流行度演化分析与预测综述

胡 颖* 胡长军 傅树深 黄建一
(北京科技大学计算机与通信工程学院 北京 100083)

摘 要：社交网络每天以爆发式的增长速率产生着大量信息，但是人们对海量信息的关注程度有限。人们关注哪些信息、对信息的关注程度如何随时间变化，即为信息的流行度演化问题。流行度演化反映了人们的关注点和信息的流动与传播。建模与预测网络信息的流行度演化有助于信息传播和人类行为的研究、辅助舆情监控、并带来极大的应用和商业价值。近几年，研究人员在该方面取得了丰硕的研究成果，但尚缺乏对这些成果进行梳理、总结的综述。该文系统地回顾网络信息流行度演化的主要工作，对分析与预测方法、模型、发展脉络进行梳理。首先从定性和定量方面阐述了流行度演化的特点；介绍如何量化影响流行度演化的众多因素，并对它们进行分类、总结；然后将已有的建模和预测方法归纳为 3 类：基于早期流行度、基于影响因素、基于级联传播，从原理、典型成果、特点比较、适用范围等方面对这 3 类方法进行评述；最后根据目前模型和方法的特点以及现实需求，指出了未来流行度演化的研究方向。

关键词：社交网络；流行度演化；预测；信息传播；网络信息
中图分类号：TP393; TP391 **文献标识码**：A **文章编号**：1009-5896(2017)04-0805-12
DOI: 10.11999/JEIT160743

Survey on Popularity Evolution Analysis and Prediction

HU Ying HU Changjun FU Shushen HUANG Jianyi
(*Department of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*)

Abstract: Online social network is generating information at an explosive rate. Information competes with each other for people's limite attention. How people's attention to information evolves over time is referred to as the problem of popularity evolution. Popularity evolution reflects what people focus on and how information flow and diffuse. Popularity evolution prediction of online information helps the studies of information diffusion and human behaviors, assists public opinion monitoring, and brings high application value and commercial value. In recent years, researchers have gained great research achievements. However, there is still a lack of survey which reviews and summarizes existing work. This paper systematically reviews main work of popularity evolution analysis and prediction, and gives summarization to the existing methods and models. First, insight into understanding popularity evolution patterns from qualitative and quantitative perspectives is provided. How to measure factors affecting popularity evolution and to classify them in taxonomy are introduced. Third, the methods of modeling and predicting popularity evolution are categorized into three classes: previous-popularity-based, factor-based, and diffusion-based. These three classes from the following aspects are elaborated: theory, representative work, characteristic comparison, and application scope. Finally, the paper is concluded and future research directions are given according to existing work and current demands.

Key words: Social network; Popularity evolution; Prediction; Information diffusion; Online information

1 引言

因为在线社交网络的迅速普及，每时每刻都有大量的网络信息(online information)，如视频、微博、图片等，被生产出来并得以传播。人们每天面临着海量信息，而人们的注意力又是有限的，研究表明[1,2]，大部分信息都未获得关注度，仅有少量信息受到人们的关注。那么，哪些信息会受到关注，关注程度[1−3]如何随时间变化，即为流行度演化(popularity evolution)问题。流行度演化反映了人们

收稿日期：2016-07-14；改回日期：2016-12-30；网络出版：2017-02-24
*通信作者：胡颖 huyingustb@163.com
基金项目：国家 973 规划项目(2013CB329601)
Foundation Item: The National 973 Program of China (2013CB329601)

的关注点以及信息在时间和空间轴的流动与传播。

流行度演化分析与预测在多个领域得到广泛应用，从科学角度来说，它有助于加深对信息传播、人类行为的系统认识，理解社交现象；从应用角度来说，该问题具有广泛的应用前景，比如市场营销、网络广告、搜索引擎、推荐系统等。

流行度演化的分析与预测是一项具有挑战性的工作。流行度演化具有高度动态性[4]，因为受众多因素影响，比如外部因素：社交影响力[5,6]、网站机制[7,8]、现实事件[9,10]、时间循环(daily or weekly cycle)[11]等，内部因素：网络信息内容质量[12,13]、内容时效性[2]。这些因素混合交织在一起共同作用于流行度，很难将它们分离、单独量化。加之，社交网站涉及用户隐私的数据和算法不公开，也给建模、预测流行度演化带来了难度。

自 20 世纪 90 年代中期，在线社交网络兴起，至今产生了海量的网络信息数据，研究者开始有机会在大量真实数据的基础上对流行度演化进行分析、预测，主要涉及 3 方面的基本问题：(1)流行度演化是否存在基本的演化模式，有哪些基本模式。(2)影响流行度演化的因素有哪些，如何量化这些因素，它们如何作用于流行度演化。(3)如何对流行度演化进行建模和预测。围绕这些问题，有大量的文献，但是，尚缺乏对这些文献进行梳理的综述。本文围绕这 3 个基本问题系统地回顾了从该问题兴起至今的代表性工作、以及最新研究成果。

本文内容主要包括：第 3 节，给出和流行度演化相关的基本概念。第 4 节从定性和定量两个层面介绍流行度演化的基本模式。第 5 节从内部和外部方面，总结影响网络信息流行度增长和饱和的因素，介绍如何量化它们以及它们如何影响流行度。第 6 节回顾、讨论流行度演化的建模和预测方法。第 7 节对全文进行总结并探讨未来的研究方向。

2 文献可用性考量

本节主要介绍如何筛选参考文献、并对本综述的参考文献进行可用性考量。

本文主要从 3 个数据库选择文章：Google scholar, ACM digital library, IEEE Xplore，用“popularity evolution+social network”,“popularity prediction+social media”,“popularity prediction+online information”等作为搜索关键字。然后，对搜出来的高引用次数的文章做重点评述。

由于本文列出的参考文献较多，表 1 展示了依据 CCF ABC 类分级对主要文献进行的可用性考量。

3 基本概念

为了统一概念与符号，方便理解内容，本节对

表 1 主要参考文献可用性考量

论文	作者	年份	CCF 分级	当前引用次数
文献[40]	LERMAN K and HOGG T	2010	A	205
文献[25]	MATSUBARA Y, SAKURAI Y, PRAKASH B A, *et al.*	2012	A	129
文献[8]	BORGHOL Y, ARDON S, CARLSSON N, *et al.*	2012	A	52
文献[48]	CHENG J, ADAMIC L, DOW P A, *et al.*	2014	A	189
文献[41]	ZHAO Q, ERDOGDU M A, HE H Y, *et al.*	2015	A	30
文献[52]	WANG S, YAN Z, HU X, *et al.*	2015	A	6
文献[24]	CHENG J, ADAMIC L A, KLEINBERG J M, *et al.*	2016	A	3
文献[18]	CHA M, KWAK H, RODRIGUEZ P, *et al.*	2007	B	1260
文献[4]	YANG J and LESKOVEC J	2011	B	566
文献[49]	KUPAVSKII A, OSTROUMOVA L, UMNOV A, *et al.*	2012	B	38
文献[23]	FIGUEIREDO F, BENEVENUTO F, and ALMEIDA J M	2013	B	168
文献[50]	ARDON S, BAGCHI A, MAHANTI A, *et al.*	2013	B	24
文献[30]	PINTO H, ALMEIDA J M, and GONCALVES M A	2013	B	115
文献[44]	FIGUEIREDO F, ALMEIDA J M, GONCALVES M A, *et al.*	2016	B	3
文献[2]	WU F and HUBERMAN B A	2007	无	360
文献[19]	CRANE R and SORNETTE D	2008	无	442
文献[11]	SZABO G and HUBERMAN B A	2010	无	583
文献[26]	YU Honglin, XIE Lexing, and SANNER S	2015	无	5

流行度演化问题中一些概念和理论进行形式化描述和说明。

3.1 相关概念

流行度　给定某则网络信息 i 和某时刻 t，该信息的流行度 $y_i(t)$ 定义为人们在时刻 t 对其的关注程度。

多数工作中，研究者通常将流行度量化为人们在某时刻采取在网络信息上的积极的网络行为(观看、点赞、转发、评论)的次数。如，对于某视频，其流行度可以指它在某时刻获得的观看量；对某微博，流行度可以指它获得“点赞”的数量；对于某话题，流行度可以指它被转发的次数。流行度还可以用其他统计特性表示，比如被收藏的次数、被分享的次数。其实，选取哪个统计特性表示流行度并没有区别，因为相关研究[14,15]表明，这些统计特性之间存在很强的关系，可以互相转换。这种度量方法与我们的直觉是一致的：通常，一个视频被观看的次数越多，也就表示它越流行。

少数工作认为信息可以既会获得积极的网络行为又会获得消极的网络行为，比如 YouTube 视频的“喜爱/不喜爱”、Digg 帖子的“赞/踩”等，因此一些研究者认为只考虑积极的网络行为是片面的，还需要考虑消极的网络行为。问题是如何将 2 维的变量(积极和消极网络行为的次数)转换成 1 维的流行度。最简单的方法是用积极网络行为的数量减消极网络行为的数量[16]。Yin 等人[17]提出了更复杂的度量流行度的方法。给定某信息 i，流行度 y_i 表示方法如式(1)：

$$y_i=\begin{cases}\dfrac{p-n}{p+n+1}; & p-n\neq 0\\ \dfrac{\varepsilon}{p+n+1}; & p-n=0\end{cases}\quad p+n>\sigma \tag{1}$$

其中，σ 是流行阈值，流行度超过该阈值被认为是流行的。p 是积极网络行为的数量，n 是消极网络行为的数量，ε 是小常量。用该方法度量的流行度的值域为$[-1,1]$。积极网络行为的数量远大于消极数量时，流行度接近 1，表明该信息很流行；积极网络行为的数量和消极数量大致相等时，流行度接近 0，表明该信息具有争议性。

流行度演化　给定某则网络信息 i 和其生命长度 L_i (天/小时/分钟)，$L_i\in N^+$，其流行度演化定义为人们对该信息的关注程度的时间序列，即 $\{y_i(1), y_i(2),\cdots,y_i(L_i)\}$。

传播网络　给定某则网络信息 i 和某时刻 t，其传播网络表示为 $N_i(t)=(V_i(t),E_i(t))$。其中 $V_i(t)$ 是节点集合，是 0 到 t 时刻内，被信息 i 感染的节点。$E_i(t)$ 是边集合，如果节点 v 关注了节点 u，边 $(v\to u)$ 就被添加至 $E_i(t)$。

3.2 常用理论

幂率分布　Cha 等人[18]证明了 YouTube 视频的流行度服从幂律(power law)分布。后来，其他网络信息也被研究出基本服从这两种分布，如 Twitter 推文、Essembly 解答、Wikipedia 文章等。图 1 为 Twitter Hashtag 的最终流行度分布，当 x 轴和 y 轴做了对数处理后，这些散点基本分布在直线上，即表示服从幂率分布。幂率分布表明流行度的分布是十分不均匀的，绝大多数网络信息只能获得很少的流行度，而只有少数可以获得很多流行度。

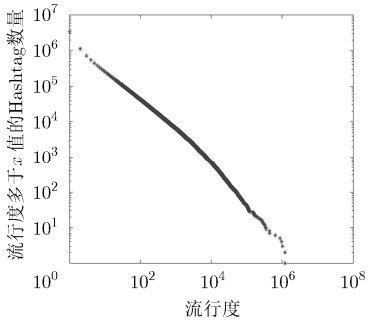

图 1　Twitter Hashtag 最终流行度分布

指数上升、幂率下降　指流行度演化的上升部分可以用指数函数更准确地拟合，而下降部分用幂率函数拟合更准确[14]。判断是指数还是幂率的方法是，将上升或下降部分的数据放在 x 轴不变、y 轴做对数处理的坐标系内，若数据点基本分布在直线上，说明符合指数；若放在 x, y 轴都做对数处理的坐标系内，数据点基本分布在直线上，说明符合幂率。研究表明，网络信息的流行度演化遵循指数上升、幂率下降定律。

“富者更富”　“富者更富”(rich-get-richer)过程导致流行度符合幂率分布，该过程也被称为优先选择(preferential attachment)过程或者耶鲁(Yule)过程。流行度的“富者更富”可以被通俗地解释成如下过程：某网络信息获得新的流行度的概率正比于该信息已有的流行度。

4　流行度演化模式

流行度演化过程中受到各种内、外部因素的影响，其态势变化多端。因此很多研究者希望回答这样的问题：是否存在一些统一的流行度演化模式？对于该问题，研究者主要研究流行度在其生命周期内的增长衰落速度[19,20]、峰[19,21]和趋势[21,22]，从而捕捉流行度时序上的特征。本节从定性和定量两个角度分析流行度演化的基本模式。

4.1 定性模式

定性模式根据演化过程中上升、下降的大致形态将流行度演化聚(分)为一些基本类别。文献[4]提出了 K-SC(K-Spectral Centroid)聚类算法。该算法和 K-means 聚类算法相似，但是它并不像 K-means 那样采用欧几里得距离计算聚类中心进行迭代，K-SC 算法采用了时间序列相似度，该相似度可以表征不同形状的、不同时间范围的、出现在不同时间点的峰。该算法根据上升下降速率的不同将流行度演化划分成了 6 类，如图 2 所示[4]：最大的一类只有 1 个峰，峰的上升较陡峭，下降较平缓，该类代表普通的网络信息的流行度演化；最小的一类在主峰后面还会跟随一些小的峰，该类代表最流行的网络信息的流行度演化。之后，Figueriedo[21]拓展了文献[4]的工作，他还发现了另一种流行度演化模式，其流行度在相当长一段时间内保持持续增长的状态，这类流行度演化是最难预测的。Figueriedo 等人[23]还研究了 3 种性质不同的视频的演化：top 视频、有版权的视频和通过关键字随机选取的视频。他们的结果表明 top 视频流行度通常会经历一个急剧上升的过程。有版权的视频的流行度基本都是在生命周期的早期阶段获得的，可以用传染病传播过程来模拟；相反，随机选取的视频在其全生命周期可以一直不断地获得流行度。

为了研究流行度演化在级联传播下除了图 2 中的 6 类外是否存在新的演化模式，Cheng 等人[24]利用 Facebook 上时间跨度为 1 年的数据，对文献[4]的工作做了进一步补充。图 2 的这 6 种模式是在流行度的短时间跨度上(数天或数月)挖掘的，当考虑更长的时间跨度时(1 年)，新的更复杂的模式被发现：流行度演化存在重复性，即流行度在长时间跨度上所经历的起伏会重复出现。该重复性可以利用具有多爆发的传染病模型进行建模。

4.2 定量模式

定量模式同样研究流行度演化过程中的上升、下降的大致形态，但是，同时还给出了拟合上升下降过程的数学模型。文献[19,20]用峰值阶段的流行度与总流行度的比值将 YouTube 视频划分为 3 类：病毒型(viral)、质量型(quality)、垃圾型(junk)。病毒型是指那些通过传染病传播(epidemic propagation)模式导致流行度具有口口相传式的增长(word-of-mouth growth)的视频；质量型的视频和病毒型的相似，但是流行度突然上升形成峰，而不是从底部平缓地增长到顶部；垃圾型是指由于某种偶然原因流行度也出现了峰，但是没能在社交网络传播开。对于这 3 种类型，作者用不同指数的 self-excited Hawkes 条件波松模型来拟合，该模型遵循幂率上升、幂率下降。后来，Matsubara 等人[25]对传染病模型进行了改进，提出了 SpikeM 模型，在 MemeTracker, Twitter 和 GoogleTrends 3 个数据集上，验证了流行度演化实际上遵循指数上升、幂率下降。

无论是定性还是定量研究流行度演化基本模式，流行度演化都大致遵循上升、下降的模式，表明流行度演化和生命个体类似，会经历爆发、顶峰、衰老的生命周期[26]。

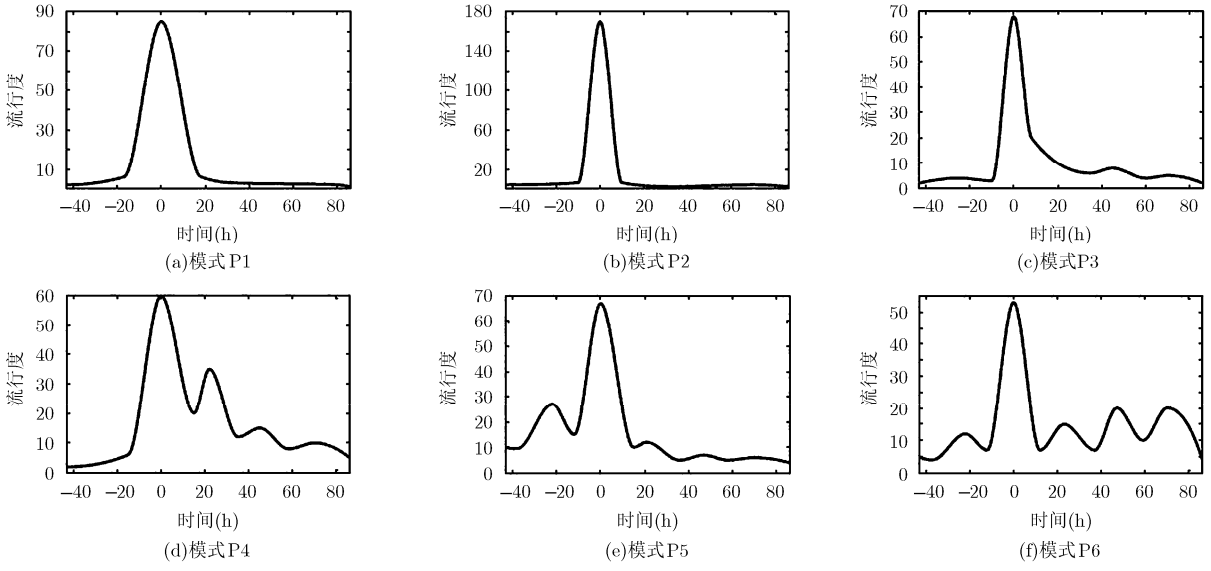

图 2 6 类流行度演化模式

5　流行度演化的影响因素

由“幂率分布”可知，只有很少一部分网络信息可以获得很多流行度，绝大多数只能获得很少的流行度；由图 2 可知，有些流行度经历顶峰后迅速衰落，有些仍会有小幅度的起伏。造成流行度演化差异的影响因素是十分重要的问题。围绕该问题的研究成果很多，概括来说，影响流行度演化的因素如图 3 所示。这里既有促使流行度增长的因素，又有促使流行度饱和的因素。下面围绕如何量化这些因素、这些因素如何作用于流行度等展开讨论。

5.1 促进流行度增长的因素

5.1.1 内部因素　Agarwal 等人[12]对博客的影响力进行量化时，发现博客的内在属性(新颖性和雄辩力)与其收到的评论数有着直接关系。如果一个博客引用很多其他文章或博客，则它不太可能是新颖的，基于上述假设，他们用外部链接的数量衡量博客的新颖性，链接数越多越不新颖；用博客的长度来衡量该博客是否具有雄辩力，博客越长越具雄辩力。他们发现博客引用外部链接数越少、博客越长，其获得的流行度越多。

Bandari 等人[13]对新闻的内在属性进行了更复杂的研究。他们从新闻中提取出 4 个特征：新闻源(news source，该新闻是哪个机构发布的，New York Times, USA Today 等)、新闻分类(category，政治类、娱乐类等)、主观性(subjectivity，新闻的主观程度和客观程度)、有名实体(named entities，新闻中是否包含有名实体，“贾斯汀•比伯”、“奥巴马”等)。他们使用不同的评分函数分别对每个特征进行量化，随后利用回归方法将新闻的最终流行度与 4 个特征的评分关联起来。他们发现这 4 个特征中，预测流行度效果最好的因素是新闻的来源，而新闻的类别在预测流行度时效果不佳。新闻中提到的有名实体对流行度的获得也有促进作用。

5.1.2 外部因素　对于文本类的网络信息，通过语义分析研究影响流行度增长的内部因素[12,13]是可行的。但是对于非文本类的网络信息，例如视频和图片，研究者们通常会从促进流行度增长的外部因素入手。

社交影响力　社交影响力[5,6]可以通过用户之间的社交活动体现出来，表现为用户的行为和思想等受他人影响发生改变的现象[5]。社交影响力出现在很多场合：现实生活中，人们倾向于听大家都喜欢听的歌，喜欢看大家都喜欢看的电影等；社交网络上，用户倾向于看点击量多的视频，浏览大家都爱浏览的帖子等。由此可见，社交影响力是网络信息获得更多流行度的潜在影响因素。

Lerman 对“用户是否会‘赞’那些被好友‘赞’过的帖子”一问这题进行了研究[7]。为了回答这个问题，他们关注某帖子初始 m 个“点赞”之后的“点赞”情况，观察它们当中有多少来自初始 m 个点赞者的粉丝。他们发现当 m 为 25 时，几乎有一半的点赞数都来自于初始 m 个点赞者的粉丝。这一数据有力地支持了用户的确会通过好友接口去看好友们喜欢的帖子的说法，且社交影响力确实对网络信息流行度有影响。

Salganik 等人[27]调研了内容质量和社交影响力对歌曲最终流行度的作用，他们人工地构造了一个调研环境。参与者可以在参考(社交影响力)或不参考以前参与者选择的情况下下载未知的歌曲。他们发现在没有社交影响力的情况下，歌曲的质量与最终流行度有必然的联系：好歌曲很少获得低流行度，差歌曲很少获得高流行度。但是当在有社交影响力的情况下，歌曲本身质量对于它们最终流行度影响不大，而其他用户的选择(社交影响力)，更容易使歌曲获得更多的流行度。

社交特征　社交特征指关于网络信息发布者的信息(例如发布者的粉丝数、关注者数、发布的网络信息数量等)。Lerman 等人对 Digg 帖子的社交特征

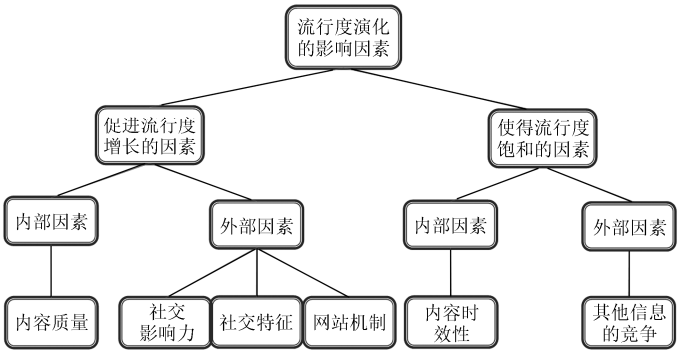

图 3 流行度演化影响因素分类图

做了大量研究。他们发现交际圈更广的用户(即他拥有更多的粉丝和关注者)发布的帖子更容易被推送至 Digg 首页，即使帖子本身并不是很有趣[7]。随后他们进行了更深入的研究，发现那些主要在发布者社交圈外传播的帖子会持续流行，而主要在发布者社交圈内传播的帖子不会很流行[28]。这个发现可以用信息传播来解释：如果一个帖子早期就开始在发布者社交圈外传播，这意味着该帖子拥有更多的级联，会有更多来自于点赞者的粉丝的点赞，随后新的点赞者又会吸引更多的粉丝的点赞，以此类推。所以“社交特征”这个因素对网络信息在早期阶段流行度的获得起了关键作用。

但是，Lerman 的工作中有一些缺陷：拥有更大交际圈用户发布的帖子会更火热，可能是因为帖子的内容更有趣，而不是因为交际圈大小对流行度的影响。为了解决单一变量的问题，Borghol 等人[8] 人工收集了一些 YouTube 视频数据集，该数据集中视频内容的有趣程度是相等的，换言之，这些视频可以被看作彼此的克隆。然后研究社交特征是怎样影响流行度的。他们分别计算克隆数据集和非克隆数据集的社交特征和流行度间的皮尔逊相关系数。发现克隆数据集的皮尔逊相关系数平方值更大。这个结果表明如果不分离社交特征和内容有趣程度这两个因素，会低估社交特征对流行度的影响。另外他们也验证了在视频的早期阶段，社交特征较其他因素对流行度增长的影响相对更大。

网站机制 另一个重要的外部因素是网站机制(推送，搜索，引入链接)。在很多社交网站上，当某网络信息在初期阶段积累了一定数量的流行度时，它会收到“奖励”：Digg 上，短时间内获得相当多点赞数的帖子会从 upcoming 板块被推送至首页；YouTube 上，流行视频会在搜索提示的列表里排名更靠前。因此，网络信息的可见度会被提升，流行度也会快速增长，导致了“富者更富”的现象。

很多网站会为网络信息提供引用(referrer，引入到网络信息的链接)。以 YouTube 视频为例：Google, Facebook，某些邮箱都会有引导用户到 YouTube 视频的链接。Cha 等人[18]发现 YouTube 所有视频中，有 47%都含有外部网站的引入链接，由引用带来的流行度可占总流行度的 90%。但是 Cha 等人仅是粗略地描述了引用对流行度增长的作用。Broxton 等人研究了引用具体是怎样影响病毒型视频流行度的。他们将视频按其是否有引用进行分类，并监视单个 YouTube 视频流行度随时间演化的情况。他们发现相较于被引用得少的病毒型视频，被引用得多的视频的流行度可以更快地到达峰值且更快地由峰值跌落，且被引用得多的视频在短期内会获得更高的观看数。

Borghol 等人[8]还认为，网络信息早期流行度也是影响流行度最重要的因素之一，因为根据“富者更富”理论，某网络信息获得新的流行度的概率正比于该信息已有的流行度。但是，其背后隐藏的因素其实是社交影响力和网站机制。人们偏向于观看点击量多的视频，这是受其他人影响，是社交影响力的作用。网站也偏向于把点击量多的视频放在显眼的位置，导致更多的人观看。

5.2 使得流行度饱和的因素

网络信息流行度会因来自其他信息的竞争[2](外部因素)和网络信息内容本身具有时效性[11](内部因素)而逐渐饱和。Wu 和 Huberman 用一个与时间独立的衰减因子 r 来表征来自其他网络信息的竞争，当 $t=0$ 时，$r=1$；当 t 趋向于无穷时，r 趋近于 0；其他情况，r 服从衰减的指数函数。他们发现在 Digg 帖子上传的前 22 h 内，r 衰减得非常快：快于幂律衰减慢于指数衰减。Szabo 和 Huberman[11]发现 Digg 帖子流行度比 YouTube 视频饱和得快，原因是很多 Digg 帖子涉及新闻、当下潮流，这些内容更具时效性。

5.3 小结

图 4 可以帮助快速理解网络信息流行度在生命周期内各个阶段的主导因素。大部分网络信息的流行度演化可以划分为 3 个阶段：初期缓慢增长阶段、中期快速增长阶段、后期逐渐饱和阶段(注意图 4 为累积流行度，对累积流行度做微分便为图 2 中的单位时间内增长的流行度演化情况。图 4 中的 3 个阶段，实际上也分别对应着图 2 的爆发、顶峰、衰老阶段)。在初期阶段，起主导因素的是社交特征，上传者的粉丝数多少起了决定性作用。随后，被网站置顶、推送的网络信息会迅速获得更多的流行度，形成“富者更富”的效应，这一阶段社交影响力也起到了重要作用。最后因为来自其他网络信息的竞争，该网络信息流行度逐渐饱和。

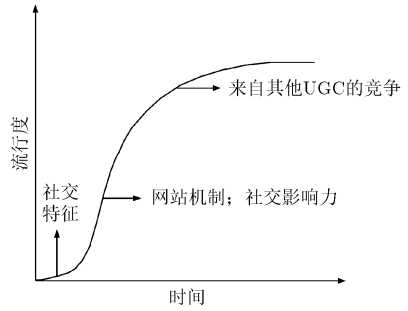

图 4 流行度演化各阶段主导因素

6　流行度演化建模与预测

研究网络信息流行度演化问题，最重要的是如何建模和预测它。预测流行度演化的主要任务是准确地估计出给定网络信息在未来某个时间点的流行度值[3,15,29−41]或者预测它变流行的可能性[42]。这方面的研究成果很多，这些成果涉及的方法可以归纳为 3 类：基于早期流行度、基于影响因素、基于级联传播。下面分别从代表理论或模型、工作原理、典型特点、适用范围等方面对这 3 类方法进行综述。

6.1 基于早期流行度的方法

研究[3,18,23,39]表明演化早期流行度和晚期流行度存在很强的关系，因此早期流行度可以作为预测因子预测未来流行度，建立回归方程[3,29,30,42]或聚、分类算法[4,21,43,44]。该类型的方法可描述为

$$y_i(t_f)=f(y_i(t_h),y_i(t_h-1),\cdots,y_i(1)) \tag{2}$$

$$t_h<t_c\le t_f\le L_i \tag{3}$$

其中，i 为给定网络信息，t_c 为当前时刻，L_i 为流行度演化生命长度，t_h 为历史时刻，t_f 为未来时刻(即待预测的时刻)，$y_i(t_f)$ 为待预测的未来流行度，$y_i(t_h)$，　$y_i(t_h-1),\cdots,y_i(1)$ 为早期流行度。

SH 模型　SH 模型是流行度演化预测的经典模型之一。文献[11]发现，如果对早期和未来流行度做对数处理，二者会存在很强的线性关系。基于这个发现，文献[11]提出了一个一元回归模型。

$$\ln y_i\left(t_f\right)=\ln r\left(t_c,t_f\right)+\ln y_i\left(t_c\right)+\xi_i\left(t_c,t_f\right) \tag{4}$$

其中，$\ln r(t_c,t_f)$ 为从样本集中训练出来的待预测时刻和当前时刻流行度的关系因子，相当于 $y=x+b$ 中的参数 b，$\xi_i(t_c,t_f)$ 为噪声，可以理解为误差。

运用 SH 模型做预测，对于那些受到短暂关注的网络信息会预测得比较准确，而对于长时间都很流行的网络信息会有很大的误差[11]。

SH 模型在预测那些具有不同流行度演化模式的网络信息也会不准确，为了克服这个缺点，Pinto 等人[30]将 SH 模型扩展成多元回归模型，以多个历史时间点的流行度作为自变量，并且考虑了流行度演化相似性的特征。该模型对不同历史时间点的流行度分配了不同的权重，这些权重可以通过最小化样本集的 mRSE(mean Relative Squared Error)训练出来。相似性则度量了待预测网络信息与样本集的相似程度。

Chang 等人[29]为了预测网络电视剧的流行度，先分析了相邻发布的网络电视剧流行度之间有很强的线性关系，基于此线性关系，作者提出了 TAR (Transfer AutoRegressive)模型，该模型认为，每集电视剧的观众(观众数量即为流行度)主要有两种群体组成：超级粉丝和随机观看者。某集电视剧的超级粉丝和随机观看者会分别依一定的概率转换成下一集电视剧超级粉丝。而每集的随机观看者的数量会随时间的推移而减少。基于这些假设，作者通过最小化样本集的 RSE(Relative Squared Error)估计出模型中的各概率初始参数。

传染病理论　除 SH 模型外，基于早期流行度方法的另一种代表性理论称为传染病理论，多数学者运用传染病理论刻画信息传播的过程，但由于流行度演化本身也是信息传播的一种体现，所以传染病理论同样适用于建模流行度演化。这里我们介绍传染病理论中的基础模型 SI(Susceptible-Infected)模型。

SI 模型假设，传播网络中每个节点有两种状态：S(易感染状态)和 I(被感染状态)，且每个 S 节点依一定的感染概率 β 被 I 节点感染，且一旦被感染永远处于感染状态。当 SI 模型应用在流行度演化上时，被感染节点数量即为流行度。SI 模型方程如式(5)：

$$\frac{\mathrm{d}\,Y(t)}{\mathrm{d}t}=\beta(N-Y(t))Y(t) \tag{5}$$

其中，$Y(t)$ 为当前时刻 t，I 节点总数量，每个 I 节点可使 $\beta(N-Y(t))$ 个 S 节点被感染，所以 $\beta(N-Y(t))Y(t)$ 即为当前时刻 t 的 I 节点增长速率，即流行度增长速率。由于 SI 模型假设感染率 β 为固定值，但实际上感染率不可能一直不变，比如对于一则新闻来说，随时间的推移，新闻的时效性降低，感染率也一定是降低的。因此，Matsubara 等人[25]在 SI 模型的基础上提出了感染率可变的 SpikeM 模型，用来建模博客的流行度演化。

基于早期流行度的方法的优点是该类方法可以应用于所有类型的网络信息，因为所有网络信息都有历史流行度。其缺点是预测不及时，必须等待历史流行度数据出来后再预测；另外，为获取历史流行度需不断地向网站发送请求，数据的获取和维护比较困难。

6.2 基于影响因素的方法

第 5 节介绍了影响网络信息流行度的因素，很多工作考虑了这些因素，运用回归建模[8,13]、分类学习[9,21]、随机过程[2,40]等对网络信息未来流行度进行预测。

Bandari 等人[13]对新闻类网络信息流行度进行预测，他们认为对于新闻这种具有时效性的网络信息，利用基于早期流行度的方法等其发布后才预测，会导致预测不及时。而新闻的文字内容在其流行度提升中起到了关键作用，利用新闻内容在其发布前就能预测。于是他们考虑了内部影响因素，定义了

新闻内容的 4 个特征：新闻源(news source，该新闻是哪个机构发布的，New York Times, USA Today 等)、新闻分类(category，政治类、娱乐类等)、主观性(subjectivity，新闻的主观程度和客观程度)、有名实体(named entities，新闻中是否包含有名实体，“贾斯汀•比伯”、“奥巴马”等)。然后将这 4 个特性作为自变量、未来流行度作为因变量，建立了对数回归方程。该方法对于预测未来流行度的精确值并不是很准确，但是在给定预测范围的情况下准确度可以达到 84%。他们还发现 4 个特征中，新闻源对未来流行度的影响最大。

He 等人[36]考虑了另外两种因素：时序因素和社交影响力。时序因素可以通过用户评论的时间戳获得，社交影响力可以通过评论的用户挖掘出来。他们将用户评论用一个双向图(U∪P, E)表示，其中 U 和 P 代表网络信息的上传者和网络信息，E 为边、代表用户评论。每条边都可以为其分配一个权值，该权值表示该评论对未来流行度的贡献。越新的评论对未来流行度的贡献越大，这样就可以用一个单调递减的指数函数来建模边的权值。基于这个双向图，他们提出了一个正则排序算法，该算法考虑了时序因素、社交影响力因素、当前流行度因素来预测网络信息未来流行度。

有些研究者认为，只预测网络信息未来某个时刻的流行度是不够的，有时候人们更关心该网络信息是否会变流行。Kong[9]等人对 Twitter 的 Hashtag 进行了学习。他们提出了两个基本任务：该 Hashtag 是否会在不久的未来流行起来；如果会流行，会多流行。为了方便预测，他们将 Hashtag 流行度的生命周期分为 4 个阶段：激活萌芽期(active)、突发增长期(bursting)、衰落期(off-burst)和消亡期(inactive)。作者考虑 Hashtag 内容因素(内容倾向性、Hashtag 长度、所属类别等)、社交特征(用户粉丝数等)、网络结构特征(节点度、图密度等)、早期流行度等因素，将这些因素作为机器学习算法或模型的输入。作者将第 1 个任务(是否会流行)看成二分分类问题，他们用支持向量机对样本集进行训练，F1-Score 作为分类准确度的评估。为解决第 2 个任务(会有多流行)，作者采用 5 个回归模型(linear regression, classification and regression tree, Gaussian process regression, support vector regression 和 neural network regression)对样本集进行拟合，用 RMSE(Root Mean Square Error)对预测结果进行评估。作者发现，对于任务 1，早期流行度是所有因素中对分类准确度影响最大的，且影响程度随着预测时间的推移而增加；对于任务 2，在激活萌芽期 6 h 前做预测，Gaussian process regression 模型效果最好，6 h 后 support vector regression 模型效果最后。

之前介绍的工作主要是对网络信息流行度增长过程的建模和预测，文献[2]考虑了流行度饱和的过程。他们认为导致网络信息流行度饱和的原因主要来自于人们容易对旧事物失去兴趣或新网络信息的竞争。他们用随机过程对网络信息流行度增长和饱和的过程进行建模：

$$Y(t)=(1+X(t))Y(t-1) \tag{6}$$

其中，$Y(t)$ 表示 t 时刻的总流行度，$X(1),X(2),\cdots$ 为服从均值为 μ，方差为 σ^2 的独立同分布的随机变量。为保证 $Y(t)$ 随时间增长，$X(t)$ 为正数。为了对饱和过程进行建模，他们又引入了一个与时间独立的衰减因子 r，当 $t=0$ 时，$r=1$；当 t 趋向于无穷 r 趋近于 0，最终表达式为

$$Y(t)=(1+r(t)X(t))Y(t-1) \tag{7}$$

由于不同类型的网络信息具有不同的影响因素，所以基于影响因素的方法通常不具通用性；但选择社交影响力和社交特征做预测具有通用性。

6.3 基于级联传播的方法

网络信息流行度演化的本质上是信息传播[45−47]的过程，而信息传播依赖于网络拓扑结构及邻居节点间的相互作用。级联传播[48,49]是一种信息传播过程，在该过程中被某信息感染的节点 u 依一定概率感染其邻接节点 v，若节点 v 被感染，它又依一定概率感染自己的邻接节点，依次类推。部分工作将级联传播应用于网络拓扑结构明显(好友关系明确)的社交网络的网络信息流行度演化预测上，如 Twitter 推文[37,48−50]，人人网视频[33]，新浪微博[39]。该类方法可描述为

$$\begin{aligned} y_i(t_f)=f(&y_i(t_h),y_i(t_h-1),\cdots,y_i(1),\\ &N_i(t_h),N_i(t_h-1),\cdots,N_i(1)) \end{aligned} \tag{8}$$

$$t_h<t_c\le t_f\le L_i \tag{9}$$

其中，$N_i(t)$ 为传播网络，$N_i(t_h),N_i(t_h-1),\cdots,N_i(1)$ 为历史传播网络。

Kupavskii 等人[49]研究了 Twitter 中 retweet 级联，他们用有向图表示 retweet 级联，图的节点是对推文进行 retweet 的用户，如果节点 B 是节点 A 的粉丝，且 A 是 B 的所有关注者中第 1 个 retweet 某推文的用户，那么就用一条从 A 指向 B 的边连接用户 A 和 B。他们用传染病模型对 retweet 级联的增长过程进行建模。为了预测某推文的 retweet 数量，即流行度，他们还考虑了初始节点的社交特征(粉丝

数、好友数等)、内容特征(推文长度、Hashtag 数等)、级联图中节点的 PageRank，并用该 PageRank 作为用户影响力的度量。利用这些特征和 GBDT (Gradient Boosted Decision Tree)模型进行机器学习，预测推文流行度。

Li 等人[33]对被用户从视频网站分享至人人网的视频流行度进行了研究，发现其流行度普遍具有早期峰值和爆发性增长。他们发现用传统的基于早期流行度的方法对该类视频流行度进行预测时准确度较低。原因是这类视频底层传播结构差异性很大，不能忽视底层传播结构对其流行度发展趋势产生的影响，且该类视频后期的流行度与前期的流行度相关性较小。于是他们提出了基于级联传播的 SoVP (Social network assisted Video Prediction)方法。SoVP 方法综合考虑了在线社交网站的底层传播结构及视频本身吸引力两点，利用用户的分享率(用户分享视频数量与其观看数量的比值)、用户 A 对于用户 B 所分享视频的观看率(用户 A 观看用户 B 所分享的视频数量与用户 B 分享的全部视频数量的比值)等对社交网络底层结构进行量化。实验表明 SoVP 方法在预测该类视频的早期峰值和爆发性增长时准确度超过基于早期流行度的方法，整体表现比较理想。但 SoVP 复杂性较高、可拓展性不强。

Bao 等人[39]考虑了网络结构特性(连接度(link density)、传播深度(diffusion depth))对新浪微博流行度的影响，对 SH 模型进行了改进，改进后的模型的 RMSE 和 MAE(Mean Absolute Error)与 SH 模型比较显著减小。他们首先测定了微博最终流行度和连接度、传播深度之间的关系。结果显示微博的最终流行度和连接度之间存在着很强的负相关性，和传播深度之间存在着很强的正相关性。这表明低连接度和高传播深度的群体更容易促进微博流行度的提升。他们还发现早期微博转发者的结构特性可以反映出网络信息的最终流行度。

自兴奋点计理论　一种可以刻画级联传播方法的理论称为自兴奋点计(self-exciting point process)理论，自兴奋点计理论的原理和波松过程类似，累计随机事件发生的次数，这里的随机事件是网络信息获得一个流行度。但和波松过程不同的是，自兴奋点计过程假设事件发生的密度(density)可以随时间变化：历史事件发生的次数的多少会影响未来的事件发生次数的多少，这刚好符合流行度增长过程中“富者更富”的特征。基于自兴奋点计理论，Zhao[47]提出了 SEISMIC(Self-Exciting Model of Information Cascades)模型。

$$\lambda_t = \lim_{\Delta t \to 0} \frac{P(y(t+\Delta t) - y(t) = 1)}{\Delta t} \tag{10}$$

其中，λ_t 为事件发生(用户转发某信息，即流行度增长)的密度，流行度增长的密度又取决于节点感染速率 p_t，每个事件(流行度增长)发生的时间 t_n，节点度 d_n 和用户反应时间 $\phi(t)$ (指用户从看到网络信息到转发该信息的时间)。

$$\lambda_t = p_t \sum_{t_n < t, n \geq 0} d_n \phi(t - t_n),\ t \geq t_0 \tag{11}$$

其中，$\sum_{t_n<t,n\geq 0} d_n\phi(t-t_n)$ 描述了用户看到某信息，该事件发生的密度，再乘以 p_t，为用户看到某信息并转发发生的密度。其中 p_t 和 ϕ 需根据样本集进行参数估计。

基于级联传播的方法于近两年备受研究者的关注，从网络信息流行度增长潜在过程的角度出发，考虑网络结构特性、级联传播关系。可以准确地预测出网络信息流行度的峰值。缺点是也必须等到历史级联出现后才能预测，可能导致预测不及时。另外，在考虑级联传播关系时，往往需要挖掘网络中两两用户之间的交互行为，导致预测算法时间复杂度差，不适用于大规模在线社交网络。

表 2 总结了 3 种网络信息流行度预测方法的原理、应用范围、优缺点。

7　总结与展望

近几年，研究人员在流行度演化分析与预测方面已经取得丰硕的理论成果：网络信息流行度服从幂律分布或者对数正态分布，具有长尾特性，可利用该特性设计缓存机制[18,39,51]；流行度演化总体符合上升、下降态势；影响因素中外部因素(社交影响力、网站机制等)对网络信息流行度演化具有更重要的影响，因此，对流行度演化进行建模和预测时应更多考虑外部因素，而非内部因素；网络信息流行度增长本质是信息传播的过程，从网络结构特性和级联传播入手，有助于预测准确度的提高。

从预测目标来看，已有工作主要是从“量”的角度出发，预测未来某个时间点网络信息的观看量、点赞量、转发量等。从预测对象来看，已有成果涉及各种类型社交网站的网络信息：人人网视频、YouTube 视频、Wikipedia 词条、Twitter 推文等，目前对于短期流行的网络信息可以取得相当高的预测准确度，而对于流行度保持长期增长的网络信息预测效果不佳[11,33]。从预测方法来看，已有研究大多基于早期流行度、影响因素或级联传播对流行度建模，运用数理统计、随机过程、机器学习、时间序列方法进行预测。从评估效果来看，研究者主要用 MSE(Mean Square Error), RAE(Relative

表 2 3 种预测方法总结

方法	原理	应用范围	优点	缺点
基于早期流行度	早期流行度和晚期流行度存在很强的关系，早期流行度可以作为预测因子预测未来流行度	所有网络信息，但对于长期流行和具有突发性峰值的网络信息预测不准确	可操作性强	可能会导致预测不及时；数据的获取和维护困难
基于影响因素	考虑影响网络信息流行度的因素	考虑不同因素的模型适用不同网络信息：考虑内容因素的适用为本类网络信息；考虑社交影响力和社交特征的适用于所有类型	预测准确度高	不同类型的网络信息具有不同的影响因素，该类方法通常不具通用性
基于级联传播	将网络结构和好友关系作为主要考虑因素	适用于网络结构明显的网络信息	考虑网络信息流行度增长的潜在过程，预测准确度高	可能会导致预测不及时。这类方法通常时间复杂度高，无法应用于大规模网络

Absolute Error), MAPE(Mean Absolute Percentage Error)和 mRSE(mean Relative Squared Error)对预测准确度进行度量，或者采用区间的方式给出预测值的范围。虽然网络信息流行度预测已经得以深入研究，我们认为至少还有以下 4 个趋势有待深入研究和探索：

(1)流行度演化的早期工作[1,53]主要单一地研究流行度在“量”上如何随时间演化。但随着各种具有强社交关系的社交媒体(Facebook, Twitter，微信等)的普及，级联传播成为信息流动和传播的主要方式。越来越多的流行度演化工作也开始结合级联传播进行研究[24,48,49]，如级联传播模式下，流行度演化是否有新的模式；网络信息通过级联传播获得的流行度(即级联的大小)是否可以预测[48,49]。所以，级联传播模式下的流行度演化是未来研究的主要方向之一。

(2)目前，流行度演化现有工作主要以时间为自变量、流行度为因变量，进行建模。主流工作的逆向问题(将时间作为因变量)，比如何时达到流行度特定值、何时流行度经历爆发等等，鲜有研究，但也是十分重要且具有应用价值的问题。因此流行度演化的逆向问题是未来的研究方向之一。

(3)对于流行度保持长期增长的网络信息，其未来流行度的预测仍是难题。因为其流行度演化掺杂了更多的不确定性因素，更容易受到外部事件的影响。目前仍没有一种通用的模型或方法可以准确地预测这类网络信息流行度。但是这类网络信息往往蕴藏着更多的商业价值，更需要舆情监管。因此流行度保持长期增长的网络信息流行度演化更值得我们建模、预测。

(4)社交网络数据是一种关系复杂、实时变化的海量数据。主流工作通常通过特征提取的方法，考虑两两用户之间的交互、单个用户影响力、用户评论等多方面因素，但这些工作往往只侧重于准确度的提升，却忽略了算法的时间复杂度，并不适用于大规模社交网络[33]。因此如何降低时间复杂度、提出适合社交网络海量数据的流行度演化模型是亟待解决的重要问题。

参 考 文 献

[1] WU F and HUBERMAN B A. Popularity, novelty and attention[C]. Proceedings of the 9th ACM Conference on Electronic Commerce, Chicago, 2008: 240–245. doi: 10.1145/1386790.1386828.

[2] WU F and HUBERMAN B A. Novelty and collective attention[J]. *Proceedings of the National Academy of Sciences*, 2007, 104(45): 17599–17601. doi: 10.1073/pnas.0704916104.

[3] HONG L, DAN O, and Davison B D. Predicting popular messages in twitter[C]. Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, 2011: 57–58. doi: 10.1145/1963192.1963222.

[4] YANG J and LESKOVEC J. Patterns of temporal variation in online media[C]. Proceedings of the 4th ACM International Conference on Web Search and Data Mining, Hong Kong, 2011: 177–186. doi: 10.1145/1935826.1935863.

[5] 吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014, 37(4): 735–751.
WU Xindong, LI Yi, and LI Lei. Influence analysis of online social networks[J]. *Chinese Journal of Computers*, 2014, 37(4): 735–751.

[6] KEMPE D, KLEINBERG J, and TARDOS É. Maximizing the spread of influence through a social network[C]. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC, 2003: 137–146. doi: 10.1145/956750.956769.

[7] LERMAN K. Social information processing in news aggregation[J]. *IEEE Internet Computing*, 2007, 11(6): 16–28. doi 10.1109/mic.2007.136.

[8] BORGHOL Y, ARDON S, CARLSSON N, *et al*. The untold

story of the clones: content-agnostic factors that impact YouTube video popularity[C]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, 2012: 1186–1194. doi 10.1145/2339530.2339717.

[9] KONG S, MEI Q, FENG L, *et al*. Predicting bursts and popularity of hashtags in real-time[C]. Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, 2014: 927–930. doi: 10.1145/2600428.2609476.

[10] HE X, GAO M, KAN M Y, *et al*. Predicting the popularity of Web 2.0 items based on user comments[C]. Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, 2014: 233–242. doi 10.1145/2600428.2609558.

[11] SZABO G and HUBERMAN B A. Predicting the popularity of online content[J]. *Communications of the ACM*, 2010, 53(8): 80–88. doi: 10.1145/1787234.1787254.

[12] AGARWAL N, LIU H, TANG L, *et al*. Identifying the influential bloggers in a community[C]. Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, 2008: 207–218. doi: 10.1145/1341531.1341559.

[13] BANDARI R, ASUR S, and HUBERMAN B A. The pulse of news in social media: Forecasting popularity[C]. Proceedings of the 6th International AAAI Conference on Web and Social Media, Dublin, 2012: 26–33.

[14] CHATZOPOULOU G, SHENG C, and FALOUTSOS M. A first step towards understanding popularity in YouTube[C]. INFOCOM IEEE Conference on Computer Communications Workshops, San Diego, 2010: 1–6. doi 10.1109/infcomw.2010.5466701.

[15] ROY S D, MEI T, ZENG W, *et al*. Towards cross-domain learning for social video popularity prediction[J]. *IEEE Transactions on Multimedia*, 2013, 15(6): 1255–1267. doi 10.1109/tmm.2013.2265079.

[16] JAMALI S and RANGWALA H. Digging Digg: Comment mining, popularity prediction, and social network analysis[C]. IEEE International Conference on Web Information Systems and Mining, Shanghai, 2009: 32–38. doi 10.1109/wism.2009.15.

[17] YIN P, LUO P, WANG M, *et al*. A straw shows which way the wind blows: ranking potentially popular items from early votes[C]. Proceedings of the 5th ACM International Conference on Web Search and Data Mining, Seattle, 2012: 623–632. doi: 10.1145/2124295.2124370.

[18] CHA M, KWAK H, RODRIGUEZ P, *et al*. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system[C]. Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, New York, 2007: 1–14. doi: 10.1145/1298306.1298309.

[19] CRANE R and SORNETTE D. Robust dynamic classes revealed by measuring the response function of a social system[J]. *Proceedings of the National Academy of Sciences*, 2008, 105(41): 15649–15653. doi: 10.1073/pnas.0803685105.

[20] CRANE R and SORNETTE D. Viral, quality, junk videos on YouTube: Separating content from noise in an information-rich environment[C]. AAAI Spring Symposium 2008: Social Information Processing, Stanford, 2008: 18–20.

[21] FIGUEIREDO F. On the prediction of popularity of trends and hits for user generated videos[C]. Proceedings of the 6th ACM International Conference on Web Search and Data Mining, San Francisco, 2013: 741–746. doi: 10.1145/2433396.2433489.

[22] ASUR S, HUBERMAN B A, SZABO G, *et al*. Trends in social media: Persistence and decay[OL]. Available at SSRN 1755748, 2011. doi: 10.2139/ssrn.1755748.

[23] FIGUEIREDO F, BENEVENUTO F, and ALMEIDA J M. The tube over time: Characterizing popularity growth of YouTube videos[C]. Proceedings of the 4th ACM International Conference on Web Search and Data Mining, Hong Kong, 2011: 745–754. doi: 10.1145/1935826.1935925.

[24] CHENG J, ADAMIC L A, KLEINBERG J M, *et al*. Do cascades recur?[C]. Proceedings of the 25th International Conference on World Wide Web. Montreal, 2016: 671–681. doi: 10.1145/2872427.2882993.

[25] MATSUBARA Y, SAKURAI Y, PRAKASH B A, *et al*. Rise and fall patterns of information diffusion: Model and implications[C]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, 2012: 6–14. doi: 10.1145/2339530.2339537.

[26] YU H, XIE L, and SANNER S. The lifecycle of a YouTube video: Phases, content and popularity[C]. 9th International AAAI Conference on Web and Social Media, Oxford, 2015.

[27] SALGANIK M J, DODDS P S, and WATTS D J. Experimental study of inequality and unpredictability in an artificial cultural market[J]. *Science*, 2006, 311(5762): 854–856. doi: 10.1126/science.1121066.

[28] LERMAN K and GALSTYAN A. Analysis of social voting patterns on digg[C]. Proceedings of the First Workshop on Online Social Networks, Seattle, 2008: 7–12. doi: 10.1145/1397735.1397738.

[29] CHANG B, ZHU H, GE Y, *et al*. Predicting the popularity of online serials with autoregressive models[C]. Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, Shanghai, 2014: 1339–1348. doi: 10.1145/2661829.2662055.

[30] PINTO H, ALMEIDA J M, and GONCALVES M A. Using early view patterns to predict the popularity of YouTube videos[C]. Proceedings of the 6th ACM International Conference on Web Search and Data Mining, San Francisco, 2013: 365–374. doi: 10.1145/2433396.2433443.

[31] TAN Z, WANG Y, ZHANG Y, *et al*. A novel time series approach for predicting the long-term popularity of online videos[J]. *IEEE Transactions on Broadcasting*, 2016, 62(2): 436–445. doi: 10.1109/TBC.2016.2540522.

[32] WU J, ZHOU Y, CHIU D M, *et al*. Modeling dynamics of online video popularity[J]. *IEEE Transactions on Multimedia*, 2016, 18(9): 1882–1895. doi: 10.1109/TMM.2016.2579600.

[33] LI H, MA X, WANG F, *et al*. On popularity prediction of videos shared in online social networks[C]. Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, San Francisco, 2013: 169–178. doi: 10.1145/2505515.2505523.

[34] TATAR A, LEGUAY J, ANTONIADIS P, *et al*. Predicting the popularity of online articles based on user comments[C]. Proceedings of the International Conference on Web Intelligence, Mining and Semantics, Sogndal, 2011: 1–8. doi: 10.1145/1988688.1988766.

[35] SIERSDORFER S, CHELARU S, NEJDL W, *et al*. How useful are your comments?: Analyzing and predicting YouTube comments and comment ratings[C]. Proceedings of the 19th International Conference on World Wide Web, Raleigh, 2010: 891–900. doi: 10.1145/1772690.1772781.

[36] HE X, GAO M, KAN M Y, *et al*. Predicting the popularity of web 2.0 items based on user comments[C]. Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, 2014: 233–242. doi: 10.1145/2600428.2609558.

[37] MA Z, SUN A, and CONG G. On predicting the popularity of newly emerging hashtags in twitter[J]. *Journal of the American Society for Information Science and Technology*, 2013, 64(7): 1399–1410. doi: 10.1002/asi.22844.

[38] KHOSLA A, DAS SARMA A, and HAMID R. What makes an image popular?[C]. Proceedings of the 23rd International Conference on World Wide Web, Seoul, 2014: 867–876. doi 10.1145/2566486.2567996.

[39] BAO P, SHEN H W, HUANG J, *et al*. Popularity prediction in microblogging network: A case study on sina weibo[C]. Proceedings of the 22nd International Conference on World Wide Web Companion, Rio de Janeiro, 2013: 177–178. doi 10.1145/2487788.2487877.

[40] LERMAN K and HOGG T. Using a model of social dynamics to predict popularity of news[C]. Proceedings of the 19th International Conference on World Wide Web, Raleigh, 2010: 621–630. doi: 10.1145/1772690.1772754.

[41] ZHAO Q, ERDOGDU MA, HE HY, *et al*. SEISMIC: A self-exciting point process model for predicting tweet popularity[C]. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 2015: 1513–1522. doi: 10.1145/2783258.2783401.

[42] LEE J G, MOON S, and SALAMATIAN K. An approach to model and predict the popularity of online contents with explanatory factors[C]. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, London, 2010, 1: 623–630. doi: 10.1109/WI-IAT.2010. 209.

[43] AHMED M, SPAGNA S, HUICI F, *et al*. A peek into the future predicting the evolution of popularity in user generated content[C]. Proceedings of the 6th ACM International Conference on Web Search and Data Mining, San Francisco, 2013: 607–616. doi: 10.1145/2433396.2433473.

[44] FIGUEIREDO F, ALMEIDA J M, GONCALVES M A, *et al*. TrendLearner: Early prediction of popularity trends of user generated content[J]. *Information Sciences*, 2016: 349–350, 172–187. doi: 10.1016/j.ins.2016.02.025.

[45] GRUHL D, GUHA R, LIBEN-NOWELL D, *et al*. Information diffusion through blogspace[C]. Proceedings of the 13th International Conference on World Wide Web, New York, 2004: 491–501. doi: 10.1145/988672.988739.

[46] YANG J and LESKOVEC J. Modeling information diffusion in implicit networks[C]. IEEE 10th International Conference on Data Mining, Sydney, 2010: 599–608. doi: 10.1109/icdm.2010.22.

[47] ZHAO J, WU J, FENG X, *et al*. Information propagation in online social networks: A tie-strength perspective[J]. *Knowledge and Information Systems*, 2012, 32(3): 589–608. doi: 10.1007/s10115-011-0445-x.

[48] CHENG J, ADAMIC L, DOW P, *et al*. Can cascades be predicted?[C]. Proceedings of the 23rd International Conference on World Wide Web, Seoul, 2014: 925–936. doi: 10.1145/2566486.2567997.

[49] KUPAVSKII A, OSTROUMOVA L, UMNOV A, *et al*. Prediction of retweet cascade size over time[C]. Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Sheraton, 2012: 2335–2338. doi: 10.1145/2396761.2398634.

[50] ARDON S, BAGCHI A, MAHANTI A, *et al*. Spatio-temporal and events based analysis of topic popularity in twitter[C]. Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, San Francisco, 2013: 219–228. doi: 10.1145/2505515.2505525.

[51] WANG S, YAN Z, HU X, *et al*. Burst time prediction in cascades[C]. 29th AAAI Conference on Artificial Intelligence, Austin, 2015: 325–331.

胡 颖： 女，1990 年生，博士生，研究方向为在线社交网络和机器学习.

胡长军： 男，1963 年生，教授，研究方向为在线社交网络、并行计算.

傅树深： 男，1992 年生，硕士生，研究方向为在线社交网络.

黄建一： 男，1989 年生，博士生，研究方向为在线社交网络.