中国科学: 信息科学 2018 年 第 48 卷 第 11 期: 1575–1588
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
社会媒体计算与自然语言处理专刊 • 论文

去中心化的微博传播动力学建模

高金华[1,2*], 刘悦[1], 程学旗[1]

1. 中国科学院网络数据科学与技术重点实验室, 中国科学院计算技术研究所, 北京 100190
2. 中国科学院大学, 北京 100190
* 通信作者. E-mail: gaojinhua@ict.ac.cn

收稿日期: 2018–04–09; 接受日期: 2018–05–22; 网络出版日期: 2018–11–14

国家重点研发计划项目 (批准号: 2017YFB0803302) 和国家自然科学基金 (批准号: 61425016, 61472400, 61572473) 资助项目

摘要 社交网络的功能正逐步由网络社交转变为社交媒体, 极大地方便了消息的传播, 也使得消息的流行度预测问题变得更具有挑战性. 传统的流行度预测方法包含基于特征的有监督学习方法和基于随机过程的传播动力学建模方法. 其中, 基于随机过程的传播动力学建模方法由于具有更好的个体预测能力, 吸引了研究人员的广泛关注. 但是, 现有的传播动力学建模方法在建模时, 都忽略了社交平台中消息传播所呈现出的去中心化特点. 本文以微博平台中消息的传播数据为基础, 分析了微博消息的去中心化传播现象, 并提出了一种叠加自增强泊松过程 (reinforced Poisson process, RPP) 模型的方法来刻画消息的传播动力学变化. 每一条信息的传播过程都被拆分为几个传播子过程的叠加, 而每个传播子过程可以用 RPP 模型很好的建模. 在真实数据集上的结果表明, 本文所提出的方法在刻画消息的传播过程以及预测消息的流行度变化等方面, 都要优于现有的方法.

关键词 社交网络, 社交媒体, 流行度预测, 去中心化, RPP 模型

1 引言

社交网络已经成为人们日常生活中重要的一部分. 随着时间的推移和互联网技术的发展, 社交网络的功能和角色也发生了很大的变化. 社交网络早期的主要功能是帮助用户在线上维护社交关系. 用户可以在平台中发布自身的状态, 同时也可以获取到自己关注的好友的状态信息. 而当下的社交网络, 更多地是在扮演着媒体的角色. 从内容上看, 用户在社交网络发布的内容, 除去自身的状态更新外, 更多地涉及当下的新闻和时事. 大部分传统媒体也都直接在社交网络平台中开通了账号, 便于第一时间发布各类消息; 从结构上看, 社交网络平台中的好友或关注机制, 更多地是在为用户提供订阅渠道. 用户通过关注或者好友机制, 从其他用户获取自己感兴趣的内容. 随着社交网络平台的媒体功能不断发

引用格式: 高金华, 刘悦, 程学旗. 去中心化的微博传播动力学建模. 中国科学: 信息科学, 2018, 48: 1575–1588, doi: 10.1360/N112018-00081
Gao J H, Liu Y, Cheng X Q. Decentralized cascade dynamics modeling (in Chinese). Sci Sin Inform, 2018, 48: 1575–1588, doi: 10.1360/N112018-00081

© 2018《中国科学》杂志社

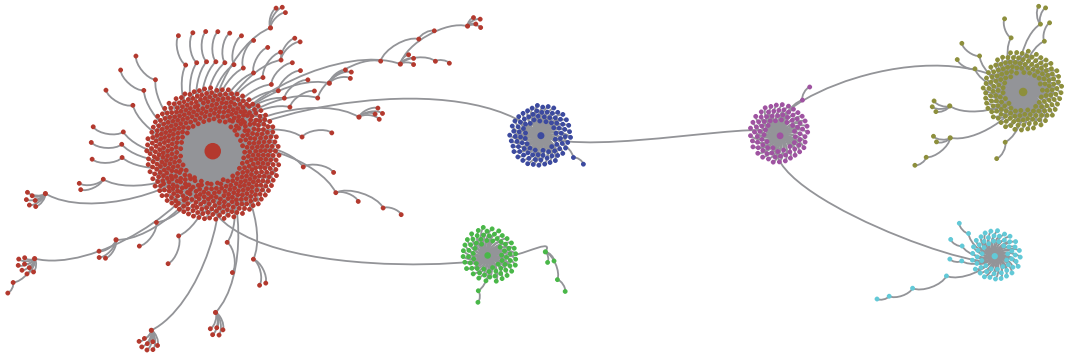

图 1　(网络版彩图) 新浪微博中消息的传播示例图

Figure 1　(Color online) An example cascade from Weibo

展, 进一步诞生了社交媒体的概念. 借助于社交网络平台的优势, 信息在社交媒体中传播地更快、更广, 但同时也为社交平台中消息传播相关的研究带来了更大的挑战.

社交媒体中信息的传播模式与传统媒体有着很大的区别. 在以门户网站、报纸、广播电视等为代表的传统媒体中, 信息的传播都是中心式结构的: 媒体是主要的发布内容者, 也是用户获取信息的主要渠道; 社交媒体中消息的传播则呈现出一种去中心化的结构: 任何用户都有可能成为传播的中心. 当用户发布消息后, 他的粉丝就可以选择转发他的消息. 在转发过程中, 部分关键转发用户在参与传播后可能会带来大量的转发, 进而使得整体的传播过程呈现出一种去中心化的结构. 图 1 展示了新浪微博中一条消息的转发示意图, 图中的每一个节点代表一条消息, 每条边表示了对应消息间的转发关系. 从图中可以看出, 在转发过程中, 有多个中间转发消息也累积了大量的转发, 使得消息的整体传播结构呈现出去中心化的特点. 这种去中心化的传播结构, 使得消息的流行度传播预测变得更具挑战性.

社交网络中消息的传播流行度预测是一个非常重要的研究问题, 在很多领域都有着重要的应用价值, 包括潜在热点内容发现 [1,2]、平台流量控制和资源分配 [3,4], 以及市场营销和广告投放等实际应用问题. 现有的流行度预测方法主要包括基于特征工程的有监督学习方法和基于随机过程的流行度动力学建模方法. 基于特征工程的有监督学习方法是将流行度预测问题形式化为分类或者回归问题, 通过特征工程的方式, 提取一系列能够指示流行度变化的特征, 利用现有的有监督学习框架, 对消息的流行度进行预测. 这类方法的核心在于发掘与流行度变化相关的特征. 现有工作中提出的用于流行度预测的特征主要包含 4 类: 时序特征 [5,6]、内容特征 [7,8]、用户特征 [9,10], 以及传播级联结构特征 [11]. 而这些特征中, 时序特征被认为是最具指示能力的特征 [12]. 这类有监督的预测方法在平均意义上具有较好的预测效果, 但是缺乏对个体的预测能力. 基于随机过程的方法通常将流行度的增长过程形式化为一个计数过程 [13], 然后借助于生存分析理论框架 [14], 来建模流行度的到达过程. 这类方法的核心在于对流行度增长过程的强度函数的刻画. 现有的模型主要包含基于自增强泊松过程 (reinforced Poisson process, RPP) 的方法和基于自激励 Hawkes 过程的方法. Shen 等 [15] 提出了基于 RPP 的模型来建模论文的引用数的增长过程. 在建模过程中, 他们考虑了影响论文引用数增长的 3 个因素: 论文本身的质量、时间衰减效应, 以及 “富者愈富” 现象. Gao 等 [16] 将 RPP 模型应用到了微博平台中, 并对 RPP 模型中的时间效应函数和激励函数进行了修正. Bao 等 [17] 引入了自激励 Hawkes 过程来建模微博平台中的流行度增长过程. 在建模过程中, 他们认为每一次对消息的转发都会为消息的流行度增长带来新的激励. 因此, 他们将消息在历史传播过程中所有转发带来的激励之和作为消息的当前流行度增长的强度. Zhao 等 [18] 进一步拓展了 Bao 等的方法. 在建模历史转发带来的激励作用时, 作者引入了转

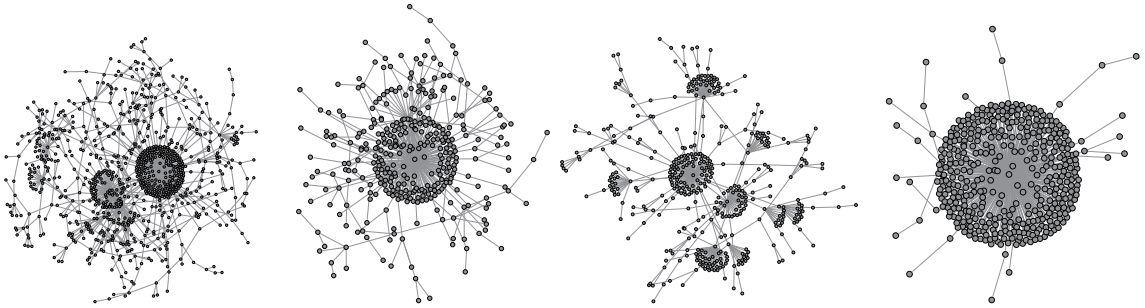

图 2 新浪微博中消息转发树示例

Figure 2 Examples of different cascade structures in Weibo. (a)∼(d) Four example cascades

发用户的粉丝数来更精确地刻画每一次转发带来的激励作用. 与基于特征工程的有监督学习方法相比, 基于随机过程的流行度动力学建模方法具有更好的个体预测能力, 在近年来也得到了研究人员的广泛关注. 此外, 还有部分基于表示学习 [19,20] 和分组策略 [21,22] 的流行度预测方法.

然而, 现有的两类基于随机过程的模型在刻画社交媒体中消息的传播过程时, 都存在着各自的缺陷. 基于 RPP 的模型将消息的传播过程形式化为一个单源传播的过程, 但是这一假设并不符合社交媒体中消息去中心化的传播模式, 而且该类模型在建模每次转发带来的激励作用时, 只是简单地统计消息的总体转发数, 而没有对转发激励的大小进行区分. 基于自激励 Hawkes 过程的模型建模了每次转发带来的激励作用, 但是它需要为每一次转发学习一个激励参数, 使得模型的参数学习在实际应用中难以实现, 需要对模型做进一步的约简, 进而影响了模型的预测能力. Gao 等 [23] 在微博传播结构建模方面进行了简单的尝试, 但是数据集太小, 数据分析也不够充分. 因此, 仍然缺乏去中心化传播结构的有效分析和刻画.

本文首先对微博平台中消息的去中心化传播过程进行了量化分析, 指出了建模去中心化传播过程的必要性. 进一步对样例传播过程进行了分析, 发现去中心化的传播过程可以视作由若干中间转发用户引发的传播子过程的叠加, 而这些传播子过程都可视作单源的传播过程, 可以利用现有的 RPP 模型较好地建模. 基于这些观察, 本文提出了一种叠加 RPP 模型的方法来刻画去中心化的传播过程. 与基于 RPP 的模型相比, 我们的模型能够很好地捕获传播过程去中心化的特点, 同时也区分了中心用户与其他转发用户带来的激励作用的差异; 与基于自激励 Hawkes 过程的模型相比, 我们的模型只需要学习中心用户的参数, 大大减少了待学习的参数个数. 在真实数据集上的实验结果表明, 本文提出的方法能够有效地提高预测效果; 同时, 我们的模型也能更好地刻画消息的去中心化传播过程.

2 数据分析

利用新浪微博平台提供的接口, 从新浪微博上随机采集了几条消息的完整传播路径, 如图 2 所示. 图中的每一个点表示一条消息, 消息之间的连边表示转发关系. 这里重点关注传播结构中的聚团现象, 因此隐去了连边的方向. 本文将消息转发树中的第 1 个节点称为源发消息, 其余节点称为转发消息. 图 2(d) 展示是典型的中心型结构, 所有转发消息基本上都是对源发消息的直接转发; 图 2(b) 也是中心型的传播结构, 只是部分传播链会比较长, 但是整个传播过程仍然只有源发消息作为中心传播节点; 图 2(a) 和 (c) 则展示了典型的去中心化结构. 除源发消息外, 部分转发消息本身也吸引了大量的传播,

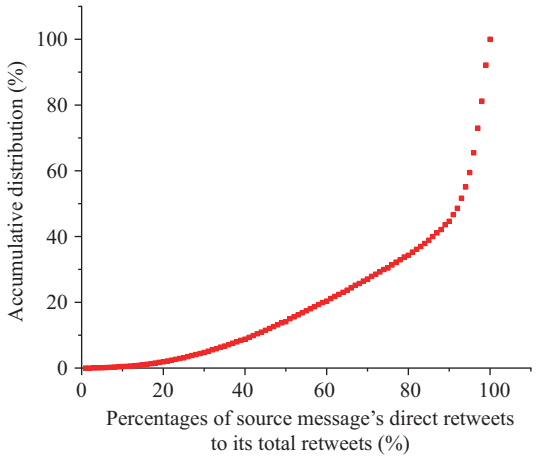

图 3 (网络版彩图) 源发消息的直接转发比例分布

Figure 3 (Color online) Distribution of ratios of source message's direct retweet number to its total retweet number

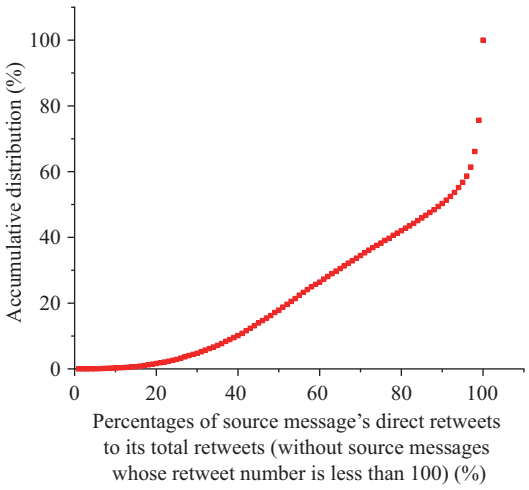

图 4 (网络版彩图) 源发消息的直接转发比例分布 (高转发数据集)

Figure 4 (Color online) Distribution of ratios of source message's direct retweet number to its total retweet number with less reteeted messages removed

从而形成了一种去中心化的传播结构. 这种去中心化的传播结构, 目前还缺乏有效的建模方法, 也是本文的建模目标.

为了进一步衡量微博中去中心化传播现象的存在比例, 我们从一个较为完整的微博数据集中, 抽取了某一天中发布的所有源发消息以及它们在发布后 24 小时内的所有转发消息. 根据消息文本内容中的 "@" 标记, 我们还原了每条源发消息在传播过程中产生的转发路径, 并将发布 1 小时后累计转发数小于 10 的消息从数据集中剔除. 最终本文的数据集中包含了近 6.5 万条源发消息以及它们的转发数据.

本文首先对源发消息的转发情况进行了统计. 对于每一条源发消息, 我们统计了它自身被直接转发的次数占最终总转发次数的比例, 结果如图 3 所示. 图中横坐标是源发消息被直接转发的次数占最终总转发次数的比例, 而纵坐标是对应比例的累积分布. 从图 3 中可以看出, 有接近 20% 的源发消息的直接转发占比低于 50%, 而直接转发占比超过 80% 的源发消息占整体数据集不到 40%. 这些数据说明将微博中消息的转发过程直接建模成以源发消息为中心的单源传播是不合理的. 为了消除低转发数消息对统计结果的影响, 我们从数据集中剔除了最终转发数小于 100 的源发消息, 并对剔除后的数据集重新进行统计, 得到的统计结果如图 4 所示. 从图中可以看出, 在高转发量数据集中, 直接转发占比较低的源发消息同样占有着不小的比例.

进一步地, 本文对中间转发消息的转发情况进行了统计. 对于每一条源发消息, 搜集了它所有的直接转发消息, 并对这些直接转发消息的转发数进行了统计. 我们将转发数最大的直接转发消息称为关键转发消息, 并对关键转发消息的转发数占源发消息总转发数的占比情况进行了统计, 得到的累积分布如图 5 所示. 这里展示的是在高转发量数据集上的统计结果. 从图中可以看出, 有 10% 的关键转发消息的转发占比超过了 50%, 有 20% 的关键转发消息带来的转发占比超过了 30%. 实际上, 微博平台中能引发大量转发的关键转发消息存在比较普遍, 例如微博平台中的意见领袖 [24, 25] 在参与到消息传播中时, 往往能引起 "二次传播" 现象. 这些统计结果说明, 消息在传播过程中, 除源发消息外还会出

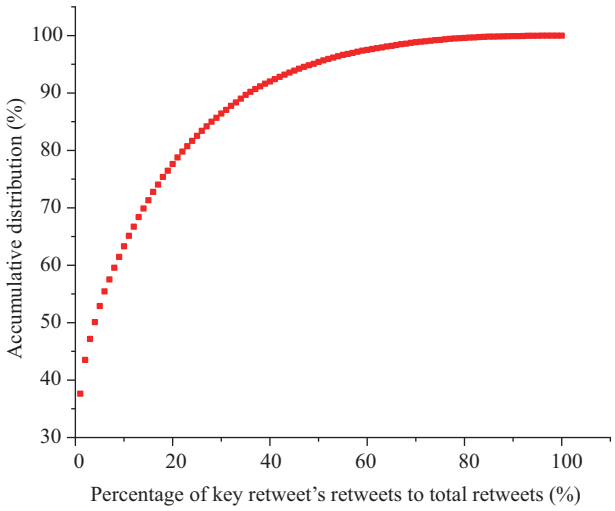

图 5 (网络版彩图) 关键转发消息的转发占比分布

Figure 5 (Color online) Distribution of ratios of key retweet's retweet number to total retweet number

现少量引起大量转发的中间转发消息, 对这些中间转发消息引起的转发情况的建模, 是刻画去中心化传播过程的关键.

综上所述, 通过对数据集中消息的转发情况的分析, 可以看出, 去中心化的传播结构在消息传播过程中比较普遍. 因此, 需要一种模型来刻画去中心化的传播过程. 进一步对各层转发消息的转发数据进行统计后发现, 部分中间转发消息会带来很高的转发量, 形成 "二次传播". 因此, 刻画这些中间转发消息所引发的转发过程, 是刻画去中心化传播过程的关键.

3 模型

3.1 问题形式化

给定一条消息 d 在观测时间窗口 $[0,T]$ 内的转发时间戳序列 $C^d=\{t_i^d|t_i^d<T,i=0,1,2,\ldots,n_d\}$, 其中 n_d 在消息 d 观测时间窗口内的转发次数, t_i^d 是消息 d 的发布时间与第 i 次转发的时间的间隔, 且满足 $0=t_0^d\leqslant t_1^d\leqslant t_2^d\leqslant\cdots\leqslant t_{n_i}^d\leqslant T$. 我们的目标是预测消息 d 在发布后任意时刻 t 的转发数.

3.2 RPP 模型

本小节先简单介绍 RPP 模型. RPP 模型是由 Shen 等 [15] 提出并被用于建模论文引用数的到达过程. RPP 模型是一个基于随机过程的模型. 它将论文引用数的增长过程形式化为一个计数过程, 提出了影响论文增长速度的几个关键影响因素, 并通过对这些关键影响因素的刻画, 来对论文增长过程的强度函数进行建模, 进而对论文的引用数作出预测. RPP 模型在建模论文的引用达到过程时, 重点考虑了 3 个影响论文引用数的因素. 首先是论文本身的质量因素. 论文的质量越高, 论文的最终引用数会越高, 而且引用数增长的速度也越快. 第 2 个因素是时间衰减效应. 类似于信息传播中的爆发衰减现象 [26], 论文引用数的增长也同样会存在爆发阶段和衰减阶段, 需要利用一个时间函数来刻画论文的引用增长速度随时间的变化过程. 最后一个因素是论文引用中的 "富者愈富" 机制 [27,28]. 在论文引用过程中, 引用数越高的论文, 越有机会获得更多新的引用. Shen 等使用了乘性模型来融合上述 3 个

因素, 最终给出了如下的论文引用强度函数:

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = x(t) = \lambda f(t;\theta) I(t), \tag{1}$$

其中, $c(t)$ 是到 t 时刻论文的总引用数, $x(t)$ 表示 t 时刻论文的引用数的增长强度, λ 用于度量论文的质量, $f(t;\theta)$ 用于刻画时间衰减效应, θ 是对应的参数. 在 RPP 模型中, 作者选用的时间衰减效应函数是 Log-normal 函数, $I(t)$ 用于刻画论文累积引用数带来的 "富者愈富" 影响. 在 RPP 模型中, $I(t)$ 函数的值就是到 t 时刻为止论文的累积引用数, 即 $I(t) = m + c(t)$, 其中 m 是作者引入的论文引用数的先验. 在给定强度函数后, 就可以利用极大似然框架来求取相应的参数. 论文的引用序列可以用 $\{t_0, t_1, t_2, \ldots, t_N\}$ 表示, 其中 t_i 是第 i 次引用的到达时刻与论文发表时刻的时间间隔, N 是论文的总引用次数. 论文引用序列的似然函数可以表示为

$$L(\lambda, \theta) = p_0(T|t_N) \prod_{i=1}^{N} p_1(t_i|t_{i-1}), \tag{2}$$

其中, $p_0(T|t_N)$ 表示时间区间 $(t_N, T]$ 内没有新的引用到达的概率. 在生存理论分析框架中, 给定强度函数 $x(t)$ 后, 时间区间 $(t_N, T]$ 内没有新的引用到达的概率可以用生存概率建模如下:

$$p_0(T|t_N) = \mathrm{e}^{-\int_{t_N}^{T} x(t_N)\mathrm{d}t} = \mathrm{e}^{-\int_{t_N}^{T} \lambda f(t;\theta) I(t_N)\mathrm{d}t}, \tag{3}$$

$p_1(t_i|t_{i-1})$ 表示了在给定前一次引用发生在 t_{i-1} 时刻的情况下, 下一次引用发生在 t_i 时刻的概率密度. 在生存理论分析框架中, 这个概率密度为

$$p_1(t_i|t_{i-1}) = x(t_{i-1})\mathrm{e}^{-\int_{t_{i-1}}^{t_i} x(t_{i-1})\mathrm{d}t} = \lambda f(t;\theta) I(t_{i-1}) \mathrm{e}^{-\int_{t_{i-1}}^{t_i} \lambda f(t;\theta) I(t_{i-1})\mathrm{d}t}. \tag{4}$$

给定上述定义后, 结合观测数据, 就可以利用极大似然框架来推断强度函数中的相应参数, 进而对论文引用数的后续变化进行预测. 在求解得到参数 $\{\lambda^*, \theta^*\}$ 后, 通过求解微分方程 (1) , 并结合边界条件 $c(T) = N$, 可以得到论文在 t 时刻的引用数的预测函数如下:

$$c(t) = (m + N)\mathrm{e}^{\lambda^* \int_T^t f(t;\theta^*)\mathrm{d}t} - m. \tag{5}$$

RPP 模型在预测论文引用数变化时, 取得了非常好的效果. 但是论文引用场景更像是一个单源传播的场景, 而消息传播则呈现出明显的去中心化结构. 因此, 本文提出了一种面向去中心化传播结构的流行度到达过程建模方法.

3.3 观察

我们挑选了一条呈现出去中心化传播结构的消息, 并对它的传播过程进行了进一步的分析和研究, 结果如图 6 所示. 从图 6(a) 中可以看到, 消息整体的传播过程可以看做几个单源传播过程的叠加. 本文将这几个单源传播过程的中心称为中心转发消息, 并将中心转发消息引发的转发称为子过程. 按照传播过程中的中心转发消息, 将图 6(a) 中的传播过程拆分为 6 个单源传播子过程. 每个子过程都可以视作一个单源传播场景, 可以使用 RPP 模型进行较好的刻画. 与 RPP 模型类似, 在建模消息的转发数增长过程时, 考虑了消息本身的质量、时间衰减效应, 以及消息当前转发数带来的激励作用. RPP 模型对 6 个子过程的建模结果如图 6(b) 所示. 从图中可以看到, RPP 模型可以很好地刻画每个中心转发消息引发的传播子过程. 最后, 将这 6 个 RPP 模型的预测结果按照时间顺序叠加, 得到的流行度变化曲线与真实的流行度变化曲线非常相近, 如图 6(c) 所示. 因此, 通过叠加多个子过程来对消息的去中心化传播过程进行建模, 是一种有效的方法.

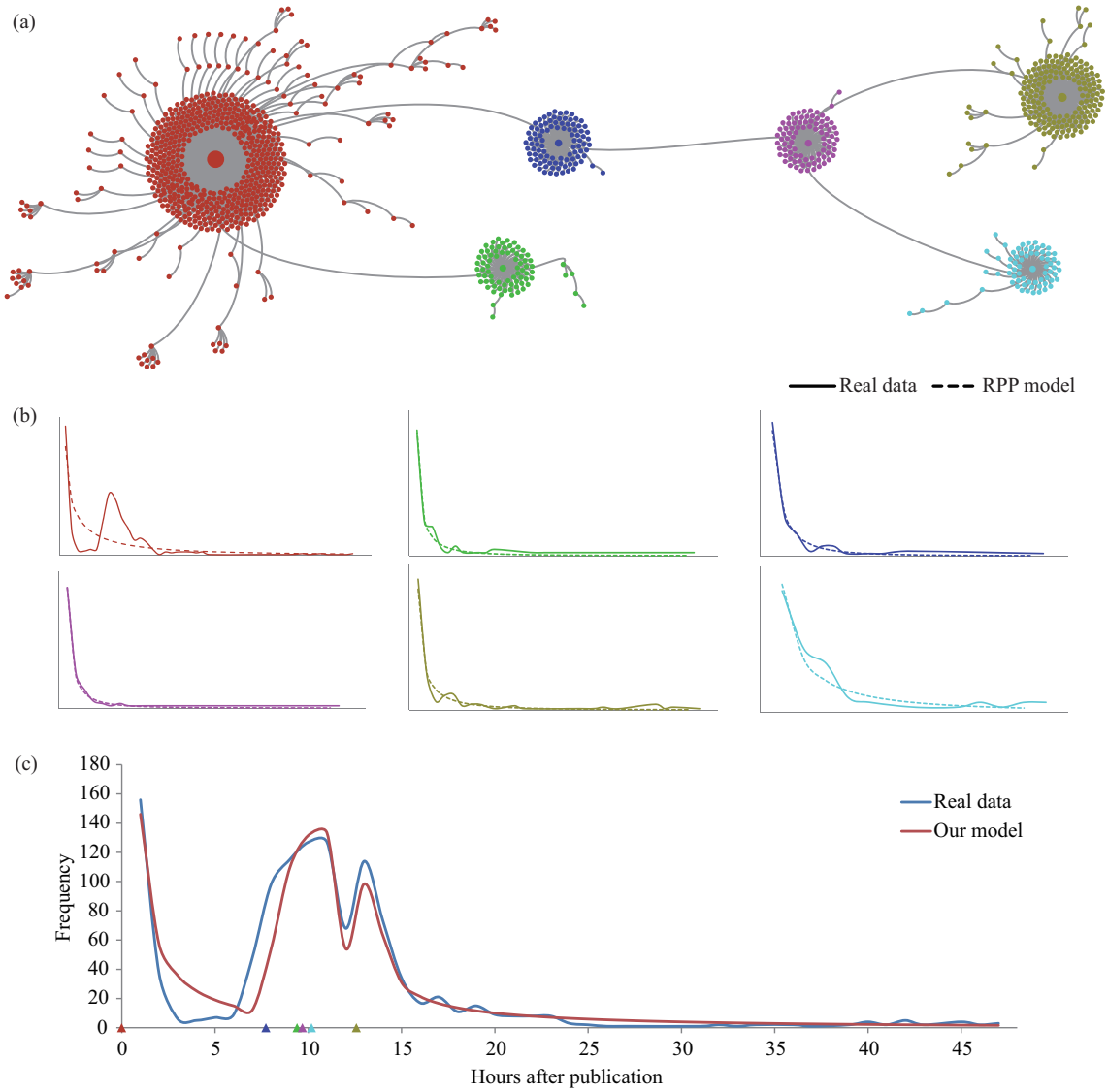

图 6 (网络版彩图) 示例消息的去中心化传播过程建模

Figure 6 (Color online) Cascade dynamics modeling for example cascade. (a) Diffusion structure of example cascade; (b) RPP modeling for the dynamics of six subprocesses; (c) aggregation of six subprocesses to fit the overall dynamics of example cascade

3.4 去中心化的微博传播动力学建模

通过上述观察, 我们发现去中心化的传播过程可以拆分为几个基于中心转发消息的单源传播过程, 而每个单源传播过程又可以用 RPP 模型很好的建模. 因此, 本文提出一种叠加 RPP 模型来建模去中心化传播过程的方法. 将消息整体的传播过程拆分为 K 个子传播过程, 其中 K 是一个超参, 用于控制传播过程中中心转发消息的个数, 它对模型的影响会在实验部分进行讨论. 对于第 l 个子过程, 用参数 $\{\lambda_l, \theta_l, \tau_l\}$ 来刻画, 其中 λ_l 表示第 l 个子过程的中心转发消息带来的激励大小, 类似于 RPP 模型中的论文质量因素; θ_l 是第 l 个子过程的时间衰减效应函数的参数. 我们认为每个子过程的衰减程度和速度是互相独立的; τ_l 表示第 l 个子过程的开始时间, 用作计算每个子过程的时间衰减效应的偏移

量. 给定上述参数的情况下, 定义微博消息的流行度增长过程的强度函数如下:

$$x(t)=\sum_{l=1}^{K}\lambda_l f(t-\tau_l;\theta_l)I(t), \tag{6}$$

其中, 对于时间衰减效应函数 $f(t)$ 和 "富者愈富" 机制 $I(t)$, 选用了和 RPP 模型相同的函数. 给定上述强度函数和参数集合 $\{\lambda,\theta,\tau\}$ 后, 我们的模型对前述时间戳序列 C^d 的似然表示如下:

$$\begin{aligned}L(\lambda,\theta,\tau)&=p_0(T|t_N)\prod_{i=1}^{N}p_1(t_i|t_{i-1})\\&=\prod_{i=1}^{n_d}\left((m+i-1)\times\left(\sum_{l=1}^{K}\{\lambda_l f(t_i-\tau_l;\theta_l)\}\right)\right)\\&\quad\times \mathrm{e}^{-(m+n)\int_0^T\sum_{l=1}^{K}\lambda_l f(t-\tau_l;\theta_l)\mathrm{d}t+\sum_{i=1}^{n_d}\int_0^{t_i}\sum_{l=1}^{K}\lambda_l f(t-\tau_l;\theta_l)\mathrm{d}t}.\end{aligned} \tag{7}$$

利用上述似然函数, 结合时间戳序列数据, 就可以求出参数 $\{\lambda^*,\theta^*,\tau^*\}$, 进而利用下述预测函数对消息的后续流行度进行预测:

$$c(t)=(m+n_d)\mathrm{e}^{\int_T^t\sum_{l=1}^{K}\lambda_l^* f(s-\tau_l^*;\theta_l^*)\mathrm{d}s}-m. \tag{8}$$

比较本文模型和 RPP 模型的强度函数, 可以发现本文模型通过叠加 RPP 模型的方式, 建模了微博的去中心化传播过程, 改进了 RPP 模型的单源传播假设, 同时本文模型还分别对每个中心转发消息带来的激励作用进行了区分; 比较本文模型的强度函数和基于自激励 Hawkes 过程的模型的强度函数, 可以发现本文模型只建模了中心转发消息带来的激励作用, 避免了后者在模型训练过程中存在的参数多导致学习困难的问题.

3.5 优化

本文采用了梯度下降的方式来优化似然函数 (7). 先对似然函数取对数, 得到

$$\begin{aligned}\ln L=&\sum_{i=1}^{n_d}\ln(m+i-1)+\sum_{i=1}^{n_d}\ln\left(\sum_{l=1}^{K}\lambda_l f(t-\tau_l;\theta_l)\right)\\&-(m+n)\int_0^T\sum_{l=1}^{K}\lambda_l f(t-\tau_l;\theta_l)\mathrm{d}t+\sum_{i=1}^{n_d}\int_0^{t_i}\sum_{l=1}^{K}\lambda_l f(t-\tau_l;\theta_l)\mathrm{d}t.\end{aligned} \tag{9}$$

借助对数函数的单调性, 可以直接优化对数似然 (9) 来求取最优参数. 首先, 需要求取对数似然函数关于各参数的梯度. 在本文模型中, 参数包括 $\{\lambda_l,\theta_l,\tau_l\}_{l=1}^K$. 其中 θ_l 是时间衰减效应函数的参数. 本文模型选取的时间衰减效应函数是 Log-normal 函数 $f(t;\mu,\sigma)=\frac{1}{t\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(\ln t-\mu)^2}{2\sigma^2}}$, 因此本文模型的实际参数包括 $\{\lambda_l,\mu_l,\sigma_l,\tau_l\}_{l=1}^K$. 对数似然函数关于各参数的梯度如下:

$$\begin{aligned}\frac{\partial\ln L}{\partial\lambda_l}&=\sum_{i=1}^{n}\frac{f_l(t_i)}{\sum_{k=1}^{K}\lambda_k f_k(t_i)}+\sum_{i=1}^{n}\int_0^{t_i}f_l(t)\mathrm{d}t-(m+n)\int_0^T f_l(t)\mathrm{d}t,\\\frac{\partial\ln L}{\partial\mu_l}&=\frac{1}{\sigma_l}\cdot\frac{\ln(t_i-\tau_l)-\mu_l}{\sigma_l}\cdot\sum_{i=1}^{n}\frac{\lambda_l f_l(t_i)}{\sum_{k=1}^{K}\lambda_k f_k(t_i)}\end{aligned} \tag{10}$$

$$
\begin{aligned}
&- \frac{1}{\sigma_l} \cdot \lambda_l \sum_{i=1}^{n} \phi\left(\frac{\ln(t_i-\tau_l)-\mu_l}{\sigma_l}\right) \\
&+ \frac{1}{\sigma_l} \cdot (m+n)\lambda_l \phi\left(\frac{\ln(T-\tau_l)-\mu_l}{\sigma_l}\right), \qquad (11)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \sigma_l} = &\frac{1}{\sigma_l} \cdot \sum_{i=1}^{n} \left\{ \frac{\lambda_l f_l(t_i)}{\sum_{k=1}^{K} \lambda_k f_k(t_i)} \left[\left(\frac{\ln(t_i-\tau_l)-\mu_l}{\sigma_l}\right)^2 - 1 \right] \right\} \\
&- \frac{1}{\sigma_l} \cdot \lambda_l \sum_{i=1}^{n} \left\{ \phi\left(\frac{\ln(t_i-\tau_l)-\mu_l}{\sigma_l}\right) \cdot \frac{\ln(t_i-\tau_l)-\mu_l}{\sigma_l} \right\} \\
&+ \frac{1}{\sigma_l} \cdot (m+n)\lambda_l \phi\left(\frac{\ln(T-\tau_l)-\mu_l}{\sigma_l}\right) \cdot \frac{\ln(T-\tau_l)-\mu_l}{\sigma_l}, \qquad (12)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \tau_l} = &\frac{1}{t_i-\tau_l} \cdot \sum_{i=1}^{n} \left\{ \frac{\lambda_l f_l(t_i)}{\sum_{k=1}^{K} \lambda_k f_k(t_i)} \left(1 + \frac{\ln(t_i-\tau_l)-\mu_l}{\sigma_l^2}\right) \right\} \\
&- \frac{1}{\sigma_l} \cdot \frac{1}{t_i-\tau_l} \cdot \lambda_l \sum_{i=1}^{n} \phi\left(\frac{\ln(t_i-\tau_l)-\mu_l}{\sigma_l}\right) \\
&+ \frac{1}{\sigma_l} \cdot \frac{1}{T-\tau_l} \cdot (m+n)\lambda_l \phi\left(\frac{\ln(T-\tau_l)-\mu_l}{\sigma_l}\right), \qquad (13)
\end{aligned}
$$

其中, $f_l(t)$ 是一种缩写, 表示使用第 l 个子过程的时间衰减效应函数, 即 $f_l(t) \equiv f(t-\tau_l;\mu_l,\sigma_l)$. 另外, 这里为了保持公式形式的简洁, 省略了时间戳表示中的上标, 即将 t_i^d 统一表示为 t_i.

本文采用了 Adadelta 方法 [29] 来自适应地调整参数的学习率. Adadelta 方法会根据参数的历史梯度来自适应地调整学习率.

4 实验

本节主要通过在真实数据集上的实验, 来对比我们提出的模型和参照模型在流行度预测问题上的表现, 并对相关实验结果进行分析. 此外, 还分析了子过程的个数对模型的预测效果的影响. 最后, 通过一些案例分析, 展示了本文提出的模型在建模去中心化传播过程时的优势.

4.1 数据集

本文将数据分析部分用到的微博转发数据集作为实验数据集. 该数据集包含了微博平台某一天中发布的所有源发消息以及它们在发布后 24 小时内的所有转发消息. 我们截取了每条消息在发出后 12 小时内的转发数据用于实验, 其中发布后一小时内的转发数据作为观察数据用于模型训练, 剩余的 11 小时内的转发数据用于预测. 为了保证模型训练的效果, 对数据集进行了进一步的筛选. 从数据集中剔除了在观察时间窗口内累积转发量小于 50 的微博消息. 最终我们的数据集包含了 20490 条消息.

4.2 参照模型和评价指标

从现有的两类流行度预测方法中分别选取了较为代表性的模型来作为参照模型. 在基于特征工程的有监督学习方法中, 选取了 Pinto 模型 [6]. Pinto 模型将观察窗口切分为等长的时间区间, 并将消息在每个时间区间内的流行度增长量作为消息的特征表示. 在基于随机过程的流行度动力学模型中, 选取了 RPP 模型 [15]. RPP 模型采用了自增强泊松过程来刻画流行度的达到过程, 并在建模过程中考虑

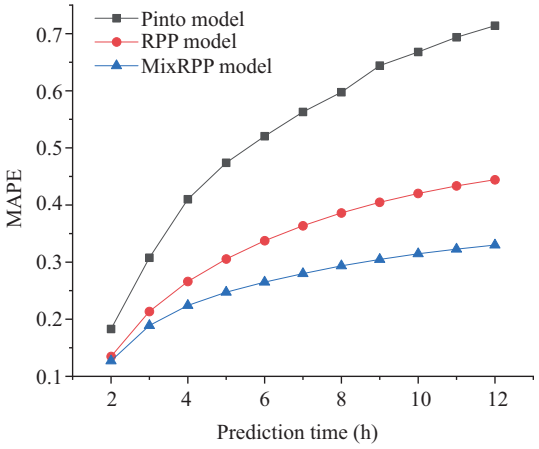

图 7 (网络版彩图) 各模型流行度预测效果对比

Figure 7 (Color online) Prediction performances of our model and baseline models

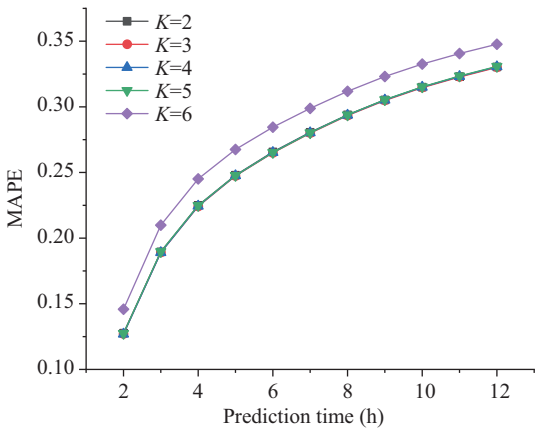

图 8 (网络版彩图) 不同 K 值下模型的预测效果对比

Figure 8 (Color online) Prediction performances under different settings of K

了消息本身的质量、时间衰减效应, 以及 "富者愈富" 机制的影响, 在论文引用数据上取得了较好的预测效果. 后续的实验分析中, 用 MixRPP 模型来指代我们提出的模型.

评价指标方面, 选择了平均绝对百分误差 (mean absolute percentage error, MAPE) 来作为模型的评价指标. MAPE 值反映了预测结果偏离真实结果的程度, 被广泛应用于回归类问题的评价之中. MAPE 值的定义如下:

$$\mathrm{MAPE}(t) = \frac{1}{m}\sum_i \left|\frac{r^i(t) - \widetilde{r^i(t)}}{r^i(t)}\right|,$$

其中, m 是测试集的大小, $r^i(t)$ 表示 t 时刻消息 i 的真实流行度, $\widetilde{r^i(t)}$ 表示模型对消息 i 在 t 时刻的流行度的预测. MAPE 的值越低, 表示模型的预测效果越好.

4.3 实验结果

本文模型和参照模型在真实数据集上的预测效果如图 7 所示. 从实验结果上可以看出, 在微博转发数预测问题上, 基于随机过程的预测方法的效果要优于基于有监督学习的方法, 这也说明了刻画转发数增长过程对于消息转发数预测问题的重要性. 对比本文模型和 RPP 模型的预测效果可以看出, 利用子过程叠加的形式来刻画微博消息的去中心化传播过程, 能够有效地提升预测效果.

4.4 实验分析

实验分析部分, 首先研究了子过程的个数 K 对模型预测效果的影响. 将 K 值从 2 遍历到 6, 并记录了不同 K 值下模型的预测效果, 结果如图 8 所示. 从结果中可以看出, 当 K 取值 2 到 5 时, 模型的预测模型没有太大变化; 当 K 值继续增大到 6 时, 模型的预测效果有了明显的下降. 这也说明虽然微博的传播过程是去中心化的, 但是能引发 "二次传播" 的节点个数并不会太多. 因此, 在模型预测效果对比实验中, 将 K 值设置为 3.

我们还从实验数据中选取了一些样例来展示本文提出的模型能够较好地刻画消息的去中心化传播过程. 图 9 中展示了两条示例微博. 其中, 图 9(a) 和 (c) 展示了 3 条走势曲线, 分别为消息在发出

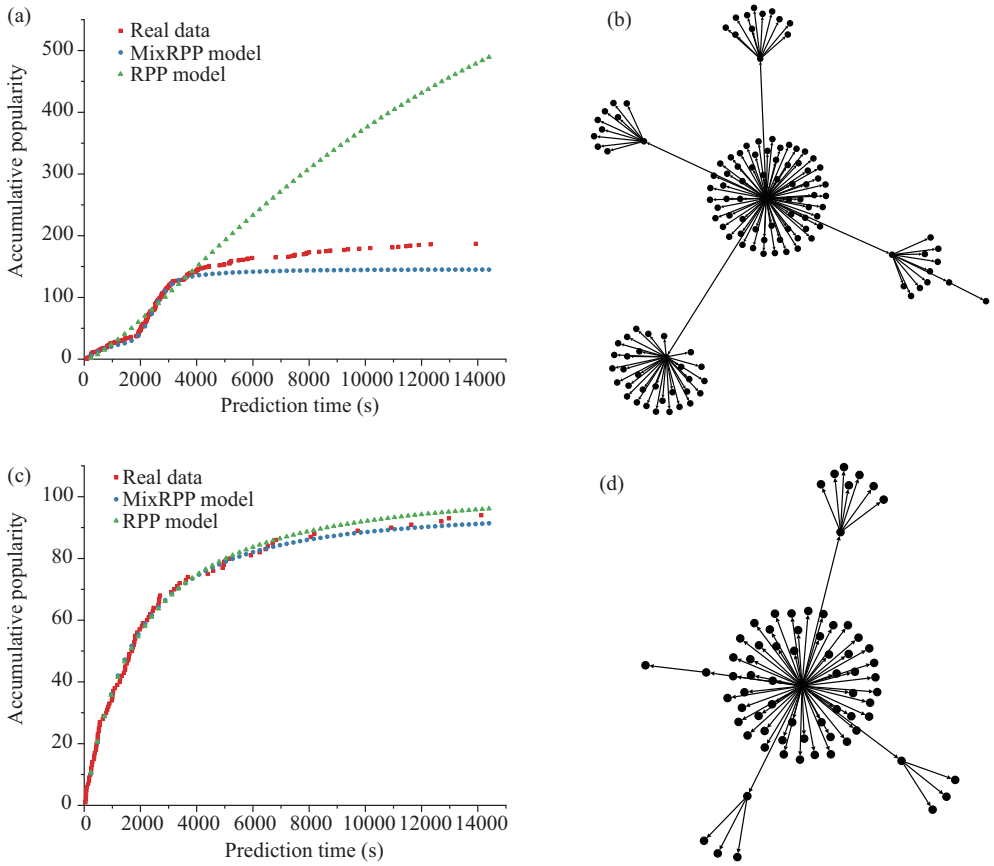

图 9 (网络版彩图) 传播样例分析

Figure 9 (Color online) Case study of two example cascades. (a) Popularity growth prediction for example cascade 1; (b) diffusion structure of example cascade 1; (c) popularity growth prediction for example cascade 2; (d) diffusion structure of example cascade 2

后 4 小时内真实的转发数增长过程以及 RPP 模型和我们提出的模型对其作出的预测; 图 9(b) 和 (d) 则展示了各消息在观测窗口内的传播级联结构. 通过观察传播级联结构可以发现, 上方示例微博的传播级联呈现出一种明显的去中心化结构, 图 9(b) 中的转发团簇对应着 "二次传播" 现象. 在对这条微博的流行度增长过程进行预测时, 可以看到 RPP 模型表现出了明显的偏差, 而且偏差随着时间的推移不断扩大; 而我们提出的模型, 无论是在观测窗口内的到达过程建模, 还是在后续时刻的流行度预测表现上, 都要明显地优于 RPP 模型. 下方示例微博整体上呈现出单源传播的结构. 从图 9(c) 的增长趋势预测图中可以看出, 在建模单源传播过程时, RPP 模型可以取得比较好的结果, 而我们提出的模型在建模单源传播结构时, 同样取得了比较好的结果.

5 总结

本文首先对微博平台中消息的传播数据进行了分析, 发现部分中间转发消息会引起 "二次传播" 现象, 从而使得消息的整体传播过程呈现出一种去中心化的结构, 而且这种去中心化的传播结构在社交媒体中广泛存在, 使得社交媒体中消息的流行度预测问题变得更有挑战性. 我们对现有的流行度预

测方法进行了总结, 并分析了现有方法在建模去中心化的消息传播过程中存在的缺陷和不足. 我们进一步分析了去中心化的传播过程, 发现整个传播过程可以被拆分为几个单源传播子过程的叠加. 基于这一观察, 提出了一种叠加 RPP 模型的方法来建模微博消息的传播过程. 与传统的方法相比, 本文方法不仅放宽了 RPP 模型的单源传播限制, 同时也解决了基于自激励 Hawkes 过程的模型存在的参数学习困难的缺陷. 本文方法重点关注于中心转发消息所引发的子传播过程的建模, 既满足去中心化结构的建模要求, 同时大大减少了参数个数. 在实际数据集上的实验结果表明, 本文的模型不仅能更好地刻画微博消息的流行度到达过程, 而且在流行度预测问题上的表现也优于现有的模型.

参考文献

1 Tatar A, Antoniadis P, Amorim M D, et al. From popularity prediction to ranking online news. Soc Netw Anal Min, 2014, 4: 174

2 Tatar A, Antoniadis P, De Amorim M D, et al. Ranking news articles based on popularity prediction. In: Proceedings of International Conference on Advances in Social Networks Analysis and Mining, Istanbul, 2012. 106–110

3 Chen X, Zhang X D. A popularity-based prediction model for web prefetching. Computer, 2003, 36: 63–70

4 Famaey J, Iterbeke F, Wauters T, et al. Towards a predictive cache replacement strategy for multimedia content. J Netw Comput Appl, 2013, 36: 219–227

5 Szabo G, Huberman B A. Predicting the popularity of online content. Commun ACM, 2010, 53: 80–88

6 Pinto H, Almeida J M, Gonçalves M A. Using early view patterns to predict the popularity of youtube videos. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining, Rome, 2013. 365–374

7 Tsagkias M, Weerkamp W, de Rijke M. Predicting the volume of comments on online news stories. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, 2009. 1765–1768

8 Bandari R, Asur S, Huberman B A. The pulse of news in social media: forecasting popularity. In: Proceedings of the 6th International Conference on Weblogs and Social Media, Dublin, 2012. 26–33

9 Kong S B, Mei Q Z, Feng L, et al. Predicting bursts and popularity of hashtags in real-time. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, 2014. 927–930

10 Cui P, Jin S F, Yu L Y, et al. Cascading outbreak prediction in networks: a data-driven approach. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, 2013. 901–909

11 Bao P, Shen H W, Huang J M, et al. Popularity prediction in microblogging network: a case study on sina weibo. In: Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, 2013. 177–178

12 Shulman B, Sharma A, Cosley D. Predictability of popularity: gaps between prediction and understanding. In: Proceedings of the 10th International Conference on Web and Social Media, Cologne, 2016. 348–357

13 Andersen P K, Borgan Ø, Hjort N L, et al. Counting process models for life history data: a review. Scand J Stat, 1985, 12: 97–158

14 Klein J P, Moeschberger M L. Survival Analysis: Techniques for Censored and Truncated Data. Berlin: Springer, 2005

15 Shen H W, Wang D S, Song C M, et al. Modeling and predicting popularity dynamics via reinforced Poisson processes. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, Québec City, 2014. 291–297

16 Gao S, Ma J, Chen Z M. Modeling and predicting retweeting dynamics on microblogging platforms. In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining, Shanghai, 2015. 107–116

17 Bao P, Shen H W, Jin X L, et al. Modeling and predicting popularity dynamics of microblogs using self-excited Hawkes processes. In: Proceedings of the 24th International Conference on World Wide Web, Florence, 2015. 9–10

18 Zhao Q Y, Erdogdu M A, He H Y, et al. Seismic: a self-exciting point process model for predicting tweet popularity. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 2015. 1513–1522

19 Li C, Ma J Q, Guo X X, et al. Deepcas: an end-to-end predictor of information cascades. In: Proceedings of the 26th International Conference on World Wide Web, Perth, 2017. 577–586

20 Cao Q, Shen H W, Cen K T, et al. Deephawkes: bridging the gap between prediction and understanding of information cascades. In: Proceedings of the 26th ACM Conference on Information and Knowledge Management, Singapore, 2017.

1149–1158
21 Ouyang S X, Li C Y, Li X M. A peek into the future: predicting the popularity of online videos. IEEE Access, 2016, 4: 3026–3033
22 Cao Q, Shen H W, Gao H, et al. Predicting the popularity of online content with group-specific models. In: Proceedings of the 26th International Conference on World Wide Web Companion, Perth, 2017. 765–766
23 Gao J H, Shen H W, Liu S H, et al. Modeling and predicting retweeting dynamics via a mixture process. In: Proceedings of the 25th International Conference Companion on World Wide Web, Montréal, 2016. 33–34
24 Iyengar R, Van den Bulte C, Valente T W. Opinion leadership and social contagion in new product diffusion. Marketing Sci, 2011, 30: 195–212
25 Staab S, Domingos P, Mika P, et al. Social networks applied. IEEE Intel Syst, 2005, 20: 80–93
26 Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system. Proc Natl Acad Sci USA, 2008, 105: 15649–15653
27 Jeong H, Néda Z, Barabási A L. Measuring preferential attachment in evolving networks. Europhys Lett, 2003, 61: 567–572
28 Wang M Y, Yu G, Yu D R. Measuring the preferential attachment mechanism in citation networks. Phy A-Stat Mech Appl, 2008, 387: 4692–4698
29 Zeiler M D. Adadelta: an adaptive learning rate method. 2012. ArXiv:1212.5701

Decentralized cascade dynamics modeling

Jinhua GAO[1,2*], Yue LIU[1] & Xueqi CHENG[1]

1. *CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190*, China*;
2. *University of Chinese Academy of Sciences, Beijing* 100190*, China*
* Corresponding author. E-mail: gaojinhua@ict.ac.cn

Abstract Social network platforms have gradually transformed into social media in recent years, which greatly eases information diffusion. Meanwhile, it also makes the task of popularity prediction for online information more challenging. Traditional popularity prediction methods include supervised methods based on feature engineering and popularity dynamics modeling methods based on stochastic processes. The latter group of methods has drawn a wide attention as it works better for predicting individual items' popularity. However, the existing methods have ignored the decentralization characteristic of diffusion cascades. In this paper, we investigate the decentralized structure of diffusion cascades, based on the Weibo dataset, and propose to model cascade dynamics by mixing a bunch of reinforced Poisson processes (RPP). The overall diffusion cascade is a mixture of several sub-cascades, each of which can be modeled by RPP. Experimental results on real world datasets demonstrate our proposed model's superiority in both characterizing popularity dynamics and predicting future popularity.

Keywords social network, social media, popularity prediction, decentralization, RPP model

Jinhua GAO was born in 1989. He is currently a Ph.D. candidate at the Institute of Computing Technology, CAS. His research interests include social computing and web data mining.

Xueqi CHENG was born 1971. He received his Ph.D. degree from the Institute of Computing Technology, CAS, Beijing, in 2006. He is currently a professor at the Institute of Computing Technology, CAS. His research interests span across network science, network information security, and network data mining.

Yue LIU was born in 1970. She received her Ph.D. degree from the Institute of Computing Technology, CAS, Beijing, in 2004. She is currently an associate professor at the Institute of Computing Technology, CAS. Her research interests mainly focus on web data search and mining.