系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
40(10) (2020, 10), 1723–1736

天涯杂谈“红会贴”热点事件分析*

张亚茹[1,2] 唐锡晋[1,2]

(1. 中国科学院数学与系统科学研究院, 北京 100190; 2. 中国科学院大学, 北京 100049)

摘要 社交媒体网站为网民们提供了在线交流的平台, 在该平台下各种话题的发言层出不穷, 获取其中的热点事件有助于了解网民们关注的重点, 挖掘潜在信息. 文章以天涯杂谈“红会贴”为数据源, 基于社会网络分析方法, 获悉了“红会贴”热点事件、网民在线回复规律, 研究表明地震捐款问题、郭美美事件等是舆情焦点, 网民双向互动的情感极性基本是相同的, 互动中先发表负面言论的数目多于先发表正面言论的数目. 由于事件的多样性、动态性, 可视化表达热点事件的演化过程尤为重要. 文章提出了结合 LDA 主题模型和 Bert 向量抽取“红会贴”故事线的方法, 由此得到“红会贴”热点事件的发展脉络, 故事线披露出了红会的信任危机.

关键词 社会网络分析, LDA, Bert 向量, 故事线.

MR(2000) 主题分类号 91D30, 05C90, 68T50

Analysis of Hot Events of the Red Cross Posts from Tianya Club

ZHANG Yaru[1,2] TANG Xijin[1,2]

(1. *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190;
2. *University of Chinese Academy of Sciences, Beijing* 100049)

Abstract Social media websites provide online communication platform for netizens. There are a large number of speeches on various topics on the platform. Obtaining hot events from social media can help to understand the focus of netizens' attention and mine potential information. Based on the Red Cross posts from Tianya Club, this paper uses the social network analysis method to learn hot events of the Red Cross posts and rules of netizens' online responses. The research shows the issues of earthquake donation and Guo Meimei event are the public focus, the emotional polarities of interactions between netizens are basically the same, and the number of negative posts first is more than that of positive posts first in the interactions. Due to the diversity and dynamic of events, it is very significant to visually express the evolution process of hot events. This paper proposes the method of extracting the

* 国家重点研发计划基金项目 (2016YFB1000902), 国家自然科学基金 (71731002, 71971190) 资助课题.
收稿日期: 2019-11-12, 收到修改稿日期: 2019-12-31.
编委: 房勇.

storylines of the Red Cross relevant posts by combining LDA topic model and Bert vector. As a result, we obtain the development of the hot events, and the storylines reveal the trust crisis of the Red Cross.

Keywords Social network analysis, LDA, bert vector, storylines.

1 引 言

近年来随着社交媒体的发展, 舆情传播更加迅速, 微博、论坛成为网民在线分享、交流的重要平台. 如何从海量数据中获取网民关注的焦点, 获悉网民对某话题事件的立场、态度成为当前社交媒体领域的研究重点. 目前相关的研究有: 探索不同阵营的群体关注的热点 [1, 2]; 研究不同社交媒体中网民关注焦点的异同 [3]; 检测区域热点话题 [4]、实时热点话题 [5]; 预测某事件流行度 [6]; 探讨某舆情事件下网民的情感极性 [7, 8] 等. 与 Web2.0 时代的 Twitter、微博数据相比, Web1.0 时代的 BBS 允许更长的发帖, 文本内容偏日常化, 这使得数据存在大量噪声, 增加了文本处理的难度. 本文以天涯论坛中天涯杂谈版块的 "红会贴" 为数据源, 采用自然语言处理方法、社会网络分析方法捕获网民谈论的热点事件, 并探析用户双向互动的基本规律, 给出热点事件分析的新视角.

故事线抽取属于事件抽取的研究范畴, 用于研究事件演化, 图论法与文本相似性度量法是故事线抽取过程中常用的两类方法. Lin 等人根据用户输入的查询关键词组, 利用动态伪相关反馈的语言模型获得相关帖子, 接着基于帖子间的时间相关性与余弦相似性构建多视角图, 并由图的最小权重支配集提取有代表性的帖子, 最后应用斯坦纳树算法将这些帖子连接起来构成故事线, 作者在 Twitter 数据集上进行了全面实验证明了该方法的有效性 [9]. 这种方法依赖于高质量的查询集. Lu 等人将事件定义为一个包含时间、地点、参与者、事件短摘要、相关帖子的五元组, 使用联合相似性度量方法建立事件间的联系, 反向索引过滤和位置敏感哈希过滤被用来提高算法的效率, 由此自动抽取微博事件演化链. 该方法依赖于事先给定的具体事件 [10]. 刘充以百度贴吧热门话题下的帖子为数据源生成实时故事线. 首先利用完全重复去重与相似度去重的组合方式过滤冗余数据, 接着基于单纯形法得到有代表性的文本集合, 并以之作为文本摘要, 最后利用斯坦纳树算法增加光滑点使得文本摘要在时间和结构上更加连贯. 该方法使用的单纯形法在处理爆发性数据流时缺乏鲁棒性 [11]. Jia 和 Tang 用 TFIDF 向量表示文本, 依据内容一致性和时间一致性生成无向图, 再利用最小权重支配集和最小生成树算法生成天涯论坛中 "刘翔事件" 故事线 [12]. Xu 和 Tang 提出风险地图法, 用 Word2vec 对关键词进行表示, 使用聚类算法生成事件簇, 根据连贯性、覆盖面、连通性三准则构建了百度新闻 "红会" 故事线 [13]. 这两种算法利用词语获取文本相似性, 没有考虑文本的语义信息. 余玉轩等基于贝叶斯网络, 将故事线视为时间、机构、人物、地点、主题、关键词的联合概率分布, 就新闻文献抽取故事线, 该研究视一个话题为一条故事线, 没有考虑事件的交叉性, 认为每个话题是完全独立的 [14]. 在构建反映 "红会贴" 热点事件的相关网络之后, 为了全面把握红会事件演化信息, 本文将从帖子关键词、文本语义两方面出发, 提出结合 LDA 主题模型和 Bert 向量来抽取主贴故事线的方法, 进而对 "红会贴" 热点事件进行展开. 文章的研究框架如图 1 所示, 我们将用网络构建部分获取的热点事件考量故事线抽取效果.

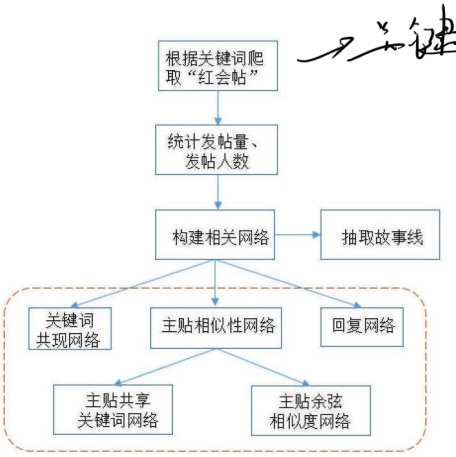

图 1 研究框架

(Figure 1 Research framework)

2 数据及分析

编制 python 爬虫程序抓取天涯杂谈版块标题包含 "红会" 、"红十字会" 且回复量大于等于 10 的帖子及相应回复, 共计 470 条主贴数据, 每条数据包含发帖人、发帖时间、标题、主贴内容、回复量、点击量、帖子编号等信息. 本节将分析帖子发布量随时间的分布, 统计每位作者的发帖量.

从帖子发布时间中匹配年月日, 统计每日发帖量, 如图 2 所示. 该图显示, 帖子主要集中在 2008 年、2011 年、2013 年, 且 2011 年的发帖比较密集, 这与现实情况吻合, 2008 年发生汶川地震, 2011 年出现郭美美事件, 2013 年发生雅安地震, 一时间将红会推上了舆论焦点. 对 470 条帖子的作者进行统计, 共计 406 位作者, 大部分作者只在天涯杂谈版块发布了一条 "红会贴", "周筱坳" 发了 12 条帖子. 发帖人数与发帖数量的关系如图 3 所示.

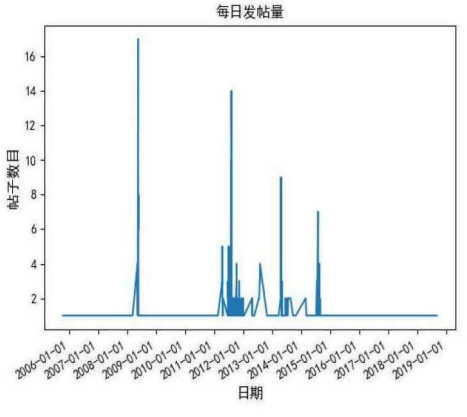

图 2 每日发帖量

(Figure 2 Number of posts per day)

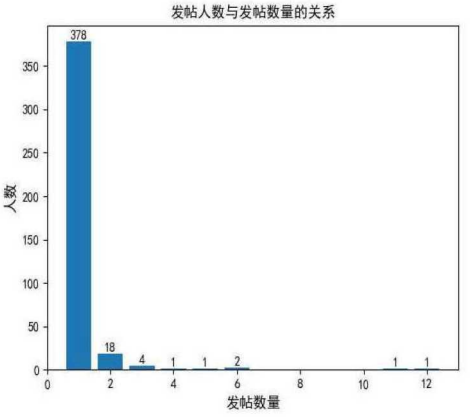

图 3 发帖人数的统计

(Figure 3 Statistics of the number of post authors)

3 关于发言及作者的相关网络构建

为了挖掘“红会贴”的潜在信息, 本节利用自然语言处理方法与可视化工具 Gephi[1] 构建不同类型的网络 (图 4、5、6、7、8), 通过分析网络特性, 多角度展示与红会相关的热点事件.

3.1 关键词共现网络

关键词共现网络以帖子中的关键词为节点, 关键词在帖子中的共现关系为连边, 使用基于模块度优化的 Louvain 算法[2] 对该网络进行社区划分, 进而获取热议关键词组. 用 TextRank 算法[3] 提取主贴文本前 20 个关键词, 选取前 8 个构造关键词共现网络. 共现次数大于等于 3 建立词语间的无向加权连边. 共计 176 个节点, 553 条边 (权重范围 3–25). 平均聚类系数为 0.73, 说明各节点的邻居节点连接紧密, 即通常几个关键词联合共现. 网络如图 4 所示. 其中节点的大小表示介数中心性的大小, 节点介数中心性指的是通过该节点的最短路径的期望数目, 用来衡量节点在网络中所起的中介作用. 边的粗细表示权重的大小. 网络中的节点被划分为 5 个社区, 有 83.52% 属于 0 社区 (矩形框右侧整个连通部分), 该社区包含了介数中心性最大的关键词: “红会”、“捐款”、“没有”、“中国”、“社会”、“问题”, 与这些词语相连的边权重较大; 矩形框中的节点属于 1 社区, 占 13.07%, 主要由 “学校”、“医院”、“孩子”、“家长”、“记者” 之类的关键词构成; 另外 3 个社区各有 2 个节点. 该网络表明 “红会贴” 的关键词主要有两大组别, 人们除了热议红会捐款问题外, 还关注有关学校、孩子的事件.

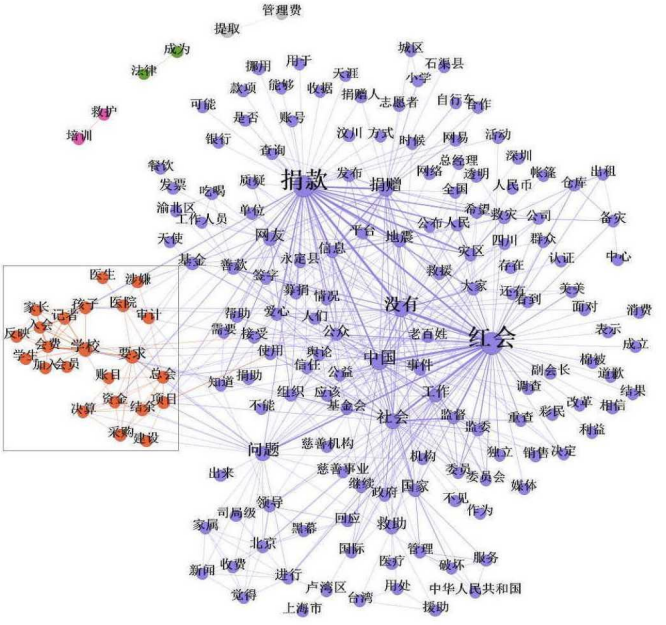

图 4 关键词共现网络

(Figure 4 Keyword co-occurrence network)

1. https://www.jianshu.com/p/86145943695a.
2. https://www.cnblogs.com/allanspark/p/4197980.html.
3. https://blog.csdn.net/wotui1842/article/details/80351386.

3.2 主贴相似性网络

3.2.1 主贴共享关键词网络

构造帖子间的共享关键词网络能够聚集相似帖子, 并获取关键性帖子. 以各主贴文本的前 20 个 TextRank 关键词为数据源, 当两贴的相同关键词数目大于等于 5 时, 建立连边. 共词网络含有 316 个节点, 952 条边, 平均聚类系数 0.475, 三角形数目 1270 个. 节点的大小表示 PageRank 的大小; 边的权重表示共词数目的多少. PageRank 由 Page 等学者首次提出, 用来在搜索引擎中根据网页超链接间的跳转为网页重要性排名, 直观解释是被很多重要页面指向的页面是重要页面[15]. 在主贴共享关键词网络中, 用它来衡量帖子在关键词涵盖方面的重要性. 一个节点的 PageRank 值可以用下面公式表示, A_{sv} 为邻接矩阵中 s、v 节点间相应的值, $d_s = \sum_t A_{st}$, $\frac{A_{sv}}{d_s}$ 为转移概率, $M(v)$ 表示节点 v 的全部邻居节点, α 为平滑系数.

$$PR(v) = \alpha \sum_{s \in M(v)} \frac{A_{sv}}{d_s} PR(s) + (1 - \alpha). \tag{1}$$

表 1 列出了网络中介数中心性、PageRank、度最大的前 5 个帖子标题, 出现 2 次以上的标题有如下 5 个: “红会的大型公益晚会是在利用白血病儿童为他们自己筹集钱财吗?” (2007-05-10)、“(关注红十字最新消息) 周润发捐 1 分成龙捐 6 毛红十字会: 10 万以下不可查” (2011-08-02)、“壹基金称红十字会捐赠信息查询数据有严重误差” (2011-08-03)、“江苏红十字会下半年接受个人捐款不到 2 万” (2011-12-11)、“[风青杨时评] 红十字会为何重新调查郭美美事件?” (2013-04-24), 这些标题揭示与红会相关的主要内容有白血病儿童公益活动、红会捐款情况、郭美美事件等.

表 1 帖子中心性分析结果

(Table 1 Centrality results of posts)

帖子标题 (ID)	介数	帖子标题 (ID)	PageRank	帖子标题 (ID)	度
[风青杨时评]红十字会为何 PK 不过壹基金? (3235440)	7024.44	红会的大型公益晚会是在利用白血病儿童为他们自己筹集钱财吗? (906094)	0.0159	[风青杨时评] 红十字会为何重新调查郭美美事件? (3242449)	33
红会的大型公益晚会是在利用白血病儿童为他们自己筹集钱财吗?(906094)	6401.47	[风青杨时评] 红十字会为何重新调查郭美美事件? (3242449)	0.0117	江苏红十字会下半年接受个人捐款不到 2 万 (2342681)	33
[风青杨时评] 红十字会为何重新调查郭美美事件? (3242449)	4759. 69	(关注红十字最新消息) 周润发捐 1 分成龙捐 6 毛红十字会: 10 万以下不可查 (2232772)	0.0113	(关注红十字最新消息) 周润发捐 1 分成龙捐 6 毛红十字会: 10 万以下不可查 (2232772)	32

续表 1 帖子中心性分析结果

(Table 1 Centrality results of posts (Continued))

帖子标题 (ID)	介数	帖子标题 (ID)	PageRank	帖子标题 (ID)	度
对中国红十字会不公正的评价 – 时下最大的 “冤假错案” (2233355)	3788.09	江苏红十字会下半年接受个人捐款不到 2 万 (2342681)	0.0112	红会: 运尸收费于情不合郭美美事件是成长烦恼 (2669807)	32
为什么红十字会不同意, 学校就不敢受捐 (2220195)	3493. 94	壹基金称红十字会捐赠信息查询数据有严重误差 (2234462)	0. 0110	壹基金称红十字会捐赠信息查询数据有严重误差 (2234462)	31

度大的节点关联的节点多, 说明该条帖子涵盖的信息丰富; 节点间的连边权重越大则说明两条帖子越相似. 因此对该网络进行如下过滤: 先查找度为 15–33 (33 为最大度) 的节点, 接着查找权重 7–20 的边, 以获取热点信息. 子网络有 32 个节点 (10.13%), 58 条边 (6.09%), 如图 5 所示. 这些节点构成的三角形数目多, 矩形框中的节点为割点, 割点右下部主要跟地震、捐款有关; 上部跟郭美美、红会社监委相关. 此外图中包含一些在原网络中度大但是连边权重小的孤立节点群, 涉及人们对 “红会为白血病儿童举办公益晚会” (2007-05-10)、“某学校不敢接受非红会善款” (2011-07-23) 的看法.

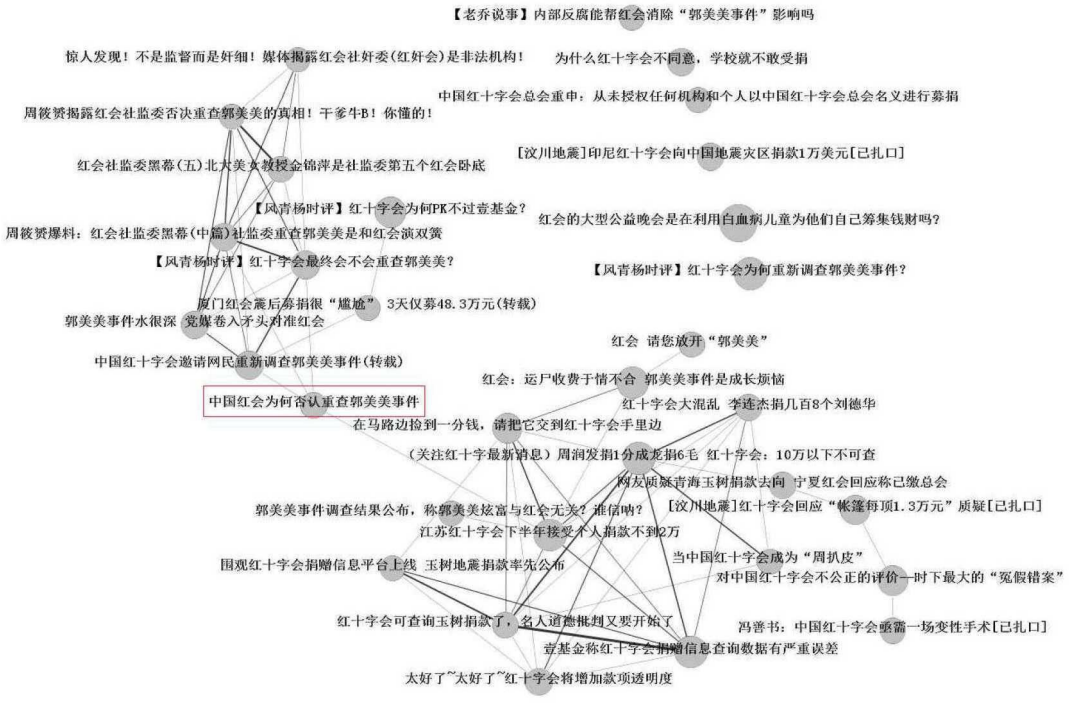

图 5 由度为 15–33 的节点、权重为 7–20 的边构成的帖子共词子网络

(Figure 5 The sub network composed of nodes with some degrees and edges with some weight)

3.2.2 主贴余弦相似度网络

主贴共享关键词网络从关键词的角度反映了帖子间的相似关系, 该部分将主贴文本向量与标题向量求平均作为帖子的向量表达, 并基于帖子间的余弦相似度构建无向加权网络, 以从文本角度分析帖子间的相似性, 进而挖掘热点事件. 主贴向量构建方法如下

1) 利用 TextRank4Sentence 函数从主贴文本中筛选出最重要的 3 个句子;

2) 用 Bert[4] (双向 Transformer 的编码器) 学习每个句子的向量表达 (768 维), 然后取平均, 阈值设为 90% 对该向量进行 PCA 降维, 最终得到 120 维的低维主贴文本向量;

3) 同样的, 利用 Bert 学习标题向量, 并降维至 120 维;

4) 将以上两低维向量的平均值作为对应帖子的向量表达.

将文本向量表达后, 求两两帖子之间的余弦相似度, 设置阈值 0.35, 即相似度大于等于 0.35 建立连边, 得到包含 141 个节点, 106 条边的网络. 网络平均聚类系数 0.542, 包含的三角形个数 32 个, 连通分量 52 个. 点与边的设置与主贴共词网络相似. 图 6 为网络的 2 个最大连通分量, 左部网民热议 “红会领导开豪车” (2011-08) 问题, 右部网民就 “中小学生被强制加入红会, 交纳会费” (2011-10) 舆情展开讨论. 图 7 是网络的 3 个次大连通分量, 依次显示了网民对 “红会出租救灾仓” (2014-08)、“红会重查郭美美” (2013-04)、“红会社监委” (2013-05) 的关注.

图 6 余弦相似度网络的 2 个最大连通分量

(Figure 6 Two largest connected components of cosine similarity network)

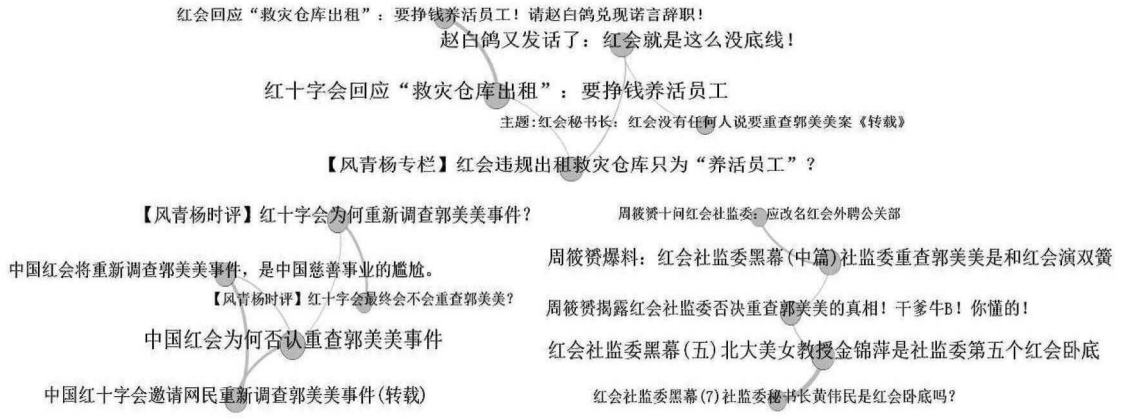

图 7 余弦相似度网络的 3 个次大连通分量

(Figure 7 Three second largest connected components of cosine similarity network)

4. https://blog.csdn.net/renyuanfang/article/details/86701148.

3.3 基于情感分析的回复网络

为了研究热帖发布后, 网民线上讨论的基本规律, 本节构造边属性为情感极性的回复网络, 分析回复现象. 在 470 条主帖中, 主贴 4497032 的回复贴最多, 根据这些回复贴中的 @ 关系以及回复时间抽取回复关系. 通常本楼的回复内容在最后一个层级, 因此抽取上述回复贴中最后一层级的文本作为回复内容, 并删除空回复和符号回复.

数据集 BosonNLP[5] 是基于微博、新闻、论坛等数据来源构建的情感词典, 包含情感词及其打分, 基于该情感词典以及程度副词词典, 否定词列表, 对回复内容进行情感分析. 处理过程如下

1) 对回复内容进行分句;

2) 遍历句中所有情感词, 每个情感词初始权重设为 1, 判断当前情感词与下一个情感词之间是否有程度副词和否定词 (第一个情感词则对其之前的内容进行判断), 否定词权重为 -1, 程度副词权重共有三种取值: 1.5、2、3, 将权重累乘来更新权重, 情感词分值乘以权重作为该情感词在句中的分值, 所有情感词句中分值相加即为句子的情感值;

3) 将所有句子的分数累加作为该回复内容的情感值, 取其符号作为最终的情感极性, 共计三种取值 1、-1、0.

获取用户间的指向关系以及回复内容的情感极性后, 构建回复网络, 每条边代表两用户间的一条回复, 属性即为情感极性. 该网络包含 368 个节点, 724 条边, 负极性边 455 条, 正极性边 242 条, 中性边 27 条, 负极性边占据总边数的 62.85%. 回复时间始于 2014 年 7 月 21 日止于 2014 年 8 月 5 日.

图 8 为过滤图: 先筛选出双向边, 接着在此基础上, 取度大于等于 2 的节点 (2- 核心节点), 即删除孤立节点. 共计 52 个节点 (14.13%), 273 条边 (37.71%, 包含了双向互动的所有连边), 44 组互动关系. 节点大小表示度的大小, 深色边表示负极性回复, 浅色边中除 “横扫草原的黑风暴” 与 “20032012”、“远方 2013” 与 “ly13” 间包含中性回复, 其余浅色边表示正极性回复. 该图显示一般双向互动的情感极性是一样的. 对 44 组互动关系开始与结束时的情感极性进行统计,

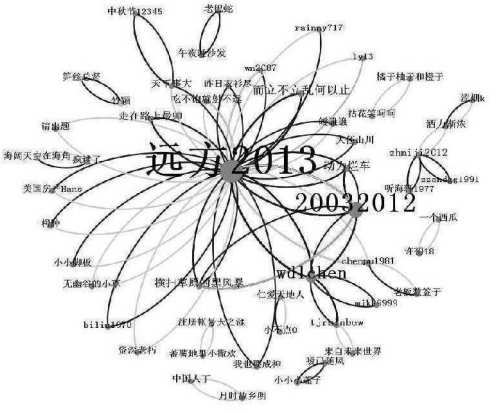

图 8 回复网络中的双向互动

(Figure 8 Two-way interactions in the reply network)

5. https://bosonnlp.com/dev/resource.

得到负 - 负、正 - 正、负 - 正、正 - 负、正 - 中、负 - 中的数目分别为 17, 10, 8,7, 1, 1, 有 26 组是先进行负向发言的, 18 组是先进行正向发言的. 在包含 “远方 2013” 的 23 组互动中, 只有 “远方 2013”–“ly13”、“远方 2013”–“天下事大” 这两组是 “远方 2013” 先与他人建立互动关系的, 由此可见 “远方 2013” 的初次发言吸引了网民们的眼球, 以致引起其他网民与之建立互动关系.

Li 和 Tang 等人 [16] 研究了天涯杂谈帖子的点击量与回复时间间隔的统计学规律, 发现群体新颖性和回复时间间隔均具有指数衰减的特性, 但与 Twitter 数据集比较, Twitter 的关注水平下降得更迅速. 本文将从双向互动持续时间的角度进一步探讨论坛中的回复行为. 图 9 对 44 组互动关系的持续时间进行了统计, 0 表示互动在一天内结束, 1 表示互动在 2 天内结束, 以此类推. 大部分的互动在一天内完成, 但仍有持续时间较长的互动, 这也反映了网民对 BBS 帖子关注的持久性.

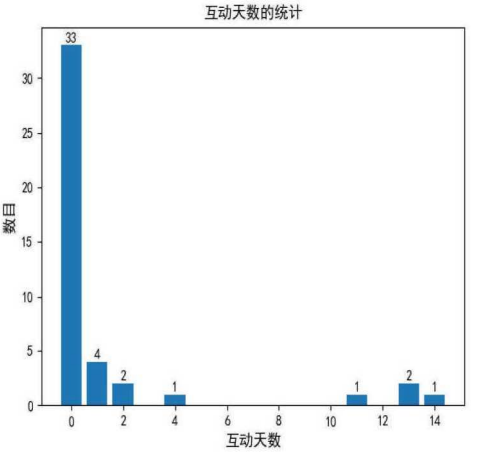

图 9 互动天数的统计

(Figure 9 Statistics of interaction days)

4 故事线抽取

第 3 节利用社会网络分析方法初步探析了 “红会贴” 的热点事件, 获取了用户发言的基本规律. 本节将从全局角度分析 “红会贴” 涉及的焦点问题, 通过抽取故事线概括帖子的主要内容、描述事件发展过程.

4.1 帖子话题分析

LDA 主题模型是一种基于概率的话题分析方法, 认为一篇文档中每个位置的单词按照多项分布选择某个主题, 再从这个主题中依据另一个多项分布选择某个词语, 即一篇文档代表了一些主题构成的一个概率分布, 而一个主题代表了很多词语构成的一个概率分布. 假设有 M 篇文档, K 个主题, $\vec{\alpha}$、$\vec{\beta}$ 为 Dirichlet 分布的先验参数, N_i 表示第 i 篇文档下的单词总数, 每个单词位置 $W_{i,j}, i = 1, 2, \cdots, M, j = 1, 2, \cdots, N_i$. LDA 主题模型的生成过程如下

1) 根据 $\vec{\theta_i} \sim Dir(\vec{\alpha}), i = 1, 2, \cdots, M$, 抽取第 i 篇文档下的主题分布;

2) 根据 $\vec{\varphi}_k \sim Dir(\vec{\beta}), k = 1, 2, \cdots, K$, 抽取第 k 个话题下的词语分布;

3) 由 $z_{i,j} \sim Mul(\vec{\theta}_i)$ 为每个单词选择主题;

4) 由 $w_{i,j} \sim Mul(\vec{\varphi}_{z_{i,j}})$ 为对应主题选择一个词语;

5) 循环执行上述步骤得到每个主题对应的词语.

文档的话题分布 $\vec{\theta}_i, i = 1, 2, \cdots, M$ 和话题下的词语分布 $\vec{\varphi}_k, k = 1, 2, \cdots, K$ 为需要估计的参数, 通常采用 Gibbs 采样法学习 LDA 模型参数, 从而能够得到每篇文档在各个话题下的概率, 选择概率最大的话题作为文档所在话题.

本文使用 LDA 主题模型对不同时间段的帖子进行话题分析, 从而聚集相同话题的帖子, 为后续故事线抽取做准备. 以主贴文本切词作为语料, 将 470 个数据集均分到 10 个时间段, 在话题个数区间 3–10 遍历, 根据最小困惑度选择每个时间段的最佳话题个数进行 LDA 话题分析. 图 10 显示总的话题数目为 64 个, 整体困惑度比较小. 进而可以得到每个话题下包含的帖子、每个话题的前 20 个关键词.

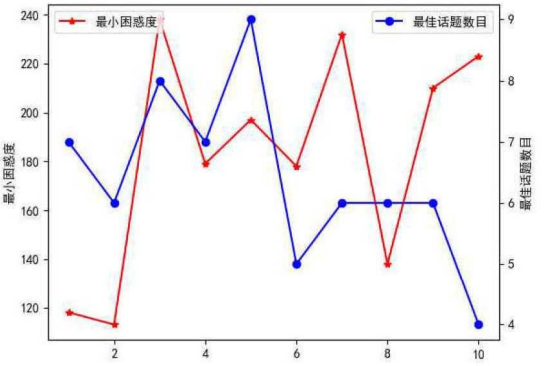

图 10 LDA 在各时间段的最小困惑度以及最佳话题数目

(Figure 10 The minimum perplexities and the best topic numbers of LDA in each time period)

4.2 故事线抽取算法

将每个话题下主贴文本的 Bert 向量求平均作为该话题的向量表达, 设置阈值, 当相邻时间段的两个话题的余弦相似度超过阈值, 建立有向边, 以将不同时间段的话题串接起来. 之后基于有向边寻找长度大于等于 5 的通路径, 由此得到 13 条候选故事线, 如图 11 所示, 共计 27 个节点, 25 条边, 06 表示第 0 个时间段的第 6 个话题.

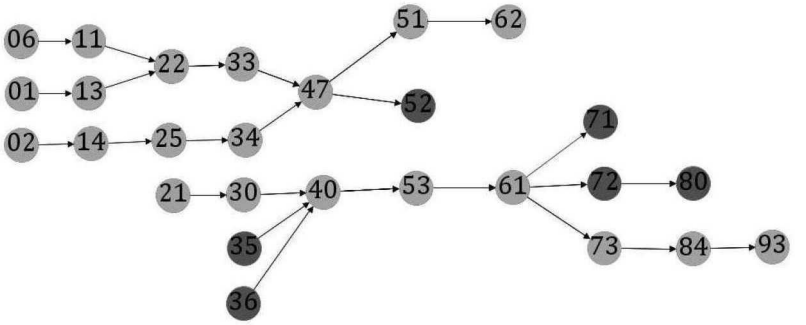

图 11 13 条候选故事线与 4 条最终故事线

(Figure 11 Thirteen candidate storylines and four final storylines)

图 11 显示故事线的连通性能 (故事线的交叉节点数目) 较好, 因此选择最终的故事线时, 我们主要考虑连贯性和覆盖面两个指标 [13]. 具体策略如下

1) 每步从候选路径中选择最长的路径 (覆盖面);

2) 如果路径长度相同, 选择路径上边的平均相似度大的那一条 (平均相似度各不相同) (连贯性), 直至故事线上的节点数目达到候选故事线节点数目的 75% 以上.

图 11 中的浅色节点及其形成的路径表示最终的故事线. 共计 21 个节点, 19 条边, 4 条线.

图 12 标注了每个节点有代表性的关键词或具体事件, 时间注明的是每一时间段的开始时间与结束时间. 故事线的时间段按照年份可划分为三段, 依次反映红会在 2008 年的汶川地震、2011 年的郭美美事件、2013 年的雅安地震以及相应年份的其它热点事件上的舆情演化. 四条线的含义如下: 第一条线: 组织给某县级学校捐款, 学校称只接受县红会的捐赠, 接着公众、记者进行查询, 红会领导做出回应澄清事实, 企业得以为该所学校做公益. 接着四川雅安发生地震, 在实施救助的同时, 民间爆出红会社监委的黑幕, 红会副会长赵白鸽接受采访就问题给出官方回应. 之后发现红会违规出租救灾仓库, 红会就此再次做出回应称为解决员工工资. 第二条线: 红会拒绝公布账目, 网易因此终止与其合作. 接着网爆重庆红会用救灾善款吃喝, 红会否认; 上海红会工作人员餐费超出标准, 红会责令其退回超额款项. 之后新闻媒体传播郭美美事件以及红会领导开豪车事件, 民众质疑、红会辟谣, 而郭美美接受采访并声明 “不想因为自己让百姓对红会失去信心”. 继信任危机后红会发起有关自然灾害的捐款, 社会民众的参与度逐渐降低. 第三条线: 汶川地震发生, 民众质疑红会提供的帐篷价格, 并不满于红会将部分款项留作本地慈善基金、根据国际惯例收取物资管理费. 接着公众关注点转移到红会对患白血病孩子以及温州动车事故遇难者的捐赠上, 之后转向第二条线后续部分. 第四条线: 网站登录汶川地震消息, 全民提供物资, 进行捐款, 但民众认为相比于捐款, 捐赠物资更保险一点, 红会表明会将款项用于灾区救助. 之后与第三条线汇合. 根据故事线的内容, 我们把第一条线称为 “回应线”, 第二、三、四条线称为 “信任危机线”.

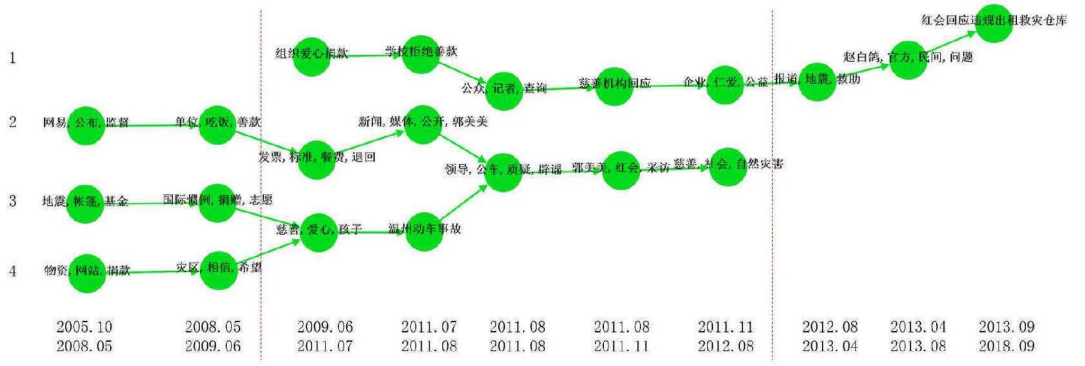

图 12 从天涯杂谈 “红会贴” 抽取的 4 条故事线

(Figure 12 Four storylines extracted from the Red Cross posts of Tianya Club)

将故事线涉及的热点事件与社会网络分析结果进行比较以评价覆盖面, 图 5–7 共包含 8 个热点事件, 图 12 涵盖其中 6 个, 覆盖率为 0.75. 利用话题关键词集计算每条线相邻时间段的 Jaccard 相似度, 再取平均作为该条线的连贯性, 4 条故事线的平均连贯性为 0.177, 相当于故事线中每相邻时间段话题关键词集 (20 个关键词) 中有 6 个词语是相同的. 此外, 各条线分别覆盖的帖子数目为 61、50、40、64, 四条线共覆盖 162 条 (34.47%) 帖子. 该故事线不仅从全局角度探析了 “红会贴” 热点事件, 而且将事件进行了展开, 明确了事件的来龙去脉.

5 结束语

本文以天涯杂谈 “红会贴” 为数据源, 通过网络构建与故事线抽取两方面工作, 获取了热点事件并自动生成了事件动态演化图谱. 在网络构建部分, 关键词共现网络划分结果揭示出网民除了热议红会捐款问题外, 还关注有关学校、孩子的事件. 通过分析主贴相似性网络的节点中心性和连通分量, 文章从主贴关键词角度和帖子文本角度展现出了 “红会帖” 的舆论热点: 郭美美事件、地震捐款问题、有关孩子的舆情事件、红会领导开豪车等. 针对某条热帖的回复贴构建的边属性为情感极性的回复网络表明活跃的人只占少部分, 双向互动的情感极性基本是相同的, 负向先行发言多于正向先行发言, 互动大部分能够在一天中完成, 这一定程度上反映出了论坛中网民的回复规律. 故事线抽取部分首先利用 LDA 主题模型对不同时间段的帖子进行话题分析, 从而聚集相同话题的帖子, 接着用 Bert 向量表达主贴文本, 并将每个话题下主贴文本的 Bert 向量求平均作为该话题的向量表达, 最后以连通性、连贯性、覆盖面为评价指标, 串接相邻时间段的话题构建故事线. 故事线对上述舆情热点进行了展开, 披露出红会的信任危机.

本文提供了一种从过载信息中捕获网络热点事件的新思路, 它既利用了中心性、联通分量等网络属性又结合了文本相似性度量, 能够推广到各种社交媒体平台中以获取网民关注焦点, 了解网民诉求, 管理网络舆情等. 由于事件的动态性, 监测事件的演化、跟踪事件的发展尤为重要, 并且相对于单一事件, 人们更关注于关联事件的演化. 文章提出的故事线抽取方法, 有效地展示了与红会相关的热点事件的演化过程. 红会作为一个大型慈善机构, 关乎着社会救助能否及时到位, 关系着民生发展, 与红会相关的舆情始终抓人眼球, 如新冠肺炎疫情期间, 武汉红会被爆办事不利, 物资分配不公, 对红会舆情的追踪仍然是今后的一项

重要研究工作. 未来也期望将该故事线抽取算法拓展到其它语料中, 以达到普适性应用, 例如搜集网络平台中关于某企业的热点言论, 建立企业舆情故事线, 以使企业及时感知自身风险、转换模式、遏制谣言等; 搜集微博中热点话题下的碎片化信息, 建立实时事件演化链, 以使网民能够便捷地获取关联内容, 理清事件发展. 本文仅结合社会网络分析结果对故事线进行了评测, 缺乏对比实验, 今后将尝试使用多种方法提取故事线, 比较实验结果, 全面把握事件演化过程. 此外, 本文工作主要集中于对在线社交媒体事件的分析, 对于网民在线行为的分析有待深入. 未来将重点研究网民回复意愿、转载行为, 这对促进正面事件的传播与制止不正当言论的扩散极为重要.

参 考 文 献

[1] Wang C, Tang X J. Stance analysis for debates on traditional Chinese medicine at Tianya Forum. Proceedings of the Conference on Lecture Notes in Computer Science, 2016, 321–332.

[2] Becatti C, Caldarelli G, Lambiotte R, et al. Extracting significant signal of news consumption from social networks: The case of Twitter in Italian political elections. *Palgrave Communications*, 2019, **5**(1): 91.

[3] 贾玉改, 唐锡晋. 在线群体对社会风险事件关注焦点研究. 系统科学与数学, 2017, **37**(11): 2178–2191.
(Jia Y G, Tang X J. Exploring the focus of online community on societal risk event. *Journal of Systems Science and Mathematical Sciences*, 2017, **37**(11): 2178–2191.)

[4] 曹丽娜, 唐锡晋. BBS 话题的地理分布分析. 系统科学与数学, 2016, **36**(5): 671–682.
(Cao L N, Tang X J. Analysis of topics distribution in geography based on BBS. *Journal of Systems Science and Mathematical Sciences*, 2016, **36**(5): 671–682.)

[5] Choi H J, Park C H. Emerging topic detection in Twitter stream based on high utility pattern mining. *Expert Systems with Applications*, 2019, **115**: 27–36.

[6] Moniz N, Torgo L. A review on web content popularity prediction: Issues and open challenges. *Online Social Networks and Media*, 2019, **12**: 1–20.

[7] 许诺, 唐锡晋. 基于天涯论坛球迷情感分析与行为挖掘. 系统科学与数学, 2017, **37**(9): 1915–1929.
(Xu N, Tang X J. The sentiment analysis and behaviors mining of posts from fan's home board of Tianya Forum. *Journal of Systems Science and Mathematical Sciences*, 2017, **37**(9): 1915–1929.)

[8] 李泉, 李萌, 成洪权, 等. 基于文本聚类与情感分析的群租房微博舆情量化研究. 图书情报研究, 2019, **12**(1): 82–89, 105.
(Li Q, Li M, Cheng H Q, et al. Public opinions of group leasing in Chinese social media: A research based on text cluster and sentiment analysis. *Journal of Library and Information Research*, 2019, **12**(1): 82–89, 105.)

[9] Lin C, Lin C, Li J X, et al. Generating event storylines from Microblogs. Proceedings of the Conference on Information and Knowledge Management, 2012, 175–184.

[10] Lu Z Y, Yu W R, Zhang R C, et al. Discovering event evolution chain in Microblog. 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, 2015, 635–640.

[11] 刘充. 基于中文社区的实时故事线脉络生成. 硕士论文. 厦门大学, 厦门, 2017.
(Liu C. Generating real time event storylines base on Chinese community. Master Thesis. Xiamen University, Xiamen, 2017.)

[12] Jia Y G, Tang X J. Generating storyline with societal risk from Tianya Club. *Journal of Systems Science and Information*, 2017, **5**(6): 524–536.

[13] Xu N, Tang X J. Generating risk maps for evolution analysis of societal risk events. Proceedings of the Conference on Knowledge and Systems Sciences, 2018, 115–128.

[14] 佘玉轩, 熊赟. 基于贝叶斯网络的故事线挖掘算法. 计算机工程, 2018, **44**(3): 55–59.
(She Y X, Xiong Y. Storyline mining algorithm based on bayesian network. *Journal of Computer Engineering*, 2018, **44**(3): 55–59.)

[15] Brin S, Page L, Motwami R, et al. The PageRank citation ranking: Bringing order to the web. Technical Report, Computer Science Department, Stanford University, 1998.

[16] Li Z P, Tang X J, Zhou H J, et al. An empirical investigation and theoretic modeling for the collective online visiting behaviors. *Physica A: Statistical Mechanics and Its Applications*, 2018, **503**: 969–980.