

论坛中民生话题的识别与分析

周雪妍^{1,2}, 杨 静², 林泽鸿¹

(1 哈尔滨学院 工学院, 黑龙江 哈尔滨 150086; 2 哈尔滨工程大学
计算机科学与技术学院, 黑龙江 哈尔滨 150000)

摘 要: 对贴近生活的民生话题进行识别与分析. 民生话题的识别主要结合 LDA 模型和相似度模型并在时间轴上进行刻画. 通过相邻时间片话题相似度检测, 进而得到持续时间超过阈值期数的民生话题. 此方法在真实数据集验证了有效性.

关键词: 论坛; 民生话题; 社会网络分析

中图分类号: TP393

文献标识码: A

文章编号: 1000—7180(2014)10—0005—04

Identification and Analysis the Livelihood Topic in Forum

ZHOU Xue-yan^{1, 2} YANG Jing² LIN Ze-hong¹

(1 College of Engineering, Harbin University, Harbin 150086, China;

2 College of Computer Science and Technology, Harbin Engineering University, Harbin 150000, China)

Abstract: The identification and analysis the livelihood topic in forum is an important research direction. As a real-time public information platform, forum topics are divided into sudden topics and persistent ones which are close to the life of the people's livelihood, and our research mainly focuses on the analysis of livelihood topics. Identifying livelihood topic mainly combines the LDA model and similarity model on the timeline, through the correlation analysis between the adjacent time slice, gets the persistent livelihood topic which has time slices more than threshold periods. The methods in real data sets validate the effectiveness.

Key words: forum; livelihood topic; social network analysis

1 引言

论坛是 Internet 上的一种电子信息服务系统. 论坛上聚集了许多愿意分享经验、信息和思想的用户, 这些用户通过注册的唯一 ID 来浏览他人发布的信息并发布自己的消息而形成一个线程^[1]. 论坛交流已经成为当今社会交流思想和信息传播的一个重要途径. 论坛上的话题主要分为两类: (1) 突发事件相关话题, 特点是持续时间短, 讨论激烈; (2) 民生话题, 特点是持续时间较长, 一般为贴近生活的话题. 已有的研究主要集中于前者, 如论坛热点话题的发

现及预测研究^[2-3]、突发事件后不实消息传播的研究^[4]、基于舆情的敏感新信息搜索方法等^[5]. 后者的研究相对较少, 同时由于民生话题是舆情分析的难点, 其相关研究具有实际意义. 因此, 本文算法主要目的是提取时间轴上的民生话题.

识别民生话题算法结合时间维度, 应用 LDA 话题模型^[6]和短文本相似度评估模型^[7]进行提取. 民生话题的提取首先要保证聚类结果话题相关, 算法采用 LDA 生成话题模型, 短文本相似度模型筛选并聚集相关帖子; 另一方面本文采用相邻时间片话题交叉匹配的方法保证话题在时间轴

收稿日期: 2013—11—30; **修回日期:** 2014—01—28

基金项目: 国家自然科学基金(61370083, 61073043); 高等学校博士学科点专项科研基金(20122304110012); 哈尔滨市青年后备人才专项资金项目(2014RFQXJ081); 黑龙江省教育科学研究青年专项课题(GBD1213045); 教育厅人文社科项目(12542083); 哈尔滨师范大学人文社科预研项目(SYB2012-02); 哈尔滨学院学科发展青年基金(HUYF2013-011)

上的持续性. 在得到了民生话题对应的帖子、参与者及其回复关系后, 就可以应用社会网络分析进一步得到如关键人物分析、社区划分、话题演化等方面的研究.

2 相似度评估模型

Quan^[7]提出了基于话题模型的短文本相似度计算方法, 该方法是利用话题模型来修正短文本的特征向量, 最终仍然通过计算向量之间的夹角余弦来表示文本之间的语义相关性. 本文算法参考其向量化模型, 采用最小值阈值的方法, 本文方法在不降低结果准确性的同时, 能够相对减少计算开销.

相似度评估模型主要针对相邻两个时间段的话题进行分析, 假设前一期话题 $T_{\text{former}} = \{t_1, \dots, t_i, \dots, t_n\}$, 对应的话题向量为 $t_i = \{(t_{i1}, p_{i1}), \dots, (t_{ij}, p_{ij}), \dots, (t_{iN}, p_{iN})\}$; 后一期话题 $T_{\text{later}} = \{t_1, \dots, t_k, \dots, t_m\}$, 对应的话题向量为 $t_k = \{(t_{k1}, p_{k1}), \dots, (t_{kl}, p_{kl}), \dots, (t_{kM}, p_{kM})\}$. 要找到相邻时间片上相似度高的话题, 需要进行 $n \times m$ 次相似度计算, 即两个时间段的每一个话题都需要和另一个时间段的所有话题进行相似度评估.

$$r_{i,d} = \sum_{\text{word} \in t_i \cap d} \min p(\text{word})$$

$$s_{i,k} = \sum_{\text{word} \in t_i \cap t_k} \min(p(\text{word})) \quad (1)$$

式中, $r_{i,d}$ 表示标题向量 d 与话题 t_i 的相关度, 如果相关度大于阈值 σ_2 , 即认为该帖子是与话题相关. $s_{i,k}$ 表示话题向量 t_i 与 t_k 的相关度, 它等于两个话题中同时出现的某个词汇的概率较小值的总和. 当 $s_{i,k}$ 大于设定阈值 σ_1 时, 算法认为这两个话题是相似的. 当同一话题持续超过若干期, 可以认定为民生话题.

同时, 在一个时间段内也需要应用相似度来衡量一个话题 t_i 的规模, 这时只需要应用 t_i 的关键词与向量化的帖子标题 d 进行匹配, 所有匹配成功的概率总和即为相关度^[8-9].

3 论坛民生话题识别方法

由于网络数据的时间特色, 把用户交互的数据按相等的时间 T 来划分, 并假设这些时间段是独立的或者小部分重合的, 在本文实验中, 假设数据是按半个月(15天)进行划分的. 对于每一个时间段的数据进行话题建模, 并通过相邻时间段话题相似度比较使得跨越多个区域的民生话题浮现出来.

得到民生话题算法的主要思想如图1所示. 首

先, 用户交互数据按时间区间进行划分, 并使用LDA得到时间段内的话题模型. 然后, 相邻时间段的话题数据进行相似度计算, 把跨越多个时间段的相关度高的话题定义为民生话题^[10].

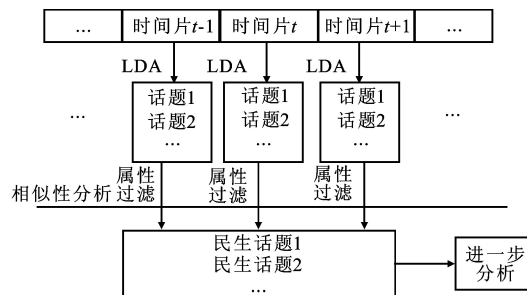


图1 民生话题提取算法

算法的具体步骤如下所示.

Step 1. 应用LDA话题模型分时间段进行话题识别, 并统计话题相关的帖子数、参与用户数、回复率、热度等; 对于时间段 t , 得到 $N = \text{TOPIC}_i(t)$ 个经过属性过滤的话题, 每个话题包含关键词列表及对应概率. 其中属性过滤满足以下条件之一的话题被保留:

- (1) 标题向量与该话题向量相似度大于阈值 σ_2 的帖子总数大于等于10;
- (2) 参与该话题的总用户数大于等于该期活跃用户的10%;
- (3) 热度(点击数除以该期活跃用户数)大于等于10%;
- (4) 回复率(总参与人数除以总点击数)大于等于30%;

其中, 阈值 σ_2 在实验中取值为0.05, 即如果一个帖子标题向量含有某话题的一个主关键词, 则可以确定这个帖子是该话题的相关帖子.

Step 2. 识别民生话题. 通过计算相邻时间段话题的相似度来得到延续数期的民生话题. 在本文实验中, 话题相关度阈值 $\sigma_1 = 0.09$. 对于这样的问题, TOPIC_i , 如果它在 $j, j+1, \dots, j+s$ 这几个时间段有定义, 则这个话题的帖子数 POST_i 和用户数 USER_i 分别为

$$\text{POST}_i = \sum_{t=j}^{j+s} \text{POST}_i(t)$$

$$\text{USER}_i = \bigcup_{t=j}^{j+s} \text{USER}_i(t) \quad (2)$$

突发事件的帖子时间跨度小, 网络演化结构相对简单, 不属于本文关注的民生话题. 突发事件一般在时间轴上持续时间不足3期, 因为有些突发事件可能正好发生在两个时间段交汇处, 因此, 定义民生话题最好持续 $t_{\text{req}} (s > t_{\text{req}})$, 实验中 $t_{\text{req}} = 3$.

4 实验

4.1 数据集

本文数据来源于中国知名论坛天涯论坛,获取数据为天涯杂谈自2011年1月至2013年10月的数据,包含325 288个用户、102 756个主帖和4 524 756条回复.提取民生话题的实验在以上数据集的一个子集(2012年10月到2013年10月)上进行.在这些注册用户中,有12 701人在2011年到2013年时间段内至少发表了一个帖子,3 724人发表至少2个帖子,并有573人发表多于5个帖子.单贴平均回复数(只考虑至少有5个回复的帖子)是62.91.2011年用户共发表帖子10 324个和评论400 571个(38.8评论/帖子),2012年发表帖子31 146个和评论1 326 819个(42.6评论/帖子),2013年发表帖子61 286个和评论2 797 366个(45.6评论/帖子).

4.2 话题的识别

对给定的数据以发帖时间预处理,应用LDA话题模型进行话题识别,其中超参数 α 和 β 分别设为50/ Z 和0.01,话题数 Z 设为50,Gibbs抽样迭代次数设为1 000.每个时间片所得到的50个话题显然有一部分不是本文关注的民生问题,因此通过属性过滤进行分析处理,将得到每个月数目不同的话题数如表1所示.

表1 各期话题数

期数	话题个数	期数	话题个数
1	25	13	18
2	20	14	26
3	31	15	15
4	24	16	30
5	27	17	15
6	16	18	17
7	13	19	27
8	15	20	31
9	10	21	38
10	20	22	22
11	19	23	33
12	20	24	24

得到的总话题数为536个,平均每期22个,其中第9期最少仅有10个,第21期最多为37个.首先分析每个话题的相对大小,图2中给出了话题相关帖子数量的统计信息.为了提取民生话题之前不流失更多的有效数据,这里相似度参数取值 σ_2 取0.05,即一个帖子标题如果含有某话题的一个主关

键词即被保留.可见88%的话题的规模都集中具有61到150个相关帖子.

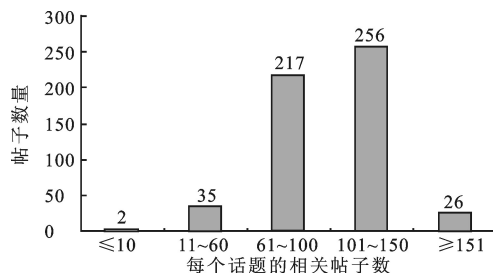


图2 话题相关的帖子数量统计图

4.3 民生话题的识别

接下来的分析主要是识别民生话题,这些话题需要持续存在超过一个给定的期数限制.民生话题的数量受相似度取值的影响比较大,图3中给出了两者的关系.一般情况下,话题的主关键词一般频率比重为0.05左右,所以当相似度确定在0.1时就意味着有两个主关键词一致,这时基本可以肯定两个话题是讨论的一个问题.实验也证明,图中是一个重要的拐点发生在 $\sigma_1=0.09$ 时,对应得到了16个民生话题,经手工验证和排查发现准确性较高,其话题质量高.

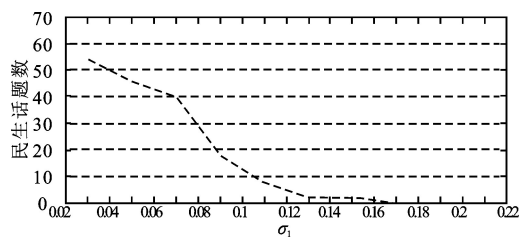


图3 民生话题的数量和相似度参数的关系

得到的16个民生话题及其相关的帖子4 216个.这大大缩小了进一步对民生问题进行分析的数据规模.平均每个民生话题有相关帖子263个,按最短持续3期进行计算,平均每87个,低于一般保留话题的规模,这也体现了民生话题并不具有突发性话题的高发帖率、高点击率和高回复率,其突出特点是持续周期长.

4.4 话题持续时间分析

图4中给出了民生话题持续的时间分布,其中12个(75%)民生话题持续3期(3个月),这是也是本实验认定持续话题期数下限.同时4个话题持续超过4期.民生话题的分布是比较均匀的,只有2013年5月(13期)有4个民生话题同时存在.分析数据发现,在这个时间段为毕业季和高考的时间,同时与青春类电影如《致青春》的上映时间重合,使得

此类话题持续高温并不断演化,但是话题的演化问题超出了本文研究范围.同时《中国最强音》(X Factor)和《中国梦之声》(Chinese Idol)等全民歌唱进取真人秀节目的热播也是这是时间段民生话题的贡献者.

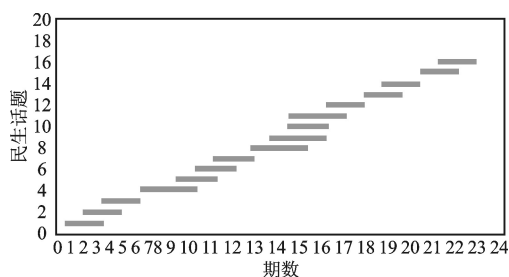


图4 民生话题持续时间的分布图

同时所得到的民生话题是具有较高的相异性的,因为相同时间段内重叠的话题不多,而时间上相隔较远的话题虽然有可能相似,但显然是两个不同的事件.如每年毕业、高考、就业这个民生话题都会重复上演但有每年不同,这类话题的演化分析不在本文研究范围之内.所以本算法所提取的持续话题是具有很好的多样性的.

5 结束语

本文根据论坛数据的特点^[11],设计并实现了民生话题的提取算法.为了区别民生话题和热点话题,本文按时间片对数据进行了划分,并通过话题模型 LDA 对各个时间片分别进行话题提取并属性过滤,进一步应用相似度模型对相邻时间片的话题进行交叉匹配,得到持续火热的民生话题.实验结果揭示了大多数的民生话题都持续 3 期左右,在数量上远远小于突发性话题,且时间分布上较均匀.通过手工验证得到的民生的话题一般为社会大众话题,进一步分析这些话题的参与者的社区组成、核心人物及思想派别,将对社会舆情分析很有益处.

参考文献:

[1] Lu D, Lixin D. Sentiment analysis in chinese BBS [C]// Intelligence Computation and Evolutionary Computation. Berlin Heidelberg, Springer 2013: 869-873.

- [2] Liu Hong, Bi Wei Li. Hot topic detection research of internet public opinion based on affinity propagation clustering[C]// Computer, Informatics, Cybernetics and Applications. Netherlands, Springer, 2012: 261-269.
- [3] Zhang Beibei. A time-varying propagation model of hot topic on BBS sites and Blog networks[J]. Information Sciences, 2012(87): 15-32.
- [4] Zhang F, G Y Si, Pi Luo. Study on rumor spreading model based on evolution game[J]. Journal of System Simulation, 2011, 23(9): 1772-1775.
- [5] 王文, 王树锋, 朱锡芳. 一种基于舆情敏感新信息的发现和搜索方法[J]. 微电子学与计算机, 2013, 30(12): 113-116.
- [6] Blei David M, Andrew Y Ng, Michael I Jordan. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003(3): 993-1022.
- [7] Quan X J, Liu G, Lu Z, et al. Short text similarity based on probabilistic topics[J]. Knowledge and Information Systems, 2010, 25(3): 473-491.
- [8] Ding Fei. Modeling reading and replying activities in a BBS social network[J]. Journal of Electronic Science and Technology, 2010, 8(4): 300-307.
- [9] Naruse K, Kubo M. Lognormal distribution of BBS articles and its social and generative mechanism[C]// Proc. of the 2006 IEEE/WIC/ACM International Conf. on Web Intelligence. HongKong, 2006: 103-112.
- [10] Geng Changxin. Bursty hot-words detection for campus BBS[J]. Telkomnika Indonesian Journal of Electrical Engineering, 2013(11): 6.
- [11] 胡局新, 鞠训光. 基于贝叶斯推理和 TFIDF 算法的中文关键词智能抽取[J]. 微电子学与计算机, 2012, 29(9): 197-200.

作者简介:

周雪妍 女, (1981—), 博士研究生, 讲师. 研究方向为数据挖掘、社会计算、语义社会网络.

杨 静 女, (1962—), 教授, 博士生导师. 研究方向为数据与知识工程、社会计算.

林泽鸿 男, (1981—), 硕士, 讲师. 研究方向为数据挖掘、机器人.