软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(2):247−263 [doi: 10.13328/j.cnki.jos.004944]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

微博信息传播预测研究综述*

李 洋, 陈毅恒, 刘 挺

(哈尔滨工业大学 计算机科学与技术学院 社会计算与信息检索研究中心,黑龙江 哈尔滨 150001)

通讯作者: 刘挺, E-mail: tliu@ir.hit.edu.cn

摘 要: 微博已经逐渐成为人们获取信息、分享信息的重要社会媒体,深刻影响并改变了信息的传播方式.针对微博信息传播预测问题展开综述.该研究对舆情监控、微博营销、个性化推荐具有重要意义.首先概述微博信息传播过程,通过介绍微博信息传播的定性研究工作,揭示微博信息传播的特点;接着,从以信息为中心、以用户为中心以及以信息和用户为中心这 3 个角度介绍微博信息传播预测相关研究工作,对应的主要研究任务分别是微博信息流行度预测、用户传播行为预测和微博信息传播路径预测;继而介绍可用于微博信息传播预测研究的公开数据资源;最后,展望微博信息传播预测研究的问题与挑战.

关键词: 微博;信息传播预测;信息流行度;传播行为;信息传播路径

中图法分类号: TP18

中文引用格式: 李洋,陈毅恒,刘挺.微博信息传播预测研究综述.软件学报,2016,27(2):247−263. http://www.jos.org.cn/1000-9825/4944.htm

英文引用格式: Li Y, Chen YH, Liu T. Survey on predicting information propagation in microblogs. Ruan Jian Xue Bao/Journal of Software, 2016,27(2):247−263 (in Chinese). http://www.jos.org.cn/1000-9825/4944.htm

Survey on Predicting Information Propagation in Microblogs

LI Yang, CHEN Yi-Heng, LIU Ting

(Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Microblogs have gradually become popular platforms for users to acquire and share information with the public, which have brought a profound impact on information propagation. This paper presents a survey of predicting information propagation in microblogs. It is important to public opinion monitoring, online marketing and personalized recommendation. The paper first introduces the mechanism of information propagation, and reveals the characteristics of information propagation in microblogs through a brief overview of the qualitative research. Then, representative work is reviewed for the prediction of information propagation from three aspects including information centered prediction, user centered prediction and information-user centered prediction. The three corresponding tasks are predicting the popularity of information, predicting individual spread behaviors and predicting the path of information dissemination respectively. Next, the publicly available data sets for information propagation in microblogs are summarized. Finally, the key challenges are discussed to suggest the future research directions.

Key words: microblog; prediction of information propagation; popularity of information; spread behavior; path of information dissemination

微博是一种允许用户及时更新简短文本并公开发布的微型博客形式.作为一种新兴的信息传播载体,微博在国内外得到了广泛的应用[1].目前,国外最流行的微博服务网站是 Twitter.美国统计脑研究中心(Statistic Brain

* 基金项目: 国家重点基础研究发展计划(973)(2014CB340503); 国家自然科学基金(61472107, 61202277)

Foundation item: National Program on Key Basic Research Project (973) (2014CB340503); National Natural Science Foundation of China (61472107, 61202277)

收稿时间: 2015-05-08; 修改时间: 2015-10-17; 采用时间: 2015-11-15

Research Institute)于 2014 年 7 月发布的统计结果显示:Twitter 活跃用户数量多达 6.45 亿,平均每天产生大约 5 800 万条信息(http://www.statisticbrain.com).国内则以新浪微博和腾讯微博为主要代表.

信息(知识)等通过交互行为而到达个体的过程叫做信息传播[2].微博的广泛流行不仅导致信息呈爆炸式增长,而且为互联网信息传播的方式带来了巨大的变革.过去,人们只能通过搜索、浏览等方式从少数信息源获得信息.在微博中,信息内容主要通过人与人之间建立的"关注-被关注"网络进行传播.人与人之间的互联、人与信息之间的互联高度融合,人人参与到信息的产生与传播过程,这种传播方式使得一条信息能够在短时间内传播到数百万计的用户[3,4].然而,大量的用户生成信息(user generated content,简称 UGC)也带来了诸如信息过载、虚假信息泛滥等问题,微博信息传播预测的研究为解决这些问题提供了可能.微博信息传播预测是指在掌握现有信息传播形态的基础上,依照一定的方法和规律对未来的信息传播趋势进行测算,以预先了解信息传播的最终过程和结果[5],为信息传播的干预提供依据,从而辅助决策、趋利避害.例如,研究用户的在线行为以及传播行为规律将有助于网络公司准确地把握用户的偏好,并将可能感兴趣的话题信息、其他用户或者用户社群推荐给该用户[6,7];企业可根据信息传播规律寻找潜在客户,进行定位化的广告宣传,以较低的费用将新产品推广至整个网络,从而产生较大的社会影响和商业价值;通过预测信息的传播范围和用户的观点态度,政府部门可以准确地判断舆论的热点问题,以便及时采取科学的控制和引导[6].

近年来,基于社会媒体的研究工作受到广泛关注[8,9].微博信息传播预测是一个具有很大科学价值和应用价值的研究课题.该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,吸引了大量的复杂网络、信息检索和自然语言处理等计算机领域学者们的关注和重视.本文主要侧重于自然语言处理领域的相关研究.微博信息传播预测中的两个主要元素是信息和用户,信息与用户之间存在相互作用[2].目前,根据预测任务的侧重点不同,可概括为以信息为中心、以用户为中心和以信息和用户为中心这 3 个方面.

1) 以信息为中心的预测研究忽略个体的传播行为,只关注信息的整体传播趋势,如传播范围、传播周期等特性,从而为舆情监控提供了可能.其主要任务是微博信息流行度预测,预测对象是微博信息,具体可分为针对某一条微博信息或特定的微博信息集合,例如含相同主题、超链接、标签等.
2) 以用户为中心的预测研究以用户的兴趣和行为建模为基础,分析用户是否会参与某信息的传播,从而为个性化推荐提供了可能.其主要任务是用户传播行为预测,预测对象是用户.

以上两个角度的研究工作并非完全独立,对用户传播行为的预测可辅助信息流行度的预测.

3) 以信息和用户为中心的预测研究既关注信息的整体传播情况,又关注网络中所有个体对信息的传播行为.其主要任务是通过分析个体的传播行为或传播概率预测信息的传播路径,预测对象是信息和用户.

本文首先概述微博信息传播过程,揭示微博信息传播规律;其次,从 3 个角度介绍微博信息传播预测的主要研究工作,即,以信息为中心的微博信息流行度预测、以用户为中心的用户传播行为预测(以转发为例)、以信息和用户为中心的微博信息传播路径预测;接着,总结公开数据资源;最后,展望此研究的问题与挑战.

1 微博信息传播

相对于微博而言,社交网站[10,11]、博客[12]等信息传播的研究起步较早.作为一种新型社会媒体,微博特有的短文本性、弱关系性和即时性等特点在信息传播中发挥了重要作用.本节首先举例说明微博信息的传播过程,然后通过综述近年来微博信息传播的定性研究工作,从传播内容和传播形态两方面总结微博信息传播特点.

1.1 微博信息传播过程

在微博中,信息通过用户之间的传播行为,沿着用户之间的关联关系构成的社会网络进行传播.通常用图 $G=\{V,E\}$来表示此社会网络,其中,V是节点集合,即该网络中的用户集合;$E\subseteq V\times V$是图 G中的边集合,表示用户之间的关注关系.微博允许用户之间存在单向关注关系,边 e_{ij}表示用户 v_i关注了用户 v_j,因此,G 是有向图.在微博中,用户可直接获取到其所有关注用户发布的微博消息,微博信息传播的主要途径是用户的转发行为[6],转发行为使得信息从发布者开始,按照用户之间的关注关系组成的路径进行层级式传播[13,14].图 1 描述了在用户转发

行为的作用下,微博信息传播的过程.节点 *A* 表示信息发布者,节点 *B*~节点 *E* 表示信息转发者,节点 *F* 和节点 *I* 表示边界用户,虚线箭头方向代表信息传播方向.通常来说,转发行为建立在关注关系基础之上.节点 *B* 和节点 *A* 之间有关注关系,用户 *B* 关注用户 *A*,则 *B* 能够接收到 *A* 的即时更新信息(*B* 是 *A* 的"粉丝",*A* 是 *B* 的"关注者").图 1 中,*A* 发布了某微博信息,*A* 的粉丝 *B*,*C*,*D* 对此信息转发,虽然用户 *A* 与用户 *E* 之间没有直接关注关系,但由于 *E* 与 *B* 的关注关系,经过 *B* 的传播行为,由 *A* 发出的信息仍然可以传播到用户 *E*,*E* 若将此条信息继续转发,信息又可传播至 *I*.同理,信息也可传播至 *F*.*I* 和 *F* 不再继续产生转发行为,则信息在此条路径上的传播停止.

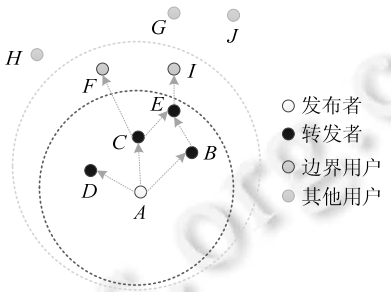


Fig.1 Retweet behavior and information propagation in microblogs

图 1 微博中的转发行为与信息传播

1.2 微博信息传播特点

1.2.1 传播内容特点

微博的短文本性和即时性决定了微博信息内容与其他社会媒体的差异:短文本性缩短了用户获取信息和发布信息的时间,加快了信息传播的速度;即时性主要体现为信息更新速度快,信息内容时效性更强,这一点在 Kwak 等人[14]对 Twitter 全网信息的统计分析中得到了证明,超过一半的转发行为发生在微博发布的 1 小时内,75%的转发行为发生在 1 天内,只有 10%的转发行为发生在 1 个月以后.因此,微博信息传播内容的特点可概括为时效性强、以日常讨论和时事信息为主[1,15−19]、具有一定的地域性差异[20−23].通过对微博信息的内容进行大规模的统计分析以及与传统媒体的对比,研究者发现:微博信息传播内容主要以日常生活的讨论和时事信息的分享为主[1],微博对时事信息的传播与传统媒体保持一致,对于那些经常被传统媒体忽视的地方性事件具有更强的传播能力[17−19].一些研究工作通过实证分析的方法研究特定主题的信息内容(以突发性事件、灾难事件为主)在微博上的传播特点,如 2009 年红河洪水[21,22]、2010 年玉树地震等[23].作者通过事件关键词抽取事件发生时间段内的微博及其转发微博,发现此类信息的传播具有地域性差别.普通用户在信息传播的参与度上不如事件发生地的用户活跃,普通用户更关注事件内容的概要信息而非细节,他们通常转发新闻标题或链接,以表达自己的感情;而事件发生地当地用户所发布或转发的信息则更加具体、详细.

1.2.2 传播形态特点

相对于关系驱动的社交网站,微博是一种信息驱动的弱关系型网络.微博的弱关系性体现于其允许用户之间存在单向关联关系,这种单向关联关系使微博网络中节点的互连性更好,网络结构更加紧密.微博信息传播形态的主要特点是传播速度快、传播路径长.通过对微博与社交网站如 Facebook,Digg 等信息传播的大量对比分析,研究者们得出了一致性结论:微博相对于社交网站有更多的单向关联,通过这些单向关联,微博信息传播路径更长、传播速度更快,具有小世界网络性质,符合六度分隔理论[24−26].Yi 等人[24]还对新浪微博 500 个热点信息的传播形态做了更细致的分析,将其概括为 7 种主要传播形态:波纹式模型、蒲公英式模型、菌落式模型、烟花式模型、蜂巢式模型、双子星式模型、随机引爆式模型,其中以波纹式和蒲公英式最为常见.

在大量数据采样和实证分析的基础上,研究者得出微博信息传播内容和传播形态的特点,微博信息传播与其他社会媒体的信息传播存在诸多的区别.这种差异性决定了微博在社会媒体时代信息传播中举足轻重的地位.因此,对微博信息传播的研究停留在浅层分析是远远不够的.从下一节起,本文将介绍微博信息传播预测的

相关研究工作.

2 微博信息流行度预测

自社会媒体出现以来,用户生成内容大量涌现,信息过载等诸多问题随之而来,因此,信息流行度预测研究备受关注.信息流行度是指信息在社交网络中最终的传播过程和结果,通常与信息的形式和传播方式有关.比如,视频分享网站上,视频信息的流行度通常由浏览数、分享数来衡量[10,27,28];新闻平台中的信息流行度则由新闻评论数来表示[29].微博信息流行度可以多种形式呈现,可概括为传播范围[30−34]和传播周期[35,36]两个角度.

2.1 微博信息流行度定义

传播范围从空间的角度衡量微博流行度,关注信息整体传播趋势.微博的总转发数[30]、总评论数[31]和总浏览数[32,33]是描述信息传播范围最常见的信息流行度表示.值得强调的是,微博信息的转发数和浏览数是两种独立的流行度定义,即使在转发数相同时,由于转发节点出度不同,微博信息的浏览数也有所不同;尽管两者之间存在正相关关系,但是这种相关关系并不强.

传播周期从时间的角度衡量微博信息的流行度,关注信息在网络中的传播速度以及持续传播时间.传播速度指自微博信息发布之时起至最早的转发行为所经历的时间[34].微博的生命周期指微博信息自发布之时起到不再受到任何关注(转发、评论等)为止的时间间隔[35,36].微博具有极强的时效性,普通微博信息的生命周期通常都很短,因此,生命周期常用作含有特定主题(含有相同标签或事件)的微博集合的流行度衡量标准.比如热点话题信息,持续讨论的时间越长,其流行度越高.

2.2 预测方法

微博信息流行度预测是依照一定的方法和规律对信息未来的流行度进行测算.从信息类型的角度看,主要包括普通微博或相同主题的微博集合(含特定标签、特定链接或特定事件).从模型角度看,目前,微博信息流行度预测的研究方法以基于传染病模型和分类或回归模型的预测方法为主.

- 传染病模型源于早期信息扩散理论.信息扩散理论主要包括羊群效应、信息级联、创新扩散理论以及传染病模型[2].早在社会媒体出现之前,信息传播就是复杂网络传播动力学研究的热点问题,比如要研究创新(innovation)、传染病(epidemic)和产品(product)在真实社会网络中的扩散,他们提出许多实证和模型来解释这些信息传播的特征和机制.但由于获得真实数据的困难,该研究只停留在理论层面.社会媒体的出现改变了这一现状,用户之间信息的传播由无形变成了有形,因此,信息扩散理论被应用于博客、微博等信息流行度预测研究之中,其中以传染病模型和信息级联模型最为典型.
- 分类或回归模型主要通过统计机器学习的方法预测微博信息的流行度,重点研究的问题是信息流行度的影响因素以及各个因素的重要性、如何将影响因素表示成特征、如何利用抽取的特征训练机器学习模型,对新信息的流行度进行有效的预测等.

2.2.1 传染病模型

传染病模型是对疾病在人群中的表现和分布形式进行数学建模的方法,典型的传染病模型包括 SI (susceptible-infected),SIS(susceptible-infected-susceptible),SIR(susceptible-infected-recovered)以及 SIRS (susceptible-infected-recovered-susceptible),如图 2 所示.其主要思想是:将人群中的个体按照其所处的状态进行分类,关注每类状态下个体数量比例的演化,处在每个状态的个体比例通过微分方程求解.基本状态包括易感染状态 S、感染状态 I 以及恢复状态 R.首先,状态为易感染状态 S 的个体以固定的概率β变成感染状态 I.在 SIS 模型中,状态为 I 的个体会以概率γ再变成易感染状态 S;在 SIR 模型中,状态为 I 的个体会以概率γ恢复,并保持在恢复状态 R;而在 SIRS 型中,个体在恢复之后又以概率λ变成易感染状态.

信息传播与传染病传播具有一定的相似性,信息可视为传染病,信息传播过程相当于疾病感染过程.许多研究者直接将 SIR 模型[37,38]或 SIS 模型[39]应用于微博信息传播的建模与流行度预测,两者主要区别在于 SIS 模型假设用户会多次转发同一信息.Liu 等人[37]基于 SIR 模型构造 Twitter 的谣言传播模型,该模型将用户分成信息

传播者、信息未知者和不受影响者这 3 类,在仿真实验中,模型还结合了微博网络的弱关系性以及传播者的影响力计算传播概率β.还有一些学者结合微博信息传播的具体问题,在经典传染病模型的基础上提出了改进模型.如:

- Xiong 等人[40]对用户的浏览行为和转发行为做了细致的区分,提出了 SCIR 模型(susceptible,infected,contacted,refractory),即,用户在浏览信息后即处于接触状态 C,用户若产生转发行为,则变为 I 状态;若不转发,则变为 R 状态.
- Wu 等人[41]人根据微博中用户发表、浏览、回复和转发微博的基本行为,将微博的信息交流分成信息发布、信息接收、信息加工、信息传播这 4 个阶段,并考虑信息丢失(被用户忽略的信息熵),提出竞争窗口模型,能够对信息传播动态的建模.
- Yang 等人[42]在 SIS 模型的基础上提出了线性影响模型(linear influence model),根据当前微博信息的流行度预测未来某一时间的流行度.模型以微博信息的传播受各节点影响力支配为假设,建立每个节点的影响函数,此函数用以量化该节点对后续被激活节点的影响力,某时间处于活跃状态节点的影响力之和即为此时刻信息的流行度.

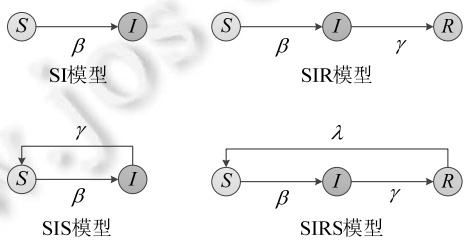


Fig.2 Four typical epidemic models

图 2 4 种典型传染病模型

2.2.2 分类或回归模型

分类或回归模型主要从分析影响信息传播关键因素的角度出发,利用统计机器学习模型对微博信息流行度进行预测.将待预测的微博信息表示成一组基于影响因素的特征,把微博信息流行度预测问题转化为分类[43−45]或者回归问题[46],通过大量的已知数据训练出机器学习模型对未知信息进行预测.基于分类模型的方法通常是将微博信息按照流行程度划分为几个等级,然后进行多元分类,目前没有统一的分类标准[43,44].基于回归模型的方法试图找到影响因素与微博信息流行度之间的相关关系,进而使用线性回归或非线性回归模型进行微博信息流行度的预测.无论是基于分类模型还是基于回归模型,这类预测方法的本质问题都是对于微博信息流行度影响因素的分析,即,特征的选择与表示.目前,常用的特征主要有以下 3 类.

(1) 发布者影响力

社会影响力是一个用户影响他人的观点、情感和行为的现象[47].Bhattacharya 等人[36]对 Twitter 上 12 个主要新闻媒体的公共账号所发布的新闻进行了传播分析,发现信息源是非常重要的影响因素,权威媒体 BBC 和 NYTimes 的微博信息的流行度明显高于其他.Bakshy 等人[46]通过对 Twitter 上 160 万用户及其微博信息流行度进行分析,观察到最流行微博的发布者具有较大粉丝数量,他们的其他微博也普遍具有较高流行度.他们基于回归树模型进行微博流行度预测,发现平均被转发数、最大被转发数等特征与用户微博转发量呈现相关性,从而证明用户影响力能够影响微博的传播.Zhang 等人[48]通过对新浪微博大量数据进行分析,也得到类似的结论:微博流行度与用户粉丝数呈正相关性,即,粉丝数越多,微博越易被转发.但这种相关性并不是很大,这与微博复杂的信息传播机制有关.Mao 等人[49]认为:不能只简单地根据用户历史上的一些统计数据,如被转发数量、被提及数量等来估计用户在未来传播信息、造成社会影响、引发其他用户互动的能力.他们通过进行相关性分析,发现用户传播信息能力的大小随着时间的推移是会较为显著地改变的.还有一些研究工作[43,46,50]直接使用用户的上下文特征表征用户影响力因素,如粉丝数、双向好友数、聚集系数以及 PageRank 值[51].然而,微博网络中用

户的影响力不仅取决于其网络结构,而且也与用户自身兴趣有关.Weng 等人对新加坡 Twitter 用户进行了统计,他们发现,多数用户与他们的大部分粉丝之间有互相关注的关系,证实 Twitter 用户之间的关注关系存在同质性的特点.因此,仅使用粉丝数或 PageRank 值这种基于网络结构的特征衡量用户影响力是不准确的,他们在 PageRank 基础上提出了融合网络结构和用户兴趣的 TwitterRank 算法,构建主题影响力模型[52].

(2) 文本内容

微博的内容可决定微博信息的流行度,比如,如果某人在一条微博中谈论当前网络中备受争议的话题,则这条微博信息将比此人同一时间发布的其他微博信息更有可能引发其好友的传播行为[53].Hong 等人[43]使用主题模型学习微博文本内容的主题分布,将其作为文本特征预测微博是否流行.Ma 等人[54]和 Tsur 等人[55]发现,微博内容涉及时事热点话题时更容易引发用户的传播.Jenders 等人[56]利用自然语言处理技术深入分析微博的文本内容,如命名实体、情感倾向等,发现内容特征与用户特征可以互补,结合使用能够提升预测的准确度.Tan 等人[57]则单纯地从语言表达的角度分析了微博内容对转发量的影响.为了排除其他因素(用户影响力、文本主题)的干扰,在实验中选取了来自于同一发布者且含有相同 URL 的微博作为对照组,对照组中的微博主题内容相同、表达方式不同,证明不同措辞和语言习惯对微博转发量的影响.通常,一条微博不仅含有文本信息,还附有超链接、标签、图片、音频、视频等多媒体信息.Suh 等人[30]利用主成分分析的方法,发现超链接和标签数量与微博信息是否被转发具有较强的相关性.Zhao 等人[58]发现,含有多媒体内容的微博比普通微博有更大的转发量和更长的生命周期.Can 等人[59]研究视觉提示对转发量的影响,利用微博中图片的灰度信息以及 GIST 特征作为视觉特征.

(3) 用户活跃时间

微博信息更新极快,用户的活跃时间有限,因此由信息过载引发的“信息丢失”现象也较为严重.比如,同样内容的微博信息,发布在上午 11 时将会比发布在晚上 11 时更容易被转发,因为用户在白天的活跃度更高[45,60].Liu[61]通过分析不同官方账号粉丝用户转发时间的分布,发现相似的用户转发行为遵循相似的时间模式,如 McDonald's 和 iTunesMovies 的微博分别在中午 12 点和下午 18 点左右转发频率最大.他们通过对用户转发行为建模,发现用户的转发行为具有特定的时间分布.因此,用户的活跃时间因素也是影响微博信息流行度的主要因素.对用户活跃时间的分析,有助于在线营销以及信息传播预测.

此外,还有一部分工作[44,62,63]利用微博信息初期传播路径特征,如传播深度、传播广度、连接紧密度等,预测未来的信息流行度.

2.3 方法比较

前面介绍了微博信息流行度预测的研究工作,不同的预测方法基于不同的模型,预测内容也有所不同,现在进行总结与对比.

- 基于传染病模型的预测方法

传染病模型的方法通过对信息传播过程的分析和提炼,利用数学方法对信息传播规律做出抽象表示.该方法利用动力学演化方程组刻画不同类型节点随时间的演化关系,具有数学的严密性,可以描述信息整体的传播规律.但假设个体之间的连接关系是随机的,忽略用户之间的网络结构以及参与信息传播的个体之间的差异.此类方法侧重于传播过程中个体在几个状态之间的重新分配,关注整体传播情况,因此只适用于传播范围的估计,比如某时刻处于某状态个体的比例.

- 基于分类或回归模型的预测方法

基于分类或回归模型的方法关注信息传播的影响因素,基于影响因素的分析构建并选择特征,从而训练基于分类或回归的预测模型.无论是分类还是回归模型,其关键都在于解释各个特征与信息流行度的相关性.因为不依赖于网络结构,这类模型可预测信息传播范围、传播周期,也无法预测信息传播路径.

除了两类主要的预测方法之外,部分研究者将经济学分析方法用于微博信息流行度预测,比如基于时间序列模型的预测方法[64]、基于泊松过程的预测方法[5].这类方法基于随机过程理论和数理统计学方法,研究随机数据序列所遵从的统计规律,进一步推测未来的发展趋势,以用于解决实际问题.由于信息传播也具有随时间变化

的规律性,因此,此类模型也可用于微博信息流行度的预测.特别地,通过对信息传播路径的预测也可以估计其传播范围.由于这类方法既考虑用户行为又考虑信息主体,因此本文将其归为以信息和用户为中心的信息传播路径预测研究,将在第 4 节重点介绍.

3 用户转发行为预测

微博信息流行度预测的研究工作主要是以信息为中心,分析预测信息的传播范围、传播周期等特性,只关注信息的宏观传播趋势,不关注微观个体的传播行为.然而,个体的传播行为影响着信息整体的传播趋势.因此,一些研究者以用户为中心,从微观角度对用户的传播行为进行建模与预测.本节将对这类研究工作展开介绍.具体来说,预测用户是否会参与某信息的传播.在微博中,用户对信息的传播行为主要指转发行为[30,65].本节将以微博中用户的转发行为为例,介绍微博用户传播行为分析与预测的研究工作.

3.1 转发行为影响因素

首先,对微博用户转发行为预测问题进行形式化定义.已知$p\in P$表示微博的发布者,$u\in U$表示微博的接收用户,$v\in V$表示微博信息.假设发布者 p 在 t 时间发布了微博 v,则每个微博可以用四元组来表示$\langle u,v,p,t\rangle$.令 r_{uvp} 表示用户 u 对该微博的转发态度.用户转发行为预测是在给定$\langle u,v,p,t\rangle$的情况下,求解 r_{uvp},即,在 $t'(t'>t)$时刻,核心用户 u 是否会转发用户 p 在 t 时刻发布的微博信息 v.其中,r_{uvp} 可以由多种形式来表示,比如布尔值、相对顺序、概率值等.用户的转发行为是多种因素共同作用的结果[66],这使得看似简单的问题变得具有挑战.Boyd 等人[67]利用定性分析的方法研究 Twitter 上用户的转发行为,通过对收集到的转发信息进行归类总结,列出用户转发行为产生的原因.本文综合以往研究内容,将用户转发行为产生的影响因素概括为两类——信息内容因素和群体影响因素.信息内容因素主要包括信息内容自身特点以及信息内容与用户兴趣的吻合程度:前者包括信息内容的流行程度(是否热门话题)、信息内容的丰富性(是否含有多媒体、图片等),后者指用户是否对此类信息感兴趣.群体影响因素主要包括信息发布者对用户的影响以及其他信息转发者对用户的影响.

3.2 预测方法

用户转发行为预测是指通过一定的手段学习用户的兴趣和行为规律,从而对未知的用户转发行为进行预测.按照预测基本假设的不同,用户转发行为预测方法可分为基于用户过往行为的预测、基于用户文本兴趣的预测、基于用户所受群体影响的预测以及基于混合特征学习的预测.主要使用的模型包括:协同过滤模型、主题模型、因子图模型以及分类模型等.

3.2.1 基于用户过往行为

基于用户过往行为的预测方法依据用户在预测时间点前的过往行为,预测用户未来的行为.该方法认为:用户的兴趣短时间内不会改变,用户转发微博的行为受用户兴趣所驱动[68].因此,可充分利用已知的用户偏好或行为,预测未知的用户偏好或行为.用户过去转发了某些微博信息,则很可能还会对类似内容的微博信息感兴趣.

基于以上假设,Zaman 等人[6]最先利用协同过滤模型预测用户的转发行为.协同过滤的概念来源于推荐系统,目前应用最广泛的是矩阵分解技术,其核心思想是:假设用户的兴趣只受少数几个因素的影响,因此将稀疏且高维的“用户-物品”矩阵分解为两个低维矩阵,通过用户对物品的评分信息来学习用户特征矩阵 $u_i\in R^K$ 和物品特征矩阵 $v_j\in R^K$,最后重构低维矩阵预测用户对物品的评分 $\hat{r}_{ij}=u_i^Tv_j$.类似地,用户与微博信息可构建用户-信息矩阵,在此矩阵中,元素的值为 1 表示用户转发该微博信息.然而,不同于传统的商品推荐,用户转发信息的数据集中,由于新的信息不断出现,用户-信息矩阵是非常稀疏的,存在较严重的冷启动问题.因此,后续基于用户过往行为的预测研究工作致力于在传统协同过滤模型基础上融入丰富的特征,如用户属性特征、微博信息特征以及传播结构特征等[68−70].其中,文献[68,69]对微博信息内容进行了关键词和主题抽取,把用户-信息矩阵转化为用户-关键词矩阵或用户-主题矩阵,在一定程度上缓解了数据稀疏导致的新信息冷启动问题.

3.2.2 基于用户文本兴趣

基于用户文本兴趣的预测方法认为,用户对某信息的转发行为源于用户对微博文本内容的兴趣.此类方法

将用户历史微博信息视为该用户的伪文档,通过对用户进行文本兴趣建模,预测用户对未知信息的转发行为.

许多研究通过词袋模型(bag-of-words)对用户文本以及微博信息进行向量表示,然后计算文本相似度,相似度越大,用户转发该信息的可能性越大[43,65].另一方面,以隐含狄利克雷分布(latent Dirichlet allocation,简称LDA)[71]为代表的主题模型及其变型被广泛应用于社会媒体用户文本兴趣建模任务中.在 LDA 模型基础上,Xu等人[66]提出混合的潜在主题模型来解释用户转发行为的产生过程,该模型假设用户转发微博的文本来源于用户自身的兴趣分布、热点主题分布或其好友的主题分布.因此,用户微博中的每个词的生成方式取决于隐变量 x,x 服从每个用户的多项分布 λ.最后,将用户转发行为预测问题转化成求解每条文本生成概率问题:

$$p(v)=\frac{1}{N_v}\sum_{w\in v}p(w).$$

类似的工作还有文献[61].文中提出,用户的转发行为遵循 3W 模式〈When,Who,What〉,引入隐变量 c 表示该模式,用户的转发行为是基于特定的转发模式 c 以及文本主题 z 共同作用的结果,即

$$p(w\mid c)=\sum_z p(w\mid z)p(z\mid c).$$

Zhang 等人[72]认为,用户的文本兴趣是随时间变化的,因此提出了基于分层狄利克雷过程(hierarchical Dirichlet process HDP)的非参数贝叶斯模型[73].该模型不仅能够对用户兴趣进行动态的主题建模,还融合了其他影响因素,如用户所受其他用户的影响.

3.2.3 基于用户所受群体影响

基于用户所受群体影响的预测方法的基本假设是,用户转发行为的产生主要由于其所受群体的影响.一般而言,用户转发行为所受的群体影响分为两个方面:一方面来自于信息发布者的影响,另一方面来自于群体中其他人的转发行为的影响[72,74].为了验证用户的行为主要受其亲密好友的影响这一假设,Zhang 等人[74]利用准实验设计的方法设置对照组和干扰组,验证了用户之间局部影响力的存在.基于局部影响力和全局影响力的逻辑回归模型,有助于提升用户转发行为预测的结果.实验结果还表明:用户转发某微博的可能性与其好友中转发该微博的人数成正比,而与这些好友形成的社交圈数成反比.

因子图(factor graph)是对函数因子分解的表示图,在社会网络建模中得到了广泛应用[75−77].它将多变量函数描述为二分图,每一个因子图都包含两类节点——变量节点(variable node)和因子节点(factor node),边只连接不同类型的节点.如果函数的因子受某些变量的影响,那么该因子节点和这些变量节点之间则建立边.Yang 等人[75]提出了基于因子图模型的监督学习框架,对用户转发行为进行预测,将用户与微博信息发布者以及该信息转发轨迹中其他转发者之间的文本兴趣、拟社会交互等因素等作为因子.Bian 等人[78]除了考虑用户所受群体影响之外,还考虑了微博文本信息流行度的影响.在此以文献[78]为例,说明利用因子图模型对用户转发行为建模和求解的过程.首先假设用户转发行为来自于 3 种影响因子:兴趣驱动的影响 $f_I(p,u,v)$、社会关系驱动的影响 $f_S(p,u)$以及内容流行驱动的影响 $f_E(v)$.模型的优化目标为最大化似然函数:

$$P(Z)=\prod_{(p,u,v)\in A,z=1}f_I(p,u,v)\times\prod_{(p,u,\cdot)\in A,z=2}f_S(p,u)\times\prod_{(\cdot,\cdot,v)\in A,z=3}f_E(v).$$

其中,A 为训练集中所有转发行为集合,$Z=\{z_1,z_2,\ldots,z_{|A|}\}$对应每个转发行为的隐变量.

通常,通过和积算法(sum-product algorithm)求解各个变量的边缘分布.假设 $r_{p,u,v}\in\{-1,1\}$表示用户 u 是否转发由 p 发布的微博 v,$R=\{r_{p,u,v}\}$为待预测的用户转发行为集合,G 为当前网络拓扑结构,则预测问题可转化为:在给定 G 和 A 的情况下,求解所有待预测转发行为的最大联合条件概率问题 $P_\theta(R|G,A)$[78].

3.2.4 基于混合特征学习

基于混合特征学习的预测方法将转发行为预测视为二元分类问题,分析影响用户转发行为的因素并作为特征,然后选择适当的分类器训练分类模型.常见的特征可概括为独立特征和关系特征.独立特征指核心用户、微博以及微博发布者各自的特征;关系特征指三者之间的相互作用特征,如用户与微博发布者之间的社会关系、用户对微博内容的感兴趣程度以及微博发布者在该信息主题的权威度.如果微博发布者与核心用户有较亲密的社会关系,那么核心用户对微博发布者的信息更容易产生转发行为,社会关系特征可体现于两者是否为双

向好友、两人历史上微博互转的频度等.此类方法的关键在于各种特征的选择和组合.在对比各种因素对用户转发行为的影响时,常用的方法是基于特征递增法(add-one-feature-in)或特征排除法(leave-one-feature-out)[65]设计分组对比实验以及准实验设计的方法.Xu 等人[65]将用户转发行为的影响因素划分为基于社会关系的特性、基于内容的特征、基于发布者的特征,训练多种分类器(决策树、SVM(support vector machine)、逻辑回归),并使用特征排除法对比了各类特征的有效性,并说明社会关系特征相对于其他特征更加重要.此外,许多研究工作[79−81]也得出了类似的结论.比如,Luo 等人[81]使用基于 Pointwise 的排序学习方法[82]预测用户的转发行为,通过构建基于二元分类的排序函数对某微博的可能转发者进行 top-K 排序,发现:如果用户与微博发布者有较多的历史交互、相似的文本兴趣、相似的活跃时间,则该用户更容易产生转发行为.

3.3 方法比较

前文以转发行为为例介绍了微博用户转发行为预测的研究工作,4 种方法预测的基本假设不同,现将其进行总结与对比.

- 基于用户过往行为的预测方法

假设用户的转发行为反映用户的兴趣,依据用户在预测时间点前的过往行为预测用户未来的行为.这类方法主要使用的模型是协同过滤模型,该模型能够挖掘用户兴趣,利用已知的用户偏好或行为预测未知的用户信息偏好或行为.但是由于微博信息时效性强,新信息不断产生,因此,此类方法面临较严重的新信息冷启动问题.融入用户属性特征、微博文本特征等可缓解冷启动问题.

- 基于用户文本兴趣的预测方法

假设用户对某信息的转发行为主要源于用户对微博文本内容的兴趣,通过用户的过往微博文本信息对用户进行文本建模,从而预测用户对信息的转发行为.这类方法在用户拥有一定数量的微博文本信息时效果较好;但对于文本内容较少的用户,很难学到其真正感兴趣的内容.

- 基于用户所受群体影响的预测方法

假设用户转发行为的产生源于所受群体的影响,包括信息发布者的影响和其他信息转发者的影响.这类方法中较多使用因子图模型,除用户之间的相互影响外,因子图模型还可建模其他影响因素,如内容流行度的影响等[78].

- 基于混合特征学习的预测方法

将转发行为预测视为二元分类问题,认为用户转发行为是多种因素作用的结果.分析影响用户转发行为的因素并将其表示为特征,然后选择适当的分类器训练分类模型.这种方法最为简单、直观,模型解释性弱,依赖于特征的选择与组合.

4 微博信息传播路径预测

前面两节分别介绍了以信息为中心和以用户为中心的微博信息传播预测研究工作以信息为中心的信息流行度预测不关注用户之间的传播行为,以用户为中心的用户转发行为预测不关注信息整体的传播情况.本节的研究内容既关注用户之间的转发行为又关注信息整体的传播路径,即通过对用户转发概率或转发行为的预测进行微博信息传播路径的预测.

4.1 微博信息传播路径

相对于流行度,传播路径更细粒度地从空间角度描述微博信息的传播情况[34],此处,我们沿用第 1.1 节中图 1 的例子,举例说明什么是微博信息传播路径.用户 A 发布了某信息,A 的粉丝 B,C,D 以及 B 的粉丝 E 均产生了转发行为,假设每个用户仅有 1 次转发行为.以信息发布者 A 为根节点,通过追溯用户的转发行为,可构建信息传播的树型结构,即信息传播路径,如图 3 所示.

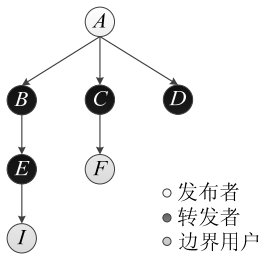


Fig.3 A tree structure of information propagation

图 3 信息传播的树型结构其他

4.2 预测方法

微博信息传播路径预测的目标是:预测信息最终经过哪些节点,哪些节点继续传播,到哪些节点终止.核心思想是:已知用户之间的网络结构,根据节点之间的关系推测信息传播的可能性.主要预测方法包括独立级联模型、线性阈值模型、分类模型以及博弈论模型.

4.2.1 独立级联模型

独立级联模型[83]和线性阈值模型[84]是信息级联模型的代表,信息级联模型定义了社交网络中信息通过用户之间的影响传播的方式与机制,即,传播群体中的个体通过网络结构相连,个体的决定只能受他们直接邻居(好友)的决定所影响.与传染病模型不同,信息级联模型考虑底层网络结构,通过网络中的节点以及节点之间的影响力构建信息传播过程,从而预测信息的传播路径.

独立级联模型假设节点在给定时间点存在两种状态,即,活跃态(active)或不活跃态(inactive).初始时,有一组节点处于活跃态,各个活跃节点将同步地影响与之相连的不活跃节点,使其转为活跃态.t 时刻,活跃节点 u 可以独立地以概率 $p_{u,v}$ 激活处于非活跃状态的邻居节点 v,若 $p_{u,v}$ 大于某随机生成的概率值 $rand$,则 v 被激活.不管 u 是否成功激活 v,u 对 v 都不再产生激活行为.

独立级联模型与线性阈值模型通常将激活概率或激活阈值设置为固定值(被激活节点 v 入度的倒数)或固定分布,这种设置忽略了不同用户之间的差异,在一定程度上限制了模型的灵活性,不符合微博信息传播的真实情形.因此,在利用信息级联模型进行微博流行度预测研究时,研究者通常会根据实际情况在级联模型基础上调整激活概率或激活阈值的设置方式.Zhu 等人[85]利用独立级联模型预测新浪微博中事件相关信息的传播,他们认为:用户接收大量信息,并以一定的概率阅读到某信息,基于高斯分布可为用户构建概率阅读模型.即,用户 u 在 t 时刻阅读到在 t_0 时刻推送给 u 的信息的概率为

$$p(t)=\mathrm{e}^{-\lambda\frac{(t-t_0)^2}{2\delta^2(u,t)}},$$

其中,λ为模型参数,$\delta(u,t)$为用户每页中接收的微博信息之间的最大时间间隔,此阅读概率即为独立级联模型中的激活概率.Dichkens 等人[86]基于独立级联模型预测 Twitter 中信息的转发路径,该模型假设用户之间的激活概率服从用户历史转发频率的β分布,并使用马尔科夫蒙特卡洛方法对进行参数估计.针对信息级联模型具有强同步性假设的局限性(不考虑时间因素,认为个体之间的激活行为是同步的),Guille 等人[87]提出了异步的独立级联模型.其基本思想是:将每个激活行为赋予时间片信息,用户之间的激活概率不仅依赖于用户之间的社会关系、信息文本,而且与被激活用户随时间变化的活跃度有关.由于每个激活行为都有时间片信息,因此,此模型可以预测不同时间片内的信息流行度.

4.2.2 线性阈值模型

与独立级联模型类似,线性阈值模型也假设节点在给定时间点存在两种状态,即活跃态(active)和不活跃态(inactive).初始时,有一组节点处于活跃态,各个活跃节点将同步地影响与之相连的不活跃节点,使其转为活跃态.线性阈值模型和独立级联模型的主要区别在于激活方式.在线性阈值模型中,每个非活跃状态的节点 v 都有一个激活阈值$\theta\in[0,1]$,当 v 的所有邻居节点对其的影响之和大于阈值θ时,v 才会被激活.v 的所有处于活跃状态

的邻居节点均可以参与激活 v,且可以产生多次激活行为.在微博信息传播中,将信息发布者(包括已转发者)视为活跃节点,将信息接收者视为非活跃节点,由此可见,独立级联模型是以信息发布者为中心的,线性阈值模型是以信息接收者为中心的.Galuba 等人[88]在线性阈值模型的基础上提出 At-Least-One(ALO)模型,预测 Twitter 上特定 URL 的信息流行度.该模型假设用户 u 对某信息 m 产生转发行为取决于两种因素——与时间有关的因素和与时间无关因素,并将激活概率表示成关于这两种因素的函数.文献[89]考虑了不同用户之间激活概率的差异性,结合节点间联系的强度和信息的吸引力对节点的特异性激活阈值进行计算,并且针对多个信息在社会网络中并行竞争传播的情形,提出了多信息并行竞争传播模型.

4.2.3 分类模型

基于分类模型的预测方法将信息传播路径预测问题转化为用户转发行为预测问题[90,91],此类工作也可视为独立级联模型的特殊形式,通过对单个用户传播行为的预测得到信息的整体流行度.以信息发布者为初始节点,根据用户之间的关注关系构建以该节点为根的树型结构,通过统计机器学习方法预测第 1~n 层全部粉丝用户对此信息的转发行为(通常,$1\leqslant n\leqslant 6$),从而得到信息传播的具体路径.与传统独立级联模型不同的是,用户是否被激活取决于多种因素,此类方法依赖于用户转发行为预测的准确度.

4.2.4 博弈论模型

博弈论模型将每个用户视为一个智能体,用户在接收到某信息时,会选择使自己获取最大化收益的策略.目前,已有一些研究者应用博弈论思想分析社交网络用户在信息传播时面对的成本、收益和策略选择[92],但大多数工作还停留在仿真数据层面,研究者构建某种属性的网络结构,然后利用模型在仿真的数据上的结果分析信息传播规律.此类方法可根据仿真得到的传播规律辅助预测真实的信息传播路径.

4.3 方法比较

前面介绍了微博信息传播路径预测的 4 种主要方法,下面对其进行总结与对比:

- 基于独立级联模型的预测方法

假设信息传播的过程是依靠相邻节点之间的相互影响,个体的传播行为取决于某个相邻节点对它的激活概率.

- 基于线性阈值模型的预测方法

假设个体的传播行为取决于所有相邻节点对它的影响是否超过激活阈值.

经典级联模型通常将激活概率或激活阈值设为固定值或固定分布,比如节点出度的倒数,限制了模型在真实数据中的应用.在改进的级联模型中,激活概率和激活阈值不再是固定值,而是关于信息内容、用户之间影响、时间等多种因素的函数.信息级联模型依靠微博网络结构,模型计算量大,可预测信息传播的具体路径.

- 基于分类模型的预测方法

将信息传播路径预测转化成用户转发行为的预测问题,依据关注关系,通过构建以信息发布者为根的树形传播结构,逐层预测该发布者后继节点(粉丝)的转发行为.此类方法的准确率依赖于信息转发行为预测的准确率,此方法由于将用户转发行为做分类预测,因此可能产生错误级联的问题.

- 基于博弈论的预测方法

将每个用户视为一个智能体,假设每个用户在接收到某信息时会进行利益的博弈,采取令自己获利最多的策略.此类方法多用于数据仿真研究,可根据仿真的传播规律辅助预测真实信息传播.

5 公开数据资源

前面介绍了微博信息传播预测的主要研究工作,本节将介绍可用于微博信息传播预测研究的公开数据资源,以对有兴趣的研究者有所启发.目前,可用于微博信息传播研究的公开数据资源还比较少,主要研究工作大都是基于 Twitter 或新浪微博.

Twitter 较大规模的数据资源主要有两个,分别来自斯坦福大学 SNAP(Stanford Network Analysis Project)小组以及 Haewoon Kwak[14].斯坦福 SNAP 数据集[93]的时间跨度是 2009 年 6 月 1 日~2009 年 12 月 31 日,其中覆

盖了 1 700 万 Twitter 用户以及 4.76 亿微博信息,大概占 Twitter 同时间段内全部数据的 20%~30%.Haewoon Kwak 数据集包含 4 262 个热点话题的 1.06 亿条微博以及 2009 年 7 月 Twitter 全部用户和用户关系信息[14].

新浪微博公开的大规模数据资源是 WISE2012 数据集以及清华大学唐杰老师团队发布的数据集[74].WISE2012(Web information systems engineering,http://www.wise2012.cs.ucy.ac.cy/challenge.html)发布微博竞赛任务,给定特定时间点前的涵盖 6 个事件的 33 条微博全部转发数据,参与者预测微博在发布后 30 天的转发量和阅读数.同时发布了新浪微博数据集覆盖 2009 年 8 月~2011 年 12 月的 3 亿多条微博、7 000 多万用户的关注网络.清华大学唐杰老师团队的新浪微博数据集中包含 30 万条原创微博及其 2 370 万转发微博,平均每条微博有 80 条左右转发,其中涉及 177 万微博用户以及 2 400 万关注关系[74].各个数据集的统计信息见表 1.

Table 1 Details of publicly available data sets

表 1 公开数据集详细说明

	用户数	微博数	用户关系数	完备性	数据链接地址
SNAP	1 700 万	4.76 亿	N/A	非全网数据	http://snap.stanford.edu/data/twitter7.html
H. Kwak	4.17 亿	1.06 亿	14.7 亿	全网数据	http://an.kaist.ac.kr/traces/WWW2010.html
WISE	7 000 万	3 亿	N/A	非全网数据	http://www.wise2012.cs.ucy.ac.cy/challenge.html
清华	117 万	2400 万	3.08 亿	非全网数据	http://arnetminer.org/billboard/Influencelocality

6 总结与展望

本文在充分调研和深入分析的基础上,对微博信息传播预测研究进展进行了综述.首先介绍了微博信息传播的特点,然后,从以信息为中心、以用户为中心、以信息和用户为中心这 3 个角度介绍了微博信息传播预测研究中的关键问题:微博流行度预测问题、用户传播行为预测(以转发为例)以及微博信息传播路径预测.最后介绍了微博信息传播研究的公开数据资源.微博是一类典型的社交网络服务,它体现了当前社交网络服务的海量性、实时性和个性化等特点.在微博信息传播预测研究中,尚有许多值得深入探索的问题.在本文的最后,基于大量的调研和近几年来的研究经验,提出了该研究中一些值得进一步挖掘的研究点,希望对本领域的其他研究者有所启发.

(1) 动态预测

目前,绝大部分的信息传播预测都是以静态网络拓扑结构、静态的用户行为为基础的,但是在现实中,微博信息传播速度极快,不断有新用户和新信息产生,无论是微博的流行度还是用户之间的关系网络,亦或是用户自身的行为和兴趣,都是随时间动态变化的.如何构建适应微博流行度演变的动态时间序列模型,如何将微博网络动态变化特征添加到信息传播模型中以及运用流数据对用户行为进行建模,都是值得深入挖掘的问题.一成不变的预测模型已经无法适应实时预测的需要.一种可能的方案是应用在线学机器学习(online learning)技术.在线学习的目的是正确预测训练实例的标注,当一次预测完成时,其正确结果便可获得,这一结果可直接用来修正模型.将在线学习应用到微博信息传播预测中,每加入一个新实例,可根据其预测结果与真实结果更新模型假设,从而不断调整预测模型,以适应微博信息传播的多变性,实现实时的动态预测.

(2) 信息竞争

目前已有的信息传播预测研究都是基于信息在传播过程中彼此独立的假设,然而由于微博信息量的爆炸式增长,严重的信息过载会导致信息之间的竞争.因此,在预测微博信息传播时应考虑信息之间的相互影响,比如,同一时间点内是否有其他信息在同样的网络中甚至同样的节点之间传播,当前信息与其他信息在传播中相互的作用是互惠还是互斥,当前信息相对于其他信息是否存在传播优势等因素.

(3) 内容分析

微博内容通常简短且口语化、较碎片化严重,因此,传统文本内容分析技术很难理解信息的主题.所以在分析微博信息的内容特征时,不但需要研究更准确的特征关键字抽取方法,而且要结合该信息的上下文情境,这些上下文包括该条信息发布前后的邻居信息、该信息的评论或是来自于其他用户的相关信息等.

总之,微博信息传播预测不仅兼具科学价值与应用价值,而且是一个充满机遇与挑战的研究方向.随着以微博为代表的社会媒体时代的快速发展,未来必定会吸引更多学者的关注和研究.

致谢 在此,向对本文的研究工作提供帮助的老师和同学表示感谢,特别感谢丁效、赵妍妍、付博、赵森栋等给本文提出的宝贵意见.

References:

[1] Java A, Song XD, Finin T, Tseng B. Why we Twitter: Understanding micro-blogging usage and communities. In: Zhang HZ, *et al*., eds. Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. New York: ACM Press, 2007. 56−65. [doi: 10.1145/1348549.1348556]

[2] Zafarani R, Abbasi MA, Liu H. Social Media Mining: An Introduction. New York: Cambridge University Press, 2014. 197−198.

[3] Zhao D, Rosson MB. How and why people Twitter: The role that micro-blogging plays in informal communication at work. In: Teasley S, Havn E, eds. Proc. of the GROUP 2009. New York: ACM Press, 2009. 243−252. [doi: 10.1145/1531674.1531710]

[4] Huberman BA, Romero DM, Wu F. Social networks that matter: Twitter under the microscope. First Monday, 2009,14(1):1−9.

[5] Gao S, Ma J, Chen ZM. Modeling and predicting retweeting dynamics on microblogging platforms. In: Cheng XQ, Li H, eds. Proc. of the WSDM 2015. New York: ACM Press, 2015. 107−116. [doi: 10.1145/2684822.2685303]

[6] Zaman TR, Herbrich R, Gael JV, Stern D. Predicting information spreading in Twitter. In: Proc. of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds. New York: ACM Press, 2010. 17599−17601.

[7] Feng W, Wang J. Retweet or not? Personalized tweet re-ranking. In: Leonardi S, Panconesi A, eds. Proc. of the WSDM 2013. New York: ACM Press, 2013. 577−586. [doi: 10.1145/2433396.2433470]

[8] Liu T, Ding X, Zhao SD, Duan JW. Prediction technology based on social media. Communications of the CCF, 2015,11(3):26−33. (in Chinese).

[9] Li D, Xu ZM, Li S, Liu T, Wang XW. A survey on information diffusion in social networks. Chinese Journal of Computers, 2014, 37(1):189−206 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2014.00189]

[10] Szabo G, Huberman BA. Predicting the popularity of online content. Communications of the ACM, 2010,53(8):80−88. [doi: 10.1145/1787234.1787254]

[11] Lerman K, Hogg T. Using a model of social dynamics to predict popularity of news. In: Rappa M, Jones P, eds. Proc. of the WWW 2010. New York: ACM Press, 2010. 621−630. [doi: 10.1145/1772690.1772754]

[12] Leskovec J, Mcglohon M, Faloutsos C, Glance N, Hurst M. Patterns of cascading behavior in large blog graphs. In: Proc. of the SDM 2007. Philadelphia: SIAM, 2007. 551−556. [doi: 10.1137/1.9781611972771.60]

[13] Webberley W, Allen S, Whitaker R. Retweeting: A study of message-forwarding in Twitter. In: Proc. of the Workshop on Mobile and Online Social Networks (MOSN 2011). Washington: IEEE Computer Society, 2011. 13−18. [doi: 10.1109/MOSN.2011.6060787]

[14] Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? In: Rappa M, Jones P, eds. Proc. of the WWW 2010. New York: ACM Press, 2010. 591−600. [doi: 10.1145/1772690.1772751]

[15] Honey C, Herring SC. Beyond microblogging: Conversation and collaboration via Twitter. In: Proc. of the 2nd Hawaii Int'l Conf. on System Sciences (HICSS 2009). Washington: IEEE Computer Society, 2009. 1−10. [doi: 10.1109/HICSS.2009.89]

[16] Yu L, Asur S, Huberman BA. What trends in Chinese social media. arXiv:1107.3522, 2011.

[17] Petrovic S, Osborne M, Mccreadie R, Macdonald C, Ounis I, Shrimpton L. Can Twitter replace newswire for breaking news? In: Proc. of the ICWSM 2013. Menlo Park: AAAI Press, 2013. 713−716.

[18] Subasic L, Berendt B. Peddling or creating? Investigating the role of Twitter in news reporting. In: Clough P, *et al*., eds. Proc. of the ECIR 2011. Berlin, Heidelberg: Springer-Verlag, 2011. 207−213. [doi: 10.1007/978-3-642-20161-5_21]

[19] Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan HF, Li XM. Comparing Twitter and traditional media using topic models. In: Clough P, *et al*., eds. Proc. of the ECIR 2011. Berlin, Heidelberg: Springer-Verlag, 2011. 338−349. [doi: 10.1007/978-3-642-20161-5_34]

[20] Starbird K, Palen L. Pass it on?: Retweeting in mass emergency. In: Proc. of the Conf. on Information Systems for Crisis Response and Management. 2010.

[21] Vieweg S, Hughes AL, Starbird K, Palen L. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In: Mynatt E, ed. Proc. of the CHI 2010. New York: ACM Press, 2010. 1079−1088. [doi: 10.1145/1753326.1753486]

[22] Starbird K, Palen L, Hughes AL, Vieweg S. Chatter on the red: What hazards threat reveals about the social life of microblogged information. In: Inkpen K, Gutwin C, eds. Proc. of the CSCW 2010. New York: ACM Press, 2010. 241−250. [doi: 10.1145/1718918.1718965]

[23] Qu Y, Huang C, Zhang PY, Zhang, J. Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake. In: Chairs G, Tang JC, Wang J, eds. Proc. of the CSCW 2011. New York: ACM Press, 2011. 25−34. [doi: 10.1145/1958824.1958830]

[24] Yi CQ, Bao YY, Xue YB, Jiang JC. Research on mechanism of large-scale information dissemination based on sina Weibo. Journal of Frontiers of Computer Science and Technology, 2013,7(6):551−561 (in Chinese with English abstract).

[25] Lerman K, Ghosh R. Information contagion: An empirical study of the spread of news on digg and Twitter social networks. In: Proc. of the ICWSM 2010. Menlo Park: AAAI Press, 2010. 90−97.

[26] Guo ZB, Li ZT, Tu H. Sina microblog: An information-driven online social network. In: Proc. of the 2011 Int'l Conf. on Cyberworlds (CW). Washington: IEEE Computer Society, 2011. 160−167. [doi: 10.1109/CW.2011.12]

[27] He XN, Gao M, Kan M Y, Liu YQ, Sugiyama K. Predicting the popularity of Web 2.0 items based on user comments. In: Chairs G, Trotman A, eds. Proc. of the SIGIR 2014. New York: ACM Press, 2014. 233−242. [doi: 10.1145/2600428.2609558]

[28] Figueiredo F, Benevenuto F, Almeida JM. The tube over time: characterizing popularity growth of youtube videos. In: King I, ed. Proc. of the WSDM 2011. New York: ACM Press, 2011. 745−754. [doi: 10.1145/1935826.1935925]

[29] Kong QC, Mao WJ. Predicting popularity of forum threads based on dynamic evolution. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2767−2776 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4730.htm [doi: 10.13328/j.cnki.jos.004730]

[30] Suh B, Hong LC, Pirolli P, Chi EH. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: Proc. of the Social Computing (SocialCom 2010). Washington: IEEE Computer Society, 2010. 177−184. [doi: 10.1109/SocialCom.2010.33]

[31] Artzi Y, Pantel P, Gamon M. Predicting responses to microblog posts. In: Chu-Carroll J, ed. Proc. of the NAACL 2012. Stroudsburg: Association for Computational Linguistics, 2012. 602−606.

[32] Kupavskii A, Umnov A, Gusev G, Serdyukov P. Predicting the audience size of a Tweet. In: Proc. of the ICWSM 2013. Menlo Park: AAAI Press, 2013. 693−696.

[33] Ma HX, Qian WN, Xia F, He XF, Xu J, Zhou AY. Towards modeling popularity of microblogs. Frontiers of Computer Science, 2013,7(2):171−184. [doi: 10.1007/s11704-013-3901-9]

[34] Yang J, Counts S. Predicting the speed, scale, and range of information diffusion in Twitter. In: Proc. of the ICWSM 2010. Menlo Park: AAAI Press, 2010. 355−358.

[35] Kong S, Feng L, Sun G, Luo K. Predicting lifespans of popular tweets in microblog. In: Hersh W, ed. Proc. of the SIGIR 2012. New York: ACM Press, 2012. 1129−1130. [doi: 10.1145/2348283.2348503]

[36] Bhattacharya D, Ram S. Sharing news articles using 140 characters: A diffusion analysis on Twitter. In: Proc. of the ASONAM 2012. Washington: IEEE Computer Society, 2012. 966−971. [doi: 10.1109/ASONAM.2012.170]

[37] Liu DC, Chen X. Rumor propagation in online social networks like Twitter—A simulation study. In: Proc. of the 3rd Int'l Conf. on Multimedia Information Networking and Security (MINES). Washington: IEEE Computer Society, 2011. 278−282. [doi: 10.1109/MINES.2011.109]

[38] Xu XD, Xiao YT, Zhu SR. Simulation investigation of rumor propagation in microblogging community. Computer Engineering, 2011,37(10):272−274 (in Chinese with English abstract).

[39] Wang H, Li YP, Feng ZN, Feng L. ReTweeting analysis and prediction in microblogs: An epidemic inspired approach. China Communications, 2013,10(3):13−24. [doi: 10.1109/CC.2013.6488827]

[40] Xiong F, Liu Y, Zhang ZJ, Zhu J, Zhang Y. An information diffusion model based on retweeting mechanism for online social media. Physics Letters A, 2012,376(30-31):2103−2108. [doi: 10.1016/j.physleta.2012.05.021]

[41] Wu M, Guo J, Zhang C, Xie JJ. Social media communication model research bases on Sina-Weibo. In: Proc. of the 6th Int'l Conf. on Intelligent Systems and Knowledge Engineering. Berlin, Heidelberg: Springer-Verlag, 2011. 445−454. [doi: 10.1007/978-3-642-25661-5_57]

[42] Yang J, Leskovec J. Modeling information diffusion in implicit networks. In: Proc. of the ICDM 2010. Washington: IEEE Computer Society, 2010. 599−608. [doi: 10.1109/ICDM.2010.22]

[43] Hong LJ, Dan O, Davison BD. Predicting popular messages in Twitter. In: Sadagopan S, Ramamritham K, Kumar A, Ravindra MP, eds. Proc. of the WWW 2011 Companion. New York: ACM Press, 2011. 57−58. [doi: 10.1145/1963192.1963222]

[44] Gao S, Ma J, Chen ZM. Popularity prediction in microblogging network. In: Chen L, *et al*., eds. Proc. of the 16th Asia-Pacific Web Conf. Berlin, Heidelberg: Springer-Verlag, 2014. 379−390. [doi: 10.1007/978-3-319-11116-2_33]

[45] Kupavskii A, Ostroumova L, Umnov A, Usachev S, Serdyukov P, Gusev G, Kustarev A. Prediction of retweet cascade size over time. In: Chen XW, ed. Proc. of the CIKM 2012. New York: ACM Press, 2012. 2335−2338. [doi: 10.1145/2396761.2398634]

[46] Bakshy E, Hofman JM, Mason WA, WattsDJ. Everyone's an influencer: Quantifying influence on Twitter. In: King I, ed. Proc. of the WSDM 2011. New York: ACM Press, 2011. 65−74. [doi: 10.1145/1935826.1935845]

[47] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks. In: Li Y, ed. Proc. of the SIGKDD 2008. New York: ACM Press, 2018. 7−15. [doi: 10.1145/1401890.1401897]

[48] Zhang S, Xu K, Li HT. Measurement and analysis of information propagation in online social networks like microblog. Journal of Xi'an Jiaotong University, 2013,47(2):124−130 (in Chinese with English abstract). [doi: 10.7652/xjtuxb201302021]

[49] Mao JX, Liu YQ, Zhang M, Ma SP. Social influence analysis for micro-blog user based on user behavior. Chinese Journal of Computers, 2014,37(4):791−798 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2014.00791]

[50] Gao S, Ma J, Chen ZM. Effective and effortless features for popularity prediction in microblogging network. In: Chung CW, ed. Proc. of the WWW Companion 2014. New York: ACM Press, 2014. 269−270. [doi: 10.1145/2567948.2577312]

[51] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. Technical Report, No.1999-66, Stanford InfoLab, 1999.

[52] Weng JS, Lim EP, Jiang J, He Q. Twitterrank: Finding topic-sensitive influential twitterers. In: Davison BD, Suel T, eds. Proc. of the WSDM 2010. New York: ACM Press, 2010. 261−270. [doi: 10.1145/1718487.1718520]

[53] Wu S, Tan C, Kleinberg JM, Macy MW. Does bad news go away faster? In: Proc. of the ICWSM 2011. Menlo Park: AAAI Press, 2011.

[54] Ma ZY, Sun AX, Cong G. On predicting the popularity of newly emerging hashtags in Twitter. Journal of the American Society for Information Science and Technology, 2013,64(7):1399−1410. [doi: 10.1002/asi.22844]

[55] Tsur O, Rappoport A. What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. In: Adar E, Teevan J, eds. Proc. of the WSDM 2012. New York: ACM Press, 2012. 643−652. [doi: 10.1145/2124295.2124320]

[56] Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral tweets. In: Schwabe D, Almeida V, Glaser H, eds. Proc. of the WWW 2013 Companion. New York: ACM Press, 2013. 657−664.

[57] Tan CH, Lee L, Pang B. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In: Marcu D, ed. Proc. of the ACL 2014. Baltimore: Association for Computational Linguistics, 2014. 175−185.

[58] Zhao X, Zhu FD, Qian WN, Zhou AY. Impact of multimedia in Sina Weibo: Popularity and life span. In: Proc. of the Semantic Web and Web Science. Springer-Verlag, 2013. 55−65. [doi: 10.1007/978-1-4614-6880-6_5]

[59] Can EF, Oktay H, Manmatha R. Predicting retweet count using visual cues. In: He Q, Iyengar A, eds. Proc. of the CIKM 2013. New York: ACM Press, 2013. 1481−1484. [doi: 10.1145/2505515.2507824]

[60] Bae Y, Ryu P, Kim H. Predicting the lifespan and retweet times of tweets based on multiple feature analysis. ETRI Journal, 2014, 36(3):418−428. [doi: 10.4218/etrij.14.0113.0657]

[61] Liu GN, Fu YJ, Xu T, Xiong H, Chen GQ. Discovering temporal retweeting patterns for social media marketing campaigns. In: Fan JP, Pei J, eds. Proc. of the ICDM 2014. Washington: IEEE Computer Society, 2014. 905−910. [doi: 10.1109/ICDM.2014.48]

[62] Bao P, Shen HW, Huang JM, Cheng XQ. Popularity prediction in microblogging network: A case study on Sina Weibo. In: Schwabe D, Almeida V, Glaser H, eds. Proc. of the WWW 2013 Companion. New York: ACM Press, 2013. 177−178.

[63] Zaman T, Fox EB, Bradlow ET. A Bayesian approach for predicting the popularity of tweets. arXiv:1304.6777v3, 2014. [doi: 10.1214/14-AOAS741]

[64] Lu R, Yang Q. Trend analysis of news topics on Twitter. In: Proc. of the ICWSM 2012. Menlo Park: AAAI Press, 2012. 327−332.

[65] Xu Z, Yang Q. Analyzing user retweet behavior on Twitter. In: Proc. of the ASONAM 2012. Washington: IEEE Computer Society, 2012. 46−50. [doi: 10.1109/ASONAM.2012.18]

[66] Xu ZH, Zhang Y, Wu Y, Yang Q. Modeling user posting behavior on social media. In: Hersh W, ed. Proc. of the SIGIR 2012. New York: ACM Press, 2012. 545−554. [doi: 10.1145/2348283.2348358]

[67] Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In: Proc. of the 3rd Hawaii Int'l Conf. on System Sciences (HICSS 2010). Washington: IEEE Computer Society, 2010. 1−10. [doi: 10.1109/HICSS.2010.412]

[68] Chen KL, Chen TQ, Zheng GQ, Ou J, Yao EP, Yu Y. Collaborative personalized tweet recommendation. In: Hersh W, ed. Proc. of the SIGIR 2012. New York: ACM Press, 2012. 661−670. [doi: 10.1145/2348283.2348372]

[69] Pan Y, Cong F, Chen K, Yu Y. Diffusion-Aware personalized social update recommendation. In: Yang Q, King I, Li Q, eds. Proc. of the 7th ACM Conf. on Recommender Systems. New York: ACM Press, 2013. 69−76. [doi: 10.1145/2507157.2507177]

[70] Hong LJ, Doumith AS, Davison BD. Co-Factorization machines: Modeling user interests and predicting individual decisions in Twitter. In: Leonardi S, Panconesi A, eds. Proc. of the WSDM 2013. New York: ACM Press, 2013. 557−566. [doi: 10.1145/2433396.2433467]

[71] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[72] Zhang Q, Gong Y, Guo Y, Huang XJ. Retweet behavior prediction using hierarchical dirichlet process. In: Proc. of the AAAI 2015. Menlo Park: AAAI Press, 2015.

[73] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical dirichlet processes. Journal of the American Statistical Association, 2006, 101(476):1566−1581. [doi: 10.1198/016214506000000302]

[74] Zhang J, Liu B, Tang J, Chen T, Li JZ. Social influence locality for modeling retweeting behaviors. In: Rossi F, ed. Proc. of the IJCAI 2013. Menlo Park: AAAI Press, 2013. 2761−2767.

[75] Yang Z. Predictive models in social network analysis [MS. Thesis]. Beijing: Tsinghua University, 2011 (in Chinese with English abstract).

[76] Tang J, Sun J, Wang C, Yang Z. Social influence analysis in large-scale networks. In: Elder J, Fogelman FS, eds. Proc. of the SIGKDD 2009. New York: ACM Press, 2009. 807−816. [doi: 10.1145/1557019.1557108]

[77] Tan C, Tang J, Sun J, Lin Q, Wang FJ. Social action tracking via noise tolerant time-varying factor graphs. In: Rao B, Krishnapuram B, eds. Proc. of the SIGKDD 2010. New York: ACM Press, 2010. 1049−1058. [doi: 10.1145/1835804.1835936]

[78] Bian J, Yang Y, Chua TS. Predicting trending messages and diffusion participants in microblogging network. In: Geva S, Trotman A, eds. Proc. of the SIGIR 2014. New York: ACM Press, 2014. 537−546. [doi: 10.1145/2600428.2609616]

[79] Hoang TA, Lim EP. Retweeting: An act of viral users, susceptible users, or viral topics? In: Proc. of the SDM 2013. Philadelphia: SIAM, 2013. 569−577.

[80] Song GH, Li ZT, Tu H. Forward or ignore: User behavior analysis and prediction on microblogging. In: Proc. of the Advanced Research in Applied Artificial Intelligence. Berlin, Heidelberg: Springer-Verlag, 2012. 231−241. [doi: 10.1007/978-3-642-31087-4_25]

[81] Luo ZC, Osborne M, Tang JT, Wang T. Who will retweet me? Finding retweeters in Twitter. In: Jones G, Sheridan P, eds. Proc. of the SIGIR 2013. New York: ACM Press, 2013. 869−872. [doi: 10.1145/2484028.2484158]

[82] Li H. Learning to rank for information retrieval and natural language processing. Synthesis Lectures on Human Language Technologies, 2011,4(1):1−113. [doi: 10.2200/S00348ED1V01Y201104HLT012]

[83] Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. In: Proc. of the SIGKDD 2003. New York: ACM Press, 2003. 137−146. [doi: 10.1145/956750.956769]

[84] Young HP. The diffusion of innovations in social networks. In: Proc. of the Economy as an Evolving Complex System III: Current Perspectives and Future Directions. Oxford University Press, 2006. 267−282.

[85] Zhu X, Jia Y, Nie YP, Qu M. Event propagation analysis on microblog. Journal of Computer Research and Development, 2015, 52(2):437−444 (in Chinese with English abstract).

[86] Dickens L, Molloy I, Lobo J, Cheng PC, Russo A. Learning stochastic models of information flow. In: Proc. of the ICDE 2012. Washington: IEEE Computer Society, 2012. 570−581. [doi: 10.1109/ICDE.2012.103]

[87] Guille A, Hacid H. A predictive model for the temporal dynamics of information diffusion in online social networks. In: Mille A, Gandon F, Misselis J, eds. Proc. of the WWW 2012 Companion. New York: ACM Press, 2012. 1145−1152. [doi: 10.1145/2187980.2188254]

[88] Galuba W, Aberer K, Chakraborty D, Despotovic Z, Kellerer W. Outtweeting the twitterers-predicting information cascades in microblogs. In: Proc. of the 3rd Wonference on Online Social Networks. Berkeley: USENIX Association, 2010.

[89] Zheng L. The research on information diffusion modeling over social networks [MS. Thesis]. Shanghai: Shanghai Jiaotong University, 2011 (in Chinese with English abstract).

[90] Cao JX, Wu JL, Shi W, Liu B, Zheng X, Luo JZ. Sina Weibo information diffusion analysis and prediction. Chinese Journal of Computers, 2014,37(4):779−790 (in Chinese with English abstract).

[91] Yang Z, Guo JY, Cai KK, Tang JZ, Li J, Zhang L, Su Z. Understanding retweeting behaviors in social networks. In: Huang J, ed. Proc. of the CIKM 2010. New York: ACM Press, 2010. 1633−1636. [doi: 10.1145/1871437.1871691]

[92] Montanari A, Saberi A. The spread of innovations in social networks. Proc. of the National Academy of Sciences, 2010,107(47): 20196−20201. [doi: 10.1073/pnas.1004098107]

[93] Yang J, Leskovec J. Patterns of temporal variation in online media. In: King I, ed. Proc. of the WSDM 2011. New York: ACM Press, 2011. 177−186. [doi: 10.1145/1935826.1935863]

附中文参考文献:

[8] 刘挺,丁效,赵森栋,段俊文.基于社会媒体的预测技术.中国计算机学会通讯,2015,11(3):26−33.

[9] 李栋,徐志明,李生,刘挺,王秀文.在线社会网络中信息扩散.计算机学报,2014,37(1):189−206. [doi: 10.3724/SP.J.1016.2014.00189]

[24] 易成岐,鲍媛媛,薛一波,姜京池.新浪微博的大规模信息传播规律研究.计算机科学与探索,2013,7(6):551−561.

[29] 孔庆超,毛文吉.基于动态演化的讨论帖流行度预测.软件学报,2014,25(12):2767−2776. http://www.jos.org.cn/1000-9825/4730.htm [doi: 10.13328/j.cnki.jos.004730]

[38] 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.

[48] 张赛,徐恪,李海涛.微博类社交网络中信息传播的测量与分析.西安交通大学学报,2013,47(2):124−130. [doi: 10.7652/xjtuxb201302021]

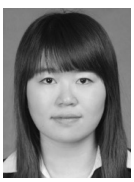
[49] 毛佳昕,刘奕群,张敏,马少平.基于用户行为的微博用户社会影响力分析.计算机学报,2014,37(4):791−798. [doi: 10.3724/SP.J.1016.2014.00791]

[75] 杨子.社会网络分析中的预测模型[硕士学位论文].北京:清华大学,2011.

[85] 朱湘,贾焰,聂原平,曲铭.基于微博的事件传播分析.计算机研究与发展,2015,52(2):437−444.

[89] 郑蕾.面向社会网络的信息传播模型研究[硕士学位论文].上海:上海交通大学,2011.

[90] 曹玖新,吴江林,石伟,刘波,郑啸,罗军舟.新浪微博网信息传播分析与预测.计算机学报,2014,37(4):779−790.



李洋(1987－),女,黑龙江哈尔滨人,博士生,主要研究领域为社会计算,信息检索,自然语言处理.

刘挺(1972－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为社会计算,信息检索,自然语言处理.

陈毅恒(1979－),男,博士,讲师,CCF 专业会员,主要研究领域为社会计算,信息检索,自然语言处理.