



一种融入公众情感投入分析的微博话题发现与细分方法

琚春华^{1,2}, 鲍福光^{1,3}, 戴俊彦²

(1. 浙江工商大学现代商贸研究中心, 浙江 杭州 310018;

2. 浙江工商大学计算机与信息工程学院, 浙江 杭州 310018;

3. 浙江工商大学工商管理学院, 浙江 杭州 310018)

摘要: 为了提升微博话题发现效率以及发现质量问题, 提出了一种融入公众情感投入分析的微博话题快速发现与细分方法, 促使话题演化, 进而产生新话题及其情感变化趋势。首先, 基于情感词典和 TFDF 值在历史语料库中挖掘常用情感词并构建情感词库; 其次, 快速抽取情感文本, 结合 Sigmoid 函数检测情感投入密集期, 保证话题事件挖掘的质量; 最后, 通过改进的模糊 C-均值聚类算法在新的微博数据中发现高质量话题。实验结果表明, 本文方法能够有效提升移动环境下的话题发现效率及质量。

关键词: 情感词; 微博; 话题发现; NE-FCM

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016158

Discovery and segmentation method in micro-blog topics based on public emotional engagement analysis

JU Chunhua^{1,2}, BAO Fuguang^{1,3}, DAI Junyan²

1. School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

2. Contemporary Business and Trade Research Center of Zhejiang Gongshang University, Hangzhou 310018, China

3. School of Business Administration, Zhejiang Gongshang University, Hangzhou 310018, China

Abstract: To improve the discovery efficiency and quality of micro-blog topic, a method of rapid discovery and segmentation in micro-blog topics based on public emotional engagement analysis was proposed, it would prompt evolution of the topics, then generate new topics and gain emotional change trend. Firstly, common emotional words were mined from corpus to build emotional thesaurus based on emotional word dictionary and TFDF. Then, emotional text was extracted quickly and sigmoid function was utilized to detect the intensive period of emotional engagement, ensuring the validity of topic mining. Besides, an improved adaptive FCM was used to cluster and discover topics. The experimental results show that this method can enhance the efficiency and quality of topic discovery in mobile environment.

Key words: emotional word, micro-blog, topic discovery, NE-FCM

收稿日期: 2016-03-09; **修回日期:** 2016-06-03

基金项目: 国家自然科学基金资助项目(No.71571162); 国家科技支撑计划基金资助项目(No.2014BAH24F06); 浙江省哲学社会科学规划课题(No.16NDJC188YB); 浙江省自然科学基金资助项目(No.LY14F020002); 教育部人文社会科学重点研究基地项目资助(No.14JJD630011, No. 13JDSM16YB)

Foundation Items: The National Natural Science Foundation of China(No.71571162), The National Key Technology R & D Program of China(No.2014BAH24F06), Zhejiang Province Philosophy Social Sciences Planning Project(No.16NDJC188YB), Natural Science Foundation of Zhejiang Province of China(No.LY14F020002), Key Research Institutes of Social Sciences and Humanities Ministry of Education(No.14JJD630011, No.13JDSM16YB)



1 引言

微博作为代表性的移动社交应用,允许人们即时分享最新消息和想法。2013年,新浪微博注册用户已超过5.36亿户,微博内容涵盖了社会生活的各个方面,公众不仅仅是在网上冲浪,同时也成为了波浪的制造者。而据参考文献[1]所述,合理采用新浪微博API所爬取的数据将有较大的价值。研究微博的使用状况发现,微博作为一种社交工具在移动互联网时代正慢慢承担着短信、博客、即时通信等功能,提供的服务种类也日趋多样化,总体来说微博成为了用户表达自身感受,分享各种信息的主要途径。微博文本在这一过程中作为一种用户情感的微观实例,以短文本的形式传递出用户对话题的情感信息,如对新闻和当前事态的评论等。

通过对信息进行话题的自动识别和已知话题的持续跟踪,帮助人们发现网络中讨论的热点,一直是自然语言处理领域的研究重点^[2]。以微博为代表的社交工具不同于传统媒体,在移动互联网时代具有数据量大、文本较短、产生速度快和非结构化等特点,加大了其话题发现的难度^[3]。情感投入分析是目前分析Web文本的一种重要方向和方法。同时,互联网公众在网络上发表的语言状态情感存在一种“涌现和传播演化”现象,针对上述现象,本文提出了一种融入公众情感投入分析的微博话题快速发现与细分方法。其中,基于情感词计算文本情感投入是一种有效的方法,其主要思想是应用情感词汇在文本中的出现情况来预测和衡量文本情感投入以及公众的各类态度和趋势,从而使话题分化,衍生出新的话题并引导网友的讨论,如参考文献[4,5]。网友对“#话题#”进行讨论,发表自己的评论,包含了自己的观点,促使老话题衍生新话题,促使话题演化,进而产生新话题及其情感变化趋势。

2 相关研究

随着移动互联网的迅猛发展,海量信息的挖掘方法逐渐受到研究人员的关注。社交应用内的话题信息发现是新形势下的重要研究方向,目的在于帮助人们应对信息过载问题,从而提升处理效率。传统的话题发现方法可以追溯到VSM(vector space model,向量空间模型)的信息数据映射。但使用VSM是建立在特征向量维数稳定的基础之上,一旦各文本特征维数相差较大就会影响最终的计算效果。

针对这一现象,孙宏纲等人^[6]利用知网词库,提出了一种VSM扩展的解决方法。Kaleel S B等人^[7]提出了一种基于LSH(location sensitive hash,位置敏感散列)函数的话题事件检测算法,采用两次LSH分别获取网络数据中的独立事件和交叉事件。但使用上述方法进行话题挖掘,在文本特征处理上将耗费大量的时间,不适用于移动互联网下海量信息的话题发现。

参考文献[8]提出了基于情感符号的在线突发事件检测方法,通过已有的微博情感符号抽取相应文本以满足实时处理要求。O'Connor等人通过采用Opinion Finder中的主观词汇对微博进行情感标记,并将结果同手工测得的指数进行联系,发现消费者信心指数和政治情感指数都与从微博中计算出来的情感相关联^[5]。杨小平等人^[9]利用微博表情符号对微博文本进行情感倾向标注,构建情感词典。冯时等人^[10]利用句法进行博文的情感分析,发现在普通主题搜索的基础上进行情感倾向分析,将有助于主题趋势的理解。应晶等人^[11]认为公众在表达观点时,往往会用情感词来突显,而这些情感词会随着话题的周期变化而变化。因此,通过情感词典构建,分析微博或Web文本情感倾向,对当前热门话题发现及其变化趋势有着重要的作用。由此得知,从公众情感角度分析文本类数据拥有一定的理论基础且能够有很好的扩展性。

本文在传统话题发现研究基础之上进行了改进,通过情感词结合微博特性挖掘情感密集期,约简了文本集。设计了一种融入公众情感投入分析的微博话题快速发现与细分方法,采用名词性实体改进话题聚类算法,增强话题发现效率及质量。

3 融入公众情感投入的微博话题快速发现模型

3.1 基本定义

定义1 情感词库: $S=\langle s_1, s_2, \dots, s_n \rangle$, s_i 表示情感词。

定义2 微博文本集: $D=\langle d_1, d_2, \dots, d_n \rangle$, 其中, $d_i=\{w_1, w_2, \dots, w_n\}$, w_i 表示文本 d_i 的特征项。时间 T 内的文本可表示为 $D_T=\{d_i\}$ 。

定义3 情感文本: $D_S=\{d_{1s}, d_{2s}, \dots, d_{ns}\}$ 表示为存在情感词的文本集合, d_{is} 即情感文本, S 表示情感词库。

定义4 话题集:在时间 T 内,基于情感投入检测到的话题集表示为: $C_T=\{c_{1T}, c_{2T}, \dots, c_{nT}\}$, 其中,话题 $c_{iT}=\{d_{1i}, d_{2i}, \dots,$

d_n 表示由一系列相应文本组成的话题。

3.2 模型框架

本文主要目的在于通过微博情感投入密集期的检测达到约简文本集、有效提升话题发现效率及质量的目的,并以此为基础增强移动互联网环境下的话题掌控力。为此,需解决以下几个问题:

- 如何构造适宜的微博情感词库;
- 如何检测情感投入密集期及快速抽取情感文本;
- 如何基于微博短文本特征提升发现话题价值。

基于上述问题,本文提出的模型框架如图1所示,主要由微博情感词库构建、情感投入密集期检测、融入情感投入的微博话题发现三大模块组成。其中,模块1基于知网的HowNet情感词典、中国台湾大学的Ntusc情感词典和大连理工大学的情感词汇本体库构建初始情感词库;微博影响力代表着文本在话题发现中的参考价值,模块2中结合微博影响力并采用Sigmoid函数检测情感投入密集期;模糊C-均值算法是众多模糊聚类算法中应用最成功的算法^[12],模块3通过改进的FCM算法,设计了一种基于名词性实体的模糊C-均值算法NE-FCM。

3.3 微博情感词库构建

情感词库的构建过程中面临两个问题:词库情感词在微博语言环境中需具有一定的适用性,即出现概率;基于情感词库抽取情感投入较多的文本,匹配时间往往过长。因此,本文基于大规模微博语料库和三大著名情感词集,运用TF和DF算法相结合的TFDF值以及双字散列索引表实现具体情感词库的构建,词库满足语料库和情感词集变化而动态更新的需求。语料库的采集,利用中国爬盟所

提供的WeiboCrawlerApp爬取了600万条新浪微博,每条微博作为一个文本单位。

在文本特征选择和权重计算领域,TFIDF算法因其计算简单、较高的准确率和召回率受到广泛应用^[13,14]。逆向文件频率(IDF)是指某文本集 D 的特征词 w_i ,根据其在文本 d_i 中出现的频率赋予相应权重,而给予只在少数文档中出现的特殊词较高的权重,显然无法适用于微博情感词库的构建需求。

因此本文在对三大情感词集进行冗余处理后,首先基于语料库过滤非常用低频特征,即计算情感词的最大值,再乘以文本频数 DF ,记为 s_i 的TFDF值。在实际微博环境中,由于大量推广类信息的影响,增加了部分情感词的 TF 值,故本文采用增加 DF 值的方式提高微博情感词库的代表性。经过上述步骤,本文从600万条微博语料库中共挖掘得到1231个适用情感词。词库构建后很容易以顺序表的方式存储在硬件设备中,但海量数据的查找匹配效率会成为制约其发展的重要因素。在微博情感词库中各长度情感词的统计见表1。

表1 情感词长度统计

情感词字数	1	2	3	4	5	6
比例	2.7%	49.68%	9.67%	31.75%	3.7%	0.8%

从表1可以发现,微博情感词库中长度为2的情感词约占据了一半,一般情感词长度越长所占比例越小。基于以上事实,采用双字散列索引^[15]的数据结构,对于最大匹配和全切分法,其处理速度比以往的逐字二分提高了57.5%和60.5%。情感词索引结构如图2所示。

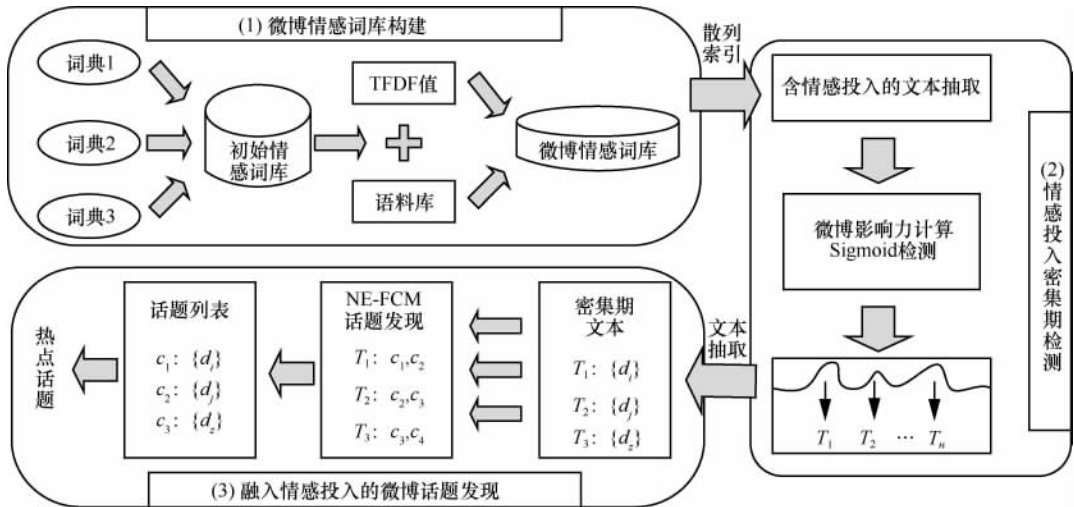


图1 融入公众情感投入分析的微博话题快速发现模型

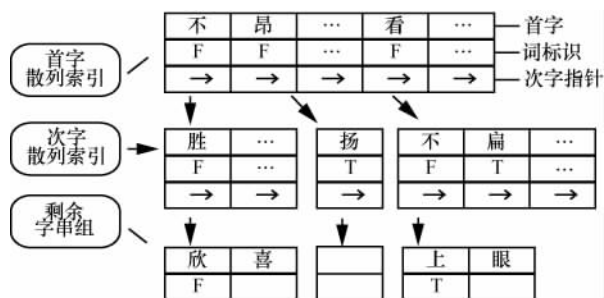


图2 情感词散列索引示意

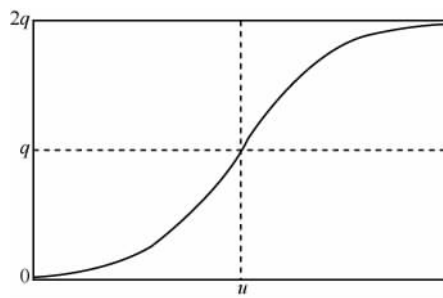


图3 Sigmoid 函数

3.4 情感投入密集期检测

情感投入密集期检测是基于已有情感词库对微博短文本进行情感词匹配,挖掘出微博用户情感投入的密集期,并将文本按密集期进行归类。对于微博文本集 D ,根据微博情感词库 S 及双字散列索引结构快速抽取情感文本 d_s ,算法如下。

算法1 情感文本抽取

输入 微博文本集 D ,微博情感词库 S 。

输出 情感文本集 D_s

(1) $\forall d \in D$, 设定文本标记 $\text{flag} = \text{false}$;

(2) For A in d :

If 首字散列索引 $a \neq \text{null}$: $\text{flag} = \text{true}$

若 a 的指针 q_1 为空, continue ;

否则得到以 A 字起始的次字散列索引 b ;

在 b 中通过散列定位到字 B , 由指针 q_2 得到以 AB 起始的剩余字串组 L ;

按正向最大匹配规则从 L 中依次匹配,取得情感词 s ;

If 遍历结束: $\text{flag} = \text{true}$, 将文本 d 加入 D_s ;

(3) 重复步骤(1)、(2),直至所有文本分类完成。

伴随着话题热度的提升,用户微博文本中采用情感词表达自身观点的比例会明显增加,出现情感投入的密集期。本文对经过算法1抽取得到的情感文本集 D_s 进行情感密集期挖掘。对于抽象的微博文本情感投入,以情感词作为其标准度量是现今公认的有效方法。而对于微博环境,信息流的传播过程中高影响力微博对公众情感表达有直接导向作用,例如明星微博往往会引起涌现情况的发生。

为此,本文引入测算微博影响力的转发 $R(d)$ 和评论 $M(d)$ 指标以及统计得到的文本情感词数 $\text{Num}(d)$, 规范化求和得到 $f(d)$ 。Sigmoid 函数对数据细微变化敏感,并可以抑制高频次商业微博对数值结果的影响。其函数图像如图3所示。

本文采用 Sigmoid 函数构造密集期度量函数。设时间 T 内文本数为 N_T , 搜寻不同时段,选用整体均值作为度量标准,称时间 T 为公众情感投入的密集期,设 $f(d)$ 的中值为 q 、均值为 u , 若:

$$\frac{1}{N_T} \sum \frac{2q}{1 + e^{-(f(d_i) - u)}} \geq \text{mean} \quad (1)$$

从而情感文本集 D_s 根据不同情感密集期 T 而被划分为几个不相交的子集。

3.5 微博话题发现

对情感密集期内的文本集 D_T , 采用改进的模糊 C-均值聚类方法 NE-FCM 发现微博话题。由于真实语言环境的复杂性,特征词隶属于各聚类对象之间的界限往往不是很清晰。一种处理广泛存在不确定性的模糊集合论,对中文语境下的模糊概念划分具有较好的处理效果,其中,FCM 算法通过不断迭代优化目标函数 $J(U, C)$, 得到样本点 d_i 对所有类中心的隶属度矩阵 $U[u_{ij}]$, 从而决定样本点的类属 c , 以达到对数据样本自动分类的目的。

根据模糊聚类算法基本思想,本文中微博文本集 $D_T = \langle d_1, d_2, \dots, d_n \rangle$ 各文本以特征词项的方式表示,即 $d_i = \{w_1, w_2, \dots, w_n\}$ 。在每个情感密集期内,采用 TFIDF 算法计算每个特征项 w_i , 作为对其重要程度的评价依据。隶属度矩阵 $U[u_{ij}]$ 将 D_T 聚类成模糊划分 $c_1, c_2, \dots, c_n, u_{ij}$ 表示样本点 d_j 对类 c_i 的隶属度并满足 $\sum_{i=1}^c u_{ij} = 1, u_{ij} \in [0, 1]$, 则模糊 C-均值聚类算法的目标函数为:

$$J(U, C)_{\min} = \left\{ \sum_{j=1}^n \sum_{i=1}^c u_{ij}^m \text{dist}_{ij}^2 \right\}_{\min} = \sum_{j=1}^n \left[\left\{ \sum_{i=1}^c u_{ij}^m \text{dist}_{ij}^2 \right\}_{\min} \right] \quad (2)$$

其中, m 为模糊度, dist_{ij} 表示第 j 个样本到类 c_i 的欧式距离。

由上述可见,算法采用类内平均加权误差的方法不断优化目标函数 $J(U, C)$, 一旦改变量小于阈值 ε 或达到最大迭代次数则停止。对任一初始聚类中心 C^0 , 由式(2)可知求

属度矩阵 U 中各列独立,依据拉格朗日乘子法计算一阶式后可知:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\text{dist}_{kj}}{\text{dist}_{ij}} \right)^{\frac{2}{1-m}}} \quad (3)$$

$$c_i = \frac{\sum_{j=1}^n u_{ij}^m d_j}{\sum_{j=1}^n u_{ij}^m} \quad (4)$$

得到当目标函数 $J(U, C)$ 有解时,隶属度 u_{ij} 及聚类中心 c_i 满足的必要条件,即式(3)、式(4)。因此可知 FCM 算法将聚类结果 C 看作由初始聚类中心 C^0 出发的一种映射,不断迭代。微博文本话题集中,各类样本间数目往往相差较大,极易发生收敛到局部极小点的情况。

在对大量以微博为代表的短文本进行分析的基础上,发现以人名、地名、时间等为代表的名词性实体在文本中拥有较强的代表性,选用名词性实体较多的点作为聚类中心会具有更好的话题发现效果,从而提出了一种基于名词性实体的模糊 C-均值聚类算法(NE-FCM)。基于 NLPPIR 词性标识系统,在名词性特征词中计算 DF 值进行筛选,避免多余计算消耗。从而将文本的词空间划分为名词性实体集及一般特征项集,如:

$$d = \{e_1, e_2, \dots, e_n\} \cup \{w_1, w_2, \dots, w_n\} \quad (5)$$

采用欧式距离计算名词性实体空间和特征项空间距离,定义新的文本点 d_i 和 d_j 之间的距离为:

$$\text{dist}(d_i, d_j) = \sqrt{\beta \cdot \sum (e_i - e_j)^2 + (1 - \beta) \cdot \sum (w_i - w_j)^2} \quad (6)$$

其中, β 为柔性参数,且 $\beta < 0.5$ 。包含的相同名词性实体越多,两者之间距离越小。

为避免算法依赖初始聚类中心的缺陷,本文转变随机初始值为满足名词性实体代表性的有目的的初始聚类中心选择,具体步骤如下:

步骤 1 计算任意文本点 d_i 和 d_j 之间的距离,生成点距离矩阵 $Dist$,选取拥有最短距离的两文本点的中间值作为 c_1^0 ;

步骤 2 选定距离阈值 α ,依据 $Dist$ 矩阵从与 C_1^0 两点距离都大于阈值 α 的文本点中选择 c_2^0 ;

步骤 3 如上所述,依据 $Dist$ 矩阵在余下文本点集中寻找与已确定类属的点距离都大于阈值 α 的点,并以此确定初始聚类中心 c_i^0 。

可以看到,本文方法通过不断搜索距离矩阵 $Dist$,避

免了大量因为计算距离产生的时间消耗。虽然牺牲了部分精确度,但在后续迭代过程中完全允许类似初始值的选取方式。基本满足了在不同名词性实体表征空间内的聚类需求。文本特征矩阵往往具有较高的维数,本文采用 PCA 主成分依次降维。

采用 NE-FCM 算法的基本步骤如下。

算法 2 基于名词性实体的模糊 C-均值聚类算法(NE-FCM)

输入 情感密集期 T 内相关文档 D_T , 最大迭代次数 $iter$, 聚类数 C_n , 阈值 ε 。

输出 聚类 $C_T = \{c_i\}$ 。

(1) $\forall d \in D$, 形成文档 d 的特征词项划分;

(2) 依据式(6),聚类数 C_n 和上述步骤选出初始聚类中心 C^0 ;

(3) 计算目标函数 $J(U, C)$, 利用矩阵范数比较相邻两次隶属度 U , 若小于 ε 或达到最大迭代次数 $iter$, 则算法停止;

(4) 重新计算隶属度矩阵 U 及聚类中心 c_i , 重复步骤(3)。

对情感密集期 T 内的文本集 D_T 经过 NE-FCM 算法后,各文本被分到不同的类 c , 由于同一话题时间延续的不确定性,本文采用话题相似性度量的方式进行合并。

4 实验及结果分析

4.1 实验数据分析

验证本文提出的融入公众情感投入分析的微博话题快速发现方法的有效性,语料库构建的数据为 2013 年 7 月采集得到的 600 万条新浪微博。为保证研究内容意义,使用在 2013 年 11 月 1-21 日内包含“二胎”关键字的 95 404 条有效微博作为数据集,在以上数据集中进行实验,所有数据都通过新浪微博 API 获得。本文实验操作环境为 Windows7 64 位, Intel Pentium4, 4 GB 内存的 PC。采用 Python 作为数据处理工具。本文通过 Python 接口,运用张华平博士发布的 NLPPIR 汉语分词系统,并去除代词、语气助词等高频出现但无实际意义的停用词,提升算法性能。

表 2 是以天为单位的数据集中情感文本的分布情况,可以发现每日微博文本中超 50% 的文本会运用情感词表达,而总体来说数据集中平均 78.45% 的微博有用情感词表达自身情感或观点的习惯。这也从侧面证明了本文使用情感词作为公众情感投入的衡量以及数据集重要约简指标的有效性。



表 2 情感文本比例

日期	情感文本比例	总数
11月1日	77.81%	1 798
11月2日	85.03%	1 456
11月3日	80.71%	1 068
11月4日	66.86%	2 718
11月5日	80.86%	1 588
11月6日	74.20%	1 241
11月7日	79.82%	1 070
11月8日	84.74%	1 134
11月9日	81.37%	1 737
11月10日	91.57%	997
11月11日	83.97%	2 070
11月12日	58.23%	5 562
11月13日	67.86%	5 858
11月14日	67.00%	4 453
11月15日	82.38%	9 130
11月16日	74.71%	12 448
11月17日	79.67%	8 732
11月18日	80.85%	8 511
11月19日	84.76%	9 087
11月20日	79.46%	7 852
11月21日	85.64%	5 646
平均	78.45%	

进一步验证了公众情感投入中情感词的使用规律,设定时间窗口 T 为 0.5 天,并以图 4 所示内容说明。图 4 中曲线分别展示了各时间窗口内情感词与文本数量,其

Pearson 相关指数为 0.91,一方面说明本文情感词库构建方法的有效性;另一方面说明了结合情感词作为微博短文本的密集期发现指标符合数据潜在规律,对于数据集进行话题发现有较强的指导价值。

图 4 中 14 条柱体分别标注了运用本文方法挖掘得到的情感密集期以及其时间窗口序号。微博话题的高速传播期一般在 2~3 天,可以看到密集期基本涵盖了所有数据时段,过滤数据的同时仍拥有较高的代表性。

4.2 实验结果及分析

本文采用上述融入公众情感投入分析的微博话题发现方法框架进行对比实验,设定算法最大迭代次数为 1 000,阈值 $\varepsilon=10^{-6}$,模糊度 $m=2$ 。在各时间窗口内分别设定聚类数为 25,构成最终的话题列表。在 Singlepass、FCM 算法中不进行名词性实体标识,聚类各时间窗口内数据集;本文 NE-FCM 算法聚类情感密集期内的数据集。由于实验数据量大,人工分类所有微博话题将耗费大量的时间,本文采用如下方式对实验结果进行评估。

(1) 发现时间

发现时间是指数据集中话题发现的时间消耗。时间越短越能体现相应方法的优越性。从表 3 中可以看出,本文算法平均在 1.8 min。其中,最快检测时间小于 1 min,平均检测时间接近 Singlepass 检测时间。

(2) 准确率

准确率为算法检测出的话题集中相关文档数与话题

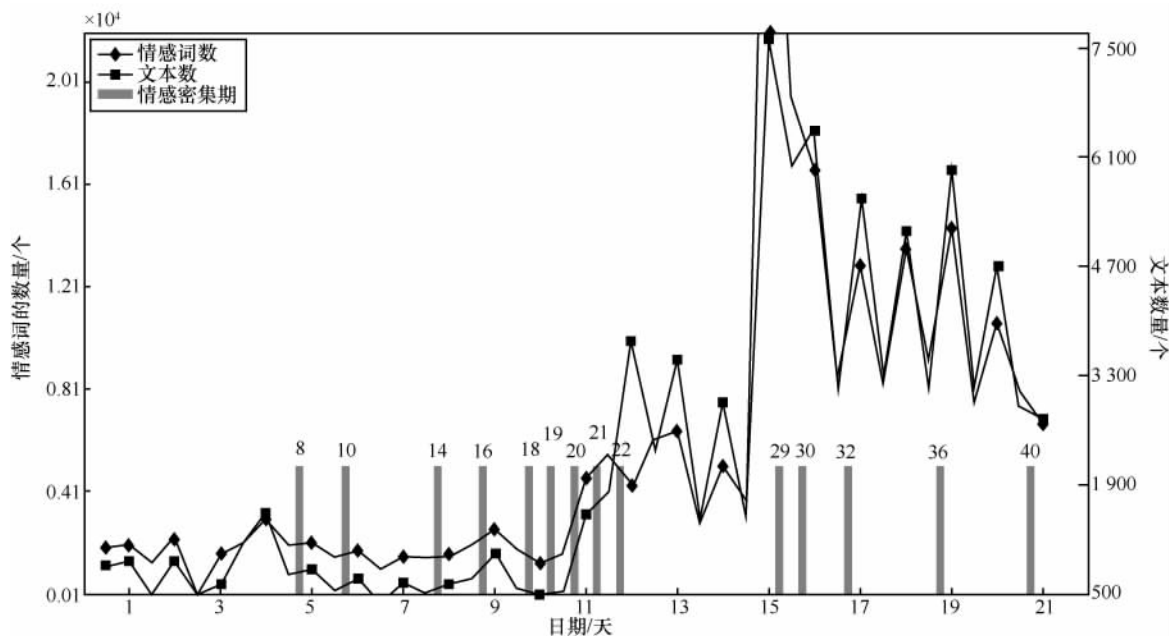


图 4 数据分布情况

表 3 检测时间比较

比较项	Singlepass 算法	FCM 算法	NE-FCM 算法
最短时间/s	61	82	52
最长时间/s	213	361	287
平均时间/s	96	193	103

文档总数的比例,是衡量话题发现精度的重要指标。对于检测出的话题列表集合,抽取其中 10 个话题计算其准确率,人工判定微博文本是否属于此话题。其结果见表 4。

表 4 算法准确率对比

比较项	Singlepass 算法	FCM 算法	NE-FCM 算法
最高准确率	90.3%	86.3%	91.7%
平均准确率	62.7%	73.8%	84.8%

由表 4 可以看出,本文算法具有相对较好的准确率,平均准确率超过 84%,能够适应微博环境下的话题发现要求。与传统的 Singlepass、FCM 话题发现方法相比,本文通过情感密集期的挖掘与名词性实体的标注可以避免大量的非目标文本的干扰,有效提升话题发现准确率。

(3) 命中率

命中率为算法检测出的话题占参考话题的比例,是衡量算法发现话题能力的重要指标。为确保参考事件的完备性,参阅了新浪微博风云榜以及各大主流网站当时有关的新闻报道,人工标注“二胎”相关话题作为参考话题,包括“国家放开单独二胎政策”、“马伊琍文章怀二胎”以及“山东长岛放开二胎人口负增长”在内的参考话题共计 31 个。分别选取各时间窗口中文本集准确率 Top12、Top16、Top20、Top24 的类,合并得到最终话题列表。话题命中率比较结果如图 5 所示。

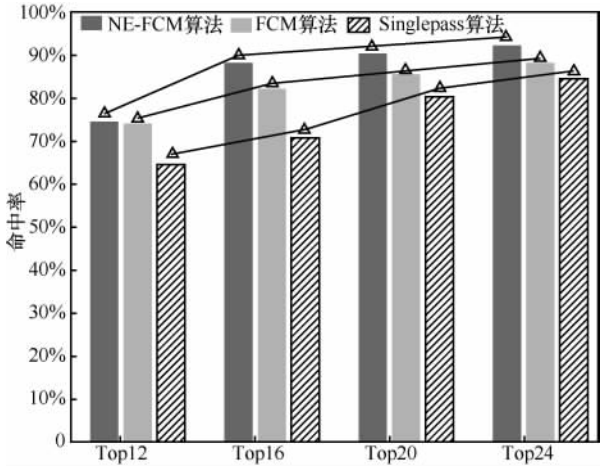


图 5 话题命中率对比

从图 5 看出,3 种算法都可以检测出大部分微博话题,NE-FCM 算法通过对情感密集期内的数据集操作,明显拥有更高的话题发现效率。虽然 FCM 方法在 Top12 拥有较好的命中率,但随着合并类数的增多,无法避免冗余数据引起的话题模糊问题,导致命中率增长缓慢,而 Singlepass 算法则受制于低准确率的影响。本文方法通过情感密集期的选择,约简数据集的同时保证了话题发现的完备性,NE-FCM 算法在不同范围内选择初始聚类点,保证了算法话题发现的稳定性,结合较高的话题准确率,在较小合并类值的条件下,已经达到较高的话题命中率。

(4) 话题发现质量分析

本文通过设定情感密集期约简数据集,大大降低了处理数据的规模,其目的在于提升微博话题发现效率以及发现质量。发现话题的质量主要可由准确率和命中率组成,从图 6 可以看出,随着情感密集期窗口数的增加,准确率逐渐降低,命中率迅速升高。

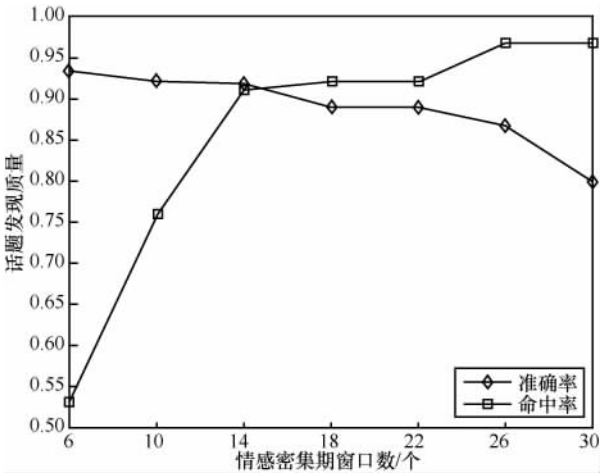


图 6 话题发现质量分析

因此合理选择情感密集期窗口数不仅可以大幅降低数据处理规模,提升话题发现效率,也有助于话题发现质量的提高,降低研究人员工作的复杂度。进一步,在本文方法得到的相关话题文本集中取 Top6 的情感词。从表 5 中可以看出,公众对话题 1 多采用改革、重大等情感词,表明对政策类话题的高关注度以及重视程度;对话题 2 采用支持、如愿以偿等情感词,表明对这一话题人物的支持及祝福;而对话题 3,公众更多表达了对结果的惊讶以及对事件原因的评论。表 5 表明,通过高质量话题可以较清晰地分析得到公众对相关话题的情感态度。



表 5 话题情感词

话题	情感词 Top6
“国家放开单独二胎政策”	改革、完善、决定、压力、关注、重大
“马伊琍文章怀二胎”	关注、恋爱、阳光、娱乐、支持、如愿以偿
“山东长岛放开二胎人口负增长”	免费、烦恼、创新、大胆、意想不到、贵

5 结束语

移动互联网社交应用的快速发展,加大了对社会热点话题挖掘的需求。本文提出了一种融入公众情感投入分析的微博话题快速发现与细分方法。采用构建情感词库的方式适应微博语言环境,挖掘情感密集期,从而提升微博话题发现效率以及发现质量。实验证明,本文提出的方法在处理以微博为代表的海量短文本方面有较强的准确率、命中率和实用性,进而可以发现广大用户的话题情感态度和兴趣特征,构建用户话题模型,广泛应用在信息服务业和商业等领域的推荐,为信息服务推荐奠定良好基础,也是下一步研究的重点。

参考文献:

- [1] 陈舜华, 王晓彤, 郝志峰. 基于微博 API 的分布式抓取技术[J]. 电信科学, 2013, 29(8): 146-149.
CHEN S H, WANG X T, HAO Z F. A distributed data-crawling technology for microblog API[J]. Telecommunications Science, 2013, 29(8): 146-149.
- [2] 张晓艳, 王挺. 话题发现与追踪技术研究 [J]. 计算机科学与探索, 2009, 3(4): 347-357.
ZHANG X Y, WANG T. Research of technologies on topic detection and tracking [J]. Journal of Frontiers of Computer Science & Technology, 2009, 3(4): 347-357.
- [3] MCANDREW A J, MOSHFEGHI Y, JOSE J M. Building a large-scale corpus for evaluating event detection on Twitter[C]// The 22nd ACM International Conference on Information & Knowledge Management, October 27-November 1, 2013, San Francisco, USA. New York: ACM Press, 2013: 409-418.
- [4] 李生琦, 田巧燕, 汤承. 基于《<知网>》词汇语义相关度计算的消歧方法[J]. 情报学报, 2009, 28(5): 706-711.
LI S Q, TIAN Q Y, TANG C. Disambiguating method for computing relevancy based on HowNet semantic knowledge[J]. Journal of the China Society for Scientific Andtechnical Information, 2009, 28(5): 706-711.
- [5] O' CONNOR B, BALASUB R, ROUTLEDGE B R, et al. From tweets to polls: linking text sentiment to public opinion time series[C]//The Fourth International AAAI Conference on Weblogs and Social Media, May 23-26, 2010, Washington, DC, USA. Palo Alto: AAAI Press, 2010: 122-129.
- [6] 孙宏纲, 陆余良, 刘金红, 等. 基于 HowNet 的 VSM 模型扩展在文本分类中的应用研究[J]. 中文信息学报, 2007, 21(6): 101-108.
SUN H G, LU Y L, LIU J H, et al. A study of the application of VSM expansion in text categorization based on HowNet [J]. Journal of Chinese Information Processing, 2007, 21(6): 101-108.
- [7] KALEEL S B, ABHARI A. Cluster-discovery of Twitter messages for event detection and trending [J]. Journal of Computational Science, 2015(6): 47-57.
- [8] 张鲁民, 贾焰, 周斌, 等. 一种基于情感符号的在线突发事件检测方法[J]. 计算机学报, 2013, 36(8): 1659-1667.
ZHANG L M, JIA Y, ZHOU B, et al. Online bursty events detection based on emoticons[J]. Chinese Journal of Computers, 2013, 36(8): 1659-1667.
- [9] 桂斌, 杨小平, 张中夏, 等. 基于微博表情符号的情感词典构建研究[J]. 北京理工大学学报, 2014(5): 537-541.
GUI B, YANG X P, ZHANG Z X, et al. Research on building lexicon for sentiment analysis based on the Chinese microblogging [J]. Journal of Beijing Institute of Technology, 2014(5): 537-541.
- [10] 冯时, 付永陈, 阳锋, 等. 基于依存句法的博文情感倾向分析研究[J]. 计算机研究与发展, 2012(11): 2395-2406.
FENG S, FU Y C, YANG F, et al. Blog sentiment orientation analysis based on dependency parsing [J]. Journal of Computer Research and Development, 2012(11): 2395-2406.
- [11] 陈旻, 朱凡微, 吴明晖, 等. 观点挖掘综述[J]. 浙江大学学报(工学版), 2014(8): 1461-1472.
CHEN M, ZHU F W, WU M H, et al. Survey of opinion mining[J]. Journal of Zhejiang University(Engineering Science), 2014 (8): 1461-1472.
- [12] 齐森, 张化祥. 改进的模糊 C-均值聚类算法研究 [J]. 计算机工程与应用, 2009, 45(20): 133-135.
QI M, ZHANG H X. Research on modified fuzzy C-means clustering algorithm[J]. Computer Engineering and Applications, 2009, 45(20): 133-135.
- [13] 范云满, 马建霞. 基于 LDA 与新兴主题特征分析的新兴主题探测研究[J]. 情报学报, 2014, 33(7): 698-711.
FAN Y M, MA J X. Detection of emerging topics based on LDA and feature analysis of emerging topics [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33 (7): 698-711.
- [14] 贺亮, 李芳. 基于话题模型的科技文献话题发现和趋势分析[J]. 中文信息学报, 2012, 26(2): 109-115.
HE L, LI F. Topic discovery and trend analysis in scientific literature based on topic model [J]. Journal of Chinese Information Processing, 2012, 26(2): 109-115.

- [15] 李庆虎, 陈玉健, 孙家广. 一种中文分词词典新机制——双字哈希机制[J]. 中文信息学报, 2003, 17(4): 13-18.

LI Q H, CHEN Y J, SUN J G. A new dictionary mechanism for Chinese word segmentation[J]. Journal of Chinese Information Processing, 2003, 17(4): 13-18.

[作者简介]



琚春华(1962-),男,博士,浙江工商大学教授、博士生导师、校长助理,计算机与信息工程学院院长,主要研究方向为智能信息处理、数据挖掘、电子商务与物流优化等。



鲍福光(1986-),男,浙江工商大学博士生,主要研究方向为智能信息处理、数据挖掘和供应链协同合作。



戴俊彦(1990-),男,浙江工商大学硕士生,主要研究方向为数据挖掘、智能信息处理等。