软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(6):1473-1485 [doi: 10.13328/j.cnki.jos.004627]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

微博消息传播中意见领袖影响力建模研究*

王晨旭[1], 管晓宏[1,2], 秦 涛[1], 周亚东[1]

[1](西安交通大学 智能网络与网络安全教育部重点实验室,陕西 西安 710049)

[2](清华大学 自动化系 智能与网络化系统研究中心,北京 100084)

通讯作者: 秦涛, E-mail: tqin@sei.xjtu.edu.cn, http://gr.xjtu.edu.cn/web/qin.tao/home

摘 要: 在微博网络中,消息的传播与意见领袖的影响力密切相关.然而,意见领袖在消息动态传播过程中所表现出的影响力难以量化衡量,这对意见领袖影响力的评估和消息传播趋势的预测带来了巨大挑战.针对这一问题,提出一种基于消息传播的微博意见领袖影响力建模与测量分析方法.在分析消息传播模式的基础上,采用动态有向图描述消息在微博网络中的传播过程;发现该过程可近似分解为由各个意见领袖所驱动的子过程,根据对意见领袖影响力属性特征的分析发现,该子过程可以由指数截断的幂律衰减函数来描述.对模型中各个参数进行估计,可以定量地衡量意见领袖在消息传播过程中的初始影响力、影响力衰减指数及其影响力持续时间等指标.结合新浪微博数据的分析结果显示:消息的传播范围与传播过程中参与传播的意见领袖的数量呈弱相关;虽然意见领袖的初始影响力与其粉丝数量的大小正相关,但影响力衰减指数的大小以及影响力持续时间的长短与粉丝数量几乎无关.最后,采用所建模型对真实微博消息的传播趋势进行预测,结果表明,所提模型能够较好地对热门消息的传播趋势进行预测,这对微博中公众舆论的控制及广告定点投放具有重要意义.

关键词: 微博;意见领袖;影响力;消息传播

中图法分类号: TP311

中文引用格式: 王晨旭,管晓宏,秦涛,周亚东.微博消息传播中意见领袖影响力建模研究.软件学报,2015,26(6):1473-1485. http://www.jos.org.cn/1000-9825/4627.htm

英文引用格式: Wang CX, Guan XH, Qin T, Zhou YD. Modeling on opinion leader's influence in microblog message propagation and its application. Ruan Jian Xue Bao/Journal of Software, 2015,26(6):1473-1485 (in Chinese). http://www.jos.org.cn/1000-9825/4627.htm

Modeling on Opinion Leader's Influence in Microblog Message Propagation and Its Application

WANG Chen-Xu[1], GUAN Xiao-Hong[1,2], QIN Tao[1], ZHOU Ya-Dong[1]

[1](MOE Key Laboratory for Intelligent and Network Security, Xi'an Jiaotong University, Xi'an 710049, China)

[2](Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University, Beijing 100084, China)

Abstract: In microblog networks, the propagation of messages is closely related to the influence of opinion leaders. However, it is hard to quantitate the influence of opinion leaders in different types of messages, which bring great challenges to evaluate the influence of opinion leaders and predict the propagation trend of a message. To solve this problem, this paper proposes a method to measure and analyze opinion leaders' influence in microblog based on the dynamical process of message propagation. By studying the spread patterns of messages, dynamical directed graph is employed to model the propagation process. The propagation of a message can be decomposed into sub-propagations raised by opinion leaders, which can be described as exponential truncated power-law decay function. By estimating the parameters in the model, the initial influence, influence decay rate and influence insistency of opinion leaders can be evaluated

* 基金项目: 国家自然科学基金(61221063, 61202392, 61103240); 国家科技支撑计划(2011BAK08B02); 苏州应用基础研究发展计划(SYG201227); 中央高校基本科研业务费专项资金; 教育部高等学校博士学科点专项科研基金(20120201120023)

收稿时间: 2013-10-28; 修改时间: 2014-02-17; 定稿时间: 2014-05-21

quantitatively. Results of the experiment based on the data collected from Sina microblog, show the total retweeted messages is weakly correlated to the number of opinion leaders in the spread process. In addition, the number of followers possessed by opinion leaders is positively correlated to the power of their initial influence. However, it exhibits no correlation with the decay rate and insistence time of the influence. The efficiency and correctness of the proposed model are validated by predicting the spread trends of hot messages in actual microblog networks, which is important to network marketing and message propagation controls.

Key words: microblog; opinion leader; influence; message propagation

微博客作为一种新兴的社交媒体,以其内容短小、信息传播速度快而有别于传统博客,自其创建以来便吸引了大量互联网用户创造并传播各式各样的信息.随着微博系统的发展,其对网络营销、社会舆论的极大影响力正受到越来越多学者的关注,而微博意见领袖影响力的建模与分析更是研究的前沿问题之一.

1 基于消息传播的微博意见领袖影响力分析方法

微博系统允许用户以转发的形式将其感兴趣的消息分享给其粉丝,一旦该消息与其粉丝产生共鸣,这些粉丝可再次转发该消息从而分享给更多的用户.研究表明:在 Twitter 中有 31%的微博为转发微博[1],而在新浪微博中转发比例则高达 47.8%[2],转发行为已成为微博网络中信息通过网络邻居节点进行传播的重要形式,也是用户分享和获取信息的重要方式.然而在微博活动中,用户对信息的订阅和传播具有很强的自主性和选择性,往往根据个人的兴趣爱好并结合信息的内容、类别与质量来决定是否继续传播一条消息.此外,微博中的绝大多数用户属于消极的信息消费者[3],他们倾向于浏览来自其他用户的消息却很少发布或转发消息.能否调动这些消极用户分享信息的积极性,便成为衡量用户影响力大小的关键因素.

目前,针对微博用户影响力的研究多集中在对网络节点的入度、转发数量以及被他人提及次数的分析上[4-6].基于 Twitter 的研究表明:在信息传播过程中,用户影响力与其粉丝数量呈弱相关[1,4,7].因此,粉丝数量大是用户具有影响力的必要条件,而非充分条件.这些研究属于对影响力的静态分析,而对意见领袖在信息传播过程中是如何动态地影响其粉丝等方面的研究还鲜有所见.事实上,用户的影响力与其身份地位以及发布的内容高度相关,用户发布的信息质量越高、吸引力与新闻性越强、用户在该领域内的权威性越高,对该消息感兴趣的人数就会越多,就能够被更多的用户转播.因此,在衡量用户的影响力时需要更加强调用户影响其粉丝传播信息的能力,而非仅仅注重其将信息传播给其受众的能力[3].针对上述问题,提出基于消息传播的微博意见领袖影响力研究方法,定量地分析意见领袖在热门消息传播中的影响力.具体步骤如下:

(1) 分析消息的传播模式.根据消息的转发时间流计算消息在每时刻的传播速度,将消息的传播速度以柱状图的形式呈现出来,从整体上把握消息的传播态势,直观地呈现消息传播趋势的变化,并对所表现出的传播模式进行分析;

(2) 构建消息传播过程的级联模型.根据用户在微博网络中的关注关系以及用户转发微博的时间戳,将消息的传播过程模型化为动态有向图,还原消息传播过程中用户之间的影响关系;

(3) 挖掘和提取消息传播过程中的意见领袖.在构建动态有向图的基础上,挖掘并提取影响力较大的用户作为意见领袖;

(4) 建模并量化分析意见领袖的影响力.对消息传播过程中意见领袖所表现出的影响力进行建模分析,评估意见领袖影响力各个特征指标的大小,定量地分析意见领袖影响力的大小.

2 相关研究

信息传播学是研究社会信息系统及其运行规律的学科,其研究重点是人与人之间的信息传播过程,进而达到控制信息传播速度与效果的目的.早期的信息传播研究主要集中在书籍、报刊、杂志以及电视广播等大众信息媒介上.Lippman 首次对公众舆论做了全景式的描述,为信息传播学的发展奠定了坚实基础[8].随后,Lazarsfield 提出了两级传播理论和意见领袖的概念,推动了信息传播学的发展[9].随着互联网的发展以及 Web 2.0 技术的广泛应用,在线社交媒体已成为人们日常生活中获取信息的重要方式,吸引了众多学者对其中的信息

传播规律进行研究.文献[1]研究了热门话题在 Twitter 网络中的传播,发现用户活跃度和粉丝数量对热门话题的传播与形成没有必然联系,相反,用户之间对传播内容所达成的共鸣会导致话题的快速传播.Wu 等人在文献[5]中证实了信息在微博网络传播过程中两级传播理论的存在.文献[6]研究了 Twitter 网络中用户社团关系的形成以及社团之间信息的分享.文献[10]研究了新浪微博中热门话题的内容,发现新浪用户与 Twitter 用户在对信息内容的喜好上有着很大的差别.文献[11]研究了图片在照片分享网站 Flickr 中的传播模式,发现图片的传播过程可以分为 3 个阶段:线性增长期、激增期和衰亡期.

信息在网络中传播时,意见领袖所表现出的影响力也是广大学者关注的焦点.Lazarsfield[9]将意见领袖定义为在人际传播网络中经常为他人提供信息、同时对他人施加影响的活跃分子.在社会网络中,影响力的定义在不同的信息传播环境下各不相同,对意见领袖的定义也略有不同.早期对互联网中意见领袖的研究集中于在线论坛[12,13],文献[12]分析了意见领袖发帖与普通用户发帖行为存在的差异,文献[13]提出了一种影响力扩散模型并将其用于在线网络论坛中意见领袖的发现.随着在线社交媒体的兴起和广泛使用,对在线社交媒体中意见领袖影响力的研究逐渐成为学者们研究的热点.文献[3]通过对 Twitter 中大量消息传播过程的研究发现:绝大多数微博用户属于消极的信息消费者,他们仅仅从关注的好友那里获取信息却很少与其粉丝分享信息,因此,作者强调:在研究微博网络中用户的影响力时,应当同时考虑用户的影响力以及该用户转发消息的积极性.文献[4,7]分别使用粉丝数量和微博转发数量来衡量用户的影响力,结果表明:粉丝数量与转发评论的数量是弱相关的,且用户的影响力在不同的衡量指标下会随时间和主题的不同而发生变化.文献[14]通过对新浪微博的研究发现,微博系统具有很强的名人效应.文献[15]从用户影响力和用户活跃度两个角度构建了微博意见领袖指标影响力体系,使用粗糙集理论对意见领袖识别问题进行了建模,发现只有很少用户可以在不同主题下同时成为意见领袖.

综上所述,国内外学者在意见领袖挖掘与影响力建模方面做了较多的研究工作,然而微博消息的传播是由用户的转发行为所驱动的动态过程,单独的分析粉丝数量等静态参数并不能准确地度量意见领袖在消息传播过程中影响力的大小.针对此问题,通过消息传播的动态转发过程对意见领袖的影响力进行建模,可以定量地分析意见领袖影响力各个特征属性的大小,从而能够准确地评估意见领袖影响力的大小,并能够对消息的传播趋势做出预测.

3 微博消息传播规律分析

3.1 微博消息传播模式分析

为了研究信息在社交网络中的传播模式,Cha 等人将 Flickr 上照片粉丝数的增长过程分为线性增长期、激增期和衰亡期,并发现这 3 种时期可以较好地刻画图片在 Flickr 网络中的传播过程[11].张赛等人在研究新浪微博中的信息传播时,将热门微博的转发生命周期分为潜伏期、激增期和衰亡期来研究微博的转发增长模式[14],并将所得结果与 Cha 等人所做工作进行对比,发现微博转发的增长模式与 Flickr 上的照片热度的增长模式有明显的不同:Flickr 中,热度处于线性增长期的照片占很大比例;而微博转发过程的增长模式多属于激增期和潜伏期,而很少处于线性增长期.事实上,微博系统中的消息传播具有很强的时效性和话题性,时效性是指最新最快的消息更容易引起用户的关注,但随着时间的流逝,消息传播的速度会快速衰减;话题性是指公众关心的话题更容易引起大多数用户的关注和转发,有时甚至可以克服时效性形成二次传播.这些因素导致消息在传播过程中会呈现出极大的波动性和突发性.

根据消息的转发时间流,将消息的传播过程定义为时间序列 $R_m=\{n_t|t=1,\ldots,T\}$,其中,n_t 为在 t 时刻单位时间内(以分钟为单位)参与转发消息 m 的用户数量,T 为观测时间(T=4320min).以时间 t 为横坐标、n_t 为纵坐标画二维柱状图,便可得到消息 m 随时间的传播趋势图.图 1 为一条具有代表性消息的传播过程,从图中可以清楚地看出消息在各个时间段内传播的快慢以及典型的激增式传播模式.消息的传播表现出了较强的时效性特征,消息的激增式传播集中在消息发布的第 1 天.

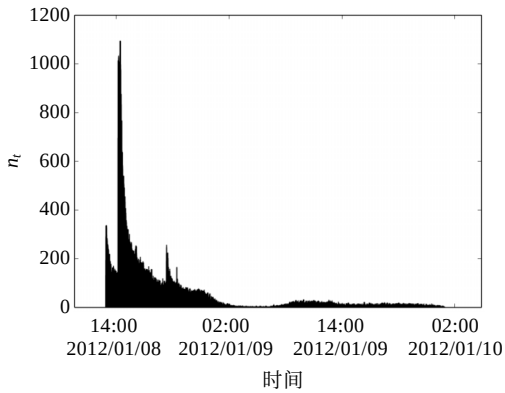

Fig.1 Propagation pattern of a typical message in microblog

图 1 微博中一条典型消息的传播模式

3.2 消息传播的级联模型

随着微博、Facebook 等在线社交媒体的兴起,消息的发布不再局限于大众传播媒介,广大普通用户逐步成为信息的源头.消息的传播过程往往是多级的,需要经过各种意见领袖的中介作用才能到达一般受众[16,17]. Wu 等人验证了在 Twitter 中信息的传播过程中,两级传播理论(two-step flow theory)的存在[5],发现在绝大多数普通用户通过名人用户的转发来获取URL信息,名人用户在URL传播过程中起着信息中继的作用.这为使用传播流模型来刻画微博网络中消息的传播提供了重要依据,从而实现对消息传播过程的重现,还原消息传播过程中用户之间的影响关系.

消息的传播流采用动态有向图 $G(V,A)$表示消息的传播,其中:V 为节点的集合,其元素为所有参与消息转发的用户;A 为带方向的边的集合,代表信息流在网络中的流向.用户转发消息的时间 t 作为节点的一个属性被保留在有向图中,从而允许研究消息传播的动态过程.集合 A 中的边根据用户之间的关注关系以及其转发消息的时间进行构建,构建方式是:用 t_*表示用户转发消息的时间,如果用户 v 关注了用户 u 且 $t_u<t_v$(用户 u 在 v 之前转发了该消息),那么就定义一条从 u 指向 v 的边,转发时间差 t_v-t_u 则代表了用户 v 对用户 u 的响应时间,并作为边的一个属性被保留在有向图中,从该属性可以判断用户对其好友响应速度的快慢.从消息传播流的构建过程可知,用户可能受到来自多个好友的影响.Bakshy 等人在研究 URL 在 Twitter 网络中的传播过程时,将用户的影响力定义为 3 种类型:最先影响力(first-influence)、最后影响力(last-influence)和均分影响力(split- influence)[18],其中,最先影响力指在所有影响某用户的好友中,最先发布该条信息的好友对其产生的影响力最大;最后影响力指在所有影响某用户的好友中,最后发布该条消息的好友对其影响力最大;均分影响力指影响某用户的所有好友均分 1 单位的影响力.根据这 3 种影响力模型的定义以及传播流 $G(V,A)$中边的响应时间,很容易构建出这 3 种影响力下的消息流模型.由于最先影响力模型容易过度估计活跃用户的影响力,且与人们对用户在消息转发时的直观认识相违背;而均分影响力模型则过于复杂,不易定量地区分用户影响力的大小;最后影响力模型认为用户在转发一条消息时仅受一个好友的影响,且用户对该好友的响应速度最快,这与人们对微博系统中好友之间进行消息转发的直观认识一致,因此,仅采用最后影响力模型对消息传播进行分析.该模型的传播流构造规则为:对 $G(V,A)$中有多条入边的节点,仅仅保留响应时间最快的边而将其他边去除,如果有多条边的响应时间相同且均最小,则随机保留其中一条.

3.3 意见领袖的挖掘

根据消息传播流最后影响力模型的构建,意见领袖定义为在消息传播过程中能够引起其粉丝大量转发的用户,其在 $G(V,A)$中表现为那些具有较大出度的节点.为了定量地识别消息传播流中的意见领袖,将消息 m 在传播过程中意见领袖的集合定义为

$$O_m=\{u|(u\in V,||\{e_{uv}|e_{uv}\in A\}||>\theta)\} \tag{1}$$

其中,$\{e_{uv}|e_{uv}\in A\}$表示节点 u 在 $G(V,A)$中的出边的集合,阈值θ的大小根据经验选取.一旦消息传播流中的意见领袖被获取,将这些意见领袖转发该条消息的时间作为该子传播过程的起始时刻,根据消息传播序列的定义,以意见领袖为中心所引发的消息传播可以表示为 $R_m^{o_i}=\{n_t\,|\,t=t_{o_i},t_{o_i}+1,...,T^{o_i}\}$.原始的消息传播序列 $R_m=\{n_t|t=1,\ldots,T\}$便可以近似分解为由这些意见领袖所驱动的消息传播之和:

$$R_m\approx\sum_{o_i\in O_m}R_m^{o_i} \tag{2}$$

图 2 为图 1 中所示消息按前 5 位意见领袖进行分解所得的消息传播图,其中,LD 为意见领袖的缩写,横坐标为时间实际 t.将消息传播模式的分析应用到传播序列 $R_m^{o_i}$ 中,便可以得到意见领袖在消息传播过程中所引起的消息传播的基本模式.由图 2 可知:以意见领袖为中心的传播均会在短时间内引起消息的大量转发,且传播速度随之急速下降,直至达到慢速传播状态.分析结果表明,图 1 中的激增式传播均是由这些意见领袖对该消息的转发所引发的.这印证了意见领袖在微博消息传播过程中作为传播引爆点的作用,证实了微博类社交媒体中两级传播模型的存在.结合新浪微博数据的实验分析结果显示,由意见领袖引发的消息传播绝大多数具有相同的模式,即:绝大多数意见领袖的转发能够立即引起其粉丝的大量转发,然后很快进入急速下降状态,直至到达慢速传播状态.该模式为分析意见领袖影响力的各个属性特征提供了可能.

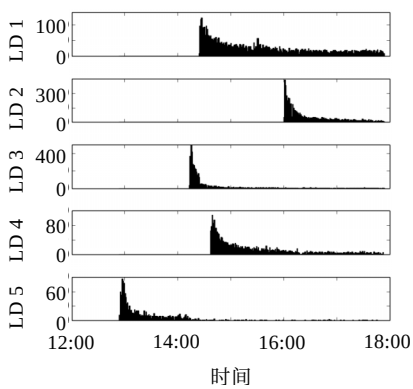

Fig.2 Message propagation patterns driven by opinion leaders

图 2 意见领袖引发的消息传播模式

4 意见领袖影响力特征分析与建模

首先假设消息在传播过程中不受消息新鲜度以及用户影响力衰减的影响,消息在传播过程中,单位时间内引发其粉丝进行转发的比例为常数,设为α;假设在消息传播期间意见领袖的粉丝数为 N 且固定不变,在 t 时刻所引起转发数的累积函数为 $y(t)$,那么单位时间内所引发的转发数为 $y'(t)$,最后假设已经转发了该消息的粉丝不会再次转发,则在 t 时刻有:

$$y'(t)=\alpha\cdot(N-y(t)) \tag{3}$$

解微分方程可得:

$$y'(t)\propto e^{-\alpha t} \tag{4}$$

在此假设条件下,意见领袖在单位时间内所引起的转发数随时间呈指数衰减.然而在消息传播初期,消息的新鲜度很高,意见领袖对其粉丝的影响力比较大,而随着时间的推移,这两者都会有所衰减,因此在单位时间内,引发其粉丝进行转发的比例将随时间变化,设为$\alpha(t)$.然而,至今仍没有一个明确的消息新鲜度和影响力随时间的衰减函数,即使能够得到$\alpha(t)$的解析表达式,求解微分方程(3)也是困难的.事实上,在意见领袖转发一条消息的初期,其粉丝看到该条消息并转发的行为可以看作是粉丝对意见领袖转发行为的反馈或响应,此时,消息传播所

形成的传播流中,边的响应时间便构成了集体层次上粉丝对意见领袖的响应时间(waiting-time).而 Radicchi 研究发现,人类在 Web 行为中集体层次上的等待时间在统计意义上服从幂律分布[19],即,单位时间内意见领袖所引起的转发数在消息传播初期为幂律衰减函数:

$$y'(t)\propto t^{-\gamma} \tag{5}$$

Yan 等人对新浪微博中用户发微博时间间隔的研究显示:用户在微博系统中的活动受兴趣和网络社会身份共同驱使,且用户发微博的时间间隔同样满足幂律分布函数[20],这与实际中所观察到的以意见领袖为中心的前期消息传播模式一致.当消息传播一段时间后,消息的新鲜度和用户的影响力衰减到一定程度时,它们可看作常数而不再变化,此时,以意见领袖为中心的消息传播便服从指数衰减.

综合上述分析,意见领袖所引发的消息传播过程可以用指数截断的幂律衰减函数来表示:

$$y'(t) = N_0 t^{-\gamma} \mathrm{e}^{-\frac{t}{\tau}} \tag{6}$$

其中,N_0,γ和τ为待估参数.在意见领袖转发初期,即,当 $t<<\tau$时,$\mathrm{e}^{-t/\tau}\approx 1$,此时消息传播的幂律衰减占主导地位,传播受指数衰减的影响很小;而当 $t>>\tau$时,由于指数衰减速度大于幂律衰减的速度,此时消息的传播由指数衰减所主导,消息的衰亡速度加快.因此,参数τ的大小决定了由意见领袖所引发的消息传播的消亡时间.事实上,对同一消息的不同意见领袖来说,N_0 可以看作意见领袖的初始影响力,N_0 越大,意见领袖的初始影响力越大;γ可以代表意见领袖影响力的衰减速度,γ越大,意见领袖影响力衰减就越快;而τ则代表意见领袖影响力的持续时间,τ越大,意见领袖影响力进入指数衰减的时间就越晚,其影响力持续时间就越长.而对同一意见领袖的不同消息来说,N_0 可以看作消息在该意见领袖粉丝中的初始欢迎度,γ可以看作消息新鲜度的衰减速率,而τ则表示消息新鲜度的持续度.因此,根据消息的传播过程中估计这些参数值,便可以比较不同用户对同一消息的影响力衰减和持续程度,从而定量地分析意见领袖的影响力.图 3 为对图 2 中意见领袖所引发消息传播的拟合结果,图中横坐标为粉丝对意见领袖的响应时间.表 1 列出了各个参数的估计值.从表 1 可知:意见领袖 2 与意见领袖 3 的初始影响力都很大,但意见领袖 2 的影响力较为持久;而意见领袖 1 的影响力则持续了近 220 分钟才进入指数衰减阶段,且其影响力衰减指数远远小于其他 4 位.将已发现的意见领袖及其影响力参数保存,便可对其以后引发消息传播的趋势做出预测.

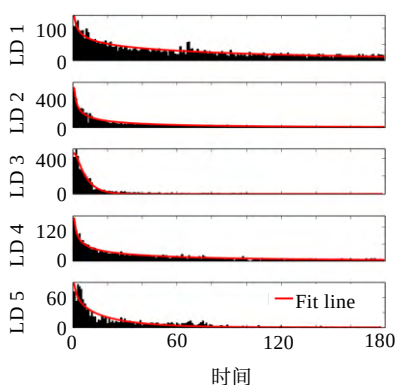

Fig.3 Fitting the influence of the opinion leaders

图 3 意见领袖影响力拟合

Table 1 Parameters estimation of opinion leaders' influence

表 1 意见领袖影响力参数估计

意见领袖	N_0	γ	τ(分钟)
LD 1	243	0.177	219
LD 2	325	0.280	72
LD 3	341	0.339	18
LD 4	106	0.255	112
LD 5	87	0.281	37

5 实验结果与分析

5.1 实验数据收集

新浪微博为开发者提供了丰富的应用程序接口(application programming interface),通过这些接口获取了从2011 年 11 月 27 日~2012 年 1 月 13 日新浪微博每日的转发微博排行榜,并爬取这些微博消息的转发时间流(timeline),共收集到 1 327 条消息以及由它们所引发的 25 593 657 个转发事件.其中,每条转发微博中包含了如下信息:1) Id:该条消息的唯一 ID;2) Created_at:消息转发时间戳;3) Text:消息的内容,少于 140 个字符;4) User:该消息的作者信息.包括用户的 ID,粉丝数、关注数、发布的微博数以及用户的简要描述等.表 2 为获取到数据的基本概要.

Table 2 Summary of the dataset

表 2 数据概要

数据名称	数据
微博消息数	1 327
总转发次数	25 593 657
最大单消息被转发次数	176 991
最小单消息被转发次数	3 290
参与用户数	8 998 624
好友关系数	175 139 393

在此数据集的基础上,爬取了所有参与转发的用户的关注列表.通过以上信息,可以构建任意一条消息的基于最后影响力模型的传播流,进而分析这些消息在网络中的传播模式,挖掘在传播过程中参与传播的意见领袖并对其影响力进行评估.

5.2 意见领袖影响力模型效用评估

根据公式(1),传播流中意见领袖的选取由节点出度的分布函数和阈值θ决定.随着阈值的增大,满足条件的意见领袖数量随之减少.将θ从 100~2 000 范围内进行变化,并对由 1 327 条消息所构建的传播流应用公式(1),图 4 为得到的意见领袖的总数量随阈值θ的变化趋势.

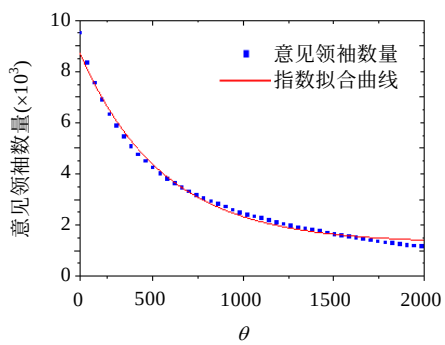

Fig.4 Number of opinion leaders as a function of threshold θ

图 4 意见领袖数量随阈值θ变化曲线

意见领袖的数量与阈值θ的关系可以用指数函数很好地拟合,所得结果如下:

$$N=1287+9282\cdot e^{-0.0022\times\theta} \tag{7}$$

拟合准确度 R^2=0.9973.随着θ的增大,意见领袖的数量呈指数衰减,并最终趋于一个稳定的值.

为了获取到足够数量的意见领袖,以便对他们的影响力进行研究但又不至于将条件设定的过于宽松,取θ=500 作为阈值,从 1 327 条微博消息中共获取到 4 270 个意见领袖.对提取到的意见领袖所引发的传播分别用公式(6)进行非线性最小二乘拟合,估计该意见领袖影响力的各个参数.公式(6)对传播过程的拟合度可以定量地

反映模型的评估性能,非线性最小二乘拟合的拟合度的计算公式为

$$R^2 = 1 - \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} y_i^2}} \tag{8}$$

其中, $\hat{y}_i$ 为拟合估计值,y_i为真实值.根据拟合度的大小,能够定量地分析模型刻画消息传播过程的准确性.图5画出了全部意见领袖所引发的传播过程拟合度的累积分布,由图5可知,70%以上传播过程的拟合度大于0.8.这表明所提模型能够适用于绝大多数由意见领袖所引发的消息传播过程,因此,模型所估计出的参数值能够反映出意见领袖各个方面的影响力.

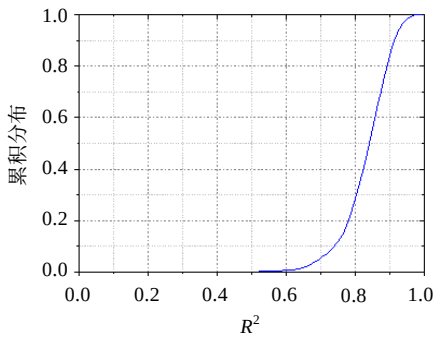

Fig.5 Cumulative distribution of fittingness

图5 拟合度的累积分布

为了进一步验证所提模型在意见领袖挖掘以及其影响力评估上的有效性,表3列出了4 270位意见领袖中平均初始影响力排名前10位的意见领袖及其各个影响力参数.此外,用户的粉丝数量[4]和被转发数[7]也通常被用来衡量用户的影响力,为了与其他意见领袖挖掘方法进行对比,表3也列出了用户的粉丝数量和该用户最近50条微博的平均被转发数量.此外,新浪微博会根据用户的活跃度、传播力和覆盖度这3大指标对名人用户的影响力进行评估,并在"名人堂"页面列出最近半年内的前100位最有影响力用户.表3最后一列给出了相应帐号在2013年7月~2014年1月之间由新浪微博官方给出的该名人用户相应的影响力排名.

Table 3 Top 10 most influential opinion leaders ordered by the averaged initial influence and their influence

表3 平均初始影响力排名前10位的意见领袖及其影响力

用户ID	用户名	$\bar{N}_0$	$\bar{\tau}$	$\bar{\gamma}$	粉丝数	平均转发数	新浪排名
1266321801	姚晨	183	241	-0.219	15 930 478	8 560	17
1195230310	何炅	172	287	-0.402	11 704 923	8 052	9
1192329374	谢娜	170	142	-0.325	13 242 926	5 107	14
1713926427	微博搞笑排行榜	165	181	-0.331	6 054 711	2 163	—
1722594714	舒淇	163	312	-0.352	7 423 279	2 456	52
1195242865	杨幂	160	205	-0.344	12 191 762	11 692	23
1282005885	蔡康永	157	287	-0.226	12 115 806	7 279	37
1704116960	小S	156	367	-0.248	13 970 797	2 891	58
1784537661	罗志祥	155	229	-0.303	9 005 761	8 737	—
1816011541	张小娴	149	196	-0.217	5 854 453	10 617	13

由表3可知:由模型得到的初始影响力排名前10位的意见领袖当中,有8位在新浪的前100位影响力名人帐号当中,表明所提模型确实能够挖掘出有影响力的意见领袖.与新浪的影响力排名相比,基于消息传播的挖掘方法能够挖掘出非认证的且粉丝数量的有限的超级传播者,如"新浪微博搞笑排行榜"等媒体用户.另外,表3还表明:根据用户的粉丝数量和平均被转发数量来衡量用户的影响力,都很有可能低估某些用户的影响力,如用户

“舒淇”和“小 S”等.与其他名人用户相比,她们的粉丝数量与平均被转发数均不占优势,但她们均有较高的初始影响力、较长的影响力持续时间以及较低的影响力衰减指数,且均被提名为新浪前 100 位最有影响力用户.再如用户“张小娴”,虽然其粉丝数量非常有限,但其影响力衰减指数却是前 10 用户当中最低的,且其平均被转发数量以及新浪排名均比较靠前.最后,与其他挖掘方法相比,所提方法不仅仅能够根据用户的初始影响力来衡量意见领袖的影响力,还可以根据用户的影响力持续时间和影响力衰减指数等其他指标来衡量用户的影响力,这为意见领袖影响力的评估提供了更丰富的参考指标.

5.3 意见领袖影响力属性分析

图 6 是意见领袖个数与消息被转发总次数的盒式图,从图中可以看出:参与消息转发的意见领袖个数越多,消息总的平均转发次数的增长趋势并不显著,且随着意见领袖个数的增多,消息的总转发次数的波动性随之变大.这表明消息的总转发次数与参与该消息转发的意见领袖数无明显的相关关系.这是因为意见领袖的判断标准并不是根据其粉丝数量的大小来衡量的,这样就可能导致一部分拥有中等粉丝数量的用户也有可能成为意见领袖,这也是随着消息传播过程中意见领袖数量的增加消息总转发次数波动性变大的主要原因.

图 7 为拟合后得到的意见领袖影响力持续时间的累积分布,由图可知:90%以上意见领袖的影响力持续时间在 30 分钟以上,而 80%以上意见领袖的持续时间小于 180 分钟,表明大多数意见领袖的影响力持续时间在 1 个小时~3 个小时之间.这表明:在消息传播过程中,意见领袖的影响力持续时间非常有限.

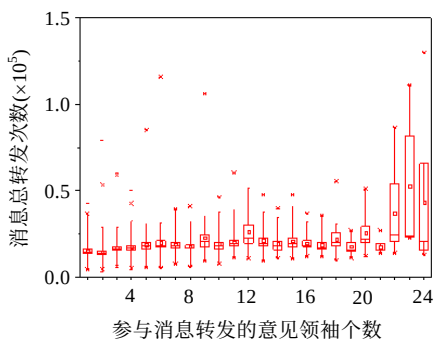

Fig.6 Relations between total reposted times and the numbers of engaged opinion leaders

图 6 消息总转发次数与意见领袖个数的关系

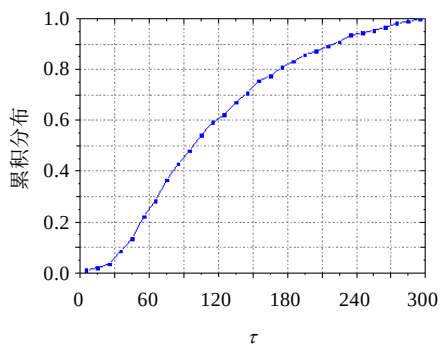

Fig.7 Cumulative distribution of the opinion leaders' influence insistency

图 7 意见领袖影响力持续性累积概率分布

基于 Twitter 的研究结果表明:粉丝数量与转发评论的数量是弱相关的,且用户的影响力在不同的衡量指标下会随时间和主题的不同而发生变化[4,7].为了评估所提模型中各个影响力衡量指标与意见领袖粉丝数的关系,对拟合所得结果与意见领袖的粉丝数做 Spearman 相关性分析,计算公式如下:

$$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{9}$$

其中,x_i 与 y_i 为对原始数据排序后的对应值.分析结果见表 4,意见领袖的初始影响力与其粉丝数量呈正相关,而意见领袖的影响力衰减系数以及持续时间与其粉丝数量的大小无明显的相关性.

Table 4 Correlation coefficients of the opinion leaders' influence and the number of followers

表 4 意见领袖影响力与粉丝数量的相关关系

影响力参数	Spearman 相关系数
N_0	0.62
γ	0.03
τ	-0.07

5.4 模型在消息传播趋势预测中的应用

消息传播趋势的预测,对社会舆论的控制以及广告的定点投放有着重要意义.热门消息的传播可近似看作是由各个意见领袖所引起的各个子传播效果的叠加,且意见领袖在消息传播过程中,其影响力可以用特定的衰减函数来描述.因此,如果能够事先知道意见领袖影响力的各个参数,就可根据该意见领袖的影响力来预测消息围绕该意见领袖所表现出的传播态势,进而对消息总的传播态势进行预测.为此,将已搜集到的消息数据集随机分为训练集和测试集两部分,其中,训练集占消息集的90%,测试集占10%.为了存储意见领袖的各个影响力参数的历史估计值,构建微博意见领袖影响力数据库,其表结构见表5.

Table 5 Database of the opinion leaders' influence

表 5 意见领袖影响力数据库

数据	格式说明
ID	数据 ID
Opinion leader	意见领袖用户 ID
N_0	初始影响力
Gamma	影响力衰减
Tau	影响力持续性
Times	意见领袖在消息中出现的次数
Mount	意见领袖在 10 分钟内引发的转发数

根据以上分析过程,对训练集中的每条消息传播过程中的意见领袖进行提取,并对其影响力参数进行估计,并将得到的影响力参数存入数据库,更新算法见算法1.

算法 1. 意见领袖数据库更新算法.

```
For:训练集中的每条消息,do:
    For:消息中的每个意见领袖,do:
        If:意见领袖存在于数据库中
            获取该意见领袖信息,计算意见领袖影响力平均值,并更新该条数据
        Else:
            直接将意见领袖的各个参数信息存入数据库
        End If
    End For
End For
```

当意见领袖在多条消息中均有参与时,需要对意见领袖的影响力求平均值,计算方法如下:

$aveValue.N_0=(oldValue.N_0\times Times+newValue.N_0)/(Times+1)$

$aveValue.Gamma=(oldValue.Gamma\times Times+newValue.Gamma)/(Times+1)$

$aveValue.Tau=(oldValue.Tau\times Times+newValue.Tau)/(Times+1)$

$aveValue.Mount=(oldValue.Mount\times Times+newValue.Mount)/(Times+1)$

$Times=Times+1$

其中,*aveValue* 为要计算的平均值;*oldValue* 为从数据库中得到的旧值;*newValue* 为意见领袖在一条新消息中的影响力估计值;*Times* 为数据库中的 *Times* 项,在更新时其与 *oldValue* 一起从数据库中获取.当训练集中的意见领袖均已获取并存储,便可以对测试集的离线消息或微博中实时消息的传播趋势进行预测,预测流程如图 8所示:实时检查新加入消息传播的用户是否为意见领袖,如果该用户为意见领袖,则从数据库中获取该意见领袖的影响力信息,等待 10 分钟后,获取意见领袖在该消息中的转发数 M_{10},根据如下公式计算意见领袖在消息中的相对初始影响力:

$$N=N_0\times M_{10}/Mount,$$

其中,N_0 为数据库中意见领袖的初始影响力,*Mount* 为数据库中意见领袖在 10 分钟内引发的转发数.意见领袖影

响力衰减指数与影响力持续时间均从数据库中直接获取.

之后,便可根据公式(6)计算该意见领袖在未来 300 分钟内的传播趋势,并与其他意见领袖引起的传播趋势预测进行叠加便可得到消息的整体传播趋势.

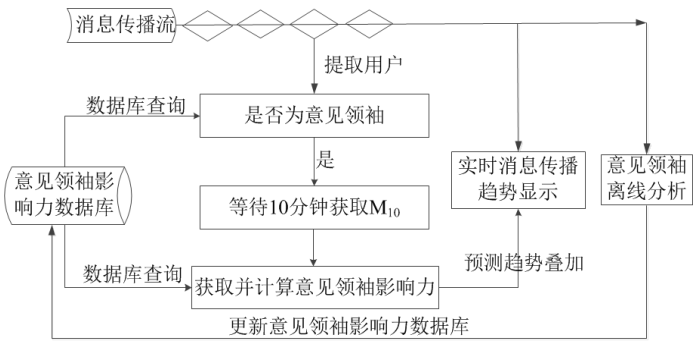

Fig.8 Framework of messages spread prediction

图 8 消息传播趋势预测框架图

为了检验预测效果,图 9 为对测试集中一条消息传播态势进行的预测.模型将消息传播的过程分解为由意见领袖所引导的消息传播子过程,并对该子过程进行预测,较好地呈现出了消息传播过程中的激增式传播态势.此外,对测试集共 132 条消息进行离线分析,共挖掘到 421 名意见领袖,其中有 398 名意见领袖已存在于由训练集得到的意见领袖影响力数据库中.对于不存在于数据库中的意见领袖,虽无法预测其引发的消息传播趋势,但随着对更多在线消息传播的收集以及更多意见领袖被挖掘出来,模型的预测能力以及准确性将不断得到增强.为了验证预测方法的准确性,图 10 画出了对测试集中 132 条消息拟合度的累积分布.由图可知,70%以上消息的拟合度大于 0.7,表明了所提方法能够适用于大多数的热门消息传播趋势的预测.同时,也有部分消息的拟合度较小,调查发现这些消息中含有未挖掘的意见领袖,从而导致最后的拟合误差较大.因此可以肯定:随着所挖掘到的意见领袖逐渐增多,消息预测的准确性也将随之提高.

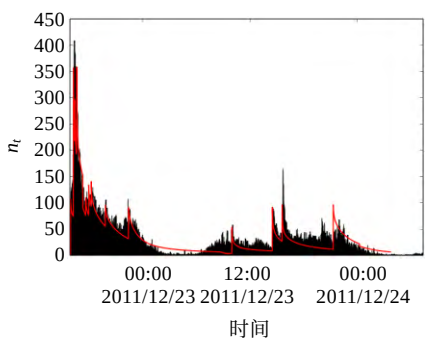

Fig.9 Growth trend prediction for one of the messages in the test dataset

图 9 对数据集中一条消息增长模式的预测

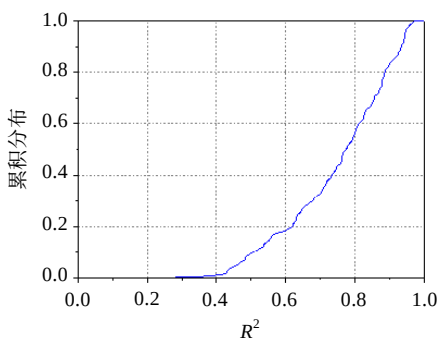

Fig.10 Cumulative distribution of the fittingness for all messages in the test dataset

图 10 测试集中各条消息的拟合度累积分布

6 结束语

为了对意见领袖在消息动态传播过程中所表现出的影响力进行量化分析,提出基于消息传播的微博意见领袖影响力研究方法,定量地分析意见领袖在消息传播过程中的影响力.在分析微博消息传播模式的基础上,将消息的传播过程用动态有向图来描述,发现消息的传播可以按照意见领袖所引发的子传播进行分解.研究发现,

由意见领袖引发的消息传播可以由指数截断的幂律衰减函数来刻画.通过估计该函数中各个参数值的大小,可以定量地评估意见领袖的初始影响力、影响力衰减系数以及影响力的持续性这 3 个重要指标.实验结果表明:该模型能够较好地刻画意见领袖在消息传播过程中所起到的作用,并能够较为准确地对热门消息的传播趋势进行预测.下一步的工作将对消息进行分类,研究意见领袖在不同消息类型中影响力的度量.

References:

[1] Asur S, Huberman BA, Szabó G, Wang C. Trends in social media: Persistence and decay. In: Adamic LA, Baeza-Yates RA, Counts S, eds. Proc. of the 5th Int'l AAAI Conf. on Weblogs and Social Media. Menlo Park: The AAAI Press, 2011. 434-437.

[2] Wang CX, Guan XH, Qin T, Li W. Who are active? An in-depth measurement on user activity characteristics in sina microblogging. In: Proc. of the GLOBECOM. Piscataway: IEEE, 2012. 2083-2088. [doi: 10.1109/GLOCOM.2012.6503423]

[3] Romero DM, Galuba W, Asur S, Huberman BA. Influence and passivity in social media. In: Proc. of the 20th Int'l Conf. Companion on World Wide Web. New York: ACM Press, 2011. 113-114. [doi: 10.1145/1963192.1963250]

[4] Cha M, Haddadi H, Benevenuto F, Gummadi KP. Measuring user influence in Twitter: The million follower fallacy. In: Proc. of the 4th Int'l AAAI Conf. on Weblogs and Social Media. Menlo Park: AAAI Press, 2010. 10-17.

[5] Wu S, Hofman JM, Mason WA, Watts DJ, Who says what to whom on Twitter. In: Srinivasan S, Ramamritham K, Kumar A, Ravindra MP, Bertino E, Kumar R, eds. Proc. of the 20th Int'l Conf. on World Wide Web. New York: ACM Press, 2011. 705-714. [doi: 10.1145/1963405.1963504]

[6] Java A, Song X, Finin T, Tseng B. Why we Twitter: Understanding microblogging usage and communities. In: Proc. of the WebKDD/SNA-KDD. New York: ACM Press, 2007. 56-65. [doi: 10.1145/1348549.1348556]

[7] Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? In: Proc. of the 19th Int'l Conf. on World Wide Web. New York: ACM Press, 2010. 591-600. [doi: 10.1145/1772690.1772751]

[8] Lippmann W. Public Opinion. 1st ed. Piscataway: Transaction Publishers, 1922.

[9] Lazarsfield P, Berelson B, Gaudet H. The People's Choice: How the Voter Makes up His Mind in A Presidential Campaign. New York: Columbia University Press, 1968.

[10] Yu L, Asur S, Huberman BA. What trends in Chinese social media. In: Proc. of the 5th SNA-KDD Workshop'11 (SNA-KDD 2011). New York: ACM Press, 2011. 81-87.

[11] Cha M, Mislove A, Gummadi KP. A measurement-driven analysis of information propagation in the flickr social network. In: Proc. of the 18th Int'l Conf. on World Wide Web. New York: ACM Press, 2009. 721-730. [doi: 10.1145/1526709.1526806]

[12] Rhee JW, Kim EM, Kim H. Exploring online opinion leadership: A validity test of the concept in the digital age. In: Proc. of the Annual Meeting of the Int'l Communication Association. TBA. Eugene: All Academic, 2007. 1-23.

[13] Fan XH, Zhao J, Fang BX, Li YX. Influence diffusion probability model and utilizing it to identify network opinion leader. Chinese Journal of Computers, 2013,36(2):360-367 (in Chinese with English abstract).

[14] 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(4):124-130 (in Chinese with English abstract).

[15] Liu ZM, Liu L. Recognition and analysis of opinion leaders in microblog public opinions. Systems Engineering, 2011,29(6):8-16 (in Chinese with English abstract).

[16] Katz E, Lazarsfield PF. Personal Influence: The Part Played by People in the Flow of Mass Communications. 1st ed. Piscataway: Transaction Publisher, 1970.

[17] Rogers EM, Shoemaker F. Diffusion of Innovations. 4th ed., New York: Free Press, 1971.

[18] Bakshy E, Hofman JM, Watts DJ, Mason WA. Identifying "influencers" on Twitter. In: Proc. of the 4th Int'l Conf. on Web Seach and Data Mining. New York: ACM Press, 2011. 1-10.

[19] Radicchi F. Human activity in the Web. Physical Review E, 2009,80(2):26118. [doi: 10.1103/PhysRevE.80.026118]

[20] Yan Q, Yi L, Wu L. Human dynamic model co-driven by interest and social identity in the MicroBlog community. Physica A, 2012, 391:1540-1545. [doi: 10.1016/j.physa.2011.08.038]

附中文参考文献:

[13] 樊兴华,赵静,方滨兴,李欲晓.影响力扩散概率模型及其用于意见领袖发现研究.计算机学报,2013,36(2):360-367.

[14] 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(4):124-130.

[15] 刘志明,刘鲁.微博网络舆情中的意见领袖识别及分析.系统工程,2001,29(6):8-16.

王晨旭(1986－),男,河南西华人,博士生,主要研究领域为在线社交网络用户行为,信息传播规律.

管晓宏(1955－),男,博士,教授,博士生导师,主要研究领域为复杂网络化系统的安全与经济性,生产制造系统的优化调度,信息物理融合系统,智能电网,网络安全.

秦涛(1982－),男,博士,副教授,主要研究领域为高速网络流量测量,线上社会用户行为监管.

周亚东(1982－),男,博士,讲师,主要研究领域为社交网络测量分析与监控.