DOI: 10.3969/j.issn.1001-3824.2013.02.002

基于统计分析的微博信息传播规律研究

于 洪,杨 显
(重庆邮电大学 计算智能重庆市重点实验室,重庆 400065)

摘 要: 在基于统计方法的基础上,发现了微博信息传播速率所具有的普遍规律及信息传播路径所具有的典型传播模式。首先,通过微博开放平台,采用应用程序接口对微博信息的转播/评论数据进行采集,分析信息传播速率随时间变化的统计特性,得出不同话题类型的微博信息传播特征具有非常相似的结论:信息发布后,其转播/评论数在短时间内达到峰值,然后便很快衰落;然后,通过可视化软件 NodeXL 对采集到的数据发现了微博信息传播的3种典型传播模式:一触即发传播模式、多级传播模式和多点触发传播模式。

关键词: 微博;统计;传播速率;传播模式
中国分类号: TP391 **文献标识码:** A **文章编号:** 1005-3824(2013)02-0006-05

0 引 言

2010年被称为中国新媒体发展中的“微博元年”,各大门户网站以及多家媒体纷纷发布自己的微博。微博,也叫微型博客(microblog),是一个以用户关系为基础,实现信息获取、分享与传播的平台,用户可使用 Wap,Web 及其它客户端应用构建个人社区,内容控制在140字以内,从而实现信息的实时分享[1]。截至2012年年底,微博用户数已达3.09亿,其中多达65.6%的用户通过手机终端使用微博,使用方式的转变使得微博在移动互联网时代潜力无限[2]。

作为第二代互联网的产物,大众观点可在微博中较为自由地表达,尤其在突发性事件报道中,微博的现场感、时效性和报道的广度是其他传统媒体不能胜任的。但是,无根据的话题信息甚至谣言也易在微博中产生并扩散。了解微博信息传播规律非常重要,有助于相关部门形成应对策略,积极地对传播过程展开疏导与指引,同时,对企业和个人的微博营销也有积极作用。

1 微博信息传播过程

作为一种非正式的迷你型博客,为用户提供娱乐休闲、生活服务的信息性分享和交流平台,微博的主要特点有:草根化和便捷性、实时感与直播性、互动与传播多维性[3-4]。

在微博中,用户与用户之间没有访问限制,用户可就自己感兴趣的微博信息发起或参与转播/评论:转播就是用户浏览微博信息,希望将内容与他人分享,使其粉丝通过自己的主页能同步看到;而评论是用户对该话题的看法。用户彼此之间通过转播/评论的形式进行互动,话题的转播/评论数量理论上应随着时间推移而增多。

但是,由于话题本身的特点,如:话题内容自身的趣味性低、时效性短以及话题很快就被新兴话题所淹没等原因,导致话题的传播在一段时间后就终止,此时转播/评论数量不再增加。此外,又由于用户偏好千差万别,如上网时段、性别、职业等因素,用户在不同时段内对不同话题的转播/评论也有所差异。

在微博中,用户对某条微博要么浏览或转播/评论,要么没有浏览或转播/评论。因此,用户是否受微博中信息的影响,可通过分析转播/评论来了解。于是,微博中信息传播的效果变化就通过转播/评论数表现出来;对微博中信息传播模式的研究,就转化为对微博转播/评论特征的认识和分析。因此,本文首先把单条微博中转播/评论数量与其增长率的关系作为主要研究对象。单条微博,指微博用户根据需要而发出的1条即时分享的信息。

2 信息传播统计特性分析

网络爬虫是一种获取数据的常规方法。爬虫以一个或多个初始网页地址(URL)作为端点,得到初始网页上的 URL,在整个抓取数据的过程中,连续不断地从当前页面上提取新的 URL 插进队列之中,直到与设定的条件匹配时才停止[5]。相对于网络

收稿日期:2013-03-09 修回日期:2013-03-29

爬虫抓取数据而言,通过微博开放的 API 接口这一方法更加便捷、高效。本文主要针对腾讯微博进行数据获取。

2.1　数据采集

通过腾讯 API 抓取数据的前提在于通过 OAuth 协议认证,该协议为用户获取资源授权提供了一个简单、安全、且开放的标准,任何服务提供商都可实现自己的 OAuth,任何软件开发商,即第三方应用都可使用 OAuth,具体授权认证流程参见文献[6]。

用户获得授权后,通过 API 请求便可获得格式为 XML 或 JOSN 的数据。XML 是一种跨平台、允许用户对自己数据进行标记和定义的强结构性语言,对于 Web 传输非常适合。因此,本文采用 XML 格式来返回数据。

由于 XML 对结构要求十分严格,而微博中的信息可能存在一些特殊的字符,因此将数据通过 API 取出后,在存入数据库进行解析前,需要对这些特殊的字符进行处理。

对微博转播/评论数据的具体采集过程参见文献[7]。

2.2　统计方法

本文先随机挑选了部分具有一定转播/评论数量的微博,以天为单位时间进行统计,分析后发现它们具有一定的规律,即: 1 条微博的 80% ~90% 转播/评论数在第一天就完成了,有时,这个比例甚至更高。图 1 为 ID = 86579112613104 的微博发出后每天的转播/评论数分布图,第一天转播/评论数占了 1 684/1 921≈87.66%。

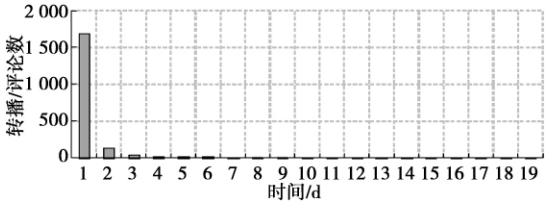


图 1　ID = 86579112613104 的微博转播/评论数分布图

因此,本文把统计数据的重点放在第一天,即: 从微博发布的时刻开始,以 1 小时为时间段,统计出每个时间段的转播/评论次数。

考虑到不同话题类型的微博其传播规律可能有所不同,不同用户在不同时间段发出的微博其传播规律也有可能不同。因此,本文根据微博用户排行榜(http://t.qq.com/rank.php)设置了 5 个话题类型: 体育类、文化类、笑话类、新闻类和追星类。每个话题类型定量收集了 8 个时间段不同用户(选取的用户排名较靠前,其影响力较强)所发出的微博及其相应的转播/评论,这 8 个时间段分别为: 7:00—9:00,9:00—11:00,11:00—14:00,14:00—16:00,16:00—18:00,18:00—20:00,20:00—22:00,22:00—24:00。微博转播/评论数在 1 000 ~10 000 不等,全部数据大小约 2 GB。

2.3　统计结果

为了对微博的转播/评论数在不同时间段有一个比较清晰的认识,我们首先将数据用公式(1)进行了归一化处理,以便统一标准; 然后,再对不同时间段、不同话题类型的微博的传播速率各自取其平均值。

$$\text{传播速率} = \frac{\text{单位时间内转播/评论数}}{\text{参与转播/评论的总数}} \quad (1)$$

首先,本文针对不同时间段、不同话题类型微博信息发布后的传播速率进行了统计。限于篇幅,图 2 仅仅以 7:00—9:00 时间段发布的微博为例,给出了 5 个话题类型的微博在其发布后 24 h 各自的平均传播速率变化曲线。

实验得到的统计数据说明: 不同时间段、不同话题类型的微博在随后 24 h 的传播速率变化趋势非常相似,说明微博信息传播速率变化趋势与其发布时间关系不大。

然后,本文再对相同时间段内不同类型微博的传播速率取平均值,得到不同类型话题在各个时间段传播速率的平均值。同样地,图 3 仍然以 7:00—9:00 时间段发布的微博为例,给出了 5 个话题类型微博在其发布后 24 小时的平均传播速率变化曲线。

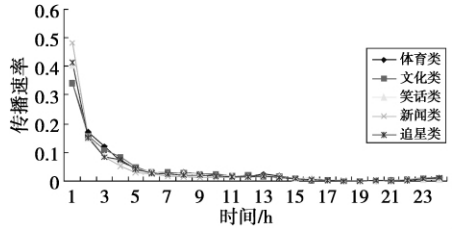


图 2　5 种话题微博发布后 24 h 平均传播速率变化趋势图

对比图 3 和图 2,明显可见,在 7:00—9:00 这个时间段内,不同话题类型微博平均传播速率趋势与各个话题类型微博发布后 24 h 各自平均传播速率变化趋势也是一致的。事实上,实验得到的统计数据说明: 不同时间段、不同话题类型微博平均传播速率趋势也是基本一致的。

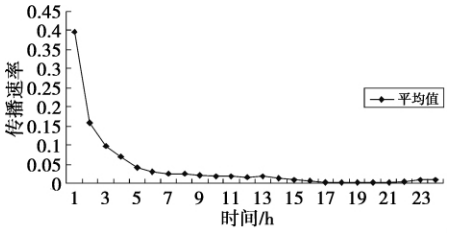


图3 不同话题类型微博发布后24 h平均传播速率变化趋势

通过对上述统计实验的分析，我们发现：不同话题类型的微博信息其传播规模虽然有所差异，但是经过归一化处理后，它们的传播特征非常相似，即：在微博发布后，短时间内微博的转播/评论数达到峰值，之后便很快地衰落下去。

出现这种传播规律可能源于微博自身的一些特点，即短小精悍、草根性强、信息共享和便捷迅速[1]。140字左右信息的简短性使得普通大众与莎士比亚位于同一水平线，低门槛的限制使得用户普遍发布信息成为可能。当1条微博信息发布后，其内容出现在该用户粉丝的广播列表中，短时间内受到大量粉丝关注，进而进行快速的转播/评论，随着时间推移，该信息逐渐被其它新兴微博信息所覆盖，粉丝能浏览到该信息的可能性就变小，能转播/评论的概率也就越小，从而阻碍了信息的传播进程，由此形成了1条先急剧增加再减小的传播速率曲线。

3 信息传播可视化分析

由上一节我们可以看出，每条微博信息传播都会形成一个特定的传播网络，为了对微博信息的具体传播路线(即信息由发布者发布后，通过发布者、发布者粉丝及其粉丝的粉丝对信息进行转播而形成的具体传播路径)有一个直观的认识，首先需要对获取到的转播/评论数据进行预处理，从而得到用户间的转播/评论关系，数据预处理的具体方法参见文献[7]；然后对微博信息的转播/评论关系通过Excel插件——NodeXL[8]进行可视化展示，并分析发现微博信息传播网络中所具有的典型传播路径模式。在给出传播路径模式前，需对微博信息传播网络进行形式化描述[9]。

微博信息传播的真实网络用有向图表示为 $G=\{V,E\}$。其中：V 表示用户节点的集合，由微博信息的原创节点 v_0 以及其它转播/评论节点 V' 共同组成：$V=\{v_0\}\cup V'$，$V'=\{V_1,\cdots,V_i,\cdots,V_n\}$，节点不但可对微博信息进行发布，还可对其他用户的微博信息进行转播/评论。其它转播/评论节点 V' 中，V_1 代表微博信息由 v_0 发布后，首先到达的节点所组成的集合，即一级传播节点集合，其传播路径长度为1；V_2 代表微博信息由 v_0 发布后并经过 V_1 中部分节点所到达的节点组成的集合，即二级传播节点集合，其传播路径长度为2；其它 V_i 含义以此类推。显然，对于 $\forall V_i,V_j,V_i\cap V_j=\phi$。$E$ 表示连接用户节点彼此间边的集合。

此外，设 $v_k\in V_k$，即 $v_1\in V_1$，$v_2\in V_2$，$\cdots$，则称 $v_0\rightarrow v_1$ 为一级传播路径；$v_0\rightarrow v_1\rightarrow v_2$ 为二级传播路径，其它以此类推。

通过上述形式化描述后，导入微博信息用户间的转播/评论关系至NodeXL中，分析发现微博信息的传播路径具有一些普遍的规律，我们将其叫作微博信息传播路径元模式：一触即发传播模式、多级传播模式或多点触发传播模式[9]。

1) 一触即发传播模式。

一触即发传播模式示意图如图4所示。

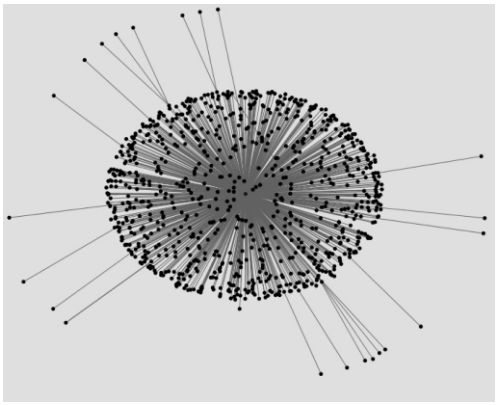


图4 一触即发传播模式示意图

一触即发传播模式通过 v_0 带领其粉丝，即 V_1 进行微博信息的转播/评论，集合 V_1 中各节点处于平等关系。信息首先通过具有一定影响力的原创节点 v_0 向其粉丝蔓延，传播路径向四周发散，且每条分支路径长度基本为1，即大部分用户对信息的传播大致都滞留在一级传播节点的位置，即 $V'\approx V_1$。其中，白色节点代表 v_0。

2) 多级传播模式。

在该传播模式中，微博信息通过原创节点 v_0 不断逐级地由内向外进行蔓延，由于 $n(n>1)$ 级传播节点集合 V' 中部分节点自己拥有一定数量的粉丝，通过这些节点的影响力从而带领其粉丝对微博信息进行进一步的转播/评论，从而使信息传播的广度和深度得以拓展。

图 5 是一个有关“爱心接力—寻人”话题的多级触发传播模式例子,显然,该条微博信息由原创节点发布后,形成了较为明显的一级传播、二级传播、三级传播,四级传播轮廓已部分形成。

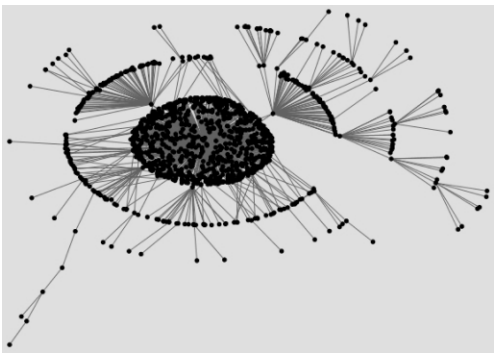
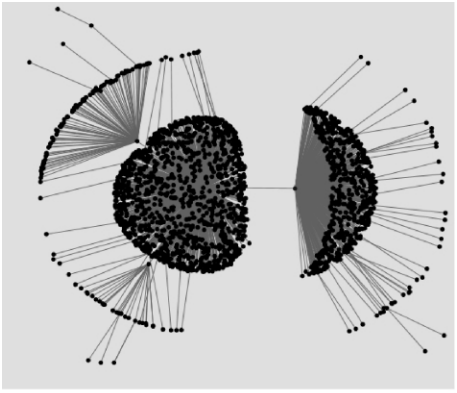


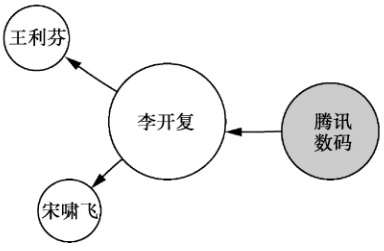
图 5　多级传播模式示意图

3) 多点触发传播模式。

在该传播模式中,微博信息由原创节点 v_0 发布后,因为 v_0 本身具有的威望会推动其直接粉丝展开信息的转播/评论,与此同时,V'中同样含有少量具有影响力的节点,这些节点的作用甚至比 v_0 还大,所以,在 $n(n>1)$ 级传播后,其效果得到加强,信息覆盖面扩大,进而形成多个节点相互呼应的传播场面。图 6 是多点触发传播模式例子及其对应网络中的关键人物(节点)转播关系。



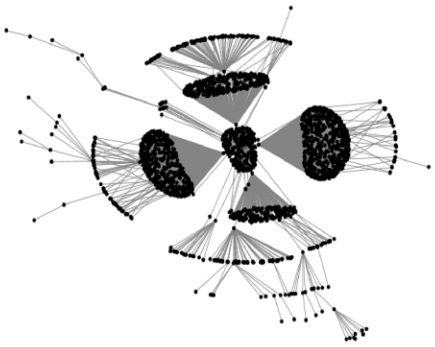
a 传播关系图



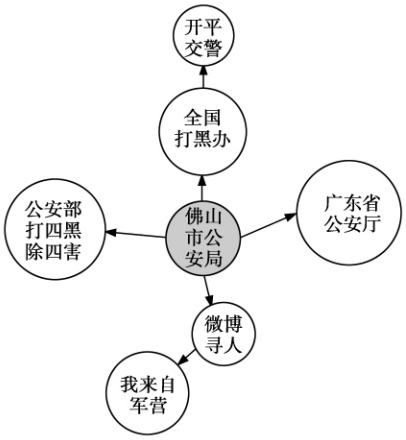
b 传播网络关键节点图

图 6　多点触发传播模式示意图

实际上,大量微博传播路径模式更多的是上述 3 种模式中的组合形式,即混合传播模式。如图 7 所示,该微博传播形式为多级传播模式和多点触发传播模式的结合。



a 传播关系图



b 传播网络关键节点图

图 7　混合传播模式示意图

4　结束语

研究微博中信息传播规律无论是对个人、企业,还是社会都意义重大,例如: 利于信息发布后的疏导与管理,有助于产品与企业的营销,便于资源的优化与分配等。

本文以统计法为基础,首先发现微博信息发布后其传播速率所具有的普遍规律: 短时间内微博的转播/评论数达到峰值,之后便很快地衰落下去; 其次,发现了微博信息传播路径所具有的 3 种典型传播模式。如何结合微博自身特点来对微博信息传播建模模拟其速率变化趋势将是下一步工作重点。

参考文献:

[1] WESTMAN S, FREUND L. Information interaction in 140 characters or less: genres on twitter [C]// IIiX ´10

Proceedings of the third symposium on Information interaction in context. New Brunswick, USA: Association for Computing Machinery, 2010: 323-326.

[2] 第31次中国互联网络发展状况统计报告[EB/OL]. [2013-03-07]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201301/P020130122600399530412.pdf, 2013.01.

[3] HAN Ruixia. The influence of microblogging on personal public participation[C]//Proceedings of the 2010 IEEE 2nd Symposium on Web Society, SWS 2010. Beijing: Association for Computing Machinery, 2010: 615-618.

[4] 懂海军,曾淑萍. 从博客到微博: 过程特征、意义建构与挑战[J]. 中国青年研究,2011(9): 89-92.

[5] CHO J, MOLINA H G, PAGE L. Efficient crawling through url ordering [J]. Computer Networks and ISND Systems(S0169-7552), 1998, 30(1-7): 161-172.

[6] 廉捷, 周欣, 曹伟. 新浪微博数据挖掘方案[J]. 清华大学学报: 自然科学版, 2011, 51(10): 1300-1305.

[7] 于洪, 杨显. 微博中节点影响力度量与传播路径模式研究[J]. 通信学报, 2012, 33(Z1): 96-102.

[8] HANSEN D, SHNEIDERMAN B, A SMITH M. Analyzing social media networks with NodeXL: insights from a connected world [EB/OL]. [2013-03-05] http://deca.cuc.edu.cn/Community/cfs-filesystemfile.ashx/_key/CommunityServer.Components.PostAttachments/00.00.01.17.38/Analyzing-Social-Media-Networks-with-NodeXL.pdf, 2012.

[9] 赵丽, 袁睿翕, 管晓宏,等. 博客网络中具有突发性的话题传播模型[J]. 软件学报,2009,20(5): 1384-1392.

作者简介:

于 洪(1972-),女,重庆人,博士,副教授,硕士生导师,主要研究方向为数据挖掘、粗糙集理论和Web智能等; 杨显(1987-),男,重庆人,硕士生,研究方向为数据挖掘。

基金项目: 国家自然科学基金资助项目(61073146,61272060); 重庆市自然科学基金资助项目(cstc2011jjA40045)。

Information propagation on microblogging using statistical analysis technique

YU Hong, YANG Xian

(Chongqing Key Lab of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China)

Abstract: Based on statistical methods, universal law and information propagation path with the typical pattern on microblogging information dissemination are found. First, through the open platform, the data of microblogging's broadcasts and comments are collected using the application programming interface. The statistical features of information propagation changing with time are analyzed, and we draw a conclusion that the characteristics of different types of topic in microblogging information propagation are very similar. That is, after the distributed information, the number of broadcasts/comments comes up to a peak in a short time, and then quickly fades out. Then, three typical modes of transmission on microblogging information propagation, which is explosive propagation mode, multi-level transmission mode, multiple point trigger propagation mode, are discovered by applying the visualization software—NodeXL.

Key words: microblogging, statistics, propagation rate, propagation mode

(责任编辑 张 诚)