社交网络营销追踪①

占 桓, 陈志德, 王 孟

(福建师范大学 数学与计算机科学学院, 福州 350007)
(福建师范大学 网络安全与密码技术福建省重点实验室, 福州 350007)

摘 要: 对社交网络营销效果追踪方面的研究尚处起步阶段. 根据社交网络营销效果追踪的需求, 设计了社交网络信息采集和数据处理平台. 该平台使用爬虫技术的方式, 在信息获取速度受限的前提下, 尽可能多的获取信息. 通过构建转发树的算法, 得到社交网络营销追踪传播图. 实验结果表明, 平台自动获取所需求的数据并处理, 绘制信息传播图, 由此找到信息传播中的引爆点及其评论信息, 即关系营销中重要关系节点及其反馈信息. 平台还可以统计出互动用户的其他相关信息, 便于社交网络营销效果的追踪.

关键词: 社交网络; 营销; 效果追踪; 传播分析; 数据可视化

Track of Social Network Marketing

ZHAN Huan, CHEN Zhi-De, WANG Meng

(College of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)
(Fujian Normal University, Network Security and Cryptography Key Laboratory of Fujian Province, Fuzhou 350007, China)

Abstract: The study on social network marketing effect tracing is still in the infancy. To meet the need of social network marketing effect, a platform is designed for social network information collection and data processing. The platform uses crawler technology to obtain information. Although the speed of information collection is limited, the platform can acquire as much information as possible. The spread diagram for social network marketing tracking will be obtained by building a forwarding tree algorithm. Experiment results show that the platform can automatically acquire and process the data, and then draw an information dissemination graph. The information dissemination graph is used to find the tipping point and comment information in the information dissemination, that is, the important relationship nodes and feedback information in relationship marketing. The platform can also obtain other related statistics information of interactive users to facilitate the tracking of social network marketing effect.

Key words: social network; marketing; track of effect; propagation analysis; data visualization

社交网络营销, 是以社交网络作为营销平台, 与客户或者潜在客户积极沟通, 商家分析消费者需求并给出相关产品和服务, 以进行品牌推广、活动策划、形象包装、产品宣传等一系列营销手段的网络活动. 社交网络营销有传播速度快、覆盖面广、目标人群集中化、影响力大、成本相对低廉、与客户互动性强的优势. 消费者在社交网络中可以品评商家的产品和服务, 商家及时得到消费者的反馈信息, 给出尽可能满足消费者需求的产品和服务, 消费者再和他人分享商家产品和服务的改进, 是一个循环互利的过程. 社交网络是一个将碎片化的用户重聚, 具有针对性的开展营销, 最终达到互利的营销效果的平台.

2009 年, 日本电通公司提出了基于网络时代特征的 AISAS 模型 (Kobayashi)[1]. AISAS 模型即: Attention(引起注意)、Interest(激发兴趣)、Search(信息搜索)、Action(购买行动)、Share(信息分享). 在全新的营销观念里, 两个具备社交网络特征的 Searh 和 Share 的出现, 突出了在互联网的 Web2.0 时代搜索和分享的重要性. 不再是 AIDMA 营销法则(Attention(注意)、Interest(兴趣)、Desire(欲望)、Memory(记忆)、Action(行动)

① 收稿时间:2014-06-16;收到修改稿时间:2014-07-30

里一味向用户进行单向的影响, 充分体现了社交网络营销里最重要的互动环节.

社交网络营销的核心是关系营销, 重点在于建立新关系, 巩固老关系. Jackson[2]明确提出关系营销概念: “关系营销是把营销活动看成是企业与消费者、供应商、分销商、竞争者、政府机构及其他公众进行互动的过程, 其核心是建立和发展企业与所有公众的良好关系”. 社交网络营销效果研究中最重要的就是得到企业所发信息转发评论的引爆点, 也就是重要的关系节点.

宁金鹏和王晓宇[3]构建了一个基于微博开放平台的在线营销系统. 通过开放 API 获取用户消息和相关信息, 重点研究了将不同开放平台的用户数据源进行融合处理, 形成统一的用户数据模型算法, 以便更好地利用各种微博平台开展营销. 李龙和李芝棠[4]提出了基于分布式的数据采集平台, 可应用于基于微博的舆情分析一级传播学和网络社会学等方面的研究. 冯典[5]、陈舜华[6]也研究了微博的信息爬取技术. 张彦超[7]等模拟了在线社交网络中信息传播过程, 并分析不同类型节点在网络中的行为规律. 仿真结果表明: 由于在线社交网络的高度连通性, 信息在网络中传播的门槛几乎为零; 初始传播节点的度越大, 信息越容易在网络中迅速传播; 中心节点具有较大的社会影响力; 具有不同度数的节点在网络中的变化趋势大体相同. SAITO K[8]和王昊翔[9]研究了社交网络中节点重要度的分析方法. 这些研究提出了一些数据采集平台, 但采集数据的目的性不强. 多采用 API 接口, 但 API 接口有越来越多的接口功能限制, 实际大规模获取数据应用中并不可取. 我们设计的信息采集平台主要采集社交网络中企业发布的每条信息的转发关系链和评论内容, 并可扩展采集消费者年龄、性别、地理位置等公开信息的功能, 目的是为分析社交网络营销效果提供数据量的支撑.

本文旨在设计一个信息收集与分析平台, 追踪每一次社交网络营销的效果, 得到用户的反馈信息. 通过构建社交网络营销信息采集与数据分析平台, 通过对每一次营销的数据进行统计, 得到企业短期和长期的消费者反馈信息, 使企业及时和消费者互动, 增强正面影响, 及时辟谣不实负面消息, 最终达到企业和消费者的双赢. 本文从社交网络信息采集平台的设计, 传播图构建算法, 可视化数据分析三个方面分析探讨.

1 平台设计

我们基于微博构建了信息采集和数据处理平台, 微博追踪过程分为信息采集和数据处理两步. 为了实现寻找关键节点能力, 我们分析了微博转发的机制, 根据实际转发机制设计了信息采集平台和构建传播树的算法.

如图 1 所示, 微博转发机制为: 微博在其首发者(A)转发页面上列出所有转发者(B、C、D、E、F、G、H、I)的转发信息. PC 网页版列表每页 20 条微博转发信息. 在每条信息里, 包括转发人转发时所写评论和转发时间, 其所有上级转发者 ID 及所写评论, 并用“//@”表示其转发层级关系.

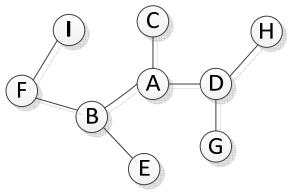


图 1 微博转发树形结构图

微博数据采集的方法主要有: 方法一, 使用模拟登录的爬虫技术获取微博页面, 再通过正则表达式匹配提取出内容信息和结构信息; 方法二, 使用开放接口 API. 但是 API 接口使用限制越来越多, 我们分析了 API 接口使用限制和功能限制, 发现不适合我们的采集任务, 于是决定采用第一种方法. 由于新浪、腾讯等开通了微博服务的公司对单个 IP 及单个用户的访问量做了限制(1000 次/小时), 不同于搜索引擎的信息采集, 单纯采用多线程编程在微博上采集数据的作用很小, 而且可能有 IP 被封的反作用. 唯一能提高效率的信息采集系统设计方法是采用分布式策略, 并用多线程技术处理各计算机的协调工作. 因此, 我们采用一台服务器为中心控制多台爬虫机获取数据的分布式系统设计方法.

分布式设计具有系统经济、可靠性高、可用性好、易于集成新系统、也易于扩展的特点. 与一个大型计算机相比, 由超级小型的计算机构建的分布式系统往往具有更高的性价比和实施灵活性. 由于是多台计算机协同处理, 分布式系统比集中式系统具有更高的可靠性和更好的可用性. 有若干台计算机崩溃, 不至于导致系统全盘崩溃. 如果系统效率还是不够高, 可以随时增加计算机, 达到扩展系统的目的来提高效率.

1.1 全局框架

全局设计上, 我们采用一台服务器进行任务分配和数据集中处理的工作, N 台爬虫机完成信息采集任务. 这样可以根据任务量的大小随时增减爬虫机, 采用中心化的设计方法, 减小爬虫机的压力, 不让爬虫机的性能成为采集数据的瓶颈. 整体设计如图 2 所示.

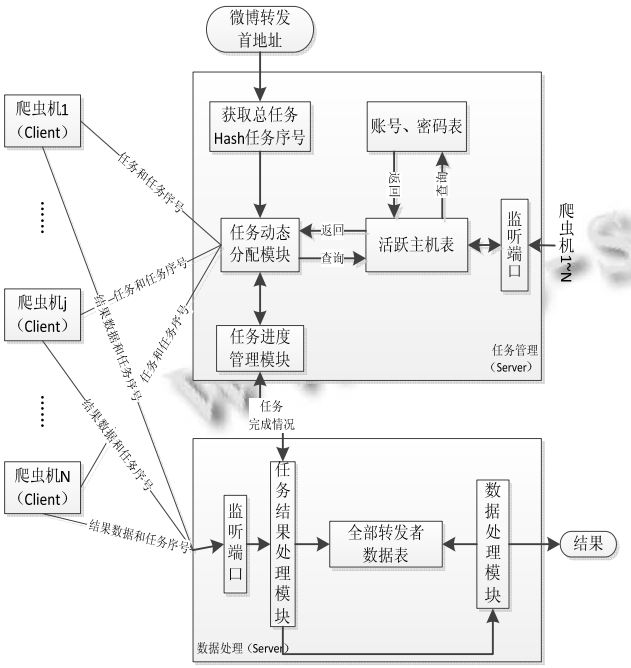


图 2　数据采集系统设计图

如图 2 所示, 我们只需配置需追踪微博的转发地址作为整个系统的入口. 在系统启动后, 监听端口程序开始工作, 监听来自爬虫机的信号, 随时刷新主机列表(hosts)中现在活跃的爬虫机信息. 根据配置的地址, 获取总任务模块获取转发评论列表的总页数(即获取所有分页的 url 地址). 获取完毕, 通知任务动态分配模块分配任务. 任务动态分配模块开始查询可用爬虫机(hosts 列表)数目并查询可用用户 ID 和密码分配给相应的主机, 然后把 url 列表切分成 N 份, 平均分配给各爬虫机. 任务分配完毕后, 继续监听 hosts 表, 如果有新的爬虫机加入, 重新分配未完成的任务(任务进度管理模块与数据处理服务器中的任务结果处理模块保持通信, 通过线程间通信的技术来实现). 这样可以在系统启动后根据需求和资源增减爬虫机数目提高效率和可靠性, 避免爬虫机的意外断电、掉线等情况造成系统崩溃. 爬虫机完成一次任务后, 开始在爬虫程序的空闲时段(为保证不超过 IP 访问量的限制, 2 次爬取任务有一段时间间隔)上传数据到服务器. 数据处理模块中的任务结果处理模块保持任务进度管理模块的通信, 完成结果入库后, 通知任务进度管理模块已完成的任务. 直到所有任务完成后任务结果处理模块开始通知数据处理模块开始任务. 完成传播图绘制、数据的统计.

1.2 爬虫机制

爬虫机采用模拟登录网页的方法获取数据. 爬虫机系统设计如图 3 所示.

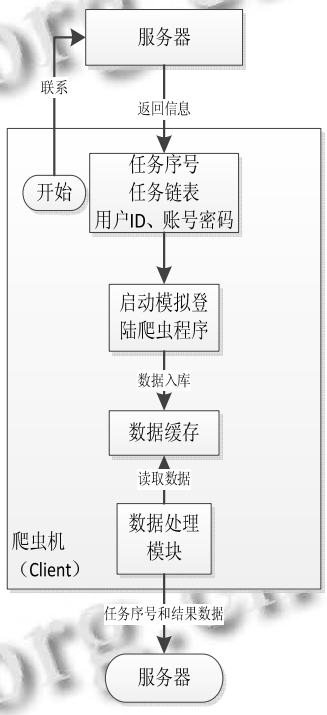


图 3　爬虫机设计图

爬虫机程序启动后, 首先和服务器中的任务管理模块取得联系, 等待任务分配, 当任务管理模块给出任务序号、url 地址链表和账号信息后, 爬虫机开始工作, 逐页获取信息. 具体实现为: 通过模拟浏览器登录微博的方法打开页面, 然后使用 HTML 网页解析和正则表达式匹配的技术逐页解析页面, 获取数据, 并保存数据到数据库. 爬虫机工作过程中, 需要注意 IP 访问量限制, 这些设置可以通过用户界面手动调节, 确保根据不同环境调整使得爬虫机效率达到高水平.

1.3 性能测试

我们对信息采集平台进行了性能测试. 采用的服务器及爬虫机信息为表 1 所示.

表 1 实验环境说明

计算机	硬件	操作系统	数据库
服务器	CPU: Xeon E3; 内存: 16G	Windows Server	MYSQL
爬虫机	CPU: i5 3470; 内存: 4G	Windows 7 64 位	MYSQL

完成一次 12457 条转发(锤子科技)[12]所用时间如图 4 所示.

图 4 任务完成日志图

我们设置平台默认采集频率为每隔 4 秒采集一页数据(我们测试的结果表明: 4 秒以下采集间隔时间会被微博反爬虫机器人识别为爬虫). 由任务完成日志记录可以看出, 在 12 台爬虫机同时工作的情况下, 我们仅仅用了 4 分 6 秒的时间, 就完成了 12467 条转发数据的采集. 在实际应用中, 企业微博转发量达到万条级以上的较低; 企业一天发布微博数量不会很大, 否则易造成粉丝反感; 一般而言, 企业所拥有的计算机远远大于 12 台; 其他信息采集可以在空闲时段完成. 基于以上四个理由, 我们认为设计的信息采集平台是可行的, 性能能满足企业的需求.

2 传播图的算法实现及分析

我们需要找出转发路径(传播路径图), 然后得到重要节点的评论信息. 由微博转发机制可知, 可以根据评论里的转发关系链设计相应的算法. 由于用户可能转发多次, 所以, 只用用户 ID 来确定转发路径是不可靠的, 我们用用户 ID 和评论作为一个整体来确定转发路径. 一般来说, 同一用户转发时所发的评论不会一样, 如果一样, 但是本质上是同一个人所说, 可以合并到一起. 所以, 我们在爬取所有转发者信息的同时也要获得转发关系链, 由转发关系链得到转发的树形结构图. 由此可以得到重要传播节点.

2.1 算法描述

我们设计 MBT(MicroBlog-Track)算法如下:

MBT 算法:

输入: 转发关系链.

输出: 树形结构图 TREE.

① 从数据库中一次性读取转发关系链, 存入列表 DATAS 中.

② 逐步处理 DATAS 列表中的转发关系链. 如果 DATAS 中的 1 个元素 data 不在 RELATIONS 中的任意一条关系链的子关系链中, 并且 data 中转发关系链的第一环不在 RELATIONS 中子关系链的任意一环中, 则保存关系链到 RELATIONS 列表中. 否则合并两个关系链成一个关系链保存到 RELATIONS 列表中.

③ 逐步处理 RELATIONS 的关系链, 构建嵌套字典 relations.

④ 如果 relations 的第一层关系不在字典 TREE 中的键 key 列表中, 则把 relations 添加进字典 TREE 中, 否则把 relations 和 TREE 中具有同样第一层关系的字典 tree 合并关系链之后添加进字典 TREE 中.

⑤ 循环处理未结束, 跳步骤③, 否则跳步骤⑥.

⑥ 处理完毕, 输出 TREE.

2.2 算法说明

步骤②中处理的是转发关系链太长被断开的情况. 因为微博限制了关系链显示的长度, 所以步骤②处理关系链的拼接. 步骤④中我们用字典这种数据结构构建传播树. 字典的数据结构为键值对, 一个键可以对应几个值, 我们把字典中的值也作为键构建了嵌套字典, 最后实现了的逻辑结构如图 1 所示. 也就是说, 最终结果为{A: C: {}, D: {G: {}, H: {}}, {B: {E: {}, F: {I: {}}}}}.

实际应用中, MBT 算法使用 Python 语言实现. 涉及到数据库的读取, 字符串的拼接、比较, 字符串转换为列表, 列表构建为字典等操作. 这些实现使用了 Python 自带的库函数, 考虑到读取源码计算时间复杂度的工作量巨大, 我们没有给出时间复杂度计算公式. 我们提供实际运行时间如图 5 所示. 其中横坐标表示转发量的大小, 纵坐标表示对应的算法实际运行时间.

我们选取的数据转发最大层数都在 7-9 层. 运行时间显示: MBT 算法复杂度为 O(n); 转发量 5 万左右时, 所用时间在 15 秒左右. 在实际应用中, 5 万条转发以上的微博量占比例比较少, 算法可用性强.

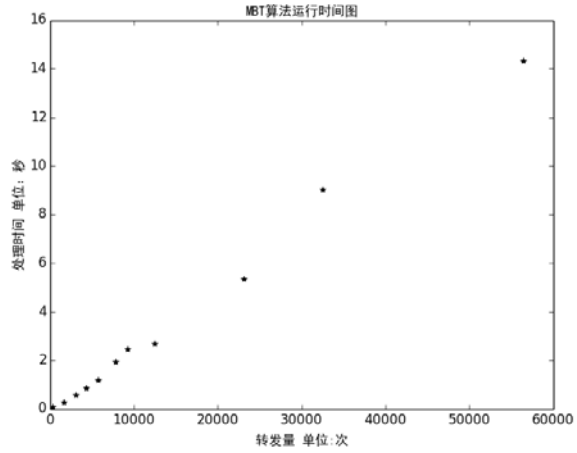


图 5　MBT 算法运行时间图

2.3 算法准确性分析

具体实现中我们用用户 ID 和评论作为一个整体来确定转发路径, 所以有一定的概率会出现不准确的可能性. 这种可能性的实际情况主要有以下三点:

① 用户会在一段时间后删除所转发的微博. 实际情况是: 影响力越大的用户, 转发微博越慎重, 删除的概率越低; 影响力越小的用户, 冲动转发微博的概率越大, 删除的可能性越高. 微博转发时间到采集时间之间的间隔越长, 删除的可能性越高. 也就是说, 信息采集的越及时, 受影响的概率越低.

② 用户在同一时间内多次转发并且评论相同, 由于评论相同, 而且是同一个用户, 我们把这种多次转发的情况合并为一次转发, 得到的结果同样具有参考价值.

③ 用户删除微博的行为我们无法控制, 删除之后, 其自身信息在转发页面会被删除. 但其信息还会保存在转发链中.

结论: 我们设计的的MBT算法在实际应用中有一定的误差, 但是可以通过及时采集的办法达到可控. 对用户删除微博的行为有一定的抵抗性, 可重现那些处于不在转发链末端的用户的转发评论信息. 基于以上两个理由, 我们认为通过MBT算法构建的传播树具有参考价值.

3　数据分析

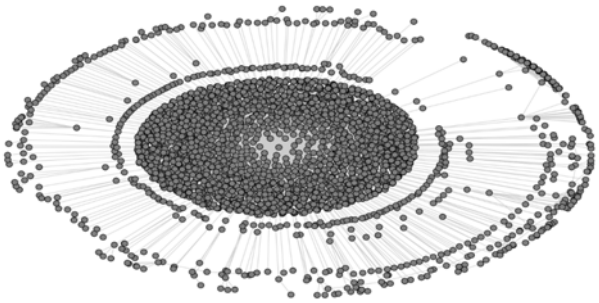
我们用知名公司的真实数据做个简单的统计分析. 分析方法为: 通过绘制出微博的网络传播图, 应用数据可视化技术来分析数据. 数据来源见参考文献. 首先我们给出相关概念的定义.

展示率: 展示率即看过某一条博文的用户数量, 展示率和转发人数和转发人的粉丝数有关.

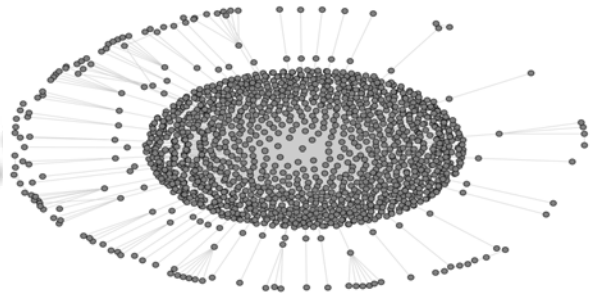
节点度: 与该节点相关联的边的条数, 也就是子级转发者的总数, 节点数越高, 一般而言, 表示展示率越高. 节点度高低可以从 TREE 中计算每个节点的子节点个数得到, 实际应用中, 我们也可以从图中直接看出那些节点拥有远大于其他节点的节点度.

传播深度: 此节点下级转发中的最大层数, 传播深度越大, 一般而言, 表示展示率越高. 实际应用中, 我们用颜色标记传播层级, 可以从图中直接看出最大传播深度.

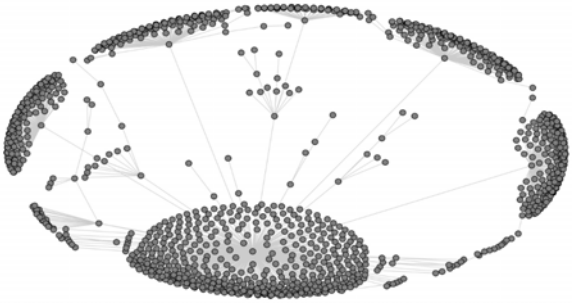
引爆点: 引爆点是一个相对的概念, 指在一条博文中拥有远大于其他节点的节点度或者传播深度. 我们可以从传播图中直接看出引爆点.



(a)　锤子科技传播图



(b)　魅族科技传播图



(c)　小米公司传播图

图 6　若干公司传播路径

从图 6 的传播路径可以看出, 锤子科技(图 6(a))[13]和魅族科技(图 6(b))[14]所发布这条博文没有引爆点的出现, 所有的传播都依赖于自身影响力. 由节点的密集程度可以看出, 锤子科技这条博文的转发量远远大于魅族科技这条博文的转发量, 每层的转发量都大于魅族科技. 这可以从图 8 所示的转发层级图得到佐证. 小米公司[15]所发微博重要的节点有 4 个, 如图 6(c)所示. 如果这四个节点不转发此条微博, 则转发次数锐减, 导致总体展示率锐减. 这表明了引爆点的作用. 我们找到的这四个引爆点分别为: 小米手机、红米手机、小米平板、神得强 Steven. 前三个 ID 是小米公司的官方运营账号, 神得强 Steven 这个 ID 为小米公司运维部经理所拥有. 进一步分析其转发评论可知: 大量转发的原因是使用了回答问题并转发得奖品的营销方法. 问题是说出最喜欢的 MIUI 功能, 从用户转发热情可以看出, MIUI 很多功能都深受“米粉”的喜欢. 我们从转发时间曲线进一步分析引爆点的作用. 转发时间曲线绘制方法为统计 10 分钟(实际系统设计时, 时间间隔可以调节)间隔时间内博文被转发的数量, 绘制转发时间曲线图如图 7 所示. 其中, 横坐标表示转发时间, 横坐标 0 表示微博发布传播十分钟后的转发量(纵坐标表示转发数量), 每隔十分钟统计一次. 即横坐标为 20 表示首发到现在历时 200 分钟.

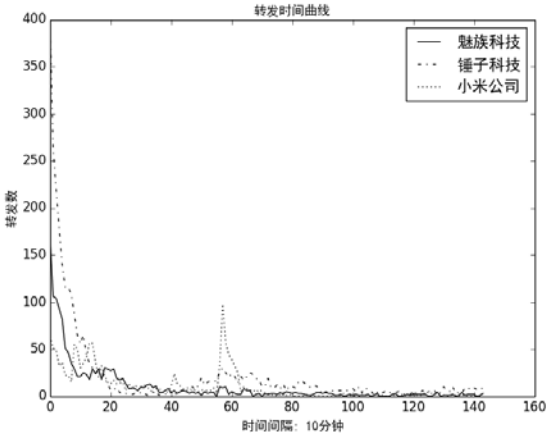


图 7 转发时间曲线

如图 7 所示, 锤子科技的博文在发布初期被大量转发, 转发量急剧下降后逐渐趋于平缓. 魅族科技总体趋于平缓. 小米公司的转发量在 6 小时后, 有一段骤升的过程, 我们按时间点找到了重要引爆点为: 小米手机、红米手机、小米平板. 这段转发骤升过程可以看出三个引爆点的作用. 而神得强 Steven 则在小米公司博文发布后 7 分钟时转发, 由图 7 的红色线条可以看出传播的初始阶段有两次上升过程. 这进一步说明了重要引爆点对于提升转发量, 提高展示率的作用.

经过对小米公司引爆点评论的分析可知: 四个引爆点评论内容为说出 MIUI 的优点, 属于正面营销评论, 而粉丝回答 MIUI 的优点也为正面营销互动.

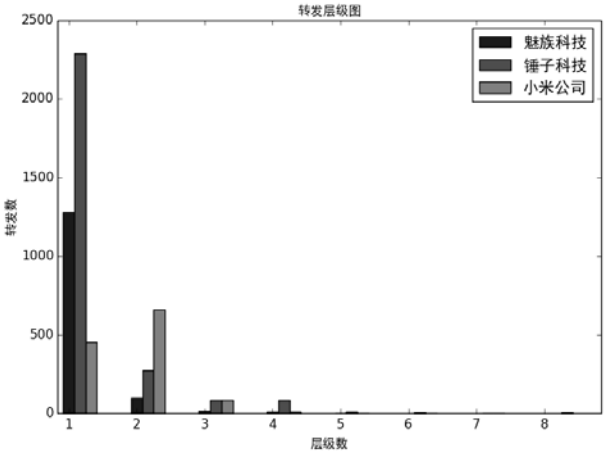


图 8 转发层级图

转发层级图可用于传播深度分析. 这三条微博的传播深度均不高, 最高为 8 层转发. 传播深度越高的用户, 其舆论导向的作用不可忽视. 实际分析得出: 只有小米公司的转发中出现了 1 次 8 层转发, 而这个节点属于回答自身粉丝问题的再次转发. 分析其评论可知: 其自身评论为: “终于听到 V6 的消息了”. 属于正面评价, 而其对粉丝的回复评论也表明其是一位忠实的“米粉”.

我们构建的信息采集平台通过自动化采集数据, 达到减少营销过程中人力资源分配的目的. 通过数据可视化的技术, 分析采集到的数据, 使营销效果一目了然, 有一定的实用性.

4 结语

本文构建了一个社交网络信息采集和数据分析平台. 根据网络营销效果分析的需求, 有针对性的采集社交网络上的数据. 我们设计了构建传播图的 MBT 算法, 通过绘制网络传播图, 使用数据可视化技术, 可以找到重要的转发节点, 获得引爆点的评论信息. 通过实际数据的分析, 得到引爆点在提升转发量、提高展示率上有巨大贡献的结论; 引爆点的评论, 对整体

舆论导向有重要作用的结论. 系统还可以做以下扩展: 获取并分析账号认证状态、性别、地域等数据. 通过持续统计客户的公开信息, 便于企业调整自己的产品和服务, 力求达到企业和消费者的互利共赢目的. 此平台有助于提高社交网络中信息传播分析和社交网络营销效果追踪的效率和准确性!并能应用于舆情分析、谣言追踪和控制等工作中.

参考文献

1 Kobayashi Y. A study of engagement in Japan. Aoyama Journal of Business, 2009, 43(4): 39–60.

2 Jackson BB. Winning and Keeping Industrial Customers: The Dynamics of Customer Relationships. Lexington, MA: Lexington Books, 1985.

3 宁金鹏,王晓宇.微博营销系统用户分析功能的设计和实现.计算机应用,2012,31(A02):233–236.

4 李龙,李芝棠,涂浩,等.一种分布式微博数据采集平台的设计与实现.广西大学学报(自然科学版),2012,36(A01): 324–328.)

5 冯典.面向微博的数据采集和分析系统的设计与实现[学位论文].北京邮电大学,2013.

6 陈舜华,王晓彤,郝志峰,等.基于微博 API 的分布式抓取技术.电信科学,2013,29(8):146–150.

7 张彦超,刘云,张海峰,等.基于在线社交网络的信息传播模型.物理学报,2011,60(5):050501-1–050501-7.

8 Saito K, Kimura M, Motoda H. Discovering influential nodes for SIS models in social networks. Discovery Science. Springer Berlin Heidelberg, 2009: 302–316.

9 王昊翔,曾珊,刘挥扬.虚拟社交网络中节点重要度分析.上海交通大学学报,2013,47(7):1055–1059.

10 Zhenhua Q, Minjun X, Zhihua N. A content tendency judgment algorithm for micro-blog platform. 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems(ICIS). IEEE. 2010, 3. 168–172.

11 Ounis I, Macdonald C, Lin J, et al. Overview of the trec-2011 microblog track. Proc. of the 20th Text REtrieval Conference (TREC 2011). 2011.

12 http://weibo.com/2968634427/B5lR9bSrh?mod=weibotime. 2014 June.

13 http://weibo.com/2968634427/B5duloqO1?mod=weibotime. 2014 June.

14 http://weibo. com/2683843043/B6EzN7PKm? mod= weibotime. 2014 June.

15 http://weibo.com/1771925961/B6mM3j44B?mod=weibotime. 2014 June.