

基于 Feeds 的社交网络活跃度分析

何 昂 阳¹ 孙 鲁 敬^{2,3} 杨 家 海^{2,3}

(汕头职业技术学院 汕头 515078)¹ (清华大学网络科学与网络空间研究院 北京 100084)²

(清华大学清华信息科学与技术国家实验室 北京 100084)³

摘 要 从用户产生和消费 Feeds 的角度分析社交网络变得不活跃的原因,通过分析人人网某大学社区用户长周期的 Feeds 行为来探讨该社区用户活跃度的变化。通过对用户活跃性周期和 Feeds 时间间隔的分析,发现越来越多的用户产生 Feeds 的活跃度在下降,并导致其他用户接收到的信息流的流速和多样性下降。社交网络用户由于各种原因离开或变得不活跃,并通过信息流对其朋友圈形成负向反馈,这可能是社交网络变得不活跃的深层原因。模拟实验表明,30%的初始不活跃用户会使得整个社区的信息流快速下降,并导致整个社区不活跃。

关键词 社交网络,活跃度分析,Feeds 行为,时间间隔分析

中图法分类号 TP399 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.031

OSN Activities Analysis Based on User Feeds

HE Chan-yang¹ SUN Lu-jing^{2,3} YANG Jia-hai^{2,3}

(Shantou Polytechnic, Shantou 515078, China)¹

(Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China)²

(Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China)³

Abstract This paper explored how one social network becomes inactive. We investigated the change of users' activeness by analyzing the activeness period and Feeds inter-event time of users in a specific OSN. Our findings reveal that these users decrease their activity frequency or depart from the social network for various reasons and increase the interval time of Feeds, resulting in the decrease of information flow in velocity and diversity for the whole community. As a result, active users will feel the inactivity from their friends and increase the probability of being inactive as a feedback, which may be the underlying reason for the inactivity of OSN. Our simulation experiment shows that when 30% of users become inactive in generating Feeds, the whole community will be affected and collapse in a short time.

Keywords OSN, Activities analysis, Feeds behaviors, Inter-event time analysis

1 引言

社交网络的发展不可避免地带来了同类竞争的问题,一些用户在不同社交网络间的迁移等行为会给另外一些社交网络带来活跃度下降的问题^[1],这需从用户行为的角度深入探讨社交网络变得不活跃的原因。

社交网络的核心目的之一就是要保持社区的活跃,留住活跃用户,提高用户粘性。对于这部分工作,现有研究主要关注社交网络繁荣的一面,如用户、社区、网络的增长情况^[2-4],而对于社交网络用户活跃度下降及其离开社交网络等行为的分析较少。从现有的文献来看,对用户不活跃的原因的研究主要是从用户的好友圈出发,分析用户好友圈的行为带来的影响以及联动效应。Wu 对用户离开社交网络的原因进行了研究,他发现用户具有“同进共退”的特性。用户通常是和朋

友一起离开社交网络,并且一个用户离开的概率可以通过其拥有的活跃好友数量和其局部关系图来预测^[5]。Dasgupta 通过研究移动电话网中用户的离开行为,发现用户离开的概率与其好友圈中已经离开的好友数相关^[6]。Richter 则进一步研究了用户离开或不活跃与其所在社区的关系,发现个人用户行为或强烈地受到社区中 leader 的影响^[7]。现有研究的立足点是分析社交圈子对用户行为的影响,研究重点在于这种影响和局部拓扑的关系^[8-10],并没有从朋友行为的角度来研究这种影响的深层作用方式。本文在此基础上进一步分析用户间行为是如何通过朋友圈影响用户的活跃状态。

社交网络用户的行为有两种:产生各种各样的 Feeds 来表达自己的兴趣、观点,并以此和朋友进行互动;浏览 Feeds 来了解朋友动态,获取有趣内容,进行交互。现有的研究工作对用户 Feeds 的主要关注点在于其行为特性的分析^[11-13],以

到稿日期:2014-09-05 返修日期:2014-11-02 本文受国家重点基础研究发展计划(2012CB315806),国家自然科学基金(61170211, 61202356, 61161140454),博士点基金(20110002110056, 20130002110058),教育部-中国移动科研基金(MCM20123041)资助。

何昂阳(1966—),男,讲师,主要研究领域为计算机网络、移动互联网等,E-mail:st8081@126.com;孙鲁敬(1988—),男,硕士,主要研究领域为网络测量、社交网络测量及用户行为分析;杨家海(1966—),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络、网络管理与测量、网络安全、云计算等,E-mail:yang@cernet.edu.cn(通信作者)。

及 Feeds 算法改进等^[14,15]。本文从 Feeds 出发,讨论其与社交网络活跃度的关系。用户产生的 Feeds 会通过关系图谱流向其好友主页,从而决定好友看到的内容。当用户离开或者变得不活跃,其产生的内容会变少,其朋友圈接收到的 Feeds 也会受影响,进而感受到圈子的不活跃而形成负向反馈,这可能是社交网络不活跃的深层原因。本文以人人网为研究对象,并以某个大学社区为例,探讨整个社区的用户发 Feeds 的活跃度是否下降,以及用户发 Feeds 活跃度的变化对整个社区用户接收到的 Feeds 流的影响,并以此来分析社交网络用户不活跃的原因。

本文第 2 节介绍采集的数据集情况,并对用户 Feeds 行为做统计分析;第 3 节分析用户产生 Feeds 行为的活跃周期和行为模式以及用户接受到的 Feeds 流的情况;第 4 节总结分析结果,并通过模拟实验来讨论社交网络不活跃的原因;最后对全文进行总结。

2 Feeds 数据集

2.1 采集方法

人人网根据学校进行社区的划分,本文分析采用的是与文献^[16]相同的基于社区的数据采集方法,这种方法采集数据的成本较低,同时能获得一个具有代表性的数据集。我们采集的数据目标是某大学社区采用默认权限用户的图谱和 Feeds 数据。对于一个用户来说,所有的 Feeds 都会以时间顺序保留在个人主页上,如果他采用了默认的权限,所有同社区的用户都可以浏览他的 Feeds。利用这一特性,使用该大学用户作为种子节点,使用广度优先搜索(BFS)采集策略完成图数据的抓取,针对每个用户使用内容爬虫抓取每个用户从注册到 2013 年 12 月 31 日所产生的所有 Feeds 数据。爬虫一共发现了大约 6 万个用户节点,其中 23207 位用户使用了默认开放策略(约 40%),本文抓取了他们所有的 Feeds 数据,约为 900 万条。使用用户的第一条 Feed 作为其注册时间的估算,并依此把用户按照注册年份进行分类,其数量的变化趋势如表 1 所列。

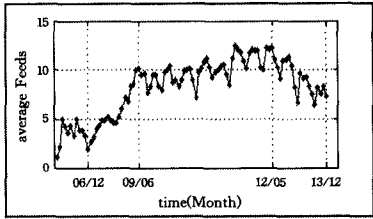
表 1 每年的用户注册数

年份	2006	2007	2008	2009	2010	2011	2012	2013
用户	1716	5117	3606	1911	3054	4048	2155	1598

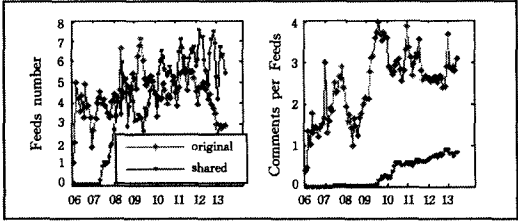
从统计数据可以看到,该大学社区中默认权限用户的数量在 2011 年之后逐渐减少了,这有可能是在 2011 年之后越来越多的人开始关注自己的隐私而选择仅好友可见的权限策略;亦或者是在 2011 年之后整个人人网面临同类社交网络的竞争,对新用户的吸引力开始下降。

2.2 Feeds 统计特性

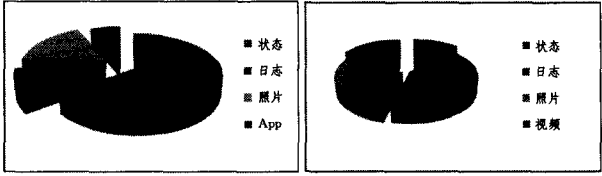
本节对用户的 Feeds 数据做了一些宏观上的统计分析。首先是整个社区每个用户平均产生的 Feeds 数量的变化趋势。图 1(a)统计了每个用户每月平均的 Feeds 数量。用户 Feeds 数量大致经历了 3 个变化区间,在 2006 年 12 月到 2009 年 6 月经历了一个比较明确的上升阶段,在 2009 年 6 月到 2012 年 5 月则相对平稳,2012 年 5 月到 2013 年 12 月呈下降趋势。



(a)用户每月的 Feeds 数量



(b)Feeds 数量及每条 Feed 的交互数



(c)原创型 Feeds

(d)分享型 Feeds

图 1 Feeds 的统计特性

接下来分析 Feeds 的原创性变化情况,原创的内容是社交网络的基石,优质的原创内容能带来大量的转发和交互,促进信息的快速流动。从图 1(b)可以看出,每个用户平均产生的原创性 Feeds 数量在 2011 年下降较快,而分享型的 Feeds 持续增多并保持相对稳定;原创和分享的 Feeds 带来的评论数目都保持着相对稳定,但是从绝对值上看,每条原创 Feed 平均能带来 6 条评论,而每条分享的 Feed 平均只能带来 2 条评论。这说明了原创性内容的重要性,社交网络应该在产品设计和运营上更多地以原创性为导向,鼓励用户创作更多优质的内容,如最近加入的短语音功能就是一次比较有成效的尝试。

除了分析整个社区 Feeds 的数量和原创性,我们还分析了最流行的 Feeds。按照社交网络的功能特点把 Feeds 分为状态(status)、日志(blog)、照片(photo)、视频(video)、应用(App),并分别按照原创和分享 Feeds 统计了它们所占的比例。从图 1(c)可以看出,状态是最活跃原创应用,其次是相册和日志。状态通常是短文本的自我表达,不需要太多的创作精力,因此最能激发原创性。而对于日志和相册而言,他们通常需要更多的精力创作,用户更多的是转发一些特别流行的日志和照片,所以它们在分享型 Feeds 中最流行,如图 1(d)所示。总的来说,社交网络用户更喜欢碎片化的信息创造和内容分享,而需要更多创作精力的日志和相册,通常需要依赖一些专门的公共账号来创作。

3 用户 Feeds 行为分析

3.1 用户产生 Feeds 的活跃周期分析

假设社交网络中的用户将不可避免地经历从“出生”(注册)到“消亡”(离开)的过程,在整个生命周期中其产生 Feeds 的活跃度也会相应地变化。假设用户产生 Feeds 的活跃度会经历上升期、平稳期、下降期、不活跃期。通过分析用户在最

近一段时间产生的 Feeds 量来评价用户当前的状态。对于给定的一个时间快照 T , 设用户自注册到 T 经历的时间为 D , 总共产生了 N 个 Feeds, 假如该用户在最近 20% 的生命时间内 $[0.8D, D]$ 产生了 M 个 Feeds, 则定义 $L = (M/N)/20\%$ 作为当前快照下用户活跃性的标识。如果 L 接近于 1, 那么说明用户在最近的 20% 的时间里产生了整个生命周期 20% 的 Feeds, 处于一个相对稳定的状态; 如果 L 接近于 0, 那么该用户在最近的 20% 时间内基本上没有产生 Feeds 内容, 属于不活跃的状态。以 $L=0$ 和 $L=1$ 为稳定及不活跃状态的参考值, 并取上下 0.3 的波动区间, 表 2 给出每个用户所处活跃阶段对应的 L 经验值。

表 2 L 与活跃周期

	不活跃期	下降期	稳定期	上升期
L	$[0, 0.3)$	$[0.3, 0.7)$	$[0.7, 1.3)$	$[1.3, +)$

通过分析不同时间点下处于不同活跃阶段用户的比例, 可以大致了解整个社区用户产生内容的活跃度的变化情况。按照年份统计了自 2007 年 12 月到 2013 年 12 月 7 个快照下各个活跃期用户的比例, 如图 2 所示。2011 年是一个转折点, 在 2011 年之前处于稳定期和上升期用户的比例较高且相对稳定, 2011 年之后不活跃期和下降期用户的比例逐年增加。人人网是以学校为基础的社交网络, 大学生毕业后会走向职场, 会减少他们对校园型社交网络的精力。另外一个可能的原因就是 2011 年后微博开始广泛流行起来, 人人网面临着新型社交网络的竞争。

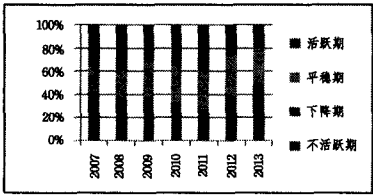


图 2 各个活跃期用户比例

3.2 用户产生 Feeds 行为模式

3.2.1 Feeds 时间间隔建模

在探讨人类活动的时间间隔分布时, 人类行为学领域存在两个经典的模型: 一个是泊松过程, 其时间间隔分布满足指数分布, 其 CDF 曲线在单对数坐标里为一条近似的直线; 另外一个是非泊松过程, 其时间间隔分布为幂律分布 $t^{-\alpha}$, 具有长尾特性, 其 CDF 曲线在双对数坐标里为近似直线。在非泊松过程下, 人类活动表现为短时间爆发和长时间静默。在发送邮件、信件回复、音乐下载、电影点播等场景都发现过活动时间间隔的长尾特性^[17,18]。

对于社交网络中用户的 Feeds 活动, 首先需要探讨其满足何种分布。文献[18]对经验数据提出了标准的拟合测试方法, 使用 P 值来测量假设的拟合模型与实际数据相符的可信度, 当 P 值小于 0.1 时假设被拒绝。我们为每个用户构建了以天为单位的 Feeds 时间间隔序列, 在拟合测试中, 80% 用户序列的 P 值大于 0.1, 即大多数用户的 Feeds 序列时间间隔分布可以使用幂律模型来拟合。

对于时间间隔分布 $t^{-\alpha}$ 来说, 参数 α 决定了其在双对数坐标中直线的斜率, 如果一个用户的 α 值较大, 那么这个用户大多数 Feeds 间隔相对较短, 这个用户更活跃的概率就越大。对于每个用户, 使用文献[18]中的方法计算其 α 值, 其分布如图 3 所示。可以看到大多数用户的 α 值落在 2 到 2.5 的区间

以内, 整个样本的 α 最小值为 1.5, 平均值为 2.3, 最大值为 3.5, 其分布形态类似于正态分布。

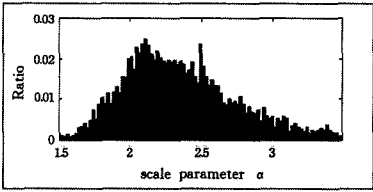


图 3 α 系数分布

3.2.2 α 系数与朋友圈的关系

在社交网络中好友通常具有结伴到来和一起离开的特性^[5]。如果一个用户拥有很多活跃的好友, 那么他也会以更大的概率在社交网络中保持活跃, 并产生更多的 Feeds。因此, 一个用户 Feeds 序列的 α 系数应该与其朋友有关。首先分析用户 Feeds 的 α 系数与其好友度数的关系。从图 4 可以看出, 当用户的朋友数量少于 200 时, 其 α 系数随着朋友数量线性增长。当用户的朋友数量超过 400 时, α 系数在均值 2.3 上下波动, 并没有线性增长, 这可能是因为随着用户拥有更多的好友, 正向的刺激就会减弱。另外, 从图中也可以看到有些用户拥有很高的节点度, 但是 α 系数却较少, 这可能是因为这些超级节点并不产生很多内容, 他们很可能是现实生活中的“明星”, 他们获得足够的关注度并不是因为其是一个出色的内容贡献者。

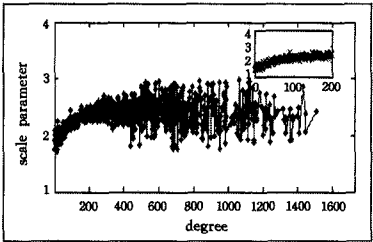


图 4 用户度数与 α 系数的关系

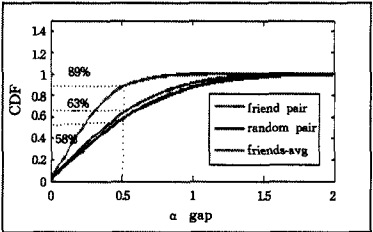


图 5 不同用户组 α 系数差值分布

朋友间的 α 系数决定着用户能够看到的内容多少, 那么用户的 α 系数不仅与朋友的个数有关, 还应与朋友的 α 系数值相关。我们把用户配对并分成两个样本, 第一个样本中的用户对全部都是朋友, 第二个样本中的用户对是由任意两个用户构成的。计算两个样本中每个用户对 α 系数的差值, 这个差值反映了朋友以及任意用户间 α 系数差的分布。同时计算用户的 α 系数与其所有好友的平均 α 系数的差, 图 5 给出了 3 种情况下 α 系数差值的 CDF 分布。朋友组的差值曲线处于随机组的上方, 但是并不明显, 均值组的曲线远远在其他两条曲线之上, 以 0.5 的差值来看, 63% 的朋友组样本的 α 系数的差值在 0.5 以内, 略大于随机组的 58%, 远小于均值组的 89%, 这说明用户的 α 系数与所有朋友的 α 系数均值比较接近, 即如果用户的朋友都有相对较大的 α 系数, 那么该用户

会接收到更多朋友的 Feeds, 他会以更大的概率产生更多的 Feeds, 我们可以称之为正向反馈效果。

3.3 用户接收 Feeds 流的分析

在社交网络中, 用户除了产生 Feeds 来自我表达以外, 还会浏览其好友的 Feeds 来消费信息。3.1 节和 3.2 节对用户的生命周期和 Feeds 的时间间隔进行了分析, 我们发现越来越多的用户的 Feeds 序列时间间隔变长了, 因此本节将讨论用户接收到的 Feeds 流, 并分析其是否与用户的活跃度有关。

对于用户接收到的信息流, 定义了信息流速和信息流多样性($v-d$ pair)两个评价指标。信息流速为用户在一天内收到的 Feeds 的数量, 如式(1)所示, v 值越大说明用户单位时间内接收到朋友的 Feeds 越多, 其作为内容消费者的需求就越容易满足。对于信息流的多样性, 我们参考了辛普森多样性系数。该系数表示群落中物种数越多, 各种个体分配越均匀, 指数越高, 群落多样性越好^[19]。借鉴该系数来表征用户信息流的多样性, 如式(2)所示, d 值越大说明用户可以接收到越多好友的信息, 其圈子就越丰富和多样化, 用户就越能感受到整个圈子好友的活跃。 F_i 为用户接收到朋友 i 的 Feeds 数量, S 为接收到的总的 Feeds 数量, T 为时间间隔(如一天), 一般来说, 活跃用户的 $v-d$ 值比不活跃用户的 $v-d$ 值要大。

$$v = (\sum_{i=1}^N F_i) / T \quad (1)$$

$$d = 1 - \sum_{i=1}^N (F_i / S)^2 \quad (2)$$

我们选择 L 值为 $[1, 3, +]$ 的用户作为活跃用户组, 选择 L 值为 $[0, 0.3]$ 的用户作为非活跃用户组。根据抓取的图谱, 聚合了这两组用户在整个生命周期中接收到的 Feeds 流数据, 并计算出每组用户平均的 $v-d$ 值, 如图 6 所示。对于活跃用户组来说, 他们接收的信息流流速最近两年达到 20 条每天, 是不活跃用户组 10 条的两倍; 对于信息流的多样性取值, 活跃组为 0.7, 非活跃组为 0.45。可以说活跃用户组在信息流流速和多样性两个指标上都远远优于非活跃组, 因此信息消费需求是用户活跃的一个必要条件, 它能反过来激发一个用户参与到社交网络中来, 并更加积极地产生 Feeds。

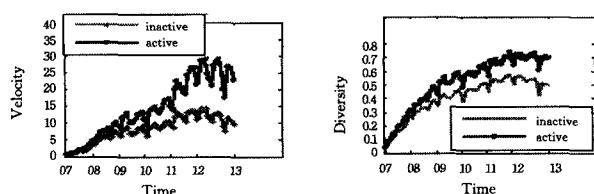


图 6 不同用户组的 $v-d$ 值

4 社交网络用户不活跃原因的探讨

前文从用户产生 Feeds 和用户浏览 Feeds 的角度对整个社区的用户活跃性进行了分析, 发现两个重要的现象: 用户产生 Feeds 的活跃度在下降; 用户接收到高 $v-d$ 的信息流是其活跃的的必要条件。本节从用户产生及接收 Feeds 的相互作用来探讨社交网络用户变得不活跃的原因。如前所述, 假设用户在社交网络中有两个基本的社交需求, 当社交网络在一定时间内不满足用户的社交需求时, 用户就会变得不活跃或者离开。

用户在社交网络中通过产生各式各样的内容来表达自己, 引起注意, 希望被朋友感知和交互; 浏览信息, 获取朋友最新状态。

有许多原因会导致上述需求不被满足: 社区变得没有吸引力, 用户不想在该社区浪费时间; 用户离开学校, 开始职业生涯, 学生时代的社区可能已经不适合; 一些更加时髦或者有吸引力的社区出现, 用户开始迁移到竞争社区; 一些原本活跃的用户渐渐感受到好朋友变得不活跃, 其圈子变得越来越冷清, 因此自己也慢慢离开。

出于这些内部或者外部的原因, 一些用户会降低他们产生内容的需求, 这种变化又会导致他们的朋友接收到的信息越来越少, 从而达不到消费信息的需求。这反过来使原本活跃的用户慢慢变得不活跃, 长时间下去将会形成恶性循环。假设整个社区中有 $p\%$ 的用户会经历上述不活跃的原因, 下面将探讨随着 p 和时间的变化, 整个社区会受到怎样的影响。

通过构建一个模拟实验来分析社交网络活跃度下降是如何产生的。在整个模拟过程中, 需要模拟用户因为一些随机原因在产生 Feeds 上变得不活跃, 以及这种不活跃如何影响到他的邻居节点。前面已经分析过大多数用户 Feeds 的时间间隔可使用 power-law model 来建模, 其 α 系数的变化能够反映用户产生 Feeds 的活跃度变化情况, 使用式(3)来产生满足 α 系数的时间间隔序列, 并在模拟过程中调整 α 系数来模拟用户活跃度的变化。

$$t = (1 - \text{rand}(1, n))^{1/(\alpha - 1)} \quad (3)$$

在整个社区中一部分用户产生 Feeds 的活跃度下降, 经过一段时间后, 会有一些用户能够明显感觉到他接收到的 Feeds 在减少, 作为反馈效果, 这部分用户将以一定的概率降低产生 Feeds 的活跃度。那么如何确定这些收到明显不活跃信号的用户? 图 6 的测量结果显示, 活跃和不活跃用户组的平均 $v-d$ 值为 $(20, 0.7)$ 和 $(10, 0.45)$, $v-d$ 比例差为 $(50\%, 40\%)$ 。当用户的 $v-d$ 值减少 $(25\%, 20\%)$ 时 $(50\%, 40\%)$ 的一半, 我们认为这些用户会接收到来自邻居的不活跃信号, 作为反馈, 这些用户将会随机降低 Feeds 的 α 值。我们把上述过程叫做一轮模拟反馈, 随着时间的推移, 整个过程是不断进行的, 整个模拟算法的步骤如下。

Step1 选择 L 为 $[1, 3, +]$ 的用户构建活跃子网, 对于每个用户, 根据其 Feeds 的时间间隔获得其 α 系数;

Step2 对于每一个用户, 使用式(3)作为其产生 Feeds 的时间间隔函数来产生 Feeds 序列, 当时间间隔达到一年的时长, 测试整个网络的 $v-d$ 值, 并以此作为基础状态;

Step3 随机在图谱中选择 $p\%$ 的节点, 按照一个随机的比例减少这些节点的 α 系数来模拟用户的随机不活跃, 使用新的 α 系数继续产生 Feeds;

Step4 每过 6 个月, 挑选出那些 $v-d$ 值最受影响的节点, 同样按随机的比例减少它们的 α 系数来模拟反馈效果。

整个反馈过程执行 4 次, 每结束一次反馈, 就测量整个拓扑里用户的平均 $v-d$ 值。在模拟程序中 p 值决定了初始不活跃用户的比例, 它对整个模拟的结果至关重要。在整个过程中使用不同的 $p(0.1, 0.2, 0.3, 0.4, 0.5)$ 来观察实验运行的结果, 如图 7 所示。看到当 p 值小于 20% 时, 整个社区的 $v-d$ 值在经历 4 轮负反馈后还能够保持在一个相对较高的水平。当 p 大于或者等于 30% 时, 局部的不活跃会很快地传递到整个社区之中, 具体来说当 p 达到 30% 时, 在经历一轮反馈之后社区的 $v-d$ 值就会接近不活跃组的水平 $(10, 0.45)$, 当经历 3 轮反馈之后, 其 $v-d$ 值已基本接近于零。如果使用

图 6 测量的阈值(20, 0.7)和(10, 0.45)来表示活跃和不活跃用户,那么大部分的用户将变得不活跃,整个社区将会崩溃。

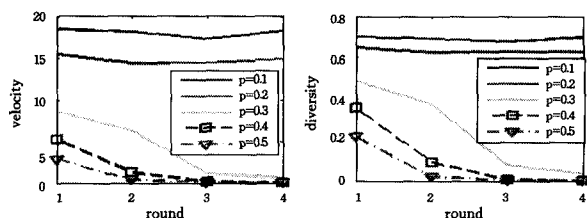


图 7 $v-d$ 值的变化

对于单个用户而言,每一轮反馈后,我们都会通过 $v-d$ 值的变化来找出最受影响的用户,图 8 示出了当 p 为(0.1, 0.2, 0.3, 0.4, 0.5)时这部分用户的节点度的 CDF 分布。可以看出当 p 值为 0.1 和 0.2 时,受影响的用户的度数基本都在 40 以下,这说明处于图谱边缘而节点度数较少的用户通常更容易受到朋友圈不活跃效应的影响。随着 p 值的增加,当 p 大于 0.3 时,不活跃用户的影响会蔓延到高节点度用户,节点度超过 60 的受影响用户达到了 20%左右。社交网络用户的不活跃和离开通常是从图谱的边沿蔓延开的,对于这部分用户,社交网络运营商可以考虑在好友推荐和信息流算法上进行定制,以保证他们在整个生命周期中都具有健康的信息流。

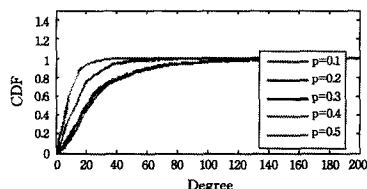


图 8 $v-d$ 变化最大的用户的 Degree 分布

模拟实验说明 30% 的不活跃用户对社交网络来说是致命的。对于实际的社交网络来说, p 值代表的是社区中由于各种原因变得不活跃用户的比例,它可以帮助我们评估用户离校或迁移带来的影响。为了研究人人网受到其他社交网络的冲击,通过另一个项目组的移动端数据获取大约 3000 名该社区用户 2014 年 1 月至 2014 年 3 月这 3 个月内使用人人、微博、微信 3 个移动 APP 的情况,计算了每个用户在 3 个社交应用上的时间来分析其社交精力的分配。图 9 示出了用户在各个 APP 上花费的社交精力百分比的 CDF 图。我们注意到,约 30% 的用户在人人网花费了 20% 左右的社交精力,很少有用户在人人网上花费超过 50% 的社交精力。移动端上最流行的是微信,甚至有 50% 的用户所有的社交精力都花在微信上,也就是说,人人网在该大学社区的 p 值远远超过了 30% 的临界点。当然,人人网有移动和网页两大入口,只看移动端会低估了它们的份额,但是随着智能手机的普及,移动端所占的比重越来越大,移动端用户社交精力分流会造成整个社区活跃性的下降。可以说,微博和微信等同类应用的有力冲击是人人网活跃度下降的主要原因之一。

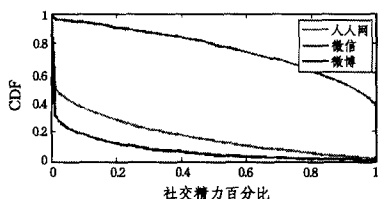


图 9 用户社交精力分布

通过对整个社区用户 Feeds 行为的分析,可以发现人人

网用户的一些不活跃的趋势。为保证社交网络持续的活跃,我们建议:1)产品设计中增加一些鼓励用户发布原创内容的元素(如短语音、短视频功能);2)对用户按照内容生产者 and 内容消费者建模,并提供相应的激励机制;3)改进推荐算法,保证用户信息流有一个相对较高的 $v-d$ 值;4)监控社区中 p 值的变化,在留住老用户的同时不断吸引新用户,让不活跃用户比例保持在 20% 以内;5)加大移动端产品的开发等。

结束语 本文通过分析用户产生的 Feeds 来研究整个社区的活跃度以及社区变得不活跃的原因。通过分析用户活跃周期,我们发现 2011 年之后,用户产生 Feeds 的活跃度在下降。在使用幂律模型对用户 Feeds 的时间间隔进行分析之后,可以发现 2011 年起用户产生 Feeds 的时间间隔在变长。另外,还分析总结了社交网络不活跃的演变过程:一部分用户因为种种原因而不愿意产生更多的 Feeds,导致整个社区信息流的质量下降,其他用户由于浏览不到有价值的信息而受到负反馈,从而导致其变得不活跃的概率增加。模拟实验显示,当 30% 的用户产生 Feeds 的活跃度下降时,整个社区在接收 Feeds 方面会受到很大的影响,并作为负向反馈加剧整个社区的不活跃。

虽然本文的分析结果只是建立在人人网的一个社区之上,但是这对基于 Feeds 的社交网络的活跃性分析是一次有益的尝试。通过对用户产生 Feeds 内容的分析,提供了一种理解社交网络的新视角,即从内容生产者和消费者模型来研究用户行为。对于互联网内容提供商而言,通过研究社区中的生产者和消费者的行为习惯,能够更有针对性地改善自己的服务,保持用户的活跃性。比如社区中新增的用户主要作为消费者存在,他们缺乏贡献内容的动力,而老用户也会因为各种原因造成贡献内容的动力下降,运营者如果能够及时监控这些问题,就能够在产品设计、Feeds 推荐算法上进行有针对性的改进。

参考文献

- [1] Kumar S, Zafarani R, Liu H. Understanding User Migration Patterns in Social Media[C]// Proc. of AAAI San Francisco, USA, Aug. 2011:1204-1209
- [2] Zhao X, Sala A, Wilson C, et al. Multi-scale dynamics in a massive online social network[C]// Proc. of the 2012 ACM SIGCOMM conference on Internet measurement. Boston, USA, 2012:171-184
- [3] Gong N Z, Xu W, Huang L, et al. Evolution of social-attribute networks: measurements, modeling, and implications using google+[C]// Proc. of the 2012 ACM conference on Internet measurement. Boston, USA, Nov. 2012:131-144
- [4] Leskovec J, Backstrom L, Kumar R, et al. Microscopic evolution of social networks[C]// Proc. of the 14th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, Aug. 2008:462-470
- [5] Wu S, Das Sarma A, Fabrikant A, et al. Arrival and departure dynamics in social networks[C]// Proc. of the Sixth ACM Int'l Conference on Web Search and Data Mining. Rome, Italy, Feb. 2013:233-243
- [6] Dasgupta K, Singh R, Viswanathan B, et al. Social ties and their relevance to churn in mobile telecom networks[C]// Proc. of the 11th int'l conference on Extending Database Technology: Advances in database technology. Nantes, France, Mar. 2008:668-677

(下转第 163 页)

- data fusion for fire detection[J]. *Information Fusion*, 2011, 12(3):150-159
- [10] Gupta V, Pandey R. Data fusion and topology control in wireless sensor networks[C]//5th International Conference on Applied Electromagnetics, Wireless and Optical Communications, Puerto de la Cruz, SPAIN, 2007:135-140
- [11] Tan Rui, Xing Guo-liang, Liu Ben-yuan, et al. Exploiting Data Fusion to Improve the Coverage of Wireless Sensor Networks[J]. *IEEE-ACM Transactions on Networking*, 2012, 20(2):157-168
- [12] Elson J, Römer K. Wireless sensor networks: A new regime for time synchronization[C]//Proceedings of the 1st Workshop on Hot Topics in Networks (HotNets-1). Princeton Univ, Princeton, 2003, 33(1):149-154
- [13] Fang Jun, Li Hong-bin. Power Constrained Distributed Estimation with Cluster-Based Sensor Collaboration[J]. *IEEE Transactions on Wireless Communications*, 2009, 8(7):3822-3832
- [14] 康健, 左宪章, 唐力伟, 等. 无线传感器网络数据融合技术[J]. *计算机科学*, 2010, 37(4):31-35
Kang Jian, Zuo Xian-zhang, Tang Li-wei, et al. Survey on Data Aggregation of Wireless Sensor Networks[J]. *Computer Science*, 2010, 37(4):31-35
- [15] Luo Hong, Tao Hui-xiang, Ma Hua-dong, et al. Data fusion with desired reliability in wireless sensor networks[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2011, 22(3):501-513
- [16] Nassar M O, Kanaan G, Awad H A H. Framework for Analysis and Improvement of Data-fusion Algorithms[C]//2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME). 2010:379-382
- [17] 焦竹青, 熊伟丽, 张林, 等. 基于接近度的多传感器数据融合方法研究[J]. *压电与声光*, 2009, 31(5):771-774
Jiao Zhu-qing, Xiong Wei-li, Zhang Lin, et al. Study on Multi-sensor Data Fusion Based on Approach Degree[J]. *Piezoelectrics & Acoustooptics*, 2009, 31(5):771-774
- [18] 宿陆, 李全龙, 徐晓飞, 等. 基于 D-S 证据理论的传感器网络数据融合算法[J]. *小型微型计算机系统*, 2006, 27(7):1321-1325
Su Lu, Li Quan-long, Xu Xiao-fei, et al. Data Fusion Algorithm for Sensor Network Based on D-S Evidence Theory[J]. *Mini-Micro Systems*, 2006, 27(7):1321-1325
- [19] Nakamura E F, Loureiro A A F, Frery A C. Information fusion for wireless sensor networks: Methods, models, and classifications[J]. *ACM Computing Surveys*, 2007, 39(3):9
- [20] 朱大奇, 于盛林. 基于 D-S 证据理论的数据融合算法及其在电路故障诊断中的应用[J]. *电子学报*, 2002, 30(2):153-155
Zhu Da-qi, Yu Sheng-lin. Data Fusion Algorithm Based on D-S Evidential Theory and Its Application for Circuit Fault Diagnosis[J]. *Acta Electronica Sinica*, 2002, 30(2):153-155
- [21] Wang H, Zhang Y, Meng L, et al. The research of fire detector based on information fusion technology[C]//2011 International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT). IEEE, 2011:3678-3681
- [22] Ghiasi S, Srivastava A, Yang Xiao-jian, et al. Optimal Energy Aware Clustering in Sensor Networks[J]. *Sensors*, 2002, 2(7):258-269
- [23] Beynon M, Curry B, Morgan P. The Dempster-Shafer theory of evidence: an alternative approach to multicriteria decision modeling[J]. *Omega*, 2000, 28(1):37-50

(上接第 153 页)

- [7] Richter Y, Yom-Tov E, Slonim N. Predicting Customer Churn in Mobile Networks through Analysis of Social Groups[C]//Proc. of SDM. Columbus, USA, Apr. 2010:732-741
- [8] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks[C]//Proc. of the 14th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, Aug. 2008:7-15
- [9] Crandall D, Cosley D, Huttenlocher D, et al. Feedback effects between similarity and social influence in online communities [C]//Proc. of the 14th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, Aug. 2008:160-168
- [10] Backstrom L, Huttenlocher D, Kleinberg J, et al. Group formation in large social networks: membership, growth, and evolution [C]//Proc. of the 12th ACM SIGKDD Int'l Conference on Knowledge discovery and data mining. Philadelphia, USA, Aug. 2006:44-54
- [11] Benevenuto F, Rodrigues T, Cha M, et al. Characterizing user behavior in online social networks[C]//Proc. of the 9th ACM SIGCOMM Conference on Internet Measurement. Chicago, USA. 2009:49-62
- [12] Schneider F, Feldmann A, Krishnamurthy B, et al. Understanding online social network usage from a network perspective[C]//Proc. of the 9th ACM SIGCOMM Conference on Internet Measurement Conference. Chicago, USA, Nov. 2009:36-48
- [13] Mislove A, Marcon M, Gummadi K, et al. Measurement and analysis of online social networks[C]//Proc. of the 7th ACM SIGCOMM Conference on Internet Measurement. San Diego, USA, Oct. 2007:29-42
- [14] Freyne J, Berkovsky S, Daly E M, et al. Social networking feeds: recommending items of interest[C]//Proc. of the Fourth ACM Conference on Recommender Systems. Barcelona, Spain, Sept. 2010:277-280
- [15] Soh P H, Lin Y C, Chen M S. Recommendation for online social feeds by exploiting user response behavior[C]//Proc. of the 22nd International Conference on World Wide Web Companion. Rio De Janeiro, Brazil, May 2013:189-198
- [16] Jiang J, Wilson C, Wang X, et al. Understanding latent interactions in online social networks[J]. *ACM Transactions on the Web (TWEB)*, 2013, 7(4):1-39
- [17] Hu H B, Han D Y. Empirical analysis of individual popularity and activity on an online music service system[J]. *Physica A: Statistical Mechanics and its Applications*, 2008, 387(23):5916-5921
- [18] Clauset A, Shalizi C R, Newman M E. Power-law distributions in empirical data[J]. *SIAM review*, 2009, 51(4):661-703
- [19] Wikipedia. Diversity index[OL]. http://en.wikipedia.org/wiki/Diversity_index