基于词共现和情感元素的突发话题检测算法①

兰 天, 郭躬德

(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: 随着自媒体的迅速发展, 微博中的舆情监控和舆情疏导成为一项重大的研究课题. 为了解决传统话题检测方法对于微博中大数据的分析往往具有复杂度高、实时性低、影响力小等问题, 提出一种基于词共现和情感分析的突发话题检测方法. 通过研究微博中情感的突发和共现关系, 从而建立情感子空间模型; 通过该模型对微博中的信息流进行分类, 最后对每个类别中的微博进行主题词提取, 实现话题检测的目的. 在 NLPIR 微博内容语料库上的实验结果表明, 该方法能够有效地从大规模微博信息中检测突发新闻, 提高突发新闻的识别率.

关键词: 话题检测; 情感; 共现关系; 微博

Bursty Topic Detection Based on Word Co-Occurrence and Emotions

LAN Tian, GUO Gong-De

(School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)
(Network Security and Cryptography key laboratory of Fujian province, Fujian Normal University, Fuzhou 350007, China)

Abstract: With the rapid development of the We-Media, monitoring and guidance of public opinion becomes a significant research subject. Traditional topic detection methods in microblog data analytics encounters the problems of high computational complexity, low real-time and recall rate. An improved algorithm based on emotions and word co-occurrence detection is proposed in this paper aiming at solving these problems. It builds a emotional subspace model through co-occurrence relation of sentiment words in hot events, and classifies the flow of information in weibos. Finally, it gets the aim of topic detection via extracting the subject in the corresponding category. The experimental results carries out on the microblog content corpus of NLPIR and verifies that this method can effectively detect news topic from the massive microblog information and realize the news topic tracking.

Key words: topic detection; emotion; cooccurrence relation; microblog

1 引言

随着 web2.0 时代带来信息技术发展的巨大变革, 自媒体平台已逐渐渗透到各个网民的生活中. 微博由其自身具备的便捷性、开放性、共享性等特点, 使之成为用户提供和分享事实和新闻的有效途径. 同时政府部门、社会名流、企业机构等的加入, 使微博成为获取重要信息的渠道.

话题检测与跟踪(Topic Detection and Tracking, TDT)作为文本挖掘的一个方向, 在微博上的运用越来越遭到人们的重视. 在国外, Takeshi[1]等提出基于 Twitter 的实时监控地震系统, 在实际应用过程中检测到了 80%以上的地震发生, 时效性超过了当地的地震告警机构. Doan[2]通过对 Twitter 上关于地震的信息进行研究, 得出微博信息对于地震预警可以起到重要的作用. 在国内, 舆情监控已经成为微博信息主要研究内容, 相关执法部门设立了相应的法规和措施. 而中文微博的文体具有一定的特殊性, 例如 140 字的限制、用语的不规范性、特殊标签的使用等, 这种特殊性使传统的文本检测方法不能很好地适用于微博文本检测. 文献[3]指出, 在微博 140 字的限制中, 中文微博的内容往往能比英文微博表达出更多的内容.

舆情导向是舆情监控的一个重要目的, 负面情绪

① 收稿时间:2015-12-09;收到修改稿时间:2016-01-15 [doi:10.15888/j.cnki.csa.005288]

的消息往往会导致严重的社会影响. 及时引导良性的微博舆论氛围, 就需要及时得出当前突发话题下网民的情绪状况、对某个事件的观点和态度等. 因此就需要对突发话题的情感进行研究, 来达到及时检测到突发情感的话题的目的. 微博中的情感信息来源主要通过表情图片、情续词两个个方面. 表情图片由微博向用户提供, 图 1 中所示的内容为微博上最常用的一组表情图片. 情绪词主要来自于人物对于喜怒哀乐的表达.

图 1　微博中最基本的情感图片组

本文针对传统本文检测算法的不足, 提出基于词共现和情感分析的突发话题检测算法, 通过共现关系研究各情感元素之间联系, 建立各情感下的空间模型, 对情感突发状况进行检测, 达到话题检测的目的.

本文的其余部分安排如下: 在第二节中, 介绍了近年来对微博话题研究的相关工作; 在第三节中, 详细介绍了本研究中提出的基于词共现和情感分析的突发话题检测算法; 在第四节中, 对于实验数据进行介绍和分析; 最后在第五节中, 总结了本次研究工作以及未来需要进行的研究.

2　相关工作

在国内, 对微博文本情感方面的研究在近几年得到了较大的重视. 文献[4]早在2002年首次利用机器学习技术进行情感分类, 并证明该方法的有效性. 该研究的提出引起众多学者对使机器学习技术进行情感分类任务较大的重视. 文献[5]在2010年提出的研究中表明, 情感词和论据词语的搭配作为情感分类的特征优于仅使用情感词作为特征. Wen Bin[6]重新定义概念的情感相似度, 有效地判定文本情感倾向性. Wang Suge[7]提出了一种基于赋权粗糙隶属度的文本情感分类方法,提高了精度和效率. 文献[8]中在 2012 年提出, 采用情绪词和表情图片对语料进行标注, 构建文本分类器的方法, 提高分类效果. 文献[9]对情感权重进行赋值, 并提出负面情绪的情感在检测中具有较大权重, 通过自查询的方法提高了查全率. 文献[10]在 2013 年提出通过情感符号和事件一样具有突发性, 并对情感模型进行构建, 通过检测情感的突发情况来达到检测突发事件的目的. 文献[11]中提出通过结合情感词属性和句法结构来确定情感值的方法, 并使情感量化表示. 文献[12]提出一种使用集中学习算法进行情感分类研究. Lu WeiSheng[13]为了克服高维数据带来的分类问题, 提出了一种词性序列作为文本特征的数据降维方法, 比传统的n-gram特征提取方法提高了分类精度, 降低了数据维度. 在文献[14]中为了克服传统情感分类方法指向不明的问题, 提出一种评价对象模型构建的方法, 同时通过特征聚合的方法进行数据降维, 有效地改善了情感分类的效果.

同时, 话题检测的方法也是微博信息研究的重要内容. 话题检测的目的主要是获取话题的主题词、话题聚类等. 文献[15,16]通过研究词汇之间形成的共现关系所表达的主题, 对话题信息进行抽取. 文献[17]通过上下文关系, 构建复合权值对主题词进行抽取, 来达到话题检测的目的.

虽然许多学者已经对微博突发话题的检测进行了研究, 但始终存在着一些较难克服的问题: 首先微博数据量庞大, 运算复杂度高, 耗时长, 难以达到实时性的效果; 其次, 话题检测的结果存在较多冗余信息, 或者话题影响度较低; 再次, 情感作为话题检测的特征, 缺少对情感元素之间的分析, 导致情感信息冗余. 因此本文就以上问题提出了相应的解决方法, 通过研究情感元素之间的共现关系, 构建情感空间, 实现数据约减; 并通过检测情感空间内突发情况来提取突发话题, 确保话题的有效性; 最后对情感空间内的微博进行聚类, 提取主题词.

3　情感模型及检测算法

微博热点话题的突发总是带来网民相应情感的变化, 从而对微博舆论产生导向作用. 然而同一情感下的表达元素可以是多样的, 使用词间共现度的方法来对时间窗口内的情感进行检测, 找到情感之间共现的规律来建立模型. 情感模型构建的主要工作包括文件流突发情感元素的检测和获取算法、突发情感的共现簇识别算法、共现簇的聚类算法. 情感模型构建基于当前时间窗口内文件流的情感元素, 在获取情感模型之后, 便可以对筛选出具有情感倾向的微博信息, 并根据情感倾向进行分类. 最后根据每一情感倾向下的

微博信息采用传统的方法进行聚类, 并获取突发话题信息.

3.1 形式化定义

定义 1. 微博文本

$$d=\{w_1,w_2,w_3,...\}$$

其中 w 表示组成微博的单词, 微博文本正是由一系列有序且有限的单词所组成.

定义 2. 时间窗口内的数据流

$$W_T=\{d_{t1},d_{t2},d_{t3},...\}$$

其中 T 代表时间窗口标号, 通常以 1 个小时作为时间窗口长度, 则 $T\in[1,24]$. t_i 表示当前文本所到达时间.

定义 3. 情感元素集

$$E^t=\{e_1,e_2,e_3,...\}$$

其中, E^t 表示时间窗口 t 内的情感元素集合, e_i 表示微博中提供的某一特定的表情图片或为情绪词语料库中的某一特定情绪词.

定义 4. 情感共现簇

$$C^t=\{e_{i1},e_{i2},e_{i3},...\}$$

其中, C^t 表示时间窗口 t 内的情感元素共现簇, 它由超过一定共现度阈值的情感元素所构成.

定义 5. 情感共现簇集

$$CSet^t=\{C_1^t,C_2^t,C_3^t,...\}$$

其中, $CSet^t$ 表示时间窗口 t 内所有情感共现簇的集合.

定义 6. 情感共现矩阵

$$X_t(e_1,e_2,...)=\begin{bmatrix} num(x_1) & x_{1,2} & x_{1,3} & \cdots & x_{1,n} \\ x_{2,1} & num(x_2) & x_{2,3} & \cdots & x_{2,n} \\ x_{3,1} & x_{3,2} & num(x_3) & \cdots & x_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & num(x_n) \end{bmatrix}$$

其中, 矩阵 X_i 表示时间窗口 t 下情感元素的共现矩阵 X, 对角线上的元素表示某一情感元素在当前时间窗口下的频数, $X_{i,j}(i\neq j)$ 表示当前时间窗口内情感元素 E_i 和 E_j 共现的频数.

3.2 文件流突发情感的检测和获取

对于时间窗口内的文件流信息进行预处理, 提取情感元素并统计, 设定参数 α 作为频数阈值, 过滤低于阈值的情感元素, 获取当前时间窗口内的主导情感元素.

在预处理过程中, 采用 ICTCLAS[19]分词系统, 它是一种中文文本分词工具, 有词性标注、命名实体识别、用户词典等功能, 并且在微博分词中具有比较好的效果. 首先需要针对微博表情图片的特定格式进行提取并在文本中移除, 否则影响分词效果; 然后在分词结果的基础上提取情绪词; 最后将相应的情感元素记录在当前时间窗口的共现矩阵中, 并增加相应的频数和情感元素之间的共现频数.

突发情感的判定需要两个参数: 情感词频率 F 和情感词增长率 G, 并且设定两个参数对于的权值 β_1 和 β_2, 其中 $\beta_1+\beta_2=1$. 通过这两个参数来衡量一个情感元素在当前时间窗口内的突发程度.

定义 7. 情感词频率

$$F_i^t=\frac{num(e_i)}{\sum num(e_i)}$$

其中, F_i^t 表示时间窗口 t 时情感元素 e_i 的频率, $num(e_i)$ 表示当前时间窗口内包含情感元素 e_i 微博的频数, $\sum num(e_i)$ 表示表示当前时间窗口内包含任一情感元素微博的频数.

定义 8. 情感词增长率

$$G_i^t=\frac{F_i^t-F_i^{t-1}}{F_i^{t-1}}$$

其中, G_i^t 表示时间窗口t时情感元素i的增长率, F_i^t 表示时间窗口 t 时情感元素 i 的频率, F_i^{t-1} 表示相邻时间窗口 t-1 时情感元素 i 的频率.

定义 9. 构造一个符合权值来评价一个情感元素突发程度

$$\omega_i^t=\beta_1\times\ln(1+F_i^t)+\beta_2\times\ln(1+G_i^t)$$

式中可以看出, 当前时间窗内一个情感元素的频率越高, 并且比过去时间窗口出现的频率越高, 该情感元素越有可能成为一个主导的情感.

定义 10. 情感元素突发阈值

$$\xi_i^t=\beta_1\times\ln(1+\overline{F_i^t})+\beta_2\times\ln(1+\overline{G_i^t})$$

其中 ξ_i^t 表示情感元素 i 在当前时间窗口 t 的突发阈值. 若 $\omega_i^t>\xi_i^t$ 则表示情感元素 i 处于突发状态.

文件流突发情感检测算法:

1) 获取时间窗口内的微博文件流 W_T;

2) 抽取文件流情感元素, 并判断是否存在情感元素集 E^t:

若不存在, 则加入情感元素集;

若存在, 则增加情感出现频数;

3) 更新情感共现矩阵 W_T;

4) 筛选阈值大于 α 的情感元素作为候选突发情

感;

5) 对候选突发情感判断突发状态, 选取大于突发阈值的情感作为突发情感.

情感共现矩阵更新方法:

1) 若情感元素集增大, 则增大矩阵维度, 矩阵维度等于情感元素个数;

2) 若情感元素 e_i 的频数变化, 则改变矩阵对角线上相应的值;

3) $\forall e_m, e_n \in d_i^t$, 则增加矩阵 X(m,n)和 X(n,m)一个单位的计数.

3.3 突发情感的共现簇识别

词间的共现现象是指两个词之间形成的统计关系. 词的共现分析是自然语言处理技术在信息检索中的成功应用之一, 它的核心思想是词与词之间的共现频率在某种程度上反映了词之间的语义关联[16]. 情感词之间的共现现象同样在某种程度上反映了它们对同一情感的表达, 因此通过情感共现簇的构建, 便可以找到表达某一情感的情感元素, 作为突发文本聚类的基础.

文献[15,16,17]使用词汇 w_x 与词汇 w_y 之间共现度 $C(w_x|w_y)$ 来表示两个词汇之间的共现关系. 由于微博数据量的庞大, 当文本信息数量超过一定时, 任意情感元素之间或多或少都会建立一点的联系, 因此需要设立一个阈值 α' 来过滤那些联系不紧密的情感元素, 这里将 α' 取值为 0.1. 当情感元素之间的共现度小于阈值时, 认为它们不相关.

文献[18]在 2006 年提出了一个求取极大完全子图的算法, 能较好地从高维矩阵中求得所有极大完全子图, 但是对空间复杂度的要求高, 无法较好地应用于微博数据构成的高维矩阵中. 在此引入文献[18]中矩阵的“逆导出子图”和“逆导出补图”两个定义及两个定理, 并对算法进行修改.

定义 11. 给定图 $G=(V,E)$, $|V|=n$. 称图 $G'=(V',E')$ 为图 G 的逆导出子图, 其中 V’是这样一种点的集合: 设 v_0 是 G 中顶点度最大的点(若此点不唯一, 任选其一), $V'\leftarrow\{v_0\}$, 且将 v_0 称为 V' 中的核心点. 对于 $\forall v_i \in V$, 如果 v_i 和 v_0 之间有边相连, $V'\leftarrow\{v_i\}\cup V'$; E' 是这样的集合: $\forall v_i, v_j \in V'$, 如果 $(x_i,x_j)\in E$, 则 $(x_i,x_j)\in E'$.

定义 12. 给定图 $G=(V,E)$, $G'=(V',E')$ 是 G 的逆导出子图. 称 $G_c'=(V'',E'')$ 为图 G 的逆导出补图, 其中 $V''=(V-V')\cup\{v_i \mid \forall v_i \in V', v_j \in V-V'$, 有 $(v_i,v_j)\in E$, E'' 是这样的集合: $\forall v_i, v_j \in V''$, 如果 $(v_i,v_j)\in E$, 则 $(v_i,v_j)\in E''$.

FMCSG 算法思想:将原图分解成若干个导出子图, 再通过子图中对极大完全子图的求解,来达到求解高维矩阵极大完全子图的问题.

下面为改进的 FMCSG 算法, 并将算法分为算法 1 和算法 2. 在算法 1 中采用递归的算法, 将原图分解为若干个图的逆导出子图; 在算法 2 中对算法 1 中的每个逆导出子图求解出极大完全子图. 在改进的算法中, 我们目的在于保证时间复杂度的前提下, 降低空间复杂度, 并求出顶点数大于 1 的极大完全子图.

算法 1.

输入: 图 G, 顶点数 m;

开始:

1)If m=0, 退出;

2)找出当前矩阵中顶点度最大的点作为核心点 v_0 及相应的顶点度 m’, 并且两点之间的相邻关系必须满足 $C(v_x,v_y)>\alpha'$,否则视为不具备相连关系.

3)If m’<2, 退出;

4)以同样的度量方法, 根据定义找出核心点 v_0 对应的逆导出子图 G’;

5)将 G’和 m’作为算法 2 的输入, 执行算法 2;

6)If m’=m, 退出;

7)以同样的度量方法, 根据定义找出逆导出子图 G”和顶点数 m”;

8)以 G”和 m”作为算法 1 的输入, 进行递归运算.

算法 1 结束.

算法 2.

输入: 图 G, 顶点数 m;

输出: Nest; //Nest 代表极大完全子图的顶点序列集, 初始值为空;

开始:

1)k=1;

2)找到核心点 v_0, 并与其他各点形成 2 阶矩阵并存入矩阵数组 MK(k).

3)for k=2 to m

将 MK(k)中的矩阵两两进行比较, 判断是否两个矩阵之间的任意点对的共现度是否大于阈值, 是则进行合并可行性判断, 通过则放入 MK(k)中并进行 MK(k)的矩阵剔除工作, 否则比较下一对矩阵. (矩阵合并可行性判断否定因素, 若合并后的矩阵已存在

MK(k)中, 或被 MK(k)中某一矩阵所包含, 则放弃. 矩阵剔除工作内容为: 若任一已有矩阵被新加入矩阵所包含, 则删除该矩阵.)

算法 2 结束.

将MK(k-1)中无法进行合并的矩阵(合并次数小于1)存入 Nest.

设图 G 的结点数为 n, 可以证明该算法和原 FMCSG 算法的时间复杂度相同为:

$$T(n)=T(n-1)*T(n-1)+O(n^2).$$

原算法的空间复杂度为 S(n)=O(n!), 但改进后空间复杂度为 S(n)=O(n^2).

3.4 共现簇的聚类和模型的生成

由于中文表达方法的多样, 同一情感可由多种情感图片和情绪词表达, 而用户的使用习惯不同, 导致某些相同情感下的情感元素共现度没有达到阈值, 因此在这里需要进行共现簇的聚类工作, 将表达相似情感的共现簇作为同一情感模型下的子空间.

共现簇内特征项的数目较少, 因此只需要采用简单的文本聚类方法.

定义 13. 情感模型

$$EM^t=\{C_{i1}^t,C_{i2}^t,C_{i3}^t,...\}$$

其中, EM^t 表示时间窗口 t 内表达某一情感的模型, 它由表达相似情感的共现簇所构成.

定义 14. 情感模型集

$$EMSet^t=\{EM_1^t,EM_2^t,EM_3^t,...\}$$

其中, $EMSet^t$ 表示时间窗口 t 内情感模型的集合.

定义 15. 任意两个共现簇 C_i^t 和 C_j^t 之间的连通关系只需满足

$$\frac{Same(C_i^t,C_j^t)}{Num(C_i^t)}>\varphi 或 \frac{Same(C_i^t,C_j^t)}{Num(C_j^t)}>\varphi$$

则认为这两个共现簇之间是连通的, 其中 $Same(C_i^t,C_j^t)$ 表示两个共现簇间相同的情感元素, $Num(C_i^t)$ 表示该簇内所含有的情感元素数量, 参数 φ 这里去 2/3.

情感模型集的自动生成算法思想: 根据无向图连通子图的求解方法, 生成若干个连通子图, 每一个连通子图便对于一个情感模型, 子图中的点为相应的共现簇.

情感模型集的自动生成算法.

输入: $CSet^t$, 情感共现簇数量 m;

输出: $EMSet^t$;

开始:

1)定义队列 L, 布尔数组 finish 初值为假;

2)if 中值都为真, 算法结束;

3)if S 为空, 建立新的情感模型 EM_i, 将 $CSet^t$ 中未被访问的任一未归入情感模型的共现簇 C_i^t 进入队列 L 中;

4)if S 不为空, 进行出队列操作, 将出队列的共现簇归入当前的情感模型 EM_i 中, finish[i]的值设为真, 并将与该共现簇连通且相应 finish 为假的其他共现簇加入队列 L 中, 返回 2).

3.5 基于词共现和情感分析的话题检测

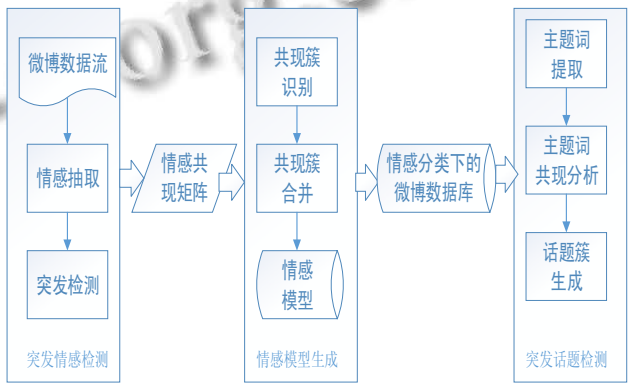

图 1 整体话题检测算法模型

图 1 中为本文提出话题检测算法的整体模型框架, 将微博根据情感分类之后, 生成各情感下的微博数据库, 对每个微博数据库进行话题检测. 因此整体话题检测算法分为 3 个部分: 突发情感检测; 情感模型生成; 突发话题检测.

在第三个步骤中, 实验中采用根据传统的词共现方法[16,17].

1) 抽取同一情感微博数据库中的主题词;

2) 根据主题词之间的共现度大小觉得是否连边;

3) 由上一步得出一个多连通图 G, 其中每一个连通分量构成一个话题簇;

4) 选出话题簇中的边数最多的主题词作为该话题的领域词汇, 形成话题.

由本文提出的检测算法模型可以看出, 该算法将相似话题下不同情感的微博也认为是不同话题, 该方法的目的是更便于找出那些直接能影响用户情绪变化的对象, 不同情感下的微博带给微博的舆论导向也不相同, 因此该方法能很好地检测出带来某一情感变化的微博话题.

4 实验结果及分析

4.1 实验数据及环境说明

本实验采用的数据库为 NLPIR[18]微博内容语料库-500 万条其中 2014 年 1 月 23 号至 2014 年 2 月 28 日连续 5 周的微博数据. 其中已剔除研究无关数据(包含空文本微博、纯字符微博、重复微博), 由图 1 所示, 研究中使用的数据共计 87601 条, 其中每日信息量的均为 2368 条. 实验中所采用的机器配置环境为: CPU 为 Intel Core i5- 4210M 2.60 GHz, 内存 4 GB, Windows7 操作系统, Eclipse 集成开发环境, JDK1.7 (Java development kit v1.7).

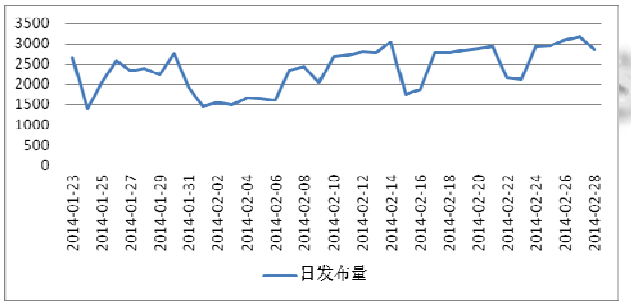

图 1　连续 5 周微博每日发布量

通过对微博数据人工收集的方法建立情绪词库, 情绪词库中的情感词共 927 个, 其中正面情绪词 420 个, 负面情绪词 507 个(数据共享链接: http://pan.baidu.com/s/1dD2V1aL). 通过微博中表情图片标签"[]"收集情感图片的方法建立动态表情图片库, 平均每日收集 135 个.

4.2 实验结果分析

实验共分为 3 个部分: 第一部分研究时间窗内, 微博数据发布与情感变化的关系, 找出情感微博与话题的关系. 第二部分将各情感对微博信息进行分类结果的比较. 第三部分将本文提出的方法与其他传统的微博话题检测方法在检测效果上进行比较.

1) 实验一: 研究情感微博与话题的关系.

首先将连续 5 周内的微博进行情感元素提取, 经过统计得到图 2 所示结果. 由图 2 所示, 微博发布曲线走势与情感微博发布曲线走势相近. 图 3 所示为微博发布量变化率与情感微博发布量变化率关系, 其中可以看出微博发布量的变化直接影响情感微博的发布量, 其中情感图片微博的发布量变化与微博发布量的变化关系最为密切, 含情绪词微博变化较为平缓. 由图 2 和图 3 可以看出, 情感图片的微博所占比重较大, 微博用户更倾向于使用表情图片来表达自己的情感态度. 图 4 中所示为每日话题微博数量, 以及相对于含情感元素的微博数量, 可见话题微博发布量的变化与情感元素关系密切, 情感元素变化能较好的用来表征话题变化. 综上可得出结论, 微博信息与情感密切相关, 话题信息更倾向于表达出某一情感内容.

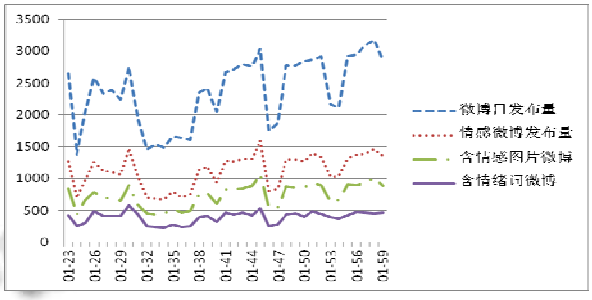

图 2　微博发布量与情感微博发布量关系

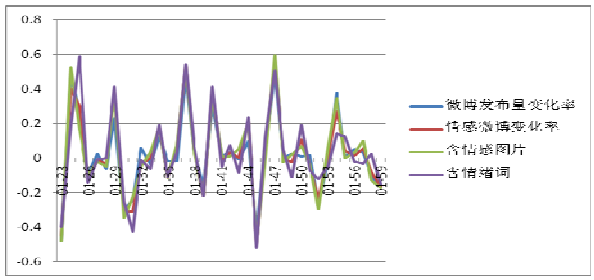

图 3　微博发布量变化率与情感微博发布量变化率关系

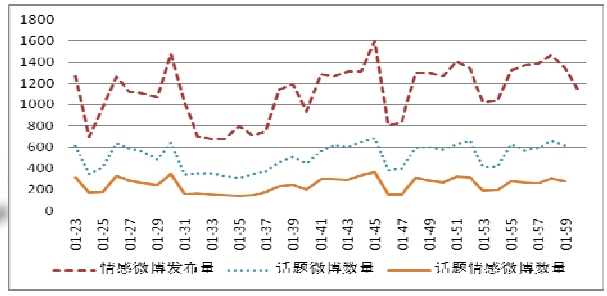

图 4　话题数量与情感微博关系

2) 实验二: 研究各情感特征对分类结果的影响.

首先分析突发话题判断参数 β_1 和 β_2 和取值. 由图 5 中, 参数 β_1 的取值对情感簇内情感词数的影响可以看出, 当 β_1 趋向 0 时, 情感词增加率对突发情感词起作用; 当 β_1 趋向 1 时, 情感词频率对突发情感词起作用; 而当 β_1 取 0.4, 相应 β_2 取 0.6 时, 情感词数可以取极大值. 因此实验中对参数 β_1 和 β_2 分别取 0.4 和 0.6.

接下来实验中将时间窗口的长度设为 60 分钟(mm). 分别采用正负情感极性、情绪词、表情图片、情感元素这些特征对微博进行情感分类, 比较类内微

博之间主题词的相似性(这里相似性用共现率来表示). 表 1 和表 2 中所示为 24 个时间窗内的平均数据结果, 情感特征采用情感元素分类, 可以使分类后类内主题词间的关系更加紧密, 且更倾向于对同一主题的表达; 同时, 情感元素的划分方法使类间的主题词分离度更高, 使不同情感表达不同的主题内容.

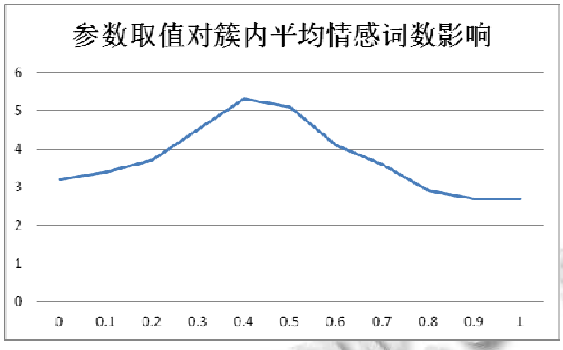

图 5 情感簇内词数与参数关系

表 1 不同情感特征与相同情感类别下的主题词的关系

	情感极性	情绪词	表情图片	情感元素
平均共现率	0.43	0.61	0.65	0.68
最高共现率	0.74	0.84	0.89	0.96

表 2 不同情感特征与不同情感类类间的主题词的关系

	情感极性	情绪词	表情图片	情感元素
平均共现率	0.26	0.12	0.11	0.11
最高共现率	0.69	0.61	0.59	0.57

3) 实验三, 不同话题检测方法检测结果对比.

其中评价指标:

话题准确率 P1: 时间窗内检测出的话题微博数作为分子, 时间窗内检测出的话题微博数作为分母. 反应算法对话题信息检测的能力.

话题覆盖率 P2: 时间窗内检测出的话题类别数作为分子, 时间窗内的总话题类别数作为分母. 反应算法检测新话题的能力.

检测时间 T: 反应时间窗口内算法对数据处理的能力, 检测时间越短表示该算法的实效性越好. (由于不同算法的预处理内容不同, 这里不计入本文预处理的时间.)

对比算法主要采用传统词同现方法[16,17]、基于情感极性的方法[9]和基于表情图片分类的方法[17], 分别记为 A1、A2 和 A3, 本文提出的方法记为 A4.

表 3 算法平均效率

	A1	A2	A3	A4
P1	0.51	0.67	0.87	0.91
P2	0.60	0.68	0.82	0.85
T	1m31s	2m7s	2m19s	1m50s

由表 3 内容可以看出, 本文提出的算法在单位时间窗口内平均可以在 2 分钟内检测出话题结果, 并且相对其他传统的话题检测算法, 有较高的话题准确率和覆盖率.

5 未来工作和总结

本文通过在情感元素共现的基础上, 建立情感共现矩阵, 通过聚类的方法形成某一特定情感的子空间模型. 利用情感子空间的划分, 对微博信息进行分类处理, 能有效地提高传统话题检测算法的实效性和精度. 在未来的工作中, 还需要考虑情感变化程度带来的影响, 主要研究在于如何找到一种能对情感类别强度量化的方法, 使得能对情感变化导致情感强度发生改变时话题的检测性能; 其次还需要考虑情感的时间段惯性带来的影响, 即在某一时段内存在固有的某一种情感, 主要研究在于如何识别这种情感, 并对其赋予一定的话题权重.

参考文献

1 Sakaki T, Okazaki M, Matsno Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. Proc. of the 19th International Conference on World Wide Web. 2010.

2 Doan S, Vo BKH, Collier N. An analysis of Twitter messages in the 2011 Tohoku Earthquake. Electronic Healthcare, 2012, 91:58–66.

3 张珊,于留宝,胡长军.基于表情图片与情感词的中文微博情感分析.计算机科学,2012,S3:146–148,176.

4 Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning Techniques. Proc. of Emnlp. 2002. 79–86.

5 周杰,林琛,李弼程.基于机器学习的网络新闻评论情感分类研究.计算机应用,2010,(4):1011–1014.

6 闻彬,何婷婷,罗乐,宋乐,王倩.基于语义理解的文本情感分类方法研究.计算机科学,2010,(6):261–264.

7 王素格,李德玉,魏英杰.基于赋权粗糙隶属度的文本情感分类方法.计算机研究与发展, 2011,(5):855–861.

8 庞磊,李寿山,周国栋.基于情绪知识的中文微博情感分类方法.计算机工程,2012,(13):156–158,162.
9 方然,苗夺谦,张志飞.一种基于情感的中文微博话题检测方法.智能系统学报,2013,(3):208–213.
10 张鲁民,贾焰,周斌,赵金辉,洪锋.一种基于情感符号的在线突发事件检测方法.计算机学报,2013,(8):1659−1667.
11 孙劲光,马志芳,孟祥福.基于情感词属性和云模型的文本情感分类方法.计算机工程,2013,(12):211–215,222.
12 刘林,刘三(女牙),刘智,铁璐.基于随机主元分析算法的BBS 情感分类研究.计算机工程,2014,(5):188–191.
13 卢伟胜,郭躬德,陈黎飞.基于词性标注序列特征提取的微博情感分类.计算机应用,2014,(10):2869–2873.
14 胡杨,戴丹,刘骊,冯旭鹏,刘利军,黄青松.基于情感角色模型的文本情感分类方法.计算机应用,2015,(5):1310–1313,1319.
15 周进华,刘贵全.基于衰减词共现图的多文档摘要研究.小型微型计算机系统,2009,(1):173–177.
16 赵文清,侯小可.基于词共现图的中文微博新闻话题识别.智能系统学报,2012,(5):444–449.
17 郑斐然,苗夺谦,张志飞,高灿.一种中文微博新闻话题检测的方法.计算机科学,2012,(1):138-141.
18 郭平,康艳荣,史晓晨.基于最大 Code 码的极大完全子图算法.计算机科学,2006,(2):188-190,200.
19 张华平.NLPIR 微博内容语料库—500 万条.自然语言处理与信息检索共享平台, http://www.nlpir.org/.
20 张华平.ICTCLAS2015 版本.自然语言处理与信息检索共享平台,http://ictclas.nlpir.org/.