在线社会网络中信息扩散

李 栋[1)] 徐志明[1)] 李 生[1)] 刘 挺[1)] 王秀文[2)]

[1)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

[2)](国家计算机网络应急技术处理协调中心 北京 100029)

摘 要 在线社会网络中信息扩散研究可以帮助网络用户获取有价值信息、帮助企业推广产品、帮助政府调控舆情,应用价值巨大.该文旨在综述在线社会网络中信息扩散研究的现状.首先详细阐述了研究背景和研究意义;随后将当前研究划分为基于理论扩散模型的研究和基于信息扩散级联的研究两类,前者包括信息扩散特性研究、信息扩散概率计算、信息扩散最大化问题和竞争性的信息扩散最大化问题,后者包括信息扩散特性研究、用户影响力计算和信息扩散预测模型,对上述各方向的研究方法和研究进展进行了概括、比较和归纳,同时对各研究方向之间的内在关联进行了深入分析;接着探讨了信息扩散动态性和在线社会网络动态性的关系;最后对该研究目前存在的问题和一些未来发展方向进行了总结.

关键词 信息扩散;在线社会网络;预测;影响力;网络动态性

中图法分类号 TP393 **DOI 号** 10.3724/SP.J.1016.2014.00189

A Survey on Information Diffusion in Online Social Networks

LI Dong[1)] XU Zhi-Ming[1)] LI Sheng[1)] LIU Ting[1)] WANG Xiu-Wen[2)]

[1)](*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

[2)](*National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing* 100029)

Abstract The research of information diffusion in online social networks can help web users to obtain valuable information, help businesses to promote their products, help governments to regulate public opinion, and the application value of the research is huge. This paper aims to make a comprehensive survey of the research of information diffusion in online social networks. Firstly, this paper elaborates background and significance of the research in detail. Secondly, we divide current field into the research based on theoretical diffusion models and the research based on information diffusion cascades, research directions of the former include information diffusion characteristic research, calculation of information diffusion probability, information diffusion maximization problem and competitive information diffusion maximization problem, research directions of the latter include information diffusion characteristic research, calculation of users' influence and predictive model of information diffusion. This paper makes a summary, comparison and generalization of research methods and progress for above research directions, and analyses deeply the relations among these different research directions. Then, we explore the relationship between information diffusion dynamic and online social network dynamic. Finally we summarize the existing problems and future directions of this research field.

Keywords information diffusion; online social networks; prediction; influence; network dynamic

收稿日期:2012-04-11;最终修改稿收到日期:2013-08-03.本课题得到国家自然科学基金(61173074)、国家"八六三"高技术研究发展计划项目基金(2011AA01A207)资助.**李 栋**,男,1987 年生,博士研究生,主要研究方向为社会计算、信息扩散.E-mail:hitlidong@hit.edu.cn.**徐志明**(通信作者),男,1967 年生,博士,教授,主要研究领域为社会计算、信息检索.E-mail:xuzm@hit.edu.cn.**李 生**,男,1943 年生,博士,教授,博士生导师,主要研究领域为自然语言处理、机器翻译和信息检索.**刘 挺**,男,1972 年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为社会计算、信息检索.**王秀文**,女,1975 年生,高级工程师,主要研究领域为社会网络分析.

1 引 言

Facebook、Twitter 等社交类网站的迅猛发展，预示着社会媒体(Social Media)成为当今互联网发展的热点和趋势，它正逐渐影响着网络用户的生活、学习和工作.社会媒体是一种在线交互媒体，具有广泛的用户参与性，用户在这种新媒体上不再是信息的被动接收者，而变成了信息的主动发布者，用户真正成为互联网的主体.社会媒体中的用户可以建立各种关系(关注、好友等)，从而产生了各种不同的虚拟的在线社会网络，这些在线社会网络支撑着用户之间的交互和信息的发布扩散.

社会媒体(Web 2.0 媒体)中用户获取信息的方式相比于传统的网页媒体 (Web 1.0 媒体)发生了巨大变化.传统的网页媒体中，用户主要通过搜索引擎从海量网页中获取自己所需信息.而在社会媒体中，搜索引擎虽然仍是重要的信息获取工具，但不再是互联网用户获取信息的最主要途径.在社会媒体中，用户是信息的主要发布者，用户如果对某方面的信息感兴趣，那么该用户需要做的是同发布这方面信息的用户建立社会链接，构建自己的社会网络，那么实时的相关信息就会通过社会网络扩散到该用户那里，可见信息在社会网络中的扩散对帮助用户获取信息起着至关重要的作用.另外，在线社会网络同现实世界中的社会网络是相互映射影响的，所以通过研究在线社会网路中的信息扩散问题，可以帮助发现现实生活中不易发现的社会现象或者社会问题.综上所述，在线社会网络中信息扩散研究不仅可以帮助用户获取信息，还可以研究现实社会问题，因此具有非常重要的研究意义.

在线社会网络中信息扩散研究具有极其广泛的应用价值，主要包括病毒式市场营销、在线广告投放、信息推荐和谣言控制等多个方面.这里需要说明的是，在线社会网络中扩散的信息，并不狭隘的限制于文本信息(微博、博客等)，还包括多媒体信息(视频、图片等)、产品信息等各类信息.信息在网络中的扩散是用户行为引起的，例如，在微博中，用户的转发行为引发了微博信息在网络中扩散，所以信息扩散的研究本质上是用户行为的研究.

信息扩散在当前学术研究中有多种称谓，主要有信息扩散(Information Diffusion)、信息传播(Information Propagation、Information Spread)、信息流动(Information Flow)等.早期，信息扩散的研究人员多是一些社会学家、传染病学家和经济学家，他们主要研究创新(Innovation)、传染病(Epidemic)和产品(Product)在真实社会网络中的扩散，但是受限于真实社会中数据获取的困难，这些研究通常采用的数据集很小，而且多是一些定性的研究.伴随着社会媒体的发展，大量丰富的在线数据可以非常方便的获取，这些在线数据不仅包括大规模的在线社会网络数据，还有海量信息在网络中扩散的数据，这为信息扩散的研究带来了新的契机，从而成为研究热点.由于信息扩散研究涉及网络结构分析、文本内容分析、大规模数据处理等多个问题，所以该研究目前吸引了大量的复杂网络、自然语言处理和信息检索等计算机领域的学者们.

近几年，国际学术界中针对在线社会网络中信息扩散的研究大量出现：(1) 计算机领域的数据挖掘、互联网技术会议(KDD、WWW、WSDM、ICDM、SIGIR、CIKM 等)中，信息扩散研究相关的文章逐年增加；(2) 高影响因子期刊 PNAS 连续刊载了数篇信息扩散相关的文章.目前，研究在线社会网络中信息扩散的研究单位和研究者，国外的主要有斯坦福大学(Jure Leskovec, Eldar Sadikov)、东北大学(Albert-laszlo Barabasi)、哈佛大学(Nicholas Christakis)、密歇根大学(Lada A Adamic, Eytan Bakshy)、康奈儿大学(Jon Kleinberg)、南加州大学(Greg Ver Steeg, Rumi Ghosh, Kristina Lerman)、卡内基梅隆大学(Christos Faloutsos, Mary McGlohon, Katia Sycara)、麻省理工学院(Wei Pan)、加州大学圣地亚哥分校(James Fowler)、微软研究院(Scott Counts, Wei Chen)；国内主要有清华大学(Jie Tang, Zi Yang, Juanzi Li)、北京航空航天大学(Jichang Zhao).

基于对信息扩散研究的大量调研，本文以是否直接研究真实的信息扩散数据(信息扩散级联)为划分基准，将目前的研究大体划分为两大类(如图 1 所示)，基于理论扩散模型的研究和基于信息扩散级联的研究，每类研究下又有相应的子方向.

这里需要特别指出的是，本文虽然按照上述分类进行介绍，但是基于理论扩散模型的研究和基于信息扩散级联的研究并非完全孤立的两类研究，下面说明一下各个研究方向之间的内在关联关系.如图 2 所示，在线社会网络可以使用一个有向或者无向图 $G=(V,E)$ 表示，图中节点代表用户，本文中节点和用户是同义.V 代表节点集合，节点有激活(浅色)和未激活(深色)两类状态，表示为 $V=V_{\text{active}} \cup$

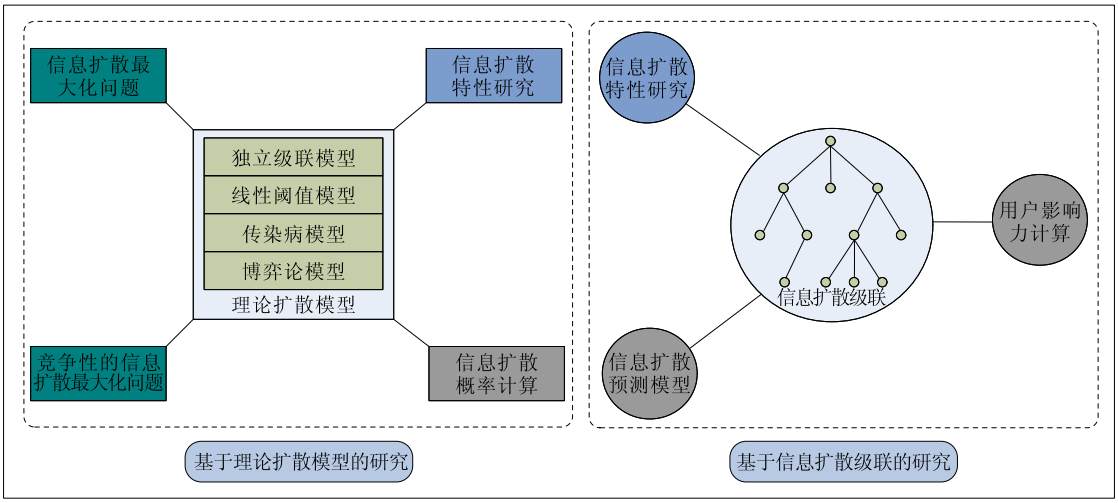

图 1　在线社会网络中信息扩散的主要研究方向

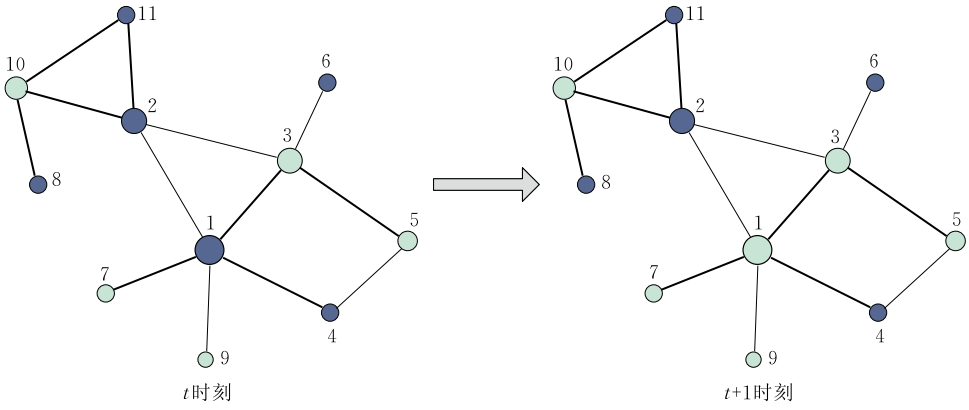

图 2　信息扩散示例

V_{inactive}，图中节点的权重代表用户影响力，对应着用户影响力计算研究；E 代表边的集合，图中边的权重代表用户间关系强度，对应着信息扩散概率计算研究. 信息扩散是一个时间动态性的过程，这里使用 V^t 代表时刻 t 节点的状态集合，信息扩散过程可描述为 $V^t \rightarrow V^{t+1}$，它是用户状态集合的跳转过程，部分未激活态的节点随着时间递进变成激活状态. 信息扩散特性研究主要研究哪些因素影响了用户状态的转换、用户状态转换的群体特征等问题，此类研究是其它研究的基础. 信息扩散模型是信息扩散研究的核心，主要任务就是充分考虑各种影响信息扩散的因素，理解、模拟并预测信息扩散过程，可以分为理论扩散模型和信息扩散预测模型，前者通常不具有预测信息扩散动态性的能力，两类模型通常会考虑用户之间的关系强度（信息扩散概率计算）对信息扩散的影响. 信息扩散最大化问题和竞争性的信息扩散最大化问题是信息扩散研究的应用性问题，当前上述两个问题的研究主要是基于理论扩散模型开展的，同时这两类问题的研究同用户影响力计算是息息相关的.

本文第 2 节首先介绍经典的理论扩散模型，随后介绍了基于理论模型的相关研究方向，此类研究主要使用理论扩散模型模拟信息在社会网络中的扩散过程来开展；第 3 节首先介绍不同的社会媒体数据集及其从中抽取信息扩散级联的方法，随后介绍基于信息扩散级联的相关研究方向，此类研究主要基于真实信息扩散数据的挖掘分析来开展；第 4 节介绍信息扩散动态性和网络动态性的关联；第 5 节总结全文并对未来的研究进行展望.

2　基于理论扩散模型的研究

基于理论扩散模型的研究不直接将真实信息扩散数据作为研究对象，而是基于一些经典的扩散模型或者这些模型的拓展，结合社会网络来开展. 过去，社会学家、传染病学家和市场学家提出了创新（Innovation）扩散模型、传染病（Epidemic）扩散模型、产品（Product）扩散模型等，这些模型虽然是针

对现实社会网络提出的，但是创新和传染病等在现实社会网络中的扩散，同信息在在线社会网络中的扩散是有相似之处的，所以这些模型在在线社会网络中也有一定的适用性. 当前，基于理论模型的研究大都使用非计算机领域学者们提出的经典模型或者这些模型的拓展. 经典扩散模型包括独立级联模型、线性阈值模型、传染病模型和博弈论模型. 基于理论扩散模型的研究的主要方向包括信息扩散特性研究、信息扩散最大化问题、竞争性的信息扩散最大化问题、信息扩散概率计算. 接下来在 2.1 节介绍经典的理论扩散模型，2.2 节介绍基于理论模型的主要研究方向和研究进展. 需要指出的是，拓展经典的理论扩散模型或者提出全新的理论扩散模型也是一个重要的研究方向，但是模型的拓展通常是为解决某个具体问题，例如研究人员将独立级联模型进行了拓展，来解决竞争性的信息扩散最大化问题. 因此理论模型的拓展或者提出通常是包含在我们上面提出的四个研究方向中，所以本文并未把理论模型作为一个单独的研究方向，而 2.1 节的主要目的是介绍一些经典的理论扩散模型，并非一个单独的方向，是 2.2 节内容的基础.

2.1 理论扩散模型

在信息扩散动态进程中，社会网络（为方便描述，文中社会网络即代表在线社会网络）中的用户是否会采纳某信息（也称作用户被信息激活或者感染），主要同社会因素和信息因素相关. 表 1 中展示了主要的几种扩散模型所考虑到的扩散因素（社会因素和信息因素）.

表 1 扩散模型的比较和分析

扩散模型	社会因素		信息因素
	单个邻居	多个邻居	
独立级联模型	是	否	否
线性阈值模型	否	是	否
传染病模型	否	否	是
博弈论模型	否	是	是

社会因素是指与用户有社会关系的邻居，邻居们的策略选择会影响到用户的行为，而用户可能某时刻单独考虑每个邻居的选择，也有可能同时考虑多个邻居的选择；信息因素是指信息本身所包含的内容及其扩散性（例如信息流行度）对用户的影响. 表 1 中展示了主要的几种扩散模型所考虑到的扩散因素，可以看出前 3 种模型都只考虑了单一的因素，而博弈论模型考虑到了多邻居的社会因素和信息因素，可见该模型是一种比较相对完善的模型. 但是，当前基于博弈论模型的研究比较少，因此基于博弈论模型的研究具有较大的研究潜力. 接下来详细介绍一下各个理论扩散模型.

独立级联模型[1-2]（Independent Cascade Model）. 在该模型中，每个初始激活节点会产生自己独立的扩散级联，级联之间是互相独立、互不干扰的. 在一个有向或无向图上，任何一条边 (u,v) 都被分配一个属于 $[0,1]$ 之间的特定值 $p_{u,v}$，这个值代表边 (u,v) 上的扩散概率. 初始时，部分节点被激活，信息从这些节点开始扩散. 在时间 t，每个当前激活的节点 u 都会以一定概率 $p_{u,v}$ 去激活它的每个邻居节点 v. 如果在时刻 t，节点 v 的多个邻居节点要激活它，这些邻居节点会随机排列地去尝试激活，所有激活尝试都在时刻 t 内完成. 无论邻居节点 u 是否成功，在随后的时刻都不能再去激活 v. 如果 v 在 t 时刻被激活，那么该节点会在 $t+1$ 时刻去激活它的邻居们. 该进程直到不再有激活行为发生而终止.

线性阈值模型[3]（Linear Threshold Model）. 在该模型中，对于网络中每个节点 v，所有同 v 相连的边的权重需要满足 $\sum_{u\in\Gamma(v)} p_{u,v}\leq 1$，$\Gamma(v)$ 代表 v 的所有邻居节点. 初始时，同独立级联模型一样，一部分节点被激活，对于任意一个节点，都被分配一个 $[0,1]$ 范围内的阈值 θ_v. 在时刻 t，一个未激活节点 v 会同时受到它的所有已激活的邻居节点的影响，如果所有邻居节点对其影响概率 p 之和大于 v 的阈值，即 $\sum_{u\in\Gamma_t(v)} p_{u,v}\geq\theta_v$，则节点 v 在 t 时刻被激活，在 $t+1$ 时刻成为激活状态. 该进程直到不再有激活行为发生而终止.

传染病模型[4-6]（Epidemics Model）. 由于传染病扩散和信息扩散的深度关联，传染病模型完全可以应用到信息扩散研究中，目前研究中使用最为广泛的传染病模型是 SIS 模型和 SIR 模型. 在 SIS 模型中，节点有两种状态：易受感染状态和已受感染状态. 在时刻 t，如果节点 v 周围有一个或者多个已受感染的节点，那么节点 v 会有大小为 p 的概率变成已受感染状态，这里的概率 p 通常是固定的，同信息本身相关，不随用户关系的变化而变化，因此我们认为传染病模型只考虑了信息因素. 一段时间之后，节点 v 可能又重新回归易感染状态. 在 SIR 模型中，节点除了有上述两种状态外，还有第 3 种免疫状态. 免疫状态是说，用户受到感染被治愈后具有了免疫性，不再被感染，也不会感染其它节点.

博弈论模型[7-8]（Game-Theoretic Model）. 该模

型的核心理念是用户根据个人利益最大化来主导自己的行为. 博弈论模型中用户是否采纳某信息，是从该用户本身利益最大的角度决定的，就是说当一个用户接触到某信息时，这个用户会根据自己可获取的最大利益来做理性的最佳选择. 在这样的模型中，对于一个用户来说，在做出行为选择时，会得到社会利益和个人利益两方面利益，社会利益是指用户做出同周围邻居相同选择时所得到的利益；而个人利益是指信息本身内容满足用户的个人偏好所带来的利益. 该模型也不同于线性阈值模型，因为用户不是基于阈值进行选择，而是衡量不同策略选择的利益大小，做出利益最大化的决定.

接下来通过实例(图 2 中节点 1)来进一步解释一下各个模型的内部机制. 在独立级联模型中，节点 3、7、9 在时刻 t 分别以一定概率去激活节点 1，如果其中一个成功，那么节点 1 在 $t+1$ 时刻将变成激活状态，如果不成功，节点 3、7、9 在以后的时刻也不能再去激活节点 1. 在线性阈值模型中，节点 3、7、9 会在时刻 t 同时去激活节点 1，当这 3 个节点的激活能力大于节点 1 的被激活阈值时，节点 1 在时刻 $t+1$ 会变成激活状态. 在传染病模型中，当节点 1 周围有节点是激活状态时，该节点就会以一定概率被激活，在时刻 $t+1$ 成为激活态，而且在一段时间之后，节点 1 可以重新回归未激活态，而前两种模型中激活态节点是不能回归未激活态的. 在博弈论模型中，节点 1 被视为一个智能个体，它会同时考虑它周围激活态节点(3、7 和 9)和未激活态节点(2 和 4)，计算选择哪种状态自己可以获取更大利益，此外还会考虑信息内容对自身带来的利益，最终做出让自身利益最大的选择.

2.2　主要研究方向

2.2.1　信息扩散特性研究

信息扩散特性主要研究影响信息扩散的因素、信息扩散同社会网络结构的关系等问题，此类研究帮助理解信息是如何在社会网络上扩散的，为其它研究提供基础性的研究结论. 信息扩散特性既可以基于理论扩散模型进行研究，也可以基于信息扩散级联开展研究. 基于理论扩散模型的信息扩散特性研究的思路通常是，使用扩散模型模拟信息在社会网络上的扩散过程，通过对比不同条件下的扩散效果，来挖掘出信息扩散的特性. 本节主要介绍基于理论模型的信息扩散特性研究，而基于信息扩散级联的扩散特性研究将在 3.2.1 节中进行介绍. 当前研究探究了以下 3 个方面对信息扩散的影响或者关联.

(1) 信息本身的扩散性.

Xu 等人[9] 基于 SIS，采用 Facebook 中的社会网络和经典的 BA 网络进行了实验，通过比较不同扩散概率(SIS 模型中的传染概率为 p)的信息的扩散情况，发现扩散概率大的信息可以感染更多的节点；许晓东等人[10] 基于另一种传染病模型 SIR，采用新浪微博的社会网络进行了实验，得到了同 Xu 等人同样的结论.

(2) 社会网络中的弱关系和结构洞.

网络中的弱关系是 Granovetter[11] 提出的，他们发现了弱关系可以帮助人找到新工作，过去有人研究弱关系和传统社会网络(例如手机通信网络[12])的关联. 随着在线社会网络的出现，Zhao 等人[13] 研究了在线社会网络中的弱关系对信息扩散的影响，关系的强弱是通过用户之间邻居的重叠数量衡量的. 他们提出了 $ID(\alpha,\beta)$ 信息扩散模型，该模型可以视为独立级联模型和传染病模型的组合，它通过设置不同 α 可以选择不同强度的边扩散，而 β 是信息本身的扩散性，对于一个节点 i，它会有 $d(i)\beta$ 个邻居被传染，$d(i)$ 是节点 i 的度. 通过使用该模型在 YouTube 和 Facebook 数据集上的模拟实验发现，社会网络中的弱关系对信息扩散有着微妙的作用，弱关系相当于一个桥，它可以把孤立的团体链接起来，打破信息扩散的局部局限性，当刻意选择弱关系作为扩散路径时不能增强信息扩散的范围，但是如果不使用弱关系则会降低信息扩散的范围.

结构洞是社会科学中的一个重要概念，最早于 1992 年被 Burt 提出[14]. 所谓结构洞就是指社会网络中的空隙，简单来说，社会网络中的人总是信任特定的人，并依赖于特定的人产生交换，当双方关系并不十分密切时，同双方关系都很密切的第 3 个人就占据了一个结构洞. 结构洞对信息在社会网络中的扩散起着重要作用，Tang 等人[15] 通过对 Twitter 数据的分析发现，1% 占据了结构洞的用户控制了 Twitter 中 25%的信息扩散. 此外他们还定义了挖掘大规模社会网络中 top-k 结构问题，并提出了两个模型来实现目标函数解决该问题，通过实验发现，识别出的结构洞可以有效地帮助团体挖掘和链接预测等应用. 此外 Li 等人[16] 提出了一个同结构洞有些类似的节点扩散度概念及其计算方法，节点扩散度可以用来衡量每个节点的扩散信息的能力，作者将节点扩散度同链接推荐问题相结合，推荐的结果可以同时增强社会媒体中的用户的设计交互和信息在网络中的扩散.

(3)社会网络整体结构.

研究者们比较同一条信息在不同社会网络中的扩散情况[9],发现社会网络的聚类系数越高,当一个聚集内的节点之间边越多,越有利于信息的扩散.许晓东等人[10]则发现网络结构越分散,信息扩散的范围越大,网络结构越聚集,信息扩散的范围越小.Montanari 等人[17]基于博弈论的扩散模型研究了该问题,却得出了基于传染病模型的研究[10]不同的结果,他们发现信息在局部联通性好的网络中扩散很快,而并非整体联通性好的网络;地理结构或者更多的确定的维度结构可以促进信息的扩散;他们还指出把新技术、新政策、政治信仰等信息的扩散简单地视为传染病传播是不够准确的.

目前基于理论模型的信息扩散特性的研究中,探究网络整体结构和信息扩散关系时,采用不同的理论扩散模型获得了不同的实验结论,具有一定争议,这就需要研究人员们确认哪种理论扩散模型更贴近真实的扩散过程.另外,目前研究采用的社会网络中规模最大的只有 100 多万节点,当前的研究结论放到上亿节点的在线社会网络中是否仍然适用存在着疑问.

2.2.2 信息扩散概率计算研究

信息扩散概率是指,信息沿着社会链接从一个用户扩散到另外一个用户的概率,可以直观理解为社会网络中边的权重,由于用户之间的社会影响力是引发信息扩散的重要原因,因此此类问题同社会影响力的计算是相关的.这里需要指出的是,笔者为方便介绍,将信息扩散预测研究放在了基于理论模型的类别中,但在该研究中不仅需要理论扩散模型,而且需要真实的信息扩散数据(扩散级联),所以该研究是即基于理论模型又基于扩散级联的研究,它是一种纽带性的研究,可以将基于理论模型的研究和基于扩散级联的研究联通起来.有些研究者[18-19]采用 WC MODEL 通过节点度来计算边的权重或者使用 TRIVALENCY MODEL 来随机分配边的权重,这些方法产生的权重过于简单和随意,并非真实信息扩散概率.此外,虽然很多研究将用户之间的信息扩散概率视为是恒定不变的,但是实际上它是时间相关和主题相关的,会随着时间或者主题的变化而变化.当前信息扩散概率计算的相关研究大体可分为以下 3 类:

(1)静态扩散概率研究.

Saito 等人[20]基于独立级联模型和真实扩散数据定义了一个似然函数最大化问题,使用 EM 算法对该问题求解获得了用户之间的信息扩散概率,这个值是静态的,不会随着时间的变化而变化.该方法虽然很简洁,但是它并不适用于大规模的数据集,Saito 等人也只使用了 10 000 多节点的博客网络进行了实验.这是因为使用 EM 算法需要不断的迭代,而每次迭代都需要更新每条边的扩散概率,显然会消耗太多的计算时间.Xiang 等人[21]基于用户描述文件的相似度和用户间的交互,提出了一个隐含变量模型来推断用户间的关系强度,此关系强度可作为用户间的信息扩散概率使用.此外,社会网络中边的权重可以从链接分析的角度进行计算,Kleinberg 等人[22]提出了大量的相关方法.

(2)时间相关扩散概率研究.

Goyal 等人[23]基于真实的信息扩散数据,提出了 3 种模型来计算信息扩散概率:第一个静态模型,计算出的用户间扩散概率是一个静态值.这个模型主要使用最大似然估计的方法,以用户 a 和 b 为例,信息从 a 流动到 b 的概率为从用户 a 扩散到 b 的信息数量同扩散到 b 的所有信息数量的比值,这个值是静态的;第二个是连续时间模型.研究人员使用了一个指数衰减模型来描述用户间具有时间动态的扩散概率.这种方法的效果是最好的,但是该模型在计算一个用户受到的联合影响力时并非增量式计算,所以在大规模数据集上应用非常困难;第三个是离散时间模型,它是连续时间模型的近似值模型,这种模型计算出的用户间扩散概率是一个离散时间函数,可以增量式计算联合扩散概率,所以适用于大数据集.研究者将这 3 种扩散概率同普遍阈值模型(General Threshold Model)相结合,对信息扩散进程进行预测,普遍阈值模型可以视为线性阈值模型的一个拓展.

(3)主题相关扩散概率研究.

研究人员对主题相关的信息扩散概率进行了研究[24-25].社会网络中的用户通常会对多个不同的主题感兴趣.而对于不同的主题,用户之间的信息扩散概率也是不同的,这个观点已经被社会学家证明过[26-27].这里举例说明一下,例如,李开复及其相关联的某用户,在互联网主题下,从李开复到该用户的信息扩散概率比较大,而在 NBA 主题下,扩散概率就会相应的变小.Tang 等人[24]使用 Topic Model[28-29]来初始化每个用户的主题分布,随后提出了一个基于主题因子模型来计算主题相关的用户之间的扩散概率.Liu 等人[25]并没有使用经典的 Topic Model 来初始化用户的主题分布,而是提出了一个

新的概率生成模型，该模型同时计算出用户的主题分布和主题相关的用户之间的扩散概率，并通过实验证明了该模型比 Tang 等人[24]的效果要好.

信息扩散概率是理论扩模型中的参数，本节介绍的研究可以理解为通过对信息扩散历史数据挖掘分析，来计算理论信息扩散模型中的参数. 当得到信息扩散概率后，理论扩散模型就脱离了其单纯的理论假设，而具有了实际应用价值. 3 类研究的应用范围各有不同，首先，对于静态扩散概率研究，它可以应用于信息扩散最大化研究(病毒式市场营销)和用户的影响力计算等；主题相关扩散概率研究增加了主题属性，它主要应用于主题相关的用户影响力计算和主题相关的团体挖掘等；时间相关扩散概率研究可以将理论扩散模型中的时刻同真实的时间相关联，把理论扩散模型和时间相关的扩散概率相结合，可以用来预测社会网络中时间相关的信息扩散进程，应用于信息推荐和突发性信息探测等. 信息扩散概率计算是纽带性的研究，将基于扩散模型和基于扩散级联两类研究联通了起来.

2.2.3　信息扩散最大化问题

信息扩散最大化问题[30]是应用性很强的研究型问题，通常也称作影响力最大化问题，下面通过一个事例来说明什么是影响力最大化问题：一个小公司推出了一款新产品，但是由于资金等问题，它只能选择一部分用户来试用这款产品(通过赠送礼品或者提供优惠的方法)，这个公司希望这些用户会喜欢这种产品，并且影响他们在社会网络中的朋友们去使用这款产品，接着，他们的朋友再影响他们朋友的朋友，类推下去，这样新产品信息就能在社会网络中逐渐的扩散开来. 影响力最大化问题就是如何选择初始节点集合，使得这些节点可以最大程度地影响社会网络中的其它节点，使信息在社会网络上可以获得最大程度的扩散. 此类研究可以广泛应用于病毒式市场营销、广告投放等，具有重要的商业价值.

通常的，影响力最大化问题可以定义为：给定一个社会网络和一个扩散模型，任务就是如何在社会网络上获取一个指定大小的节点集合，使得这个节点集合可以达到影响力最大化的效果. 影响力最大化问题的全局最优化被证明是 NP-难的问题，对于大规模的社会网络，只能采用一些优化算法获取近似的较优解. 目前解决影响力最大化问题的方法主要有以下两种：

(1) 贪心算法.

贪心算法[31-32]在对问题求解时，总是做出在当前情况下最好的选择. 使用贪心算法来解决影响力最大化问题的思路是，每次选择的节点都可以达到当前影响力最大化的效果. 假设，目前已经选出的节点集合为 A，使用的贪心算法在选取某个节点时，会把集合 A 同每个集合 A 之外的每个节点 u 结合，计算每个 $A\cup\{u\}$ 的影响力，选出产生最大影响力的那个 u，加入到集合 A 中，集合 A 初始值为空.

Kempe 等人[18]第 1 次将影响力最大化问题视为一个最优化问题，他们证明了对于独立级联模型和线性阈值模型，影响力最大化问题是一个 NP-难的问题，同时提出了贪心近似算法来求解这一问题，并证明该算法明显优于社会学中经典的基于节点度数和网络中心性的启发式算法. 他们估计节点集合影响力的方法是使用扩散模型在社会网络上模仿扩散进程，根据扩散效果来衡量的. 显然上述贪心算法需要非常多的计算时间，针对这一缺陷，Leskovec 等人[33]提出了在选择新种子节点时采用 lazy-forward 的策略来优化贪心算法，此方法虽然极大地提高了计算速度，但是在 10 000 个节点的网络中找出大小为 50 的最有影响力的节点集合仍然需要数个小时. Kimura 等人[34]对于独立级联模型和线性阈值模型提出了基于键渗流和图论的贪心算法，他们在博客的社会网络数据上进行了实验，证明了该方法在缩短计算时间上的有效性比文献[33]中的方法更加有效，但是实验中使用的博客网络只有10 000 多个节点，所处理的数据量还是比较小.

(2) 启发式算法.

社会学家提出基于节点度中心性和距离中心性[35]的启发式算法来计算节点集合的影响力，前者是按照节点的度进行排序，把高度的节点视为影响力大的节点，选择最靠前的部分节点作为结果，类似的距离中心性启发式是按照节点同其它节点之间的平均距离进行排序，距离越短说明节点影响力越大. 但是，这些算法被认为无法提供很好的扩散效果. 这是因为，首先这种启发式算法中选出的节点可能都在一个团体中，这些信息扩散就只限制在一个局部团体内，无法扩散到整个网络，另外这些启发式算法只考虑了网络结构，而没有考虑到信息在网络中扩散的动态性.

Kimura 等人[36]在独立级联模型下，提出了基于最短路径的启发式算法，但是该算法存在同贪心算法同样的问题，需要极大的计算时间. 针对该问题，Chen 等人[19]提出了一个新的启发式算法 MIA，该算法可以应用于几百万节点的网络，同时还提供了

一个可调节的参数来调节运算时间和扩散效果的平衡,该算法同文献[18,33,36]中的方法进行了比较,证明了其在处理大规模网络和扩散性能上的优越性.

综上所述,可以发现贪心算法和启发式算法各有优缺点.贪心算法的计算结果可以达到很好的扩散效果,但是需要的计算量太大,启发式算法虽然需要比较少的计算时间,但是扩散效果需要提升.因此,当社会网络规模不大,对扩散效果有较高要求时,通常可以采用贪心算法,而对于社会网络规模较大,对扩散效果要求不是特别高的时候,通常可以采用启发式算法.目前解决影响力最大化问题的最佳算法[19]也只处理了百万级规模的社会网络,但是目前在线社会网络中通常有几千万甚至几亿节点(Facebook 月活跃用户数量已突破 10 亿),如何在超大规模的网络中快速计算出固定数量的最有影响力的节点集合,应该是研究人员所关注的问题.此外,当前研究主要采用的是独立级联模型和线性阈值模型,而研究中这些模型的参数(信息扩散概率)通常都不够准确,那么这两个模型是否是研究该问题的最优模型?是否可采用其它扩散模型?采用参数化的模型是否可获得更好的研究效果?这些问题都有待解决.

2.2.4 竞争性的信息扩散最大化问题

竞争性的信息扩散最大化问题(也称作竞争性的影响力最大化问题)可以视为信息扩散最大化问题更贴近现实的拓展.信息扩散最大化问题研究的是单条信息在社会网络中的最大化扩散,而在现实世界中,网络中扩散的并非只有一条信息,通常是多条有竞争关系的信息同时在社会网络中扩散,例如运动商品信息(耐克和阿迪达斯)、新闻信息(谣言信息和可信信息)等.竞争性的信息扩散最大化问题就是研究当多条有竞争关系的信息同时在社会网络中扩散时,对于每条信息,如何从自身的角度选择初始节点集合,使得该信息得到最大化的扩散.由于几条竞争性的信息在选择初始节点时有先后顺序,所以不同次序的信息会有不同的选择策略.该问题的研究具有重要的应用价值,当社会网络中扩散的是商品信息时,该研究可以用于病毒式市场营销;当社会网络中扩散的是新闻信息时,该研究用于控制谣言信息的传播.该问题的输入同影响力最大化问题的输入类似,也是社会网络和扩散模型.

Bharathi 等人[37]将独立级联模型进行了拓展,用于模拟多条信息在社会网络中的扩散.拓展的模型中节点的状态不再只有激活和未激活两种,他们将单一的激活状态分解,每条竞争信息都对应一种状态,这样节点状态就分为未激活和多种激活状态.同时为了避免多个信息同时扩散到某一节点的现象出现,激活节点被不同信息激活时,按照独立指数分布随机分配了一个时延.基于上述扩散模型,他们从第一个选择节点的信息的角度和最后一个选择节点的信息的角度进行了研究,对于第一个选择节点的信息,当网络结构是树形时,他们提出了一种 FPTAS 算法用于选取节点,对于最后一个选择节点的信息,该问题是一个子模问题,可以使用贪心算法解决.

Carnes 等人[38]提出了基于距离模型(Distance-Based Model)和波传播模型(Wave Propagation Model),当只有一种信息在网络中扩散时,这两种模型就退化成了独立级联模型.他们首先从最后选择节点的信息的角度进行研究,发现在两种模型下,竞争性的影响力最大化问题是 NP-难的问题,通过实验分析发现贪心算法可以获得比启发式算法更好的效果,但是效果并不显著.对于第一个选择节点的信息,最佳策略就是选择度大的节点,初始激活节点比较多的信息并不一定比初始激活节点少一些的信息的扩散范围大.

Budak 等人[39]提出了 MCICM(Multi-Campaign Independent Cascade Model)和 COICM(Campaign-Oblivious Independent Cascade Model)模型.COICM 模型是 MCICM 模型的简化版,MCICM 中对于不同的信息,节点之间的扩散概率是不同的,而在 COICM 模型中是相同的.这两个模型也都是独立级联模型的拓展,同 Bharathi[37]的模型是类似的,区别在于,Bharathi 提出的模型为预防不同信息同时激活同一节点情况的发生,对每次激活行为都随机分配了一个时延,而 MCICM 和 COICM 是将信息划分了等级来处理这种情况,当两种信息同时要激活同一节点时,该节点会被优先级高的信息激活.另外该项研究同 Bharathi 和 Carnes 所研究的具体问题也略有区别.Bharathi 和 Carnes 研究的问题是信息在同一时间段内有次序先后的选择初始节点,随后这些竞争性的信息是在同一时刻开始扩散.而该研究的具体应用场景是谣言的控制,而谣言通常是先于辟谣信息在网络上扩散的,所以具体问题变成了谣言信息首先在网络上扩散,一定时延之后,为辟谣信息选择一定数量的节点来最大化控制谣言的扩散.该研究分别使用贪心算法和启发式的方法来计算节点集合,实验证明启发式的方法要优于贪心算法.实验中使用的是 Facebook 中 4 个地域相关的子网,规模只有几万个节点.

目前竞争性的信息扩散最大问题的研究还处于初级阶段，相关研究工作不多，而且实验的规模也都比较小，当前研究工作中所提出的信息扩散模型基本上都是独立级联模型的扩展，从扩展其它扩散模型（例如线性阈值模型）的角度进行研究可以作为一种研究思路，另外此类研究同影响力最大化研究密切相关，影响力最大化研究的一些方法是有一定的借鉴价值的.

3　基于信息扩散级联的研究

基于信息在社会网络中扩散路径的形状特征，通常将信息扩散的路径称为信息扩散树（Information Diffusion Trees）或者信息扩散级联（Information Diffusion Cascades）. 该类研究本质上是基于真实信息扩散数据进行研究，这里为形象表达，称作基于信息扩散级联的研究. 接下来的 3.1 节介绍了目前研究中所使用的主要数据集和从中挖掘扩散级联的方法；3.2 节介绍基于信息扩散级联的主要研究方向及其研究进展，包括信息扩散特性研究、用户影响力计算、信息扩散预测模型.

3.1　信息扩散级联

可用于信息扩散研究的数据多种多样，从早期的博客、文献引用数据、Email 到近来流行的 Twitter、Digg、Facebook 等，本节主要介绍研究中最常用的几种数据及其信息扩散级联的抽取方法，表 2 中列举了目前主流数据集的信息类型、扩散方式和开放的数据集. 这里需要指明的是，接下来介绍的这些数据集中大都包含在线社会网络，所以基于理论扩散模型的研究也可以使用这些数据. 为了让读者对信息扩散级联有一个直观的了解，图 3 中展示了我们绘制的一个信息扩散级联实例，我们采用的是新浪微博的数据，传播的信息是“中国好声音”相关的一条微博. 图 3 中所示的信息扩散级联中，左面的扩散聚簇的中心处为信息发布者，信息发布之后被其粉丝转发，再被粉丝的粉丝转发，依次转发下去，形成图示的结构. 其中有几个转发者有比较大的影响力，这些人转发后又被他们的粉丝大量转发，造成了信息扩散的“二次引爆“，形成了图 3 中右面一些较小的聚簇. 图中可以清晰地看到信息的扩散路径，该微博的转发量大，扩散级联的层数比较多，产生了比较大的影响.

表 2　信息扩散研究数据集

数据集	信息类型	扩散方式	开放数据集
博客	博客	超链接	无
微博	Hash tag、Url、微博信息	转发	http://www.wise2012.cs.ucy.ac.cy/challenge.html https://www.kddcup2012.org/c/kddcup2012-track1/data
Digg	社会化新闻	投票	http://www.isi.edu/~lerman/downloads/digg2009.html
Flickr	图片	收藏	http://socialnetworks.mpi-sws.org/data-www2009.html
Facebook	New Feed	广播	无
Email	电子邮件	转发	http://www.cs.cmu.edu/~enron/
文献数据	文章	引用	http://arnetminer.org/download
Chain Letters	信件	转发	无

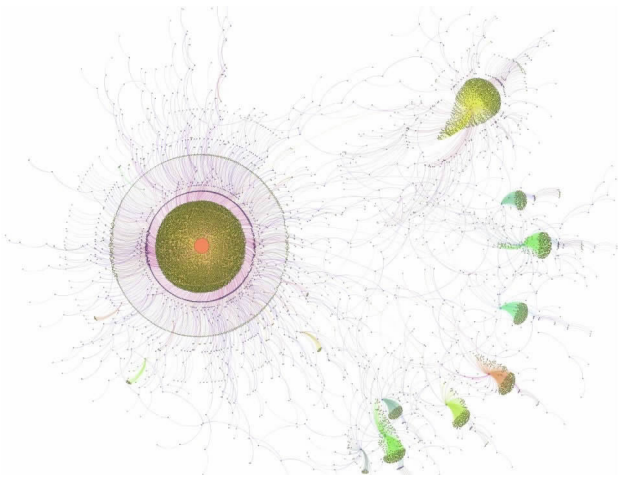

图 3　信息扩散级联实例

（1）微博. 微博是国内目前最火热的社会媒体，也是研究信息扩散的最主要的媒体. 微博媒体本身具有很强的用户参与性和信息扩散性，其单向关注的设计可以使信息在网络中达到最大程度的扩散. 目前研究主要关注 3 种信息（hash tag、url、微博信息）在社会网络中的扩散，获取这 3 种信息扩散级联的方法也不相同. 其中 hash tag 和 url 的扩散级联的获取方法是类似的，它们都是基于信息扩散的时间序列和社会网络来获得的. 例如，用户 B 和 B 的粉丝 A 的微博都包含了某个 hash tag 或者某个 url，同时用户 B 的微博发布时间比他的粉丝 A 早，那么就可以认为信息从用户 B 扩散到了用户 A. 该方法的问题是当用户A有多个关注都发布了某条

信息相关的微博时，就无法判断信息是从哪个节点流动到 A 的. 而对于某条微博信息，通常是使用转发机制来解析出扩散级联. 当用户 A 转发了用户 B 的微博时，就可以认为信息从用户 B 扩散到了用户 A. 微博内容中存在“//@:username”这样的转发标识符，可以使用这个标识符从微博内容中抽取扩散路径. 这个方法同样存在着缺陷，假设 A 有一个粉丝 B，B 又有两个粉丝 C 和 D，用户 A 和用户 D 根本没有直接关系，但是 D 直接转发了 A 的微博，这样获取的信息扩散路径是 A→D，这样就忽略了 A 和 D 之间的节点，没有这些节点，A 是无法看到 D 的微博的.

(2) Digg. Digg 是一个社会化新闻网站，用户在这个网站上可以加其他用户为好友，这样就构建了社会网络，用户可以发布自己的新闻并对其他用户发布的新闻投票，假设有两个用户 A 和 B，如果 A 对 B 的新闻投了票，那么 A 的所有好友也可以看到 B 的新闻，该投票机制类似于微博中的转发机制，使用该机制就可以获取社会化新闻的扩散路径.

(3) Flickr. Flickr 是一个图片分享网站，用户可以建立上传图片并且同其他用户建立好友关系，用户的收藏行为使得图片可以在社会网络中扩散. 假设 A 和 B 是好友，如果 A 收藏某图片的时间比 B 收藏的时间晚，那么就认为图片信息从 B 扩散到了 A.

(4) Facebook. Facebook 是世界上最大的社交网站，该网站所蕴含的社会网络中所扩散的信息被称作 New Feed，主要包括更新的用户描述信息、共享的链接和评论等，New Feed 主要依靠用户的广播行为传播，扩散路径获取方法同 Flickr 类似.

表 3 信息扩散级联整体特性研究

研究者	数据集	信息扩散级联整体特性
Leskovec[43]	博客	信息扩散级联普遍的比较宽但是比较浅，另外，统计发现 97% 的扩散级联只有 1 个节点，也就是信息没有扩散，剩余 1.8% 的扩散级联只有两个节点，最后的 1.2% 拓补结构复杂一些. 所有信息扩散级联的规模符合幂率分布，而且所有链式扩散级联和星状扩散级联的规模也是符合幂率分布的.
Kwak[40]	Twitter	95.8% 的信息扩散级联的深度只有 1，而 97.6% 的节点对符合六度分割理论，只有一小部分信息扩散级联的深度大于 6，而且没有深度大于 11 的扩散级联，信息扩散级联的规模是符合幂率分布的.
Bakshy[42]	Twitter	信息扩散级联的平均规模为 1.14，这说明大部分信息没有发生扩散，他们指出信息扩散级联的规模符合幂率分布，而信息扩散级联的深度比较接近指数分布.
Ghosh[44]	Digg	信息扩散级联的规模和深度是符合对数正态分布的.
Lerman[45]	Twitter、Digg	两个数据集的信息扩散级联规模都符合正态分布. 信息在 Digg 的社会网络中初始扩散的速度要比 Twitter 中要快，这是因为 Digg 的社会网络更稠密、内联性更好，但是信息在 Twitter 的社会网络上扩散的更远，这是因为 Digg 中信息扩散速度随着时间衰减的比较快，而 Twitter 中的信息扩散速率变化较小.
Cha[46]	Flickr	88.5% 的信息扩散级联规模不大于 5，99.8% 的信息扩散级联规模不大于 100.
Wang[42]	Email	信息扩散级联浓密而浅，信息有效扩散后就立刻停止了，95% 的扩散级联都只有 2 层，超过 4 层的不存在，扩散级联的规模和宽度都是符合幂率分布的.
Shi[47]	文献引用数据	信息扩散级联的规模和深度并不符合幂率分布
Golub[48]	Chain Letters	信息扩散级联是窄而深的.

(5) Chain Letters. Chain Letters 通常称作连锁信或者连环信，它有多种形式，当前以电子邮件形式居多，不同于普通电子邮件数据的是信件中含有引诱收信者复制再转发给其他人的信息. 用户接收到连锁信后，会受到信件中内容的引诱再次转发.

3.2 主要研究方向

3.2.1 信息扩散特性研究

对于信息扩散特性的研究，基于理论扩散模型和基于信息扩散级联的研究的区别在于，前者是使用扩散模型模拟信息在网络上的扩散，根据仿真出来的扩散效果进行研究，但是仿真是无法完全展现真实信息扩散的全部特征的，而后者是直接对真实的信息扩散数据进行分析，可以得到更完整准确的研究结果.

基于真实扩散级联的信息扩散特性研究可以分为两个角度：一个是宏观角度，从信息扩散级联的整体角度出发，主要研究信息扩散级联的整体特性；另一个是微观角度，基于单个用户或者用户之间的角度进行研究，主要研究影响信息扩散的各种因素，主要分为社会因素和信息因素.

(1) 宏观角度.

信息扩散级联的整体特性主要是指扩散级联的形状、规模和宽度等这些整体特征和分布. 表 3 呈现了不同研究者基于不同的数据集，所得出的信息扩散级联的整体特性结论. 可以发现，大部分研究[40-43]中都得出了信息扩散级联的规模和深度都符合幂率分布的结论，但是也有研究[44-45]认为是符合正态分布的，得出不同结论的原因很有可能和研究者们所使用的数据集不同有关. 由于没有标准的数据集用于研究，研究者们都是通过自己设计的采

集方法获取数据，不同的采样方法和数据规模最终导致了不同的实验结论. 在这些研究中 Ghosh 等人[44]提出了级联生成函数(Cascade Generating Function)来研究信息扩散特性，它实际是一个描述信息扩散级联的方法，可以展现信息在网络上扩散的动态细节.

(2) 微观角度.

① 社会因素.

社会因素是指社会网络中与用户存在社会关系的邻居们，用户间的社会关系及其强度是复杂多样的，包括交互强度、兴趣相似度、背景相似度等. (i) 交互强度，是指社会网络中的用户之间的行为交互的强度，具体的，在 Email 中就是用户之间的信件交流信件的数量，在微博中就是用户之间的微博转发频度. 研究人员基于 Email 数据研究了用户之间的交互性对信息扩散的影响[42]，发现信息从初始者到扩散者时通常会走弱关系边，而从扩散者到接收者时通常会走强关系边，说明信息在流动过程中有从弱关系到强关系的转向特征. 另外发现扩散者并非中心性很强的社会化中心点(Social Hubs)，而是很普通的点，研究人员对比了所有用户的度分布和扩散者的度分布，发现非常匹配，证明了这个观点. 此外，Bakshy 等人[49]针对虚拟游戏“Second Life”中的手势信息扩散进行了研究，发现两个朋友之间信息扩散的速度要比两个陌生人之间快，这里的朋友是指社会网络中有交互关系的两个用户；而且一个人扩散某条信息的概率会随着已经扩散了该信息的他的朋友数量增加而变大，对于一个有很多朋友的用户，他很少被一个特定的用户所影响而扩散信息. (ii) 兴趣相似度，一般都是把用户发布或者转发的信息汇集起来，构建用户的描述文件，然后计算用户的兴趣相似度，通常使用的方法是使用向量空间模型描述用户. 研究人员分别基于 Meme[50]和 Twitter[51]的数据进行了研究，发现当两个人的兴趣度相似越大，信息在这两个人间扩散的概率加大，反之亦然. (iii) 背景相似度，是指用户间的教育经历、工作部门、性别、年龄等信息的相似度. 研究者发现当用户在同一部门时[42]，扩散者会更快地把信息扩散给接收者，同级的用户之间信息流动比较少，两个用户的等级距离越大，信息发生扩散的可能性越大.

② 信息因素.

信息因素主要包括信息内容和信息流行度等. 当信息内容同用户的兴趣信息相似度高时，该用户参与扩散信息的概率越大，反之亦然[50-51]. Wang 等人[42]基于 Email 数据进行了更细致的研究，通过传递的 Email 信息同用户的兴趣信息相似度计算发现，从初始者到扩散者时，相似度越低，传递的几率越大，而从扩散者到接收者时，相似度越高，传递的几率越大. 可见，信息扩散过程中有明显的转向，从非专家转到专家；信息和用户的相似度的计算方法是用户之间的兴趣度的计算方法类似时，是可以互相借鉴的. 从信息流行度的角度研究发现时，流行度高的信息被用户转发的概率大，信息的流行度是某段时间内所有用户使用过的所有词的频率来估计的[51].

通过对微观角度研究的总结分析，可以发现当前研究主要针对社会因素和信息因素对信息扩散的影响. 这里的两个因素和 2.1 节中介绍理论扩散模型时的因素是相通的，上述研究可以帮助研究者们更深刻地理解信息扩散进程，为构建信息扩散预测模型提供了基础性结论. 此外，社会因素和信息因素都属于社会网络内部因素，社会网络的外部因素(例如，电视和报纸等传统媒体、现实的社会活动)同样影响着信息扩散，在构建信息扩散预测模型时可适当考虑这些外部因素.

通过上述研究可以发现，不同的应用领域(或数据集中)推动信息扩散的源动力是不同的，基于此可以将当前研究所使用的数据集分为 3 类：第一类，主要包括博客、Twitter、Digg、Flickr 和 Facebook 等，这类数据集中社会因素和信息因素同时影响着信息在网络中的扩散，信息扩散级联的规模和深度大都符合幂率或者正态分布. 第二类，主要包括 Email 和文献引用数据，这类数据集中社会因素(用户间的兴趣和的社会背景的相似性)是主要的源动力. Email 数据中的信息扩散的规模和深度是符合幂律分布的，但是信息在流动过程中有从弱关系到强关系的转向特征，这个在第一类数据中还未被发现，而文献引用数据中的信息扩散级联的规模和深度是不符合幂律分布的，同第一类数据也有所不同. 在第三类，主要代表是 Chain Letter，这类数据中信息因素是主要的源动力. Chain Letter 是一种特殊的信件，信件中包含引诱收信者转发信件的信息，而此类数据中的信息扩散级联是窄而深的，大大不同于前两类数据中的宽而浅的扩散模式.

3.2.2　用户影响力计算

社会网络中用户影响力的计算同 2.2.1 节所介绍的影响力最大化问题并非同一问题，它们的主要区别在于，影响力最大化问题是找出社会网络中最有影响力的可以使信息得以最大化扩散的一个节点

集,而这里提及的用户影响力计算是对网络中的任意单个用户节点的影响力进行衡量,可以理解为社会网络中节点权重的计算,用户影响力计算的方法可以应用到影响力最大化问题中.当前研究从社会网络结构和信息扩散级联特征两个角度来衡量节点的影响力,前者主要是进行网络拓补分析、网络中节点的出入度、接近中心性、中介中心性[52]、PageRank值和权威值(在社会网络上使用类似 PageRank 算法[53]或者 HITS 算法[54])等;后者主要是使用信息扩散级联的数量、信息扩散级联的规模、深度等特征衡量用户的影响力.接下来详细介绍一下用户影响力计算相关的研究,这些研究都是基于 Twitter 数据的.

Kwak 等人[40]使用用户的粉丝数、社会网络上用户的 PageRank 值和用户微博的转发数 3 种不同的指标来衡量单个用户的影响力,他们发现使用粉丝数和 PageRank 获取的用户影响力的结果比较相似,而使用微博转发数获取的用户影响力同前两种方法获取的影响力差别较大,这是因为前两种方法都是分析网络的拓补结构,而后者是基于信息扩散级联分析的,可见两种衡量标准是有区别的.Weng 等人[55]使用用户的粉丝数、PageRank 值和不同主题下的 PageRank 值来衡量用户的影响力.Cha 等人[56]使用了粉丝数、微博转发数和用户被提及数来计算用户的影响力,他们发现这 3 种方式获取的影响力结果有很大区别,这同 Kwak[40]得到的实验结果是相同的.文献[40-56]都使用了用户所发布信息的转发数量来计算该用户的影响力,但是信息的转发数量是一个比较粗略的衡量标准,而信息的转发树则可以直接完整地展现该信息的扩散情况.Bakshy[41]等把用户的个人属性信息(粉丝数、关注数、微博数和注册时间)和用户发布信息的扩散级联的平均规模、最大和最小规模等作为特征,使用衰减树模型[57](regression tree model)来预测用户的未来的影响力,发现过去影响力比较大,粉丝比较多的节点在未来仍然有比较大的影响力,但是这是从平均值的角度而言的,使用信息扩散级联和粉丝数来准确地预测单个个体的未来影响力是很困难的.另外他们还使用用户所产生信息本身的特征来对用户的影响力进行预测,但是由于信息本身的复杂性,这种方法并未取得提升预测效果的作用.

用户相关的数据越完全,就可以越准确地计算用户的影响力.当只拥有在线社会网络结构数据时,只能够对网络的拓扑结构分析来计算用户的影响力;而当拥有用户信息扩散历史数据时,就可以使用用户发布的信息所产生的扩散级联来衡量用户影响力.计算 Web 2.0 网站中用户影响力时,单纯的使用过去计算 Web 1.0 中网页重要度的拓补结构分析的方法,只能展现用户的一部分影响力.用户发布的信息所产生的扩散级联是衡量用户影响力的重要标准,目前研究大都只考虑了信息扩散级联规模,而扩散级联的深度和广度等特征同样重要,大多数研究忽略了这些特征.用户影响力计算的研究思路应该是网络拓补结构分析和信息扩散级联分析并重的,两者相结合才能更准确地衡量节点影响力的大小.另外,由于同一用户在不同主题下的影响力是不同的,所以还应当考虑主题因素.

3.2.3 信息扩散预测模型

通过信息扩散特性的研究,可以发现信息本身特性、用户之间的关系、在线社会网络外部因素等多个方面都影响着信息扩散.信息扩散预测模型的研究目的就是,结合这些因素对信息在社会网络中的扩散进程建模,从而达到模拟预测信息扩散动态的目的.信息扩散预测模型可以同影响力最大化问题相结合,前者是在基于信息扩散特性分析的基础上建立的具有预测能力的扩散模型,使用这类模型作为输入,应该比使用理论扩散模型得到更准确的扩散效果.

信息扩散预测模型不同于 2.1 节中的理论扩散模型,后者通常只考虑了比较单一的影响信息扩散的因素,而且这些模型是单纯的理论模拟扩散进程,模型中的参数通常是任意指定的,模型中的时刻都是理论上的时间间隔,并非真实的时间,因此理论扩散模型不具备预测信息扩散的能力.信息扩散预测模型的研究也分为宏观研究和微观研究.宏观研究是从整体的角度出发,并不对单个个体是否会扩散信息进行预测,而是对信息的扩散范围、扩散广度和扩散深度等宏观特性进行预测;微观研究是从个体的角度出发,对具体用户是否会扩散某条信息的行为进行预测,通过预测每个个体的行为也就可以得到信息在整个社会网络的扩散情况.

(1)宏观角度.

Yang 等人[51]从宏观的角度对信息扩散的预测问题进行了研究,使用特征选取结合因子图模型[58-59]对信息扩散级联的深度进行了预测,他们的方法要远优于 SVM 和 LogReg 这些分类模型,这是由于他们考虑了社会网络结构,同时使用用户的关注和粉丝的行为对该用户行为进行判断.该模型虽

然可以预测信息扩散级联的深度，但是并没有考虑到时间因素，不能预测信息的扩散速度. 此外，还有研究者[60]对信息扩散的速度、范围和距离进行预测，他们抽取了一些社会特征（用户发布信息数量、用户被提及比率等）和信息特征（是否有超链接、被转发时间等），使用这些特征和衰减模型[61]对 3 种指标进行预测，这种预测方法和 Yang[51]的方法是有些类似的，他们对各种特征对预测信息扩散的有效性进行了分析. 实验发现一个用户的被提及比率对于信息扩散级联的 3 种特征都有很强预测性，另外信息是否含有超链接和被转发时间也很重要.

不同于上述两种特征选择加模型训练[51,60]的方法，研究者[62]还提出了线性影响力模型（Linear Influence Model）用于预测信息扩散，该模型的基本思想是，在时刻 t 将要被感染的节点数量，是由在时刻 t 之前所有已被感染节点的影响力函数决定的，模型首先从全局影响力的角度建立单个节点的影响力函数，而信息扩散的整体状况就是所有单个节点影响力的和. 而关键问题就是如何估计节点的影响力函数，传统的函数估计方法是首先给定一个固定的函数模式，然后对函数中的参数进行估计，但是这种方法有着严重的缺陷，它假设了所有节点的影响力函数都是相同的格式，这种假设无法展现用户影响力的复杂性多样性. 所以他们使用了一种无参数的方法，即节点的影响力函数是一个非负值的向量，向量中的值的含义为单位时刻（小时）内该节点影响的其它节点受到感染的数量. 为估计节点的影响力函数，研究人员基于信息扩散的真实数据构造了一个非负最小方程问题[63]，该问题的计算结果就是节点的影响力函数. 节点影响力函数是时间相关的，所以可以用来预测信息扩散的速度和信息扩散的范围. 该模型没有考虑社会网络结构对信息扩散的影响，所以对没有显式社会网络的数据特别适用，例如，电子商务网站中用户购买行为信息扩散的预测.

（2）微观角度.

Yang 等人[51]从微观角度进行了研究，他们从用户的偏好、信息本身特征、信息扩散级联和时间延迟多个方面抽取了 22 个特征，采用因子图模型对用户是否转发某条信息进行了预测. Galuba 等人[64]提出了一个扩散模型，同时考虑了信息流行度、用户间社会影响力、信息扩散速度 3 个方面的因素，采用梯度衰减的方法来训练三因素相关的模型参数，并使用该模型对 Twitter 中的用户是否会涉及某条 url 进行预测. Song 等人[65]提出了一个基于时间连续马尔科夫链[66]的信息扩散模型，该模型不仅考虑到了用户之间信息扩散的概率，还考虑了信息扩散的速度. 该模型的本质就是一个状态转移速率矩阵，状态转移速率的含义是单位时间内状态转移的概率，模型的生成过程是一个状态转移速率估计计算的过程. 该扩散模型本身不能预测信息的扩散，但是研究人员基于此模型提出了两个算法，一个算法用于预测信息在固定时间段内从一个用户扩散到另一个用户的概率，另一个算法用于预测信息从一个用户扩散到另一个用户所需的时间.

赵丽等人[67]提出了一个基于节点知名度和活跃度的信息扩散预测模型，并使用博客数据进行验证. 其中，节点的知名度是在信息扩散过程中描述博客节点（博主）对其它节点的影响力的非负实数，节点的知名度受多种因素的影响，如博客日常的访问量、是否被博客首页推荐、是否被其它博客站点链接、是否被搜索引擎检索等. 节点的活跃度是描述博客节点是否经常访问其它节点并发文的活跃程度的非负实数. 该模型可用于预测节点在某个时刻是否会扩散信息，模型中考虑了内部场强、外部场强和信息扩散性三方面因素，其中内部场强表示网络内的已受感染节点对某一节点的影响，外部场强表示博客外其它的外部媒体对某一节点的影响. 信息扩散特性是由信息本身决定的，是个定值，不随时间变化而变化. 研究人员使用真实的话题扩散数据，对模型中的参数进行估计，随后使用训练好的模型对信息的扩散进行预测. 但是由于实验中网络中的节点的知名度是按照广义帕累托分布任意分配的，而不是按照真实数据计算出的，所以该模型虽然是对具体节点进行预测，但是实际应用中只能用来预测信息扩散范围和扩散速度等宏观特征. 而且这个信息扩散预测模型虽然考虑到信息本身扩散概率和外部场强，但是没有考虑到网络结构对信息扩散的影响，这应该是可以改进的地方.

以上从宏观和微观两个角度介绍了信息扩散预测模型的研究. 宏观研究是从整体的角度出发，对信息的扩散范围、扩散广度和扩散深度等宏观特性进行预测，它主要应用于突发性信息探测、舆情预测等，此类研究的方法可以在没有社会网络结构数据的情况下使用. 微观研究是对社会网络中用户的具体行为进行预测，它主要可以应用到信息推荐等方面. 此外，笔者认为，两类信息扩散预测模型研究均可以同信息扩散最大化问题相结合，应用到病毒式市场营销中. 在构建信息扩散预测模型时，最主要的

有两个问题:(1)是如何充分融合信息因素、社会因素甚至社会网络外部因素来模仿信息扩散进程;(2)是信息扩散是一个时间动态的进程,信息扩散模型中的时间如何同真实时刻关联起来.

4 扩散动态性和网络动态性

社会网络中存在两种动态性:(1)是信息在社会网络中扩散的动态性(扩散动态性);(2)是社会网络结构本身演化的动态性(网络动态性).当前计算机科学领域的学者们通常是将这两个动态进程分别研究,本文主要对前者的研究现状进行了总结和分析.社会网络结构的动态性,通常包含了社会网络中节点的增加、边的增加、取消以及边的权重变化等,它同样吸引了大量的相关研究[68-72].需要注意的是,两种动态性实际上是相互关联的,信息是在一个动态变化的社会网络中动态扩散的.

计算机科学领域的学者们所做的研究中,通常将在线社会网络的结构视为固定的.而当前有些生物学家和流行病学家考虑了网络动态性因素,对疾病在动态社会网络上的传播展开了一些研究.Blonder 等人[73]指出当疾病扩散周期和网络动态演化周期相差不大时,两种动态变化会有相互的反馈.Croft 等人[74]发现疾病在社会网络中的传播会导致病人之间的交互,说明疾病扩散对网络进化有促进作用.Volz 等人[75]发现,社会网络中的人通过不断联系带来的网络结构的进化,会导致疾病传播产生复杂的模式.同时 Volz 等人[76]还提出了 NE 模型来模拟疾病的扩散过程,同传统传染病模型不同的是,该模型是基于一个动态网络构建的.

在线社会网络中的信息扩散研究通常未考虑网络的动态性,也是有一定道理的,相比于现实社会网络中疾病的传播,在线社会网络中信息扩散的速度更快,扩散周期要小很多,同网络动态演化的周期通常存在着巨大差别,以微博媒体为例,一条微博一般在很短时间被大量用户转发,而微博中的社会网络的动态演化非常慢,因此研究者们在研究信息扩散的过程中就忽略了在线社会网络的变化.但是对于一些信息扩散较慢的媒体(例如 Flickr),考虑网络动态性可能可以获得更好的研究效果.本文中将信息扩散的研究分为了基于理论模型的研究和基于信息扩散级联的研究,这里就从这两个方面讨论一下,信息扩散研究中考虑动态网络因素的必要性、难点及研究方法同静态网络的区别.

(1)对于基于理论模型的研究.如果需要考虑网络动态性,那么最大的问题是构建类似 Volz 等人[76]提出的针对动态社会网络的理论模型,基于这种理论模型开展研究.除了使用新的理论模型外,过去研究中提出的方法是不受影响的,例如信息扩散最大化问题和竞争性的信息扩散最大化问题,这两类问题都是在网络中找出固定数量的节点集合,这类研究的目的是提出寻找节点集合的方法,无论网络是静态的还是动态的,方法都是适用的.

(2)对于基于信息扩散级联的研究.如果要考虑在线社会网络的动态性,那么面临的首要问题就是信息扩散级联的抽取,这个问题同信息扩散特性研究和节点影响力计算研究都息息相关.对于微博媒体,可以通过“//@:username”这样的转发标识符来抽取信息扩散级联,但是对于像 Flickr 这样的媒体,必须基于信息扩散记录和社会网络结构来抽取,网络是动态的,抽取难度将加大,在抓取信息扩散数据时,还要对社会网络数据不断抓取、更新并保存不同时刻的网络结构.抽取每条扩散路径时,都要使用信息扩散的时间戳到数据库中找合适时间的社会网络结构,这样做虽然可以抽取到更准确的信息扩散级联,但是无论数据抓取还是扩散级联抽取都需要更多的时间和计算,这中间就有一个准确性和计算时间的平衡.信息扩散预测模型的研究是可以适当考虑动态网络因素的,这个问题笔者在研究过程中考虑过.我们尝试提出一种信息扩散预测模型,可以对信息扩散的时间动态性进行预测,也就是预测某个具体时间段内,哪些节点会受到感染.一个人是否受到感染,是和其在社会网络中的邻居的状态相关的,由于社会网络是动态变化的,一个比较直接的思路是,假设要预测某时刻 t 的信息扩散状态,首先使用社会网络动态演化模型,对社会网络在 t 时刻的结构进行预测,随后使用预测模型在 t 时刻的网络结构上进行预测.

我们对动态在线社会网络中信息扩散的研究进行了一些总结和分析,希望对研究者们有所帮助.在信息扩散研究中是否应考虑网络动态性这个因素?哪些问题需要考虑这个因素?在网络动态的情况下研究信息扩散会存在哪些难点?其研究方法跟静态网络有何区别?这些问题都是研究者们需要注意和思考的.

5 总结与展望

在线社会网络中信息扩散研究是一个非常新的领域,目前还处于起步阶段.本文在充分调研和深入

分析当前工作的基础上，对该研究领域进行了综述，将目前工作分为基于理论模型的研究和基于信息扩散级联的研究两种，而这两种研究路线下又有一些具体的研究方向，文中详细介绍了各个方向的研究内容、研究现状及其各个方向之间的内在关联，并分析总结目前工作，给出了自己的一些观点.

在本文最后，笔者基于大量调研和自己的研究中所获得的一些经验，提出了 4 个目前该研究面临的问题和未来的研究方向，希望对本领域的研究者有所启发和帮助.

(1) 大数据处理问题. 目前大部分在线社会网络中信息扩散的研究所采用的数据量普遍偏小，特别是基于理论模型的研究. 但是在线社会网络本身数据量是巨大的，有些媒体中社会网络的节点数量是以亿为量级的，而每个节点还会每天产生信息. 在小数据量上的实验得到的结论可能并不适用于大规模数据，而小数据量下的算法或者模型也可能会因为时间或者性能的原因，无法应用于真实的大数据量下. 所以，广大研究者提出的算法或者模型应该可以有效快速地处理大规模在线社会网络数据，需要在大规模的数据上验证自己的方法.

(2) 理论模型研究和扩散级联研究相结合问题. 目前研究中所采用的理论扩散模型，多是早期社会学家或者传染病学家提出的一些经典模型或者这些模型的拓展，但是通过信息扩散特性的研究，发现信息扩散同社会改革扩散和传染病的扩散是有区别的. 相比于传统理论扩散模型，基于信息扩散级联的研究中所得到的信息扩散预测模型，应该可以更好地模拟信息扩散的内部机制，能否基于这些模型研究过去基于理论模型所研究的问题，是否可以获得更准确的研究结果.

(3) 信息扩散级联的模式挖掘问题. 不同类型的用户所发布的信息引发的扩散级联的模式是不同的，而不同类型信息产生的信息扩散级联的模式也是不同的. 如何根据信息扩散级联的模式来区分出用户的类型或者信息的种类将是非常有意义的研究. 例如可以分析谣言信息所产生的扩散级联的模式特征，根据这些特征来区分谣言和真实信息.

(4) 社会现象的挖掘问题. 社会学家分析真实的社会网络发现了六度理论、小世界理论等经典社会现象，而社会媒体中存在的虚拟的社会网络同现实世界中的社会网络是相互映射相互影响的，现实社会网络是虚拟社会网络产生的基础，虚拟社会网络是现实社会网络的延伸、补充和发展. 所以在线社会网络为我们提供了研究社会问题的平台，通过研究在线社会网路中的信息扩散问题，可以发现现实生活中不易发现的社会现象或者社会问题.

参　考　文　献

[1] Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters, 2001, 12(3): 211-223

[2] Goldenberg J, Libai B, Muller E. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. Academy of Marketing Science Review, 2001, 9(3): 1-18

[3] Granovetter M. Threshold models of collective behavior. American Journal of Sociology, 1987, 83(6): 1420-1443

[4] Hethcote, Herbert W. The mathematics of infectious diseases. SIAM Review-Society for Industrial and Applied Mathematics, 2000, 42(4): 599-653

[5] May R M, Lloyd A L. Infection dynamics on scale-free network. Physical Review E, 2001, 64(4): 066112

[6] Satorras R P, Vespignani A. Epidemic spreading in scale-free networks. Physical Review Letters, 2001, 86(14): 3200-3203

[7] Morris S. Contagion. Review of Economic Studies, 2000, 67(1): 57-78

[8] Young H P. The Diffusion of Innovation in Social Networks. Oxford: Oxford University Press, 2003

[9] Xu B, Liu L. Information diffusion through online social networks//Proceedings of the International Conference on Electrical Machines and Systems (ICEMMS 2010). Incheon, Korea, 2010: 53-56

[10] Xu Xiao-Dong, Xiao Yin-Tao, Zhu Shi-Rui. Simulation investigation of rumor propagation in microblogging community. Computer Engineering, 2011, 37(10): 272-274(in Chinese)
(许晓东，肖银涛，朱士瑞. 微博社区的谣言传播仿真研究. 计算机工程，2011，37(10)：272-274)

[11] Granovetter M S. The Strength of Weak Ties. Chicago: University of Chicago Press, 1974

[12] Onnela J P, Saramaki J, Hyvonen J, et al. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences, 2007, 104(18): 7332-7336

[13] Zhao J, Wu J. Xu K. Weak ties: Subtle role of information diffusion in online social networks. Physical Review E, 2010, 82(1): 016105

[14] Burt R S. Structural Holes: The Social Structure of Competition. Cambridge, MA: Harvard University Press, 1992

[15] Lou T, Tang J. Mining structural hole spanners through information diffusion in social networks//Proceedings of the 22nd International Conference on World Wide Web (WWW

2013). Rio de Janeiro, Brazil, 2013: 825-836

[16] Li D, Xu Z, Li S, et al. Link recommendation for promoting information diffusion in social networks//Proceedings of the 22nd International Conference on World Wide Web (WWW 2013). Rio de Janeiro, Brazil, 2013: 185-186

[17] Montanari A, Saberi A. The spread of innovations in social networks. Proceedings of the National Academy of Sciences, 2010, 107(47): 20196-20201

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

[19] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2010). Washington DC, USA, 2010: 1029-1038

[20] Saito K, Nakano R, Kimura M. Prediction of information diffusion probabilities for independent cascade model//Proceedings of the 12th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES). Zagreb, Croatia, 2008: 67-75

[21] Xiang R, Neville J, Rogati M. Modeling relationship strength in online social networks//Proceedings of the 19th International Conference on World Wide Web (WWW 2010). Raleigh, USA, 2010: 981-990

[22] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604-632

[23] Goyal A, Bonchi F, Lakshmanan L V S. Learning influence probabilities in social networks//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM 2010). New York, USA, 2010: 241-250

[24] Tang J, Sun J M, Wang C, Yang Z. Social influence analysis in large-scale networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2009). Paris, France, 2009: 807-816

[25] Liu L, Tang J, Han J W, et al. Mining topic-level influence in heterogeneous networks//Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010). Toronto, Canada, 2010: 199-208

[26] Granovetter M. The strength of weak ties. American Journal of Sociology, 1973, 78(6): 1360-1380

[27] Krackhardt D. The Strength of Strong Ties: The Importance of Philos in Networks and Organization in Book of Nitin Nohria and Robert G. Eccles (Ed.), Networks and Organizations. Cambridge: Harvard Business School Press, 1992

[28] Blei D M, Ng AY, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3(4-5): 993-1022

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

[30] Domingos P, Richardson M. Mining the network value of customers//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2001). San Francisco, USA, 2001: 57-66

[31] Nemhauser G, Wolsey L, Fisher M. An analysis of the approximations for maximizing submodular set functions. Mathematical Programming, 1978, 14(1): 265-294

[32] Cornuejols G, Fisher M, Nemhauser G. Location of bank accounts to optimize float. Management Science, 1977, 23(8): 789-810

[33] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2007). San Jose, USA, 2007: 420-429

[34] Kimura M, Saito K, Nakano R, Motoda H. Extracting influential nodes on a social network for information diffusion. Data Mining and Knowledge Discovery, 2009, 7(1): 70-97

[35] Wasserman S, Faust K. Social Network Analysis. New York: Cambridge University Press, 1994

[36] Kimura M, Saito K. Tractable models for information diffusion in social networks//Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006). Berlin, Germany, 2006: 259-271

[37] Bharathi S, Kempe D, Salek M. Competitive influence maximization in social networks//Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE 2007). San Diego, USA, 2007: 306-311

[38] Carnes T, Nagarajan C, Wild S M, Zuylen A V. Maximizing influence in a competitive social network: A follower's perspective//Proceedings of the 9th International Conference on Electronic Commerce (ICEC 2007). Minneapolis, USA, 2007: 351-360

[39] Budak C, Agrawal D, Abbadi A E. Limiting the spread of misinformation in social networks//Proceedings of the 20th International Conference of World Wide Web (WWW 2011). Hyderabad, India, 2011: 665-674

[40] Kwak H, Lee C, Park H, Moon S. What is twitter, a social network or a news media?//Proceedings of the 19th International Conference of World Wide Web (WWW 2010). Raleigh, USA, 2010: 591-600

[41] Bakshy E, Hofman J M, Mason W A, Watts D J. Everyone's an influencer: Quantifying influence on Twitter//Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM 2011). Hong Kong, China, 2011: 65-74

[42] Wang D, Wen Z, Tong H, et al. Information spreading in context//Proceedings of the 20th International Conference of World Wide Web (WWW 2011). Hyderabad, India, 2011: 735-744

[43] Leskovec J, McGlohon M, Faloutsos C, et al. Cascading behavior in large blog graphs//Proceedings of the 7th SIAM International Conference on Data Mining (SDM 2007). Minneapolis, USA, 2007: 551-556

[44] Ghosh R, Lerman K. A framework for quantitative analysis of cascades on networks//Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM 2011). Hong Kong, China, 2011: 665-674

[45] Lerman K, Ghosh R. Information contagion: An empirical study of the spread of news on digg and Twitter social networks//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM 2010). Washington, USA, 2010: 90-97

[46] Cha M, Mislove A, Gummadi P K. A measurement-driven analysis of information propagation in the flickr social network//Proceedings of the 18th International Conference of World Wide Web (WWW 2009). Madrid, Spain, 2009: 721-730

[47] Shi X, Tseng B, Adamic L A. Information Diffusion in Computer Science Citation Networks//Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM 2009). San Jose, USA, 2009

[48] Golub B, Jackson M. Using selection bias to explain the observed structure of Internet diffusions. Proceedings of the National Academy of Sciences, 2010, 107(24): 10833-10836

[49] Bakshy E, Karrer B, Adamic L A. Social influence and the diffusion of user-created content//Proceedings of the 10th ACM Conference on Electronic Commerce (EC 2009). Stanford, USA, 2009: 325-334

[50] Ienco D, Bonchi F, Castillo C. The meme ranking problem: Maximizing microblogging virality//Proceedings of the 10th IEEE International Conference on Data Mining Workshops (ICDMW 2010). Sydney, Australia, 2010: 328-335

[51] Yang Z, Guo J Y, Cai K K, et al. Understanding retweeting behaviors in social networks//Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010). Toronto, Canada, 2010: 1633-1636

[52] Sun J, Tang J. A survey of models and algorithms for social influence analysis. Social Network Data Analytics. New York: Springer US, 2011: 177-214

[53] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. Technical Report, Stanford Digital Library Technologies Project, 1998

[54] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604-632

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

[56] Cha M, Haddadi H, Benevenuto F, Gummad KP. Measuring user influence on twitter: The million follower fallacy//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM 2010). Washington, USA, 2010: 10-17

[57] Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. New York: Chapman & Hall/CRC, 1984

[58] Kschischang F, Member S, Frey BJ, Loeliger H. Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory, 2001, 47(2): 498-519

[59] Loeliger H. An introduction to factor graphs. IEEE Signal Processing Magazine, 2004, 21(1): 28-41

[60] Yang J, Counts S. Predicting the speed, scale, and range of information diffusion in Twitter//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM 2010). Washington, USA, 2010: 355-358

[61] Cox D R, Oakes D. Analysis of Survival Data. London: Chapman & Hall, 1984

[62] Yang J, Leskovec J. Modeling information diffusion in implicit networks//Proceedings of the 10th IEEE International Conference on Data Mining. Sydney, Australia, 2010: 599-608

[63] Lawson C L, Hanson R J. Solving Least Squares Problems. New Jersey: Englewood Cliffs, 1995

[64] Galuba W, Chakraborty D, Aberer K, Despotovic Z. Outtweeting the Twitterers predicting information cascades in microblogs//Proceedings of the 3rd Workshop on Online Social Networks (WOSN 2010). Boston, USA, 2010

[65] Song X D, Chi Y, Hino P, Tseng BL. Information flow modeling based on diffusion rate for prediction and ranking//Proceedings of the 16th International Conference of World Wide Web (WWW 2007). Banff, Canada, 2007: 191-200

[66] Norris J R. Markov Chains. New York: Cambridge University Press, 1997

[67] Zhao Li, Yuan Rui-Xi, Guan Xiao-Hong, Jia Qing-Shan. Bursty propagation model for incidental events in blog networks. Journal of Software, 2009, 20(5): 1384-1392(in Chinese) (赵丽，袁睿翕，管晓宏，贾庆山. 博客网络中具有突发性的话题传播模型. 软件学报，2009，20(5)：1384-1392)

[68] Snijders T A B, Van de Bunt G G, Steglich C E G. Introduction to stochastic actor-based models for network dynamics. Social Networks, 2010, 32(1): 44-60

[69] Snijders T A B, Koskinen J, Schweinberger M. Maximum likelihood estimation for social network dynamics. Annals of Applied Statistics, 2010, 4(2): 567-588

[70] Sulo R, Tanya B, Robert G. Temporal scale of processes in dynamic networks//Proceedings of the 11th IEEE International Conference on Data Mining Workshops (ICDMW 2011). Vancouver, Canada, 2011: 925-932

[71] Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 2

[72] Leskovec J, Backstrom L, Kumar R, Tomkins A. Microscopic evolution of social networks//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2008). Las Vegas, USA, 2008: 462-470

[73] Blonder B, Wey T, Dornhaus A, et al. Temporal dynamics and network analysis. Methods in Ecology and Evolution, 2012, 3(6): 958-972

[74] Croft D, Edenbrow M, Darden S, et al. Effect of gyrodactylid ectoparasites on host behaviour and social network structure in guppies Poecilia reticulata. Behavioral Ecology and Sociobiology, 2011, 65(12): 2219-2227

[75] Volz E, Meyers L A. Susceptible-infected-recovered epidemics in dynamic contact networks. Proceedings of the Royal Society of London, 2007, 274(1628): 2925-2933

[76] Volz E, Meyers L A. Epidemic thresholds in dynamic contact networks. Journal of the Royal Society Interface, 2009, 6(32): 233-241

LI Dong, born in 1987, Ph. D. candidate. His research interests include social computing and information diffusion.

XU Zhi-Ming, born in 1967, Ph. D., professor. His research interests include social computing and information retrieval.

LI Sheng, born in 1943, Ph. D., professor, Ph. D. supervisor. His research interests include natural language processing, machine translation and information retrieval.

LIU Ting, born in 1972, Ph. D., professor, Ph. D. supervisor. His research interests include social computing and information retrieval.

WANG Xiu-Wen, born in 1975, Ph. D., senior engineer. Her research interests focus on social network analysis.

Background

Along with the vast development of social media, many social web sites such as Facebook, Twitter, Sina Weibo and so on, are attracting more and more users to communicate in the new platform. At the same time, there has been tremendous interest in the research of information diffusion in online social networks. When a user observes that his neighbors adopt a piece of information, this user will be influenced so that to consider whether to adopt the information, which lead to the phenomenon of information diffusion. In fact, information diffusion is caused by user behavior, for example, users perform forward behavior to spread microblog in Sina Weibo. So information diffusion also can be regarded as user behavior diffusion.

The manner of obtaining information in social media (or called web 2.0 media) is not different from the manner in traditional web media (or called web 1.0 media). In traditional web media, web users mainly get information from massive web pages by the search engine tool. But in new social media, search engine is not the most important tool of obtaining information any longer. If the users want to get information of certain aspects, they will build social links with some other users who will release related information, and then valuable information will spread to users over social network. So, the authors should realize that the role of research of information diffusion in social networks in social media is the same as that of research of search engine in traditional web media. The research has a wide range of applications, including viral marketing, personal recommendation, public opinion regulation, and so on.

To the best their knowledge, this is the first review on the research of information diffusion in social networks. In this paper, the authors divided current study into research based on theoretical diffusion model and research based on information diffusion cascades, then introduced the main directions of each research in detail, and made a summary, comparison and analysis of research methods and progress for each research direction. The authors also summarized the existing problems and future direction of this research.

This work is supported by the Program of National Science Foundation of China (No. 61173074) and the National "863" High-Tech Research and Development Program Fund (No. 2011AA01A207). The former one aims to study large-scale social network analysis techniques, the latter one aims to develop internet language translation system.