图上发现 Sybil 簇与正常用户的分割边，而 SumUp 算法则是通过最大流理论来发现分割边。上述算法的一个共同点就是利用社区发现的思想来发现 Sybil 簇。Bimal Viswanath 总结了这 4 种算法的假设、算法步骤以及社区发现的过程，并在人工注入了 Sybil 节点的社交网络进行 Sybil 簇的发现实验，结果表明算法在 Sybil 防范上是可行的[21]。基于社交图谱来检测 Sybil 节点的最新的研究成果是 SybilRank[52]，SybileRank 使用 power-iteration 来代替 SybilGuard 等算法的随机游走过程，算法的时间复杂度更低（$n\log n$），并且基于多社区的信任节点选择方法使得 SybilRank 对多社区图谱中的 Sybil 节点也具有较高的检测精度。

在基于行为分析的 Sybil 检测领域里，Zhi Yang 等人对 RenRen 中 Sybil 用户行为特征进行了研究，并对上述算法中关于 Sybil 高度互联的假设进行了挑战[22]。作者在深入分析了 RenRenSybil 用户发送好友请求、好友请求被接受、接受好友请求等行为特征后，设计了基于好友请求及聚合系数的阈值检测器。实验发现，基于上述特性的阈值检测器与基于机器学习的检测器在检测 Sybil 用户时都具有 90%以上的正确率。通过对 Sybil 节点图谱的深入分析，作者还发现 RenRen 的 Sybil 用户并没有表现出团簇的特性，它们与真实的用户建立更多的连接，这说明在真实的社交网络中使用 SybilGuard 等检测算法可能存在一定的局限性。Sybil 用户检测算法的使用需要根据网络类型的不同和前期数据的分析来判断是否满足其核心假设，否则将会造成误报率的增加。

4.3 新权限模型的研究

针对社交网络中存在的隐私安全问题，学者们也在积极探索新的安全模型。这主要是因为已有的隐私问题尚未得到有效的解决而一些新的权限需求相继出现。主流的研究思路是开发新的权限模型并以应用集成的方式加强隐私保护。

对于隐私保护，最直接方法是使用数据加密和真实数据替换。Randy Baden 等人提出使用公私钥体系来加密数据，这样就可以保证拥有密钥的用户才能访问数据，杜绝了隐私数据被第三方获取[23]。该应用使用 ABE 属性加密机制为用户的好友和用户组（圈子）进行密钥管理。并采用 API 的方式向 Facebook 提供应用集成。测试表明 Persona 模型在 Desktoop 和 Moble device 上额外的性能开销都在能接受的范围内。Mauro Conti 等人提出使用虚假信息替换的模型来达到隐私保护。该模型让用户在网上发布自己假的信息，让朋友或好友看到自己真实的信息[53]。真实的信息存在本地，使用 xml 与自己好友交互。好友通信和信息的替换由额外开发的 Face-VPSN 浏览器插件完成。该方案的好处在于不需要作为 APP 与 Facebook 集成，具有完全的自主性，而且去中心化的结构无单点瓶颈的问题。

然而数据加密等策略并不能解决 co-author 内容泄露隐私的问题。在社交网络中用户发布的照片、日志等内容可能包含其他好友的隐私信息，当前的社交网络没有提供相应权限控制功能。Anna 等人提出基于多用户合作的安全模型，使用加分机制鼓励多个用户共同完成可能涉及多好友的照片或日记等内容的权限设置，同时该模型还能提供基于朋友距离的细粒度内容访问控制功能。

区别于传统的方法，Amirreza 等人则使用语义 Web 的概念对整个网络的数据、人物、事件进行本体建模，然后开发模型引擎，权限的设定可以采用语义表达式。它的优点是强大的语义表达式可以细化权限控制的粒度和增加权限控制的灵活性[54]，它的不足之处在于对于普通用户，该理论理解起来太难，系统需要开发强大的交互式 UI 来引导用户进行权限设置。

5 研究展望

从前文可以看出，学者们从不同的角度对社交网络进行了大量的研究工作，并在数据测量、用户行为分析、信息传播、用户隐私等领域取得了大量的研究成果。然而随着研究的不断深入以及社交网络的飞速发展，新的问题和领域也不断出现，基于作者的理解，本节总结了社交网络未来可能的研究方向，以供参考。

社交网络测量领域：

（1）数据共享机制的研究：学术界一直以一种被动的姿态通过网络爬虫等技术来获得数据，通常的结果是耗费巨大的努力却拿不到最有价值的数据。ICP 与学术界如何贡献数据，需要研究者探讨共享哪些数据、怎样共享、数据如何匿名化、接口如何开放、对爬虫的约束条件等问题。

（2）有向图采样算法研究：目前的采样算法主要关注对象是基于无向图建模的社交网络，因为有向图在经过无向图转换后也可以采用上述算法。然而诸如微博之类的社交网络，因为入度获取的不完全性决定了无向图转化思路的不可行。基于出度的有向图均匀采样将是一个研究的难点。

（3）动态采样的研究：采样算法在采样周期中认为图谱是静态的，学者们并没有就采样周期里新加入的动态节点和边对采样效果的影响进行研究。此外，能否在采样过程中引入一些时序的概念来解决此问题是一个很有挑战的研究方向。

（4）增量更新的研究：对社交网络的一次测量通常需要消耗大量的资源，然而数据更新对于很多研究又必不可少。如何在已有数据的基础上达到增量式的数据采集，如通过新边、节点预测策略来实局部更新同样是值得研究的方向。

社交网络用户行为分析领域：

（1）细分用户的行为研究：在当前的学术研究中，用户行为分析的对象默认针对所有的用户，基于细分用户的行为分析工作比较少。在同一社交网络中不同的用户类型会呈现出不同的交互行为，如认证用户、公共主页用户、学生用户，基于细分用户群体的行为分析是学术界需要探索的领域，同时这里面还牵涉到对用户划分的问题，包括如何定义活跃用户、僵尸用户、新老用户等。

(2)横向对比研究:对于用户行为的分析,学者们的研究对象通常是针对特定的社交网络如 Facebook、Twitter,而对于综合性、强媒体性、垂直性等不同侧重点的社交网络的横向对比研究相对较少,如针对国内的 RenRen、Weibo、QQ 等不同性质社区的用户行为差异分析将是未来的一个热点。

(3)行为动机研究:就当前的用户行为分析来看,主要的研究成果基本上还在数据统计层面上,如大多数学者给出了用户使用社交网络功能的情况,但对用户行为动机深层次性的研究工作仍不多见。社交网络核心目的之一就是要保持社区的活跃,留住活跃用户,提高用户粘性。对于这方面的问题,仍需要研究者结合社交关系图谱和行为数据深入探索社交关系建立、行为驱动等刺激因素。

社交网络信息传播领域:

(1)信息传播影响力的研究:给定一个营销话题,如何控制潜在用户、时间及其他相关因素才能达到相应的影响力,影响力该如何定义和评估,这些开放性的问题都需要进一步的研究。

(2)横向对比研究:与用户行为类似,不同侧重点的社交网络在信息传播的速度、范围、爆发周期等各方面都存在差异,当前在这方面的研究工作仍然缺少对多个社交网络的横向对比。

(3)信息传播与预测模型的进一步研究:目前对社交网络中信息传播模型的研究局限于对病毒传播模型和创新事物的传播模型的改进,如何发掘用户在信息传播过程中的行为模式和驱动因素,从而精确地对信息传播过程进行建模,仍是一个需要进一步研究的课题。对于信息传播的预测,目前的算法或过于简单,得出的结论也较为简单而缺少指导意义;或较为复杂(事件和数据复杂度高),难以在社交网络上大规模地应用,预测的精度也不够高。简单而高效的预测算法仍是研究者需要解决的问题。

社交网络用户隐私领域:

(1)多维度的权限模型研究:社交网络的安全策略现状不能很好地满足用户的需求,这主要体现为社交网络的权限设置引导模块不友好、权限设置的粒度不够细。如何为用户定义默认的权限策略,增加诸如不同对象、不同时间、不同距离等多维度的权限设置是一个有意义的研究方向。

(2)新权限模型的集成研究:对于学者们提出的新的权限模型,用户需要以 APP 或插件的方式进行功能扩展,笔者觉得这是限制这些安全模型应用的最大阻力。关于这方面的问题,研究者需要进一步关注权限模型在新的社交网络中推广以及与已有社交网络无缝集成的问题。

(3)高效准确的 Sybil 账户检测机制的研究:社交网络存在大量的 Sybil 账户,以 RenRen 和新浪微博为例,Sybil 用户越来越多,给正常用户的体验造成了极大的影响,然而已有的研究成果并没有大规模的应用,如何将已有的检测算法应用到这些社交网络中或者开发更高效准确的检测算法仍是亟需解决的问题。

上述 4 个方面是目前学术界研究社交网络的主要出发点,然而相关的研究工作不会局限于此,利用社交网络的社交背景与现有研究领域相结合必将成为社交网络研究的新热点之一。以微博中的主题挖掘为例,在传统的研究基础上,需要考虑社交网络环境下短文本的特性,同时还需要考虑到这些非结构化数据背后结构化的因素,例如社交关系。这些特定的元素都会催生新的研究问题,并将产生许多依托于社交网络的研究工作,如基于社交网络的网络信任机制研究[55-57]、基于社交网络的 APP 分析[58-60]、基于微博的主题发现[61]、基于社交网络的情感分析[62]、基于社交网络的社区发现[63,64]等。另一方面,社交网络巨大的数据量和访问量甚至还会促进社交网络体系架构[65]、云存储技术、大数据处理、大图引擎等基础研究领域的工作。

结束语　社交网络的研究领域具有广而深的特点,能够研究的热点很多,本文主要涉及了数据采集、用户行为、信息传播、用户隐私 4 个领域的问题,并对相关的研究成果进行了总结、分析和展望。

对于社交网络数据获取,本文从时序上总结了出现的测量技术,即基于 BFS 算法的全网测量技术、采样测量技术、点击流数据获取技术。对于用户行为分析,本文从广义用户行为和个人用户行为的角度总结了已有的工作,主要关注了社交图谱研究、活动图研究、功能性研究等问题。对于信息传播,本文从信息传播规律发现、信息传播模型建立、信息传播预测等角度对已有的工作进行了总结。对于用户隐私,本文从发现问题、解决问题、提出新方案的思路总结了学术界在定义隐私问题、提出 Sybil 检测算法、提出新权限模型等方面的工作。

最后,本文就上述 4 个领域未来的研究方向进行了探讨,希望能给相关研究者提供有意义的参考。

参 考 文 献

[1] Chau D H, Pandit S, Wang S, et al. Parallel crawling for online social networks[C]// Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 1283-1284

[2] 胡亚楠. 社交网络获取技术与实现[D]. 哈尔滨:哈尔滨工业大学, 2011
HuYa-nan. Social network data acquisition technology and implementation[D]. Harbin: HIT Unversity, 2011

[3] Gjoka M, Butts C T, Kurant M, et al. Multigraph sampling of online social networks[J]. IEEE Journal on Selected Areas in Communications, 2011, 29(9): 1893-1905

[4] Jin L, Chen Y, Hui P, et al. Albatross sampling: robust and effective hybrid vertex sampling for social graphs[C]// Proceedings of the 3rd ACM International Workshop on MobiArch. ACM, 2011: 11-16

[5] Ye S, Lang J, Wu F. Crawling online social graphs[C]// 2010 12th International Asia-Pacific Web Conference (APWEB). IEEE, 2010: 236-242

[6] Ribeiro B, Towsley D. Estimating and sampling graphs with multidimensional random walks[C]// Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. ACM, 2010: 390-403

[7] Benevenuto F, Rodrigues T, Cha M, et al. Characterizing user behavior in online social networks[C]// Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference. ACM, 2009: 49-62

[8] Schneider F, Feldmann A, Krishnamurthy B, et al. Understanding online social network usage from a network perspective[C]// Proceedings of the 9th ACM SIGCOMM Conference on Internet

Measurement. ACM,2009:35-48

[9] Jiang J,Wilson C,Wang X,et al. Understanding latent interactions in online social networks[J]. ACM Transactions on the Web (TWEB),2013,7(4):18

[10] Viswanath B,Mislove A,Cha M,et al. On the evolution of user interaction in facebook[C]// Proceedings of the 2nd ACM Workshop on Online Social Networks. ACM,2009:37-42

[11] Hansen D,Shneiderman B,Smith M A. Analyzing social media networks with NodeXL:Insights from a connected world[M]. Morgan Kaufmann,2010

[12] Leskovec J,McGlohon M,Faloutsos C,et al. Patterns of Cascading behavior in large blog graphs[C]// SDM. 2007:551-556

[13] Cha M,Mislove A,Gummadi K P. A measurement-driven analysis of information propagation in the flickr social network[C]// Proceedings of the 18th International Conference on World Wide Web. ACM,2009:721-730

[14] Galuba W, Aberer K, Chakraborty D, et al. Outtweeting the twitterers-predicting information cascades in microblogs[C]// Proceedings of the 3rd Conference on Online Social Networks. 2010:3

[15] Leskovec J,McGlohon M,Faloutsos C,et al. Cascading Behavion in Large Blog Graphs Patterns and a Model[R]. 0704. 2803. arxiv,2007

[16] Liu D,Chen X. Rumor propagation in online social networks like twitter—a simulation study[C]// 2011 Third International Conference on Multimedia Information Networking and Security (MINES). IEEE,2011:278-282

[17] Krishnamurthy B,Wills C E. Characterizing privacy in online social networks[C]// Proceedings of the First Workshop on Online Social Networks. ACM,2008:37-42

[18] Mondal M,Viswanath B,Clement A,et al. Limiting large-scale crawls of social networking sites[J]. ACM SIGCOMM Computer Communication Review,2011,41(4):398-399

[19] Huber M,Mulazzani M,Weippl E. Social networking sites security:Quo Vadis[C]// IEEE International Conference on Social Computing/IEEE International Conference on Privacy,Security,Risk and Trust. IEEE,2010:1117-1122

[20] **孙剑,朱晓妍,刘沫盟,等. 社交网络中的安全隐私问题研究[J]. 网络安全技术与应用**,2011 (10):76-79
Sun Jian,Zhu X Y,LIU M M ,et al. Privacy Issues of Social Network[J]. Network Security Technology &Application ,2011 (10):76-79

[21] Viswanath B,Post A,Gummadi K P,et al. An analysis of social network-based sybil defenses[J]. ACM SIGCOMM Computer Communication Review,2011,41(4):363-374

[22] Yang Z, Wilson C, Wang X, et al. Uncovering social network sybils in the wild[J]. ACM Transactions on Knowledge Discovery from Data (TKDD),2014,8(1):2

[23] Baden R, Bender A, Spring N, et al. Persona: an online social network with user-defined privacy[J]. ACM SIGCOMM Computer Communication Review,2009,39(4):135-146

[24] **冯典. 面向微博的数据采集和分析系统的设计与实现[D]. 北京:北京邮电大学**,2011
Feng Dian. The design and implementation of the data acquisition and analysis system for micro-blog[D]. Beijing:Beijing University of Posts and Telecommunications,2011

[25] CNpameng. Weibo-Crowdsourcing-CNpameng[EB/OL]. http://www. cnpameng. com/

[26] Chau D H,Pandit S,Wang S,et al. Parallel crawling for online social networks[C]// Proceedings of the 16th international conference on World Wide Web. ACM,2007:1283-1284

[27] **程锦佳. 基于 Hadoop 的分布式爬虫及其实现[D]. 北京:北京邮电大学**,2010
Cheng Jin-jia. Research and implementation of distributed web crawl based on Hadoop architecture[D]. Beijing:Beijing University of Posts and Telecommunications,2010

[28] **李娜娜. 云计算平台下社交网络数据获取技术研究[D]. 北京:北京邮电大学**,2013
Li Na-na. Research on data acquisition technology of social network under cloud computing platform[D]. Beijing:Beijing University of Posts and Telecommunications,2013

[29] **袁小节. 基于协议驱动与事件驱动的综合聚焦爬虫研究与实现[D]. 长沙:国防科技大学**,2009
Yuan Xiao-Jie. Research and Implementation of a Combined Focused Crawler based on Protocol-Driven and Event-Driven Crawling Techniques[D]. Changsha: National Defense Science and Technology University,2009

[30] **曾伟辉. 支持 AJAX 的网络爬虫系统设计与实现[D]. 合肥:中国科学技术大学**,2009
Zeng Wei-hui. Design and Implementation of a Web Crawler System Supported AJAX[D]. Hefei:University of Science and Technology of China,2009

[31] **夏天**. Ajax **站点数据采集研究综述[J]. 现代图书情报技术**,2010(3):52-57
Xia Tian. Overview of Research on Data Collection from Ajax Sites[J]. New Technology of Library and Information Service,2010(3):52-57

[32] Mislove A,Marcon M,Gummadi K P,et al. Measurement and analysis of online social networks[C]// Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM,2007:29-42

[33] Wang D,Li Z,Xie G. Towards unbiased sampling of online social networks[C]// 2011 IEEE International Conference on Communications (ICC). IEEE,2011:1-5

[34] Wang T,Chen Y,Zhang Z,et al. Unbiased sampling in directed social graph[J]. ACM SIGCOMM Computer Communication Review,2011,41(4):401-402

[35] Gjoka M,Kurant M,Butts C T,et al. Walking in Facebook:A case study of unbiased sampling of OSNs[C]// 2010 Proceedings IEEE INFOCOM. IEEE,2010:1-9

[36] Chun H,Kwak H,Eom Y H,et al. Comparison of online social relations in volume vs interaction:a case study of cyworld[C]// Proceedings of the 8th ACM SIGCOMM conference on Internet Measurement. ACM,2008:57-70

[37] Wilson C, Boe B, Sala A, et al. User interactions in social networks and their implications[C]// Proceedings of the 4th ACM European conference on Computer systems. ACM, 2009: 205-218

[38] Guo Z,Li Z,Tu H. Sina microblog:an information-driven online social network[C]// 2011 International Conference on Cyberworlds (CW). IEEE,2011:160-167

[39] Wu X, Wang J. How about micro-blogging service in China: analysis and mining on sina micro-blog[C]// Proceedings of 1st international symposium on From digital footprints to social and community intelligence. ACM, 2011: 37-42

[40] Saito K, Kimura M, Ohara K, et al. Selecting information diffusion models over social networks for behavioral analysis[M]// Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2010: 180-195

[41] Yang J, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter[J]. ICWSM, 2010, 10: 355-358

[42] Guille A, Hacid H. A predictive model for the temporal dynamics of information diffusion in online social networks[C]// Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012: 1145-1152

[43] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine, 1998[C]// Proceedings of the Seventh World Wide Web Conference. 2007

[44] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]// Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 261-270

[45] Saez-Trumper D, Comarela G, Almeida V, et al. Finding trendsetters in information networks[C]// Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 1014-1022

[46] Liu Y, Gummadi K P, Krishnamurthy B, et al. Analyzing facebook privacy settings: user expectations vs. reality[C]// Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011: 61-70

[47] Joshi P, Kuo C C J. Security and privacy in online social networks: A survey[C]// 2011 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2011: 1-6

[48] Yu H, Kaminsky M, Gibbons P B, et al. Sybilguard: defending against sybil attacks via social networks[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(4): 267-278

[49] Yu H, Gibbons P B, Kaminsky M, et al. Sybillimit: A near-optimal social network defense against sybil attacks[C]// IEEE Symposium on Security and Privacy, 2008. IEEE, 2008: 3-17

[50] Danezis G, Mittal P. SybilInfer: Detecting Sybil Nodes using Social Networks[C]// NDSS. 2009

[51] Tran D N, Min B, Li J, et al. Sybil-Resilient Online Content Voting[J]. NSDI. 2009, 9(1): 15-28

[52] Cao Q, Sirivianos M, Yang X, et al. Aiding the detection of fake accounts in large scale social online services[C]// Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012: 15-18

[53] Conti M, Hasani A, Crispo B. Virtual private social networks[C]// Proceedings of the first ACM conference on Data and application security and privacy. ACM, 2011: 39-50

[54] Masoumzadeh A, Joshi J. Osnac: An ontology-based access control model for social networking systems[C]// 2010 IEEE Second International Conference on Social Computing (SocialCom). IEEE, 2010: 751-759

[55] Barbian G. Assessing trust by disclosure in online social networks[C]// 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2011: 163-170

[56] **鲍捷, 程久军. 基于社交网络的群体信任算法[J]. 计算机科学**, 2012, 39(2): 38-41
Bao Jie, Cheng J J. Group Trust Algorithm based on Online Social Network[J]. Computer Science, 2012, 39(2): 38-41

[57] **乔秀全, 杨春, 李晓峰, 等. 社交网络服务中一种基于用户上下文的信任度计算方法[J]. 计算机学报**, 2011, 34(12): 2403-2413
Qiao X Q, Yang Chun, Li X F, et al. Trust Model based on user context in Online Socail Network[J]. Chinese Journal of Computers, 2011, 34(12): 2403-2413

[58] Nazir A, Raza S, Gupta D, et al. Network level footprints of facebook applications[C]// Proceedings of the 9th ACM SIGCOMM conference on Internet measurement. ACM, 2009: 63-75

[59] Nazir A, Raza S, Chuah C N. Unveiling facebook: a measurement study of social network based applications[C]// Proceedings of the 8th ACM SIGCOMM conference on Internet measurement. ACM, 2008: 43-56

[60] Gjoka M, Sirivianos M, Markopoulou A, et al. Poking facebook: characterization of osn applications[C]// Proceedings of the first workshop on Online social networks. ACM, 2008: 31-36

[61] **张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展**, 2011, 48(10): 1795-1802
Zhang C Y, Sun J L, Ding Y Q. TopicMining of WeiBo Based on MB-LDA Model[J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802

[62] Mao H, Counts S, Bollen J. Predicting financial markets: Comparing survey, news, twitter and search engine data[J]. arXiv preprint arXiv: 1112. 1051, 2011

[63] Bhat S Y, Abulaish M. A density-based approach for mining overlapping communities from social network interactions[C]// Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. ACM, 2012: 9

[64] Sachan M, Contractor D, Faruquie T A, et al. Using content and interactions for discovering communities in social networks[C]// Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 331-340

[65] Sharma R, Datta A. Supernova: Super-peers based architecture for decentralized online social networks[C]// 2012 Fourth International Conference on Communication Systems and Networks (COMSNETS). IEEE, 2012: 1-10

[66] Ye S, Wu S F. Estimating the size of online social networks[J]. International Journal of Social Computing and Cyber-Physical Systems, 2011, 1(2): 160-179

[67] Kurant M, Markopoulou A, Thiran P. On the bias of bfs (breadth first search)[C]// 2010 22nd International Teletraffic Congress (ITC). IEEE, 2010: 1-8

[68] Bonneau J, Anderson J, Danezis G. Prying data out of a social network[C]// International Conference on Advances in Social Network Analysis and Mining, 2009 (ASONAM'09). IEEE, 2009: 249-254

[69] Kurant M, Gjoka M, Butts C T, et al. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks [C]// Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems. ACM, 2011: 281-292

(**下转第 42 页**)

1)从 KNC cache 的体系结构[15]入手，深入研究 cache 预取的机制；

2)对 KNC 上的 MAO 进行性能建模[16]。

文中涉及的 BSF 和 Diffusion 3D 的 MAO 源代码及原始测试数据都已放在 Github[17]上，以供读者参考。

参 考 文 献

[1] Satish N, Kim C, Chhugani J, et al. Can traditional programming bridge the Ninja performance gap for parallel computing applications?[C]// 2012 39th Annual International Symposium on Computer Architecture (ISCA). 2012: 440-451

[2] Xue W, Yang C, Fu H, et al. Enabling and Scaling a Global Shallow-Water Atmospheric Model on Tianhe-2[C]// Proceedings of the 2014 IEEE 28th International Parallel and Distributed Processing Symposium. 2014

[3] Pennycook S J, Hughes C J, Smelyanskiy M, et al. Exploring SIMD for Molecular Dynamics, Using Intel Xeon Processors and Intel Xeon Phi Coprocessors[C]// Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing. 2013: 1085-1097

[4] Heinecke A, Vaidyanathan K, Smelyanskiy M, et al. Design and Implementation of the Linpack Benchmark for Single and Multi-node Systems Based on Intel Xeon Phi Coprocessor[C]// Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing. 2013: 126-137

[5] Krishnaiyer R, Kultursay E, Chawla P, et al. Compiler-Based Data Prefetching and Streaming Non-temporal Store Generation for the Intel(R) Xeon Phi(TM) Coprocessor[C]// Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing Workshops and PhD Forum. 2013: 1575-1586

[6] Hofmann J, Treibig J, Hager G, et al. Performance Engineering for a Medical Imaging Application on the Intel Xeon Phi Accelerator[C]// 2014 27th International Conference on Presented at the Architecture of Computing Systems (ARCS). 2014: 1-8

[7] Jeffers J, Reinders J. Intel Xeon Phi Coprocessor High Performance Programming (1st edition)[M]. Morgan Kaufmann Publishers Inc, 2013

[8] Rahman R. Intel Xeon Phi Coprocessor Architecture and Tools: The Guide for Application Developers[M]// Intel Xeon Phi Coprocessor Architecture and Tools: The Guide for Application Developers (1st edition). 2013

[9] Saini S, Jin H, Jespersen D, et al. An early performance evaluation of many integrated core architecture based SGI rackable computing system[C]// Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. 2013

[10] Hofmann J. Performance Evaluation of the Intel ManyIntegrated Core Architecture for 3D Image Reconstruction in Computed Tomography (Master Thesis)[M]. Friedrich-Alexander-University Erlangen-Nuremberg, 2010

[11] Fang J, Sips H, Zhang L, et al. Test-driving Intel Xeon Phi[C]// Proceedings of the 5th ACM/SPEC International Conference on Performance Engineering. New York, USA, 2014: 137-148

[12] SHOC-MIC benchmark[OL]. https://github.com/vetter/shoc-mic

[13] Likwid[OL]. https://code.google.com/p/likwid/

[14] PAPI[OL]. http://icl.cs.utk.edu/papi/

[15] Ramos S, Hoefler T. Modeling communication in cache-coherent SMP systems: a case-study with Xeon Phi[C]// Proceedings of the 22nd International Symposium on High-performance Parallel and Distributed Computing. New York, USA, 2013: 97

[16] Hoefler T, Gropp W, Kramer W, et al. Performance modeling for systematic performance tuning[C]// 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC). 2011: 1-12

(上接第 21 页)

[70] Valafar M, Rejaie R, Willinger W. Beyond friendship graphs: a study of user interactions in Flickr[C]// Proceedings of the 2nd ACM workshop on Online social networks. ACM, 2009: 25-30

[71] Yang J, Leskovec J. Modeling information diffusion in implicit networks[C]// 2010 IEEE 10th International Conference on Data Mining (ICDM). IEEE, 2010: 599-608

[72] Bakshy E, Rosenn I, Marlow C, et al. The role of social networks in information diffusion[C]// Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 519-528

[73] Tang S, Yuan J, Mao X, et al. Relationship classification in large scale online social networks and its impact on information propagation[C]// 2011 Proceedings IEEE INFOCOM. IEEE, 2011: 2291-2299

[74] Hansen D L, Johnson C. Veiled viral marketing: disseminating information on stigmatized illnesses via social networking sites [C]// Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. ACM, 2012: 247-254

[75] Alahakoon T, Tripathi R, Kourtellis N, et al. K-path centrality: A new centrality measure in social networks[C]// Proceedings of the 4th Workshop on Social Network Systems. ACM, 2011: 1-6

[76] Burke M, Kraut R, Marlow C. Social capital on Facebook: Differentiating uses and users[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2011: 571-580

[77] 张彦超，刘云，张海峰，等. 基于在线社交网络的信息传播模型[J]. 物理学报，2011，60(5)：60-66
Zhang Y C, Liu Yun, Zhang H F, et al. Information diffusion Model based on Online Social Network[J]. Acta Phys. Sin, 2011, 60(5): 60-66

[78] Mohaisen A, Yun A, Kim Y. Measuring the mixing time of social graphs[C]// Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. ACM, 2010: 383-389

[79] Squicciarini A C, Shehab M, Paci F. Collective privacy management in social networks[C]// Proceedings of the 18th International Conference on World Wide Web. ACM, 2009: 521-530

[80] 徐婕，康慕宁，董谷音. 基于社交网络的实时搜索引擎的排序算法研究[J]. 科学技术与工程，2011，11(28)：6879-6882
Xu Jie, Kang M N, Dong G Y. Sequencing Algorithm Based on the Social Network Real-time Search Engine[J]. Science Technology and Engineering, 2011, 11(28): 6879-6882