中国人民大学学报 2013年第5期
JOURNAL OF RENMIN UNIVERSITY OF CHINA No. 5 2013

微博中热点话题的内容特质及传播机制研究

——基于新浪微博6 025条高转发微博的数据挖掘分析

李 彪

［摘要］ 针对新浪微博两年多来高转发的6 025条热门微博，采用大数据挖掘与分析技术进行研究。热点微博话题的传播呈现出一定的规律：转发次数的分布符合幂律分布，转发信息链的长度符合指数型分布特征。新浪微博具有强弱关系同时存在于一个平台属性特征，这既不同于Twitter的社会单向度的弱关系平台，也不同于Facebook双向的强关系平台。不同类别事件在新浪微博平台中传播的信息流和时间线也有差异，可以根据不同类别热门微博的转发深度和转发宽度构建出不同类别热门的传播模式结构。微博的话语权力格局中依然存在着不平等现象，传统社会话语精英依然把持着微博话语场域的主导权。

［关键词］ 数据挖掘；热门微博；转发深度；转发宽度；话语权力格局

［作者简介］ 李彪：中国人民大学新闻学院讲师，中国人民大学新闻与社会发展研究中心研究员（北京 100872）

本文以新浪微博高转发的6 025条原创微博为研究样本，通过大数据挖掘分析技术，研究分析目前微博中的热点话题属性及传播机制。

一、研究缘起

微博作为一种新的在线社会网络形式，逐渐成为人们获取和共享信息的重要平台。据中国互联网络信息中心（CNNIC）《第31次中国互联网络发展状况统计报告》显示，截至2012年12月底，我国微博用户规模为3.09亿，较2011年年底增长了5 873万，增幅达到23.5%；手机微博用户规模达到2.02亿，高达65.6%的微博用户使用手机终端访问微博。[1]根据西方传播学的研究，一种物理属性的媒介形态被社会大众中20%以上的人群所使用，便可以称其为“大众媒介”，从这个意义上可以说，微博已成为一种大众化媒体（Mass Media）。

微博以即时性和裂变式的嵌套性等人际传播的基本属性，引发了一场“140字符的社会话语革命”。微博在整个社会话语场域中所扮演的作用也越来越重要，成为整个社会话语场域的“话语漩涡”，扮演着话语策源地、信息桥和主导者等多重角色，最大限度地解构了传统的由社会话语精英所主导的话语权力格局，将原来看似“铁板一块”的话语权力场域一分为二——官方话语场域和草根话语场域，冲击着传统的社会治理方式和社会个体存在方式，塑造了一种新的社会话

［基金项目］ 云南省院校教育合作人文社会科学项目“云南舆情监测与边疆社会稳定关系研究”（SYSX201107）

语权力面貌。

微博扮演着重要的社会话语动员角色，很多社会行动如随手拍行动、免费午餐计划等都是在微博中倡导，通过微博进行充分的社会动员进而影响到线下的；微博同时还扮演着还原社会真实、黏合社会信息碎片的重要角色，很多社会信息在微博中以碎片化的方式存在，通过微博用户的集体力量和贡献，完成社会真实的“再构建”，进而实现“社会真实的有机运动”。另外，微博还为社会信息提供意见加工、贴标签等“仪式赋予”的功能，很多信息一般以事实判断的形式进入到微博场域，微博中话语精英通过其内化的“文化地图”对其进行价值判断，以提供意见或贴标签等方式赋予其更大的社会价值意义，使之得以快速地传播开来。作为一种独立运行的社会话语场域，微博具有本身的话题偏好属性、话语扩散模型、话语权力格局等属性。因此，研究这些属性对于更好地把握微博话语场域具有重要的价值。

微博“粉丝路径”和“转发路径”的传播方式既不是传统媒体的线性传播，也不是网络媒体的网络传播，其传播速度和传播广度远远高于之前任何一种媒介产品。新浪微博每日产生 1 亿条内容。[2]在这些浩如烟海的信息中去伪存真，找到有价值或者能够展现中国微博用户信息地图的核心热点信息，成为相关研究的热点问题。

二、研究设计

（一）数据抓取

本文采用“爬虫技术”，通过新浪微博 API（Application Programming Interface）接口进行数据抓取。新浪微博与其他微博网站（如 Twitter）类似，用户之间构成有向无权网络。用户可自由添加关注的其他用户，称之为“跟随”（Followings）；也可在未经许可的情况下被其他用户关注，称之为“粉丝”（Fans）。用户发表的话题将会自动推送给该用户的所有“粉丝”；类似地，用户也可自动获知所有“跟随”所发表的话题信息，这些信息几乎都是实时更新的。为了获取新浪微博的真实用户数据，本研究编写了针对新浪微博的爬虫程序，该爬虫程序采取广度优先和随机采样策略。首先，从新浪微博“名人堂”的各个子栏目中，随机选取 10 个用户作为种子用户，加入爬虫工作列表；然后，获取这些种子用户“朋友”列表，包括“粉丝”列表和“跟随”列表。由于有些用户（比如一些名人）的“粉丝”数量很大，要获取整个网络用户信息不太现实，为此采取随机采样策略，从“朋友”列表中随机选择最多 50 名用户加入工作列表，继续爬取用户信息。采用上述策略收集的部分用户信息能较好地反映整个微博用户的情况。

（二）数据集

本文使用的数据库从 2010 年 9 月 15 日开始收集数据，目前已经收集的用户数大约有 40 万，以文本形式存储，占用空间 50G 左右。收集的信息包括两部分：（1）用户基本属性信息，如 ID、Name、Gender、VFlag、Address、Tags、Fans、Followings、Tweets；（2）用户话题内容信息，如话题内容属性、转发次数、评论次数。爬取的内容几乎涵盖了该用户的所有信息。其中 VFlag 是认证用户标识，新浪微博采取实名制等形式对知名用户进行实名认证。

（三）数据处理技术

选取以下数据作为分析对象：选取时间跨度从 2010 年 9 月至 2013 年 1 月，每条转发量在 1 000 次以上的原创微博为热点微博，在以上数据库中符合条件的微博数量总计 6 025 条，来自 2 356 位博主，每条微博平均被转发 1 836 次，总转发次数 1 108 万。

为了更好地对这些热门微博进行数据处理，将所有热门微博信息分为以下 8 个类别（见表 1）。

表 1　　热点微博的类别列表

类别	类别描述
时尚娱乐	时尚潮流，娱乐视频，闲言碎语
社会热点事件	自然灾害，公共卫生，官员腐败，社会保障
休闲心情	幽默，搞笑图片，星座，有趣的事情，智慧哲理

续前表

类别	类别描述
生活健康	生活小常识，医疗保健，环境保护
寻求帮助	发动捐赠，寻找走失人员
促销信息	广告，微博营销等
风水财运	涉及风水和财运知识，还包括转发会有好运等微博
被删除微博	管理部门或用户个人删除，显示为“该微博信息已被删除”

（四）关键定义说明

本文有两个关键概念，即转发深度和转发宽度。转发深度是指同一条微博信息传播流的环节多寡。如图1所示，原始微博经过转发者B1和转发者C1、C2等的转发，其转发深度为2级，单个话题中转发深度极值越大，其信息链条越长，说明该事件越受关注、越容易引起民众的讨论兴趣。转发宽度是指一条微博转发的信息链条中，单个信息链条结点被转发的次数，转发次数越多，转发宽度也就越大。如图1所示，从转发者B1这个节点有四个转发者进行转发，那么在转发者B1这个节点，转发宽度为4。

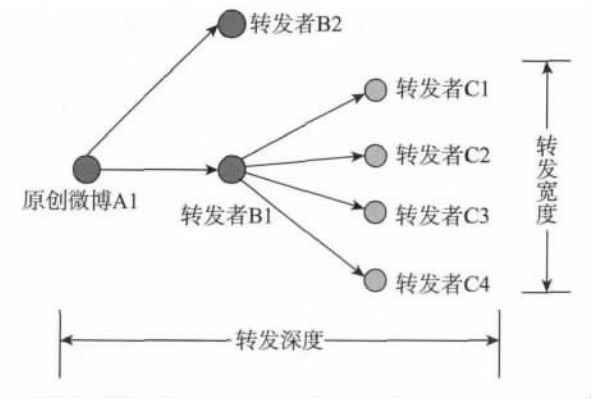

图1　微博转发深度和转发宽度示意图

需要说明，每个原创微博并不是仅仅有一个转发深度和转发宽度，可能有很多个。如图1所示，在这个原创微博A1中，总计有两个转发深度，即A1－B2的转发深度1级和A1－B1－C1的转发深度2级，其中2级是原创微博A1的极值转发深度，因此每个原创微博只有一个极值转发深度；同样，在原创微博A1中，有两个转发宽度，从A1节点产生的B1、B2的两个单位转发宽度，从B1点产生的C1、C2、C3、C4四个单位的转发宽度，同样道理，4个单位宽度是原创微博A1的极值转发宽度，因此每个原创微博只有一个极值转发宽度。

三、数据结果及分析

（一）热点话题内容分析

1. 话题内容的类别特征

通过对6 025条热点微博进行归类分析，相关结果如表2所示。

从表2可以看出，新浪微博是一个大而全的信息平台，在8个类别的热点话题中，微博用户最为关注的是“休闲心情”，占到总体的42.6%，说明目前微博用户的心理压力普遍较大，希望通过微博来获得心理的放松和安逸，另外也说明微博具有缓解压力、进行心理调节的工具属性，从这个意义上说，微博是一种“软”媒体。其次是社会热点事件，占到总体的21.9%，一定程度上佐证了微博具有媒体的属性特征。然后是时尚娱乐，占到总体的13.4%，这更多的是满足人们的娱乐、窥私等心态。这三者就占到了总体的78%。

表2　热点话题的类别分布及转发情况

	时尚娱乐	社会热点事件	休闲心情	生活健康	寻求帮助	促销信息	风水财运	被删除微博
条数	809	1 320	2 569	448	290	201	212	176
所占比例	13.4%	21.9%	42.6%	7.4%	4.8%	3.3%	3.5%	2.9%
最大转发量	13 445	35 066	18 109	7 462	29 610	33 154	10 204	23 363
平均转发量	1 721.9	2 044.4	1 606.8	1 586.7	2 572.3	2 611.6	2 172.5	2 312.8

从热点话题最大转发量上来看，社会热点事件引发的转发极值最大，凸显出微博的围观效应；其次是促销信息，由于背后有网络水军的身影，这个数值不是很准确。从热点话题的平均转发量上看，促销信息和寻求帮助信息最高，可以看出微博作为一种草根的社会化网络媒体，在社会关系的维系和拓展方面具有其他新媒体所不能比拟的价值。

2. 热点话题创作者特征分析

（1）性别特征。

通过数据统计可见，男性是热点话题的创作者主力。在所有 8 个类别的热点话题中，男性的数量都远远超过了女性，一定程度上折射出现实社会中男女之间的话语权力格局。尤其是在社会热点事件、休闲心情等类别中，说明男性依然是微博这个虚拟话语场域中的主要议程设置者和主导者。

从同一话题内性别比例分布来看，男性在促销信息、社会热点信息等类别上远远超过女性比例，是这两类信息的绝对主导者，一定程度上反映出男性积极赚钱、热心时事政治的性别特征。在生活健康、时尚娱乐等类别上女性要明显高于其他类别，也凸显出女性在微博这一虚拟社会场域中依然关注美容、娱乐等性别特征。

（2）认证特征。

通过对话题原创用户的认证特征进行分析可知，认证用户是社会热点事件、促销信息等类别的主要生产者和主导者。促销信息多是一些认证机构，而社会热点事件主要是一些加 V 的认证用户，这些人通常也拥有线下现实社会的话语权，通过认证将线下的话语权“平移”到微博话语场域。这说明微博仅在一定程度上实现了所谓的话语平权，整个社会话语场域的主导权依然被传统的社会话语精英阶层所掌握。同时也说明，社会热点事件传播过程中，这些认证用户扮演了重要的角色，因为其具有较高的社会公信力和影响力，其态度、意见乃至情绪很容易传染给草根用户，很容易引起民意的啸聚。因此，目前很多舆情热点事件的消弭都是这些社会话语精英阶层与社会管理者之间在“合意的空间”内妥协的一种结果。

非认证用户是休闲心情、生活健康、风水财运和时尚娱乐等类别的主要创作者，这些话题多是一些“鸡零狗碎”的碎片软性话题，无关“社会宏大叙事”，再次印证了微博话语权力格局中的权力结构。

（3）地域特征。

热门微博原创者所处的地域分布数据见表 3。

表 3　　热门微博原创者地域分布数据

	北京	上海	广东	海外	其他	浙江	香港	台湾	江苏
时尚娱乐	0.12	0.11	0.16	0.19	0.09	0.13	0.24	0.26	0.08
社会热点事件	0.27	0.23	0.16	0.15	0.36	0.18	0.09	0.03	0.2
休闲心情	0.4	0.38	0.43	0.42	0.36	0.33	0.46	0.58	0.44
生活健康	0.06	0.17	0.07	0.08	0.04	0.082	0.07	0.04	0.05
寻求帮助	0.04	0.037	0.032	0.035	0.09	0.081	0.03	0.02	0.037
促销信息	0.054	0.039	0.021	0.019	0.011	0.14	0.017	0	0.138
风水财运	0.022	0.024	0.088	0.021	0.024	0	0.056	0.026	0.053
被删除微博	0.025	0.032	0.021	0.063 1	0.019	0.072	0.02	0.024	0.021

北京、广东、上海三个省市是原创者主要集中区域，占到总体的 74.8%，其中北京最多，占到总体的 44.6%，这在一定程度上反映出微博场域话语权与当地经济社会发展水平有一定正相关关系。其余区域依次是海外、浙江、香港、江苏、台湾，这在一定程度上凸显出目前微博话语场域的地区分布格局。

海外、香港和台湾超过其他地区处在原创者地域前列，说明微博社会话语场域中港台、海外地域因为其文化的独特性和文化的接近性也占据

着重要位置。

将8个类别的热点话题与以上几个重点区域进行交叉分析，可见：在时尚娱乐热门微博中，台湾、香港和海外占据前三位，这和目前明星、时尚娱乐信息多来自这些地区有一定关系；社会热点事件热门微博中，其他、北京、上海占据前三位，北京是政治中心，“社会公知”人士较多，对时事政治比较关注；在休闲心情热门微博中，台湾、香港和广东占据前三位，与时尚娱乐差不多，说明在目前大中华文化圈中，大陆文化的影响力和辐射力还有待进一步提升；在生活健康热门微博中，上海地区一枝独秀，说明上海民众热爱生活、注重健康；在促销信息热门微博中，浙江和江苏所占比例最高，说明江浙民众爱做生意和营销的地域特征；在风水财运热门微博中，广东可谓一枝独秀，这与当地的文化习俗有较为密切的联系。

从上述特征分析大致可以得出以下结论：北京民众向微博话语场域输入时事政治话题，上海民众向微博话语场域贡献生活健康信息，广东民众向微博话语场域贡献风水财运信息，江浙民众则拿微博平台来做生意和营销；香港、台湾地区则向微博话语场域输入时尚、娱乐、休闲等话题，这些特征构成了目前整个微博话语场域色彩斑斓的精彩画面。

3. 热点话题转发者特征分析

（1）性别特征。

通过数据分析可见，女性是热门微博的积极转发者，从绝对数量上看，男性是社会热点事件、被删除微博等类别的主要参与转发者，其余类别则女性是主要转发者。

与上面原创者性别特征相比可以看出，男性生产热门微博，女性转发热门微博，两者分工十分明确，而男性对于社会热点事件、被删除微博等最为关注，体现出男性相较于女性更热心社会时事政治，关心社会发展。

（2）认证特征。

通过对转发用户的认证特征进行分析可知，非认证用户是所有热门微博的积极转发者，尤其是在促销信息、休闲心情等热门微博，在所有话题转发者分布上，认证用户对社会热点事件、被删除微博、寻求帮助上要超过非认证用户，体现出认证用户已经具备了“社会公知”的属性特征。

结合热门微博创作者的认证特征，可以看出目前微博话语场域的权力格局并不是所谓的话语平权特征，依然是存在多个话语“明星”的仰角特征，认证用户拥有较多的社会资本，通过自己的社会网络生产信息，非认证用户负责传播扩散信息来为这些认证用户“摇旗呐喊”，使得认证用户获取更多的社会资本，形成所谓强者越强的马太效应。

（二）热点话题传播特征

1. 转发次数分布

图2是不同热门微博转发次数的累积概率（Empirical Cumulative Distribution Function）。

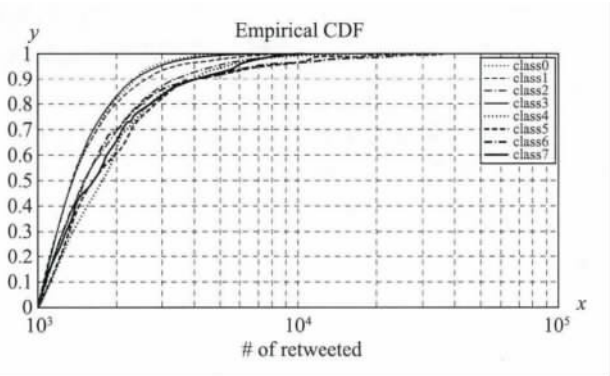

图2 不同热门微博转发次数的累积概率

图中横轴（x 轴）表示转发次数。纵轴为 y 轴，y=转发次数小于 x 次的所有微博的数量/所有微博的数量，随着 x 的增大，y 也在增大，累积概率最多为1。图2表示各类热门微博所呈现出来的转发特征基本一致，即大多数热门微博处于1 000～3 000次这个区间段内，超过3 000次的数量减少得很快。整体来看，热门微博的转发数分布符合幂律分布特征，高转发的热门微博数量不大，都集中在“长尾”段。

2. 转发深度分布

各类别话题的转发深度均值及极值见表4。

可以看出，微博热点话题的转发深度符合指数型分布。从不同类别话题的极值转发深度来看，13层级是目前所有热点微博转发的极值层级，也就是说，目前所有的话题传播的信息链条中13个环节是极值。极值转发深度较高的话题类别是风水财运、被删除微博、寻求帮助和社会

表 4　　各类别话题的转发深度均值及极值一览表

话题	转发深度均值	极值转发深度的均值	每类话题极值转发深度	结果
时尚娱乐	2.054 1	5.136 1	9	中
社会热点事件	2.323 9	6.264 4	10	高
休闲心情	2.024 3	6.063 1	8	中
生活健康	2.083 5	4.041 1	8	中
寻求帮助	2.625 4	6.615 5	11	非常高
促销信息	1.295 3	3.229 7	4	非常低
风水财运	2.507 3	6.099 3	13	高
被删除微博	2.587 5	5.791 2	12	高
所有热点微博	2.187 8	6.064 2	13	

热点事件，风水财运因为心理暗示的强制作用转发深度较深；被删除微博由于本身信息量缺乏，信息链条断裂，只有通过一个个转发的“信息碎片”拼凑才能还原事实；寻求帮助是因为转发传递社会正能量帮助别人而转发层级较高；社会热点事件由于多是一些影响力大或击中老百姓心中绷得最紧的那根弦的事件，因此转发深度也较深。

总体来看，所有热门微博的平均转发深度为2.2层左右，转发信息链条中最大转发深度的平均值约为6.1层；相较于这个标准，风水财运、被删除微博、寻求帮助和社会热点事件四类热点微博的转发深度较深；促销类转发极值转发深度和平均转发深度都最低，一定程度上可以看出，所谓的“微博营销”在没有兴趣、幽默等元素植入的前提下，只有经济利益刺激的一哄而散的捧场效应，微博营销的自身价值很值得怀疑。

3. 转发宽度分布

各类别话题的转发宽度分布见表5。总体来看，热门微博的平均转发宽度为3.8左右，其极值转发宽度的均值为81.3左右。以此为标准，促销信息微博的转发宽度均值最大，极值转发宽度均值也最大，加上其转发深度较浅，说明促销信息的传播模型是“一哄而散式”的，恰好印证了“言之无文，行而不远”的说法；而风水财运、被删除微博和社会热点事件因为信息本身比较引人注意，在单位数量人群中被转发的概率较高，因此其转发宽度较小，转发深度较深，传播模型是细长形的“面条式”结构。

表 5　　各类别话题的转发宽度均值一览表

话题	转发宽度均值	极值转发宽度的均值	结果
时尚娱乐	3.934 2	105.461 3	中
社会热点事件	2.985 3	12.732 1	低
休闲心情	3.682 1	99.582 8	中
生活健康	3.841 0	126.778 7	中
寻求帮助	2.751 8	10.333 0	低
促销信息	18.941 8	350.531 3	非常高
风水财运	2.479 9	6.451 4	低
被删除微博	2.653 0	9.771 2	低
所有热点微博	3.844 8	81.296 7	

四、结论与讨论

通过以上的数据分析，可以得出几个结论。

第一，热点微博话题的传播呈现出一定的规律：转发次数的分布符合幂律分布，转发信息链的长度符合指数式分布特征。

第二，新浪微博具有强弱关系同时存在于一个平台的属性特征，这既不同于 Twitter 的社会单向度的弱关系平台，也不同于 Facebook 双向的强关系平台。按照微博鼻祖 Twitter 的最初设计，其更多扮演的是社会的信息源角色，每个人都可以有自己的媒体，这是一种社会单向的弱关系；而 Facebook 更多的是社会关系网的嵌入，是一种双向的强关系，这种关系所吸附的社会资本更多、更牢固。因此，新浪目前打造和建立的是一个“单向＋双向”的关系平台，也可以理解为是介乎 Twitter 和 Facebook 之间的一种平台。[3](P43－46)

第三，不同类别事件在新浪微博平台中传播的信息流和时间线也有所差异。根据不同类别热门微博的转发深度和转发宽度，可以构建出不同类别热门的传播模式结构（见图 4）。

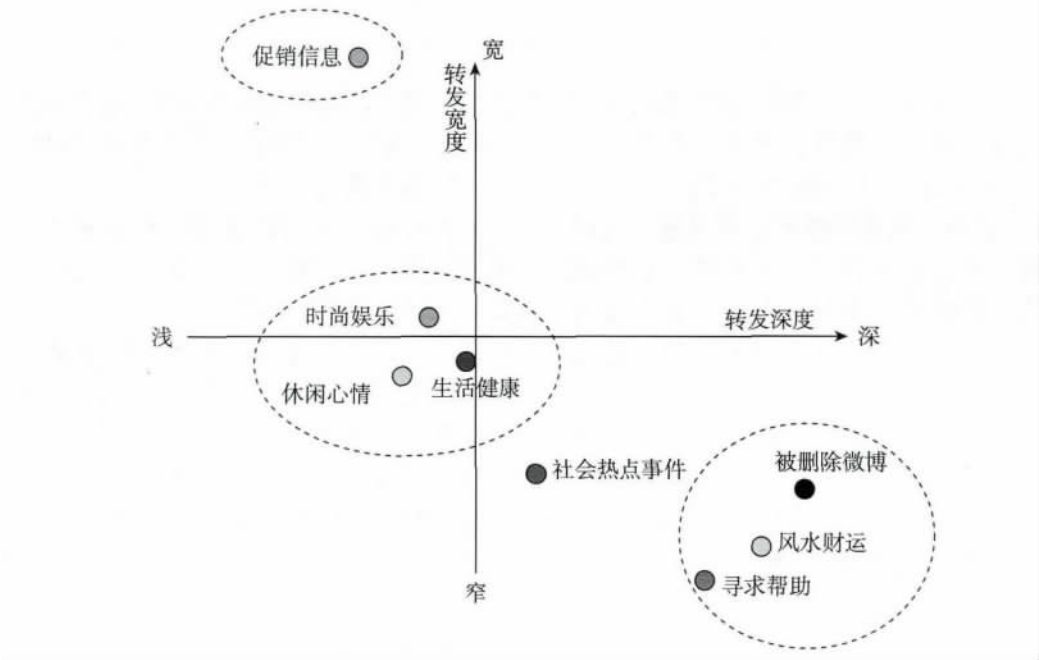

图 4　不同类别热门微博的传播模式结构

图 4 中横轴表示转发深度，纵轴表示转发宽度。可以将微博传播模式分成四种类型：深—宽传播结构、宽—浅传播结构、浅—窄传播结构、深—窄传播结构，其中时尚娱乐、促销信息属于宽—浅传播结构，这类信息的传播力有限，多是一哄而散的机制；休闲心情、生活健康信息属于浅—窄传播结构；社会热点事件、被删除微博、风水财运和寻求帮助类信息均属于深—窄传播结构，而目前看还不存在深—宽传播结构这一类别的微博信息。

如果按照传播模式进行类别归类的话，可以看出，休闲心情、生活健康和时尚娱乐信息传播结构大致相同，这一大类信息传播相对窄众，传播力不强；寻求帮助、被删除微博和风水财运微博传播结构大致相同，这一大类信息传播力较强，传播范围宽广，很容易引起民众的转发；社会热点事件单独属于一类信息，这类信息的传播力较高，覆盖的人群类别相对较广；促销信息也单独属于一类信息，其主要是因为经济利益的驱使，看似热热闹闹，但无论是影响力还是持久度都很低。

深—窄传播结构是一种效率高、传播范围广的长条形、多级传播结构，而浅—深传播结构是一种效率低、传播范围有限的扇面传播结构。

第四，微博中的话语权力格局中依然存在着不平等的现象。传统社会话语精英依然把持

微博话语场域的话语主导权，微博时代的话语平权只是“镜中花、水中月”，整个话语权力格局中依然是众星捧月式的“明星”模式，其中存在着男性微博用户主导话语权、社会话语精英群体把持社会话语权、经济社会发达地区民众掌控微博话语场域的议题设置权力等话语不平等现象。

在微博的社会话语权力格局中，“话语平权”依然“看上去很美”，微博从某种意义上带来的是“话语集权”，它通过“技术赋权”的方式让草根用户能够更多地“围观”热点事件，而其社会话语权力与新生代意见领袖依然存在不对等性，这种不对等性恰恰又是由技术决定的，微博中“关注”、“跟随”、“转发”功能，本身就是“再中心化”的过程。传统社会中金字塔形的话语结构被“投射”到微博虚拟话语场域中来，只是话语权力的主导者可能是一些“新贵”而已，“虚拟世界不再是‘像’现实世界，而是现实世界本来就有很大的‘虚拟’成分，所谓虚拟世界只不过还原了那种现实罢了”[4](P80－83)。

参考文献

[1] 中国互联网络信息中心：《第31次中国互联网络发展状况统计报告》，网易科技 http://tech.163.com/special/cnnic31/。

[2] 马海邻：《网友每日发布1亿条新浪微博》，凤凰网 http://tech.ifeng.com/internet/detail__2012__01/11/11893034__0.shtml?__from__ralated。

[3] 李彪：《微博盈利模式之惑——以新浪微博为例》，载《青年记者》，2012（16）。

[4] 魏武挥：《技术人格》，载《IT经理世界》，2012（12）。

Content Features and Spread Mechanism of Hot Topics on Micro-blog ——Based on Data Mining and Analysis of 6 025 Most Shared Weibo on Sina Micro-blog

LI Biao

(School of Journalism and Communication, Renmin University of China, Beijing 100872)

Abstract: Based on Big data mining and analysis of the 6 025 most shared weibo on Sina Micro-blog in these two years, this research has four findings: First, the spread of hot topics on Sina Micro-blog displays a regular pattern, that is, the number of sharing times follows the power-law distribution and the length of sharing chain follows the exponential distribution. Secondly, both strong ties and weak ties exist on Sina Micro-blog, and this is different from Twitter, which is a one-way weak tie platform, and different also from Facebook, which is a two-way strong tie platform. Thirdly, the spread of different kinds of topics has different information flows and timelines, and spread patterns of different kinds of topics can be built based on their sharing depth and sharing width. Fourthly, the structure of discourse power on micro-blog is still unequal and the traditional social discourse elites still control the leading discourse power on micro-blog.

Key words: data mining; hot topics; sharing depth; sharing width; structure of discourse power

（责任编辑　林　间）