系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
35(2) (2015, 2), 129–141

基于天涯论坛的 BBS 在线行为分析*

赵永亮 唐锡晋

(中国科学院数学与系统科学研究院, 北京 100190)

摘要 随着 Web 2.0 的出现以及社交网络的快速发展, 在线行为的研究日益重要. 故编制定向爬虫从 2010 年 10 月开始每日抓取天涯论坛. 文章基于抓取的 2012 年天涯杂谈板块的数据, 研究在线行为规律. 数据分析结果表明节假日及周末用户的发帖量减少; 用户的发帖行为符合日常作息规律, 有显著的日历效应; 点击量满足泊松分布与幂律分布的混合分布; 用户发帖量, 回复量和生存期均满足幂律分布. 说明只有少数的热帖具有较高的点击量或回复量和较长的生存期, 大部分的帖子缺乏关注. 提出一个帖子的热度计算公式并编制热帖推送程序. 研究发现更新帖中的热帖维持稳定. 进一步对这些热帖进行了社会风险分类.

关键词 在线行为, 天涯论坛, 幂律分布, 热帖.

MR(2000) 主题分类号 91D10, 91D30, 68P20, 90B90

ONLINE BEHAVIOR ANALYSIS BASED ON TIANYA FORUM

ZHAO Yongliang TANG Xijin

(*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190)

Abstract With the emergence of Web 2.0 and the rapid development of social networking, it is important to study online behavior. We started to download posts from Tianya Forum using spider program since October 2010. This paper analyzes the pattern of online behavior based on the posts at Tianya Zatan Board in 2012. The results show that users have fewer online activities on holidays and weekends, and users' posting behavior is in accordance with daily routine. Furthermore, the distribution of clicks follows a mixed distribution of Poisson and power-law, while the amount of user's posts, replies and survival periods satisfy power-law distribution. That is to say, only a few hot posts have high clicks or replies and long survival

* 国家重点基础研究发展计划项目 (2010CB731405), 国家自然科学基金 (71171187, 71371107, 61473284) 资助课题.

收稿日期: 2014-06-06.

编委: 房勇.

periods, while most posts are not concerned much, leading us to study hot posts. We propose one method to measure the hotness of the posts and develop a hot post push program. It is found that the very hot updated posts keep steady and their societal risk is labeled and observed.

Keywords Online behavior, Tianya Forum, power-law, hot posts.

1 引 言

Web 2.0 时代以来, 人们通过博客, 论坛, 微博等在线媒介即时表达了对社会热点, 民生问题等的大量看法. 这些看法不再仅局限于专家们的意见, 更多的是普通大众的看法. 因此, 对这些在线数据进行研究具有重要的意义. 这些在线数据是典型的大数据, 具有大数据的全体性, 混杂性, 复杂性等特点 [1], 这为对这些网络数据的挖掘增加了困难. 网络数据挖掘包括网络内容挖掘, 网络结构挖掘和网络行为挖掘等模式. 其中, 通过对网络行为的挖掘, 可以掌握用户的行为规律, 这具有重要的现实意义.

关于行为规律的研究已有很多成果. Barabási 揭示了在真实的社会网络中普遍存在的小世界与无标度特性 [2]. Li 等以人类的通信行为为例做了实证研究, 发现在人类的行为中普遍存在幂律特性 [3]. 李振鹏等对在人类行为中普遍存在的极化现象的机理进行了深入的研究 [4, 5]. 对于在线行为, 与传统的行为研究相比既有区别又有相似之处. Ari 等利用 Google 趋势对流行病进行了预测, 是在线行为分析较早的案例 [6]. Celli 等对社交网络 Friendfeed 中的在线行为进行了研究 [7]. Guan 等分析了新浪微博中热点事件的在线行为规律 [8]. Cui 等研究了网络论坛中帖子的热度计算以及热帖的回复量的分布规律 [9].

对论坛与微博的研究, 有异同之处. 微博依靠评论与转发得到关注, 而论坛帖子依靠点击与回复获得关注; 相对于微博, 论坛中的帖子一般较长, 能更加详细的叙述整个事情的来龙去脉. 因此, 本文以天涯论坛为研究对象, 研究用户的在线行为规律. 目前, 天涯社区每月覆盖用户超过 2 亿, 注册用户超过 9,000 万, 是华语圈首屈一指的网络事件与网络名人聚焦平台 [10]. 对天涯论坛中在线行为规律的研究具有重要意义, 如了解用户的行为规律, 探测网络事件的发展与演变, 引导网络事件向积极的方向发展, 有利于构建和谐社会.

下文主要分为四个部分. 第二节介绍了天涯论坛数据的获取与处理过程, 以及累积数据情况等. 第三节基于 2012 年天涯杂谈板块的数据, 对在线行为进行了分析, 研究了节假日, 周末和发帖时间对在线行为的影响, 以及用户发帖量, 点击量, 回复量与生存期的分布规律等. 第四节对热帖进行了研究, 包括热度计算, 热帖推送程序以及热帖的社会风险分类等. 第五节对全文进行了总结, 指出了本文的不足和待改进之处.

2 数据处理

本文采用网络挖掘技术对天涯论坛进行了挖掘. 编制爬虫程序从天涯论坛获取数据, 对数据作了清理和加工, 存储到数据库与 XML 文件系统中.

2.1 数据获取与处理

本文从天涯论坛获取数据, 天涯论坛包含众多板块, 各个板块的网页结构都相同. 图 1

呈现了天涯论坛的网页结构图, 对于一条帖子, 从左到右依次为帖子的标题, 作者, 点击量, 回复量与回复时间, 点击帖子的标题可以查看更加详细的信息. 各个板块的主页面均包含 80 条帖子, 这些帖子按照时间的倒序排列, 只有最新发表或者最新回复的帖子才会呈现在网页的前列, 如果想查看更久以前的帖子, 可以通过点击网页中的 “下一页” 翻页查看.

图 1 天涯论坛中某板块网页结构图

天涯论坛数据的获取与处理过程如图 2 所示. 数据的获取采用爬虫程序每日自动从天涯论坛下载. 该爬虫程序首先通过 Web 页面收集, 将天涯论坛中 “天涯杂谈”, “国际观察” 等与民生相关的版块中帖子的更新信息页面定时采集到本地机器. 之后通过 Web 页面过滤, 特征提取和信息抽取等过程, 抽取原始页面中的帖子及其更新信息, 存入数据库以及 XML 文件系统中 [11]. 在数据获取中存在两个关键问题: 第一个问题是自适应应对天涯论坛的改版, 如 2013 年 1 月天涯论坛推出了新版. 第二个问题是对缺失数据及时进行重新获取, 保证数据的完整性, 也为相关分析提供可靠的数据支撑.

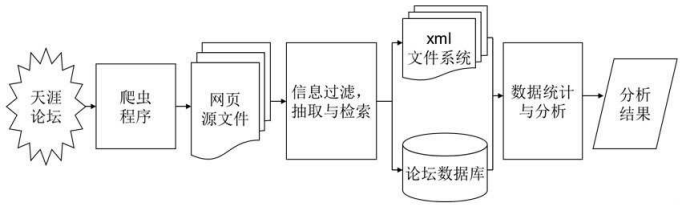

图 2 天涯论坛数据的处理过程

获取数据的存储采用 MySQL 数据库与 XML 文件系统相结合的方式, 兼有两者的优点. 对于 MySQL 数据库存储方式, 主要包含两张数据表: post 和 postupdate. 这两张数据表的结构如下所示:

post (pID, title, author, content, link, date_posted, time_posted, boardID),
postupdate (pID, date_update, time_update, clicks, replies),

其中, post 数据表存放帖子信息. pID 为自动递增主键, 唯一标识一条帖子. 其余字段分别存放帖子的标题, 作者, 首帖内容, 原始链接地址, 发表日期, 发表时间和所属板块等. postupdate

数据表存放帖子的更新信息, 以 pID 和 date_update 作为联合主键. 各字段的含义分别为帖子的主键, 帖子的更新日期, 帖子的更新时间, 点击量和回复量等.

对于 XML 文件系统存储方式, 除了存储帖子的基本信息, 还存储了每条帖子回复的具体信息. 采用 XML 文件的形式, 便于数据的管理和交换. 而且采用两种存储方式相结合的方法, 可以防止数据的丢失与破坏.

2.2 数据情况

中国科学院综合集成与知识科学小组从 2010 年 10 月开始利用爬虫程序从天涯论坛获取数据, 所抓取的板块包括天涯杂谈, 国际观察, 百姓声音与天天 315 等与民生相关的板块. 截至 2015 年 3 月, 抓取的每日新发帖的数量超过 3,000,000 条, 每日更新帖的数量超过 7,600,000 条. 在 XML 文件系统中文本数量达到 100.0 GB.

对于获取的数据, 已经做了一些研究. 张泽代等构建了数据抓取系统 [11]. 唐锡晋等应用天涯杂谈的帖子做了社会风险感知与和谐社会评测的尝试 [12]. 对帖子本身的分析也在进行, 在行为方面, 赵永亮等对用户的发帖行为规律进行了初步的分析 [13, 14]; 在内容方面, 曹丽娜等基于新发帖分析了论坛流行话题及其演化趋势 [15, 16], 陈近东等利用机器学习算法对帖子进行社会风险分类 [17].

本文对在线行为进行了深入的分析, 研究主要采用 2012 年天涯杂谈板块的所有数据. 在抓取的各个板块当中, 天涯杂谈板块是一个和网络事件以及社会舆情高度相关, 而且相对活跃的板块. 因此本文的研究使用天涯杂谈板块的数据, 这些数据包括每日的新发帖和更新帖, 其中, 新发帖指当日发表的帖子, 更新帖指在当天被更新过的帖子, 更新帖包括当日的新发帖, 而发帖量指新发帖数量或更新帖数量.

3 在线行为分析

对于天涯论坛, 个体用户的行为主要包括发帖, 点击以及回复等, 在宏观上就表现为发帖量, 点击量与回复量等定量数据. 本文将这些定量数据作为研究在线行为规律的指标. 具体的研究采用了 2012 年天涯杂谈板块的新发帖与更新帖. 研究内容包括帖子的数量统计; 节假日, 周末和发帖时间对在线行为的影响; 以及用户发帖量, 点击量, 回复量与生存期的分布等.

3.1 帖子基本情况

本文统计了每日的新发帖数量和更新帖数量, 对于缺失数据, 利用爬虫程序重新抓取; 无法重新获取的数据, 采用当月平均值代替当日值. 经过数据预处理后得到, 2012 年天涯杂谈板块的总新发帖为 409,717 条, 平均每日新发帖为 1,119 条, 其中, 每日新发帖数量的最大值为 1,927 条, 最小值为 185 条. 总更新帖为 1,195,125 条, 平均每日更新帖为 3,265 条, 其中, 每日更新帖数量的最大值为 4,755 条, 最小值为 789 条.

每月的更新帖和新发帖数量如图 3 所示. 从图 3 可以看出, 每月的新发帖数量与更新帖数量之比大约为 1:3. 除了 1 月和 2 月的新发帖和更新帖数量较少外, 从 3 月以后, 每月的新发帖和更新帖的数量整体保持稳定, 小幅度波动. 这说明天涯杂谈板块是一个用户活跃度高, 在线行为保持稳定的板块.

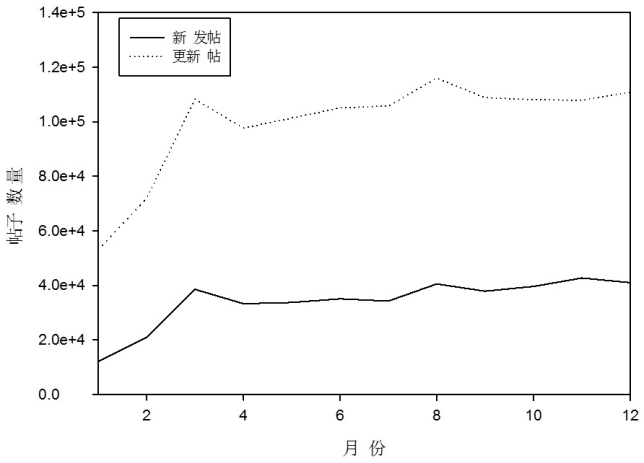

图 3 2012 年天涯杂谈板块每月新发帖和更新帖数量

3.2 节假日对在线行为的影响

为了研究节假日对在线行为的影响, 定义节假日的发帖率 r_H 如公式 (1) 所示, 发帖率反映了节假日的发帖量占平时的比重.

$$r_H = \frac{d_1}{d_2}, \tag{1}$$

其中, d_1 为节假日平均日发帖量, d_2 为节假日前后 m 天平均日发帖量, m 为节假日的长度.

利用发帖率的定义, 研究一些常见的法定节假日对在线行为的影响. 考虑的节假日包括元旦 (1.1–1.3), 春节 (1.22–1.28), 清明节 (4.2–4.4), 劳动节 (4.29–5.1) 以及中秋节和国庆节 (9.30–10.7), 由于 2012 年的中秋节和国庆节合并放假, 所以放在一起考虑. 各节假日的发帖率的计算结果如表 1 所示.

表 1 各节假日对应的发帖率

发帖率	元旦	春节	清明节	劳动节	国庆节	平均
新发帖	0.68	0.62	0.81	0.66	0.59	0.67
更新帖	0.58	0.71	0.94	0.83	0.76	0.76

从表 1 可以看出, 无论对于新发帖还是更新帖, 发帖率均小于 1, 其中最小值为 0.58. 说明节假日对在线行为有着重要的影响, 节假日的发帖量大约为非节假日的 $\frac{2}{3}$.

3.3 周末对在线行为的影响

同理, 为了研究周末对在线行为的影响, 定义周末的发帖率 r_W 如公式 (2) 所示. 其中, 工作日指周一到周五, 周末包括当周的周六和周日.

$$r_W = \frac{w}{d}, \tag{2}$$

其中, w 为周末平均日发帖量, d 为工作日平均日发帖量.

利用发帖率的定义, 研究周末对在线行为的影响. 将 2012 年每日发帖量以周为单位分开, 计算每周的发帖率, 然后取所有周的发帖率的平均值. 主要分三种情况计算发帖率, 即取所有周 (情况 1), 去除放假调休所涉及的周 (情况 2) 以及去除假期前后各 1 周 (情况 3). 各种情况下发帖率的计算结果如表 2 所示.

表 2 周末对应的发帖率

发帖率	情况 1	情况 2	情况 3	平均
新发帖	0.83	0.80	0.75	0.79
更新帖	0.90	0.89	0.86	0.88

从表 2 可以看出, 无论对于新发帖还是更新帖, 在三种情况下发帖率都小于 1, 最小值达到 0.75. 由于情况 3 完全剔除了节假日对用户发帖量的影响, 更能准确地表示周末对在线行为的影响. 这说明周末对发帖量有着重要的影响, 周末的发帖量大约为工作日的 $\frac{4}{5}$.

节假日与周末一般视为非工作时间, 用户在非工作时间的发帖率小的原因可能是, 网络用户在非工作时间内休息或休闲, 或者时间更多被现实世界中的活动所占据, 而在网上的活动相对于平时偏少, 从而导致发帖率偏低. 这种情况在对微博用户的行为进行研究时也有发现 [18].

3.4 发帖时间对在线行为的影响

为了研究发帖时间对在线行为的影响, 本文统计了一天 24 小时中每个小时的新发帖和更新帖数量, 结果如图 4 所示. 其中, 横坐标为发帖时间, 计时标准采用北京时间, 如 5 表示发帖时间为一天中的 5:00–5:59 之间.

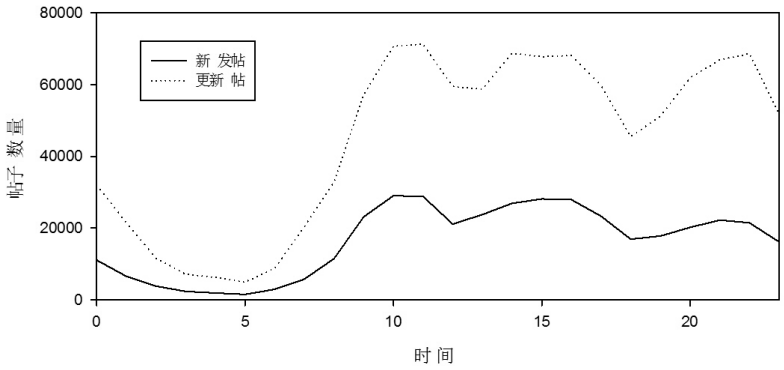

图 4 每小时的新发帖和更新帖数量

从图 4 可以看出, 新发帖数量和更新帖数量的变化基本保持一致. 发帖量在夜间达到最低, 在上午, 下午及晚上形成三个高峰, 在午餐以及晚餐的时候发帖量相对有所下降. 从发帖时间来看, 用户的发帖行为较符合人们的日常作息规律, 具有显著的日历效应.

3.5 用户发帖量的分布

用户发帖量反映了论坛用户的活跃程度. 为了研究用户发帖量的分布规律, 本文统计了发帖量相同的用户数. 统计结果表明, 随着发帖量的增加, 其对应的用户数不断减少, 但是降速不断减缓. 总共有 222,573 位用户发表了帖子, 其中仅发表一条帖子的用户达到 166,641 位, 占发帖用户总数的 74.87%; 发帖量大于 200 的用户仅有 23 位, 占发帖用户总数的 0.01%; 发帖量最多的为用户 “风云”, 发表了 2,009 条帖子. 这说明只有极少数用户的发帖量很高, 绝大多数用户的发帖量很低, 并不活跃.

为进一步探索用户发帖量的分布规律, 对发帖量及其对应的用户数取双对数, 采用最小二乘法进行线性回归, 回归结果如图 5 所示. 回归结果表明, 用户发帖量满足幂律分布, 其幂指数为 2.61, 且具有重尾特性.

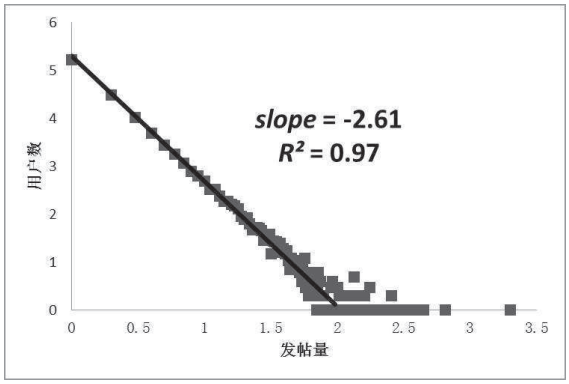

图 5 用户发帖量的分布

3.6 点击量的分布

点击量反映了用户对帖子的感兴趣程度. 为了研究点击量的分布规律, 本文统计了有相同点击量的每日新发帖的数量. 统计结果表明, 当点击量从 0 到 8 变化时, 对应的帖子数快速上升; 当点击量为 8 时, 帖子数最大, 达到 8,351 条; 当点击量大于 8 时, 随着点击量的增加, 帖子数不断下降, 但是下降的速率减慢. 点击量为 0 到 130 之间的帖子占总帖数的 80%; 当点击量大于 2,000 时, 其对应的帖子数均小于 10, 该部分帖子仅占总帖数的 1.64%. 这说明只有少数帖子的点击量很高, 绝大部分的帖子的点击量很低, 缺乏关注.

为进一步探索点击量的分布规律, 对点击量及其对应的帖子数取双对数, 采用最小二乘法进行线性回归, 回归结果如图 6 所示. 回归结果表明点击量满足泊松分布和幂律分布的混合分布. 当点击量小于 15 时, 可以用泊松分布描述; 当点击量大于 15 时, 可以用幂律分布描述, 其幂指数为 1.60, 且具有重尾特性.

点击量的分布表明在帖子刚发表时, 因为置顶容易看到, 用户就根据帖子的标题随机点击, 从而在点击量很小时呈现泊松分布的特性. 当帖子的点击量达到一定数目时, 用户就会抱着围观的心态来看帖, 用户更倾向于关注点击量高的帖子, 从而在点击量很大时呈现幂律分布的特性, 进而导致这种混合分布的产生.

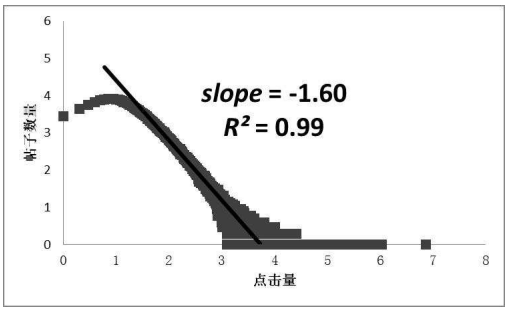

图 6 点击量的分布

3.7 回复量的分布

回复量反映了用户对帖子讨论的激烈程度. 为了研究回复量的分布规律, 本文统计了回复量相同的每日新发帖的数量. 统计结果表明, 随着回复量的增加, 其对应的帖子数不断减少, 但是降速不断减缓. 回复量为 0 到 5 之间的帖子占总帖数的 80%, 特别是回复量为 0 的帖子达到 163,714 条, 约占总帖数的 $\frac{1}{3}$. 当回复量大于 300 时, 其对应的帖子数均小于 10, 该部分帖子仅占总帖数的 0.17%. 这说明只有极少数帖子的回复量很高, 绝大部分的帖子的回复量很低, 不被用户关注与讨论, 关注很快被新帖取代.

为进一步探索回复量的分布规律, 对回复量及其对应的帖子数取双对数, 采用最小二乘法进行线性回归, 回归结果如图 7 所示. 回归结果表明, 回复量也满足幂律分布, 其幂指数为 2.07, 且具有重尾特性. 程崴等对网络论坛中回复量的分布规律进行了更细致的研究 [19].

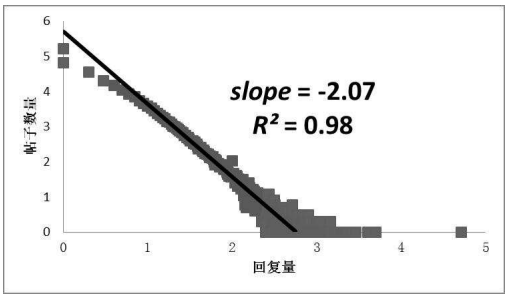

图 7 回复量的分布

3.8 生存期的分布

生存期指帖子从发帖之日起到最后一条回复之间间隔的天数. 如果生存期很短, 则表示帖子缺乏关注, 生存期反映了帖子持续讨论的时间长度. 为了探索生存期的分布规律, 本文计算了每条每日新发帖的生存期, 然后统计了生存期相同的每日新发帖数量. 为了保证每条帖子都有相同的考察时间, 以 1 年为期, 比如对于 2012 年 5 月 1 日发表的帖子, 计算其在 2013 年 5 月 1 日之前的最长生存期. 因删贴等原因, 共有有效帖子 398,003 条. 统计结果表明生存期大于 1 年的帖子总数为 15,758 条, 占所有帖子的 3.96%. 对于生存期在 1 年之内的贴子, 随着生存期的增加, 其对应的帖子数量不断减少. 生存期为 1 天的帖子达到 256,197 条, 占所有贴子的 64.37%. 说明绝大部分的贴子, 在当天发表后就沉没, 只有极少数的贴子保持回复而不沉没.

为进一步探索生存期的分布规律, 对生存期及其对应的帖子数取双对数, 采用最小二乘法进行线性回归, 回归结果如图 8 所示. 回归结果表明帖子的生存期满足幂律分布, 其幂指数为 1.45, 且具有重尾特性.

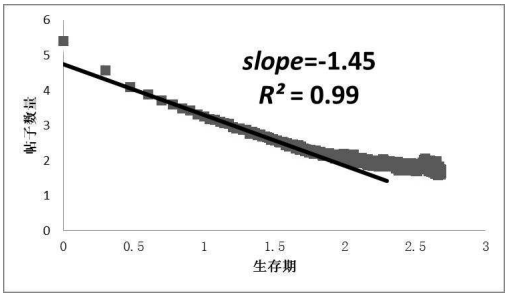

图 8 生存期的分布

点击量, 回复量以及生存期的无标度特性表明, 只有极少数的帖子的点击量或回复量很高, 绝大部分的帖子的点击量或回复量都很低; 只有极少数的帖子得到用户的持续关注, 绝大部分的帖子很快沉没. 也就是说, 用户把大部分的目光投向了极少数的帖子, 而绝大多数的帖子不被用户关注与讨论. 这极少数的帖子就是热帖, 这些热帖值得深入研究.

4 热帖分析

由上节可知, 仅有少数的热帖具有很高的点击量或回复量和很长的生命期, 这些热帖值得特别关注. 以下讨论帖子的热度计算, 并基于热度计算公式开发了每日热帖推送程序. 根据推送结果, 提取了天涯论坛保持持续更新的热帖, 并对它们进行了社会风险分类.

4.1 帖子的热度计算

对热帖进行研究的基础是帖子的热度计算. 帖子热度的衡量可以采用帖子的物理属性, 如点击量, 回复量, 回复点击比, 生存期, 回复人数等. 单一的物理属性反映了帖子的某一方面的特性, 但不够全面. 比如, 点击量高的帖子也可能是一些标题帖: 标题吸引用户点击进去, 帖子内容却都是无意义的内容, 甚至是广告. 回复量是一个比较好的指标, 但是它可能错过一些扎口的热帖, 这些热帖因扎口不能被回复而不会出现在更新帖榜的前面, 不易被人看到, 但在扎口之前也是热帖; 回复量高也可能是极少数人多次回复, 特别是广告推介等, 因此也需要考虑回复人数. 回复点击比试图寻找回复量与点击量均比较高的帖子, 但也可能是回复量与点击量均较低但两者的比例却很高的帖子. 生存期反映了帖子持续的时间长度, 但也可能是已经沉没很久的帖子又突然被回复, 导致帖子的生存期很长, 但回复量却很少.

因此, 需要采用物理属性的组合来衡量帖子的热度, 经过比较筛选, 点击量与回复量的组合是较好的选择. 本文用公式 (3) 来定义帖子 i 的热度 HD_i.

$$HD_i = \frac{w_1 c_i}{\mathrm{avg}(c)} + \frac{w_2 r_i}{\mathrm{avg}(r)}, \tag{3}$$

其中, c_i 表示第 i 条帖子的点击量; r_i 表示第 i 条帖子的回复量; $w_j (j = 1, 2)$ 为权重, 且

$\sum_j w_j = 1$, 在计算中取 $w_1 = w_2 = 0.5$, avg 表示所有帖子的平均值. 公式 (3) 试图寻找点击量与回复量都很高的帖子, 这些帖子得到关注与讨论, 就是所要的热帖.

4.2 热帖推送程序

在热度计算的基础上, 编制热帖邮件推送程序, 推送每日的热帖, 以实时观察热帖, 了解民生动态. 推送程序首先对每日所有的帖子按照热度从高到底排序, 然后提取排行前 20 的帖子推送. 推送程序每日推送的内容包括新发帖榜单和更新帖榜单, 对应于每一种榜单, 分别以点击量, 回复量, 回复点击比和热度为标准提取排行前 20 的热帖, 它们分别作为榜单的一部分. 图 9 示意了每日新发帖的点击量排行榜单, 图 10 示意了每日更新帖的热度排行榜单.

编号	标题	点击量	回复量	回复点击比	热度
1	南方日报杨兴乐记者受贿的文章太失水准(转载)(转载)	124787	0	0.00	56.39
2	网曝唐山远大职工变"临时工" "烧"完十亿资产玩破产(转载)	35782	175	0.00	25.95
3	政府大门清晨被砸，办公室惊现植物人	22478	7	0.00	10.55
4	吉林中院法官诈骗窝案"牵出"司法界"惊天假案"！(转载)	18970	52	0.00	11.48
⋮	⋮	⋮	⋮	⋮	⋮

图 9 每日新发帖的点击量排行榜示意图

编号	标题	点击量	回复量	回复点击比	热度
1	深度解析"韩寒挑战方舟子"一战究竟谁赢了？（技术帖直播）	20015465	1409585	0.07	329.30
2	从噩梦到天堂：离婚四年的成长史（性、爱、事业及其他）连载	16319489	239858	0.01	87.63
3	水泊梁山那些基情燃烧的岁月:妖言水浒之大宋盛世(笑死算自杀)1166页更新	12270456	127435	0.01	55.37
⋮	⋮	⋮	⋮	⋮	⋮

图 10 每日更新帖的热度排行榜示意图

推送程序的工作流程图如图 11 所示. 如第二节所示, 爬虫程序每日从天涯论坛抓取帖子并存储到 MySQL 数据库中, 然后推送程序从 MySQL 数据库中提取每日所有的新发帖与更新帖, 计算每条帖子的热度, 将所有帖子分别按照点击量, 回复量, 回复点击比和热度从高到底排序, 提取排行前 20 的帖子形成榜单. 系统提供了两种方式访问榜单, 第一种方式通过电子邮件发送给用户, 第二种方式通过 Web 访问.

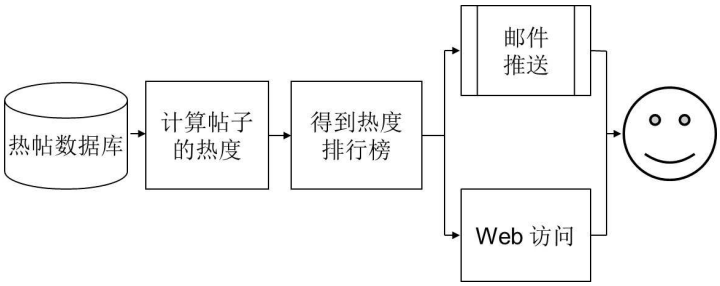

图 11 推送程序的工作流程图

4.3 热帖的社会风险分类

推送程序将每日的热帖推送给用户, 根据推送结果, 每日的新发帖中的热帖既与实时热点保持一致, 又能反映论坛的特色. 对于实时的热点问题, 几乎都能在论坛中引起热烈的讨论, 比如对于 2013 年度的热点事件李天一案, 在论坛中引发了旷日持久的讨论. 在论坛特色方面, 一些对贪污腐败的举报, 民生问题等事件也能引起热烈的讨论.

在更新帖方面, 研究发现每日的更新帖中的热帖几乎保持不变, 这是由于一些热帖几乎保持每日回复的缘故. 进一步总结发现这些热帖保持在 100 条左右, 其点击量以百万为量级, 回复量以万为量级. 按照帖子的内容, 对这些热帖的社会风险类别 [12] (共七个风险大类: 精神文明, 日常生活, 社会稳定, 政府执政, 资源环境, 国家安全和金融经济) 作人工标注, 统计结果如图 12 所示.

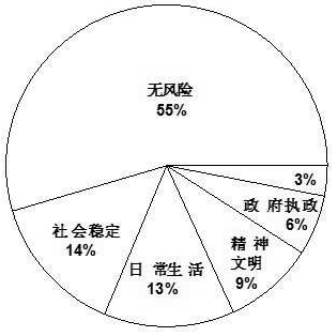

图 12 天涯杂谈板块更新帖中热帖的社会风险分类

从图 12 可以看出, 无风险的热帖最多, 占了所有热帖的 55%, 这些热帖主要是连载小说. 社会稳定类的热帖占到 14%, 包括朱令案件, 周克华案件等问题. 日常生活类的热帖占到 13%, 包括中医问题, 住房问题与工作问题等. 精神文明类的热帖占到 9%, 包括韩寒和方舟子之争, 婆媳关系等. 政府执政类的热帖占到 6%, 包括腐败问题, 新闻联播问题等. 其它依次为国家安全类, 金融经济类与资源环境类, 这些分类的贴子数量较少, 共占了 3 %.

这些更新帖中的热帖的社会风险比例与每日新发帖中的热帖的社会风险比例不同, 在新发帖中的热帖当中, 无风险的帖子仅占 10% 左右, 而在更新贴中达到了 55%, 无风险的热帖多为连载小说, 反映了有风险的热帖, 特别是政府执政和社会稳定类的热帖容易很快引起反响, 激起国民情绪, 但或许由于地域限制, 很快失去关注, 也或许因网管等人为行动而移除公众视线之外. 在社会风险分类中, 各个类别是不均衡的, 其中日常生活, 精神文明, 政府执政与社会稳定占了主要的部分, 这些主要类别也是网民最关注的问题, 其它风险类别的帖子较少.

5 结束语

互联网的飞速发展累积了海量的用户在线数据, 从在线数据中探测群体行为规律, 无论对于及时掌握用户的兴趣或行为的变化规律, 还是对于网络舆情监控与分析, 网络商业运营等都具有重要的理论和实践意义.

本文利用网络挖掘技术获取了天涯论坛中的帖子并对在线行为进行了分析. 首先介绍了天涯论坛数据的处理过程, 包括数据的获取和处理, 数据情况等. 然后基于 2012 年天涯杂谈板块的数据, 对在线行为进行了分析, 研究了节假日, 周末和发帖时间对在线行为的影响, 以及点击量, 回复量与生存期的分布规律等. 研究结果表明用户在非工作时间内的发帖较少, 发帖具有显著的日历效应; 用户发帖量, 点击量, 回复量与生存期均在不同的程度上满足幂律分布; 极少数的帖子得到了用户的大部分关注, 而绝大多数的帖子不被用户关注与讨论, 这促使了对这极少数热帖的研究. 为此本文提出帖子的热度计算公式, 并编制了热帖推送程序以便及时考察热帖的行为与内容.

本文的研究也存在一些不足和待改进之处. 随着数据量的不断增大, 需要基于 Hadoop 的数据处理平台的支持, 来提高数据分析的效率, 进行更大规模的数据分析. 对于在线行为的分析, 还需要更深入的研究. 本文只是对天涯杂谈板块的规律进行了研究, 对其他的板块, 包括国际观察, 百姓声音等板块还需要进行对比分析. 而且, 对于行为的研究需要细化, 比如, 一条热帖就是一个研讨过程, 这种在线研讨与传统的群体研讨有着本质的区别, 对其中的行为规律需要深入的分析 [20]. 更为本质的, 对天涯论坛数据的分析, 除了对于行为的分析外, 还应对帖子内容进行分析, 比如, 实现论坛热点事件的自动发现, 以及鉴别论坛热点问题中参与者的观点态度, 这些都是很有趣的问题, 当然也需要大量的工作.

参 考 文 献

[1] 维克托 · 迈克 - 舍恩伯格, 肯尼迪 · 库克耶. 大数据时代 — 生活, 工作与思维的大变革. 浙江人民出版社, 2013.

[2] Barabási A L. The architecture of complexity. *Control Systems, IEEE*, 2007, **27**(4): 33–42.

[3] Li N N, Zhang N, Zhou T. Empirical analysis on temporal statistics of human correspondence patterns. *Physica A: Statistical Mechanics and Its Applications*, 2008, **387**(25): 6391–6394.

[4] Li Z P, Tang X J. From global polarization to local social mechanisms: A study based on ABM and empirical data analysis. In Bai Q, Ren F, Zhang M, et al. eds., *Smart Modeling and Simulation for Complex Systems*, Studies in Computational Intelligence Volume 564, Springer, 2015, 29–40.

[5] 李振鹏, 唐锡晋. 集体行动的阈值模型. 系统科学与数学, 2014, **34**(5): 550–564.

[6] Ari S, Alison S, Kate G, et al. The utility of Google trends for epidemiological. *Geospatial Health*, 2010, **4**(2): 135–137.

[7] Celli F, Di Lascio F M L, Magnani M, Pacelli B, Rossi L. Social network data and practices: the case of Friendfeed. *Advances in Social Computing*, Springer Berlin Heidelberg, 2010, 346–353.

[8] Guan W, Gao H, Yang M, et al. Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events. *Physica A: Statistical Mechanics and Its Applications*, 2014, **395**: 340–351.

[9] Cui L J, He H, Liu W. Research on hot issues and evolutionary trends in network forums. *International Journal of* $\boldsymbol{u}$*- and* $\boldsymbol{e}$*-Service, Science and Technology*, 2013, **6**(2): 89–98.

[10] Introduction of Tianya Forum. http://help.tianya.cn/about/history/2011/06/02/166666.shtml.

[11] 张泽代, 唐锡晋. 面向天涯论坛的 Web 挖掘的初步研究. 系统科学与管理科学新理论, 新方法, 新技术及应用 (第十一届全国青年系统科学与管理科学学术会议暨第七届物流系统工程学术研讨会论文集), 武汉理工大学出版社, 2011, 199–204.

[12] Tang X J. Exploring on-line societal risk perception for harmonious society measurement. *Journal of Systems Science and Systems Engineering*, 2013, **22**(4): 469–486.

[13] 赵永亮, 唐锡晋. 基于天涯论坛的用户发帖行为规律研究. 大数据时代管理科学与系统科学的机遇与挑战 (第十二届全国青年管理科学与系统科学学术会议论文集), 厦门大学出版社, 2013, 305–313.

[14] Zhao Y L, Tang X J. A preliminary research of pattern of users' behavior based on Tianya Forum. Proceedings of the 14th International Symposium on Knowledge and Systems Sciences, Japan: JAIST Press, 2013, 139–145.

[15] Cao L N, Tang X J. Topics and trends of the online public concerns based on Tianya Forum. *Journal of Systems Science and Systems Engineering*, 2014, **23**(2): 212–230.

[16] 曹丽娜, 唐锡晋. 基于主题模型的 BBS 话题演化趋势分析. 管理科学学报, 2014, **17**(11): 109–121.

[17] Chen J D, Tang X J. Exploring societal risk classification of the posts of Tianya Club. *International Journal of Knowledge and Systems Science*, 2014, **5**(1): 36–48.

[18] Guo Z, Li Z, Tu H, Li L. Characterizing user behavior in Weibo. Mobile, Ubiquitous, and Intelligent Computing (MUSIC), 2012 Third FTRA International Conference on, IEEE, 2012, 60–65.

[19] 程葳, 钟华, 孙娇华. 网络论坛中发帖行为复杂性研究. 系统工程学报, 2009, **24**(4): 385–391.

[20] Zhao Y L, Tang X J. In-depth analysis of online hot discussion about TCM. Proceedings of the 15th International Symposium on Knowledge and Systems Sciences, Japan: JAIST Press, 2014, 275–283.