一种融合用户关系的自适应微博话题跟踪方法

柏文言[1,2],张 闯[1],徐克付[1],张志明[3]
(1. 中国科学院信息工程研究所,北京 100093;2. 北京邮电大学计算机学院,北京 100876;
3. 北京英孚泰克信息技术股份有限公司,北京 100089)

摘 要: 针对微博口语化、文本短小等特点以及现有研究的不足,本文提出了一种融合用户关系的自适应微博话题跟踪方法. 首先,在当前跟踪的时间窗内,推文被映射到特征空间,并作为候选推文集合. 然后,针对推文的分布特点以及话题跟踪的目的,变换推文特征空间. 在此基础上,利用改进的 K-means 聚类算法对候选推文集合进行二元聚类,从而划分出相关推文集合,即当前话题目标模型. 本文通过 Twitter 平台获取数据进行实验,实验结果表明,该方法能够实时地跟踪话题热度的变化以及焦点的演变,并提高了微博中话题跟踪的稳定性. 该方法为用户推荐、舆情分析等领域提供了有效的支撑.

关键词: 微博; 话题跟踪; 自适应; 用户关系; 极坐标; K-means 算法
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)06-1375-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.06.014

A Self-Adaptive Microblog Topic Tracking Method by User Relationship

BAI Wen-yan[1,2], ZHANG Chuang[1], XU Ke-fu[1], ZHANG Zhi-ming[3]
(1. *Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093, *China*;
2. *School of Computer Science, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
3. *Information Technologies Co, (Beijing) Ltd, Beijing* 100089, *China*)

Abstract: Considering the colloquial, short text and other characteristics of microblog and deficiencies in research of it, this article proposes a self-adaptive topic tracking method of microblog by user relationship. First of all, during the tracking time window, the candidate tweet set is mapped into feature space. Secondly, aiming at the characteristic of tweet distribution and the purpose of topic tracking, the paper converts the tweets' feature space. Based on this operation, a binary clustering on tweets set can be constructed by improved K-means clustering algorithm. The yielded relative collection is the target model of the current topic. The experiments with the data extracted from Twitter, show that this method can track down the trend of hot topics and the evolution of focuses in real time, and improve the stability of topic tracking in microblog. This method serves well for user recommendation and public opinion analysis.

Key words: microblog; topic tracking; self-adaptive; user relationship; polar coordinates; K-means algorithm

1 引言

随着微博的快速发展和广泛使用,在微博空间中每天都会产生海量的数据信息,这些信息包括文字、图片以及视频链接等. 在微博的网络空间中,微博用户往往更加关注热点事件的实时进展情况. 从而,在这种实时更新的宏大信息流中,用户对于微博话题的动态跟踪有着迫切的需求. 话题跟踪(Topic Tracking, TT)是信息跟踪研究领域的一个实例,其旨在根据给定的话题信息,在后续未知数据流中跟踪已知话题的所有相关报道信息,得到话题的进展和演变情况. 与传统的网页新闻和博客相比,微博具有内容短小、原创性、口语化以及实时性等特点,所以针对微博的话题跟踪,传统的长文跟踪方法已经不

收稿日期:2015-07-24;修回日期:2016-01-08;责任编辑:覃怀银
基金项目: 国家自然科学基金(No. 61602474)

适用.因此需要在传统话题跟踪方法的基础上,从微博的特点出发,研究如何解决微博话题跟踪中短文本处理、时变性等难点问题,从而实现在错综复杂的数据流中获得话题的相关信息,满足用户对话题进行持续跟踪的需求.

2 相关工作

针对微博的话题跟踪的特点,许多学者突破了长文本话题跟踪的技术,在微博话题跟踪领域展开了新的研究.在国外,一些学者对英文的微博进行了研究,其中主要是基于 Twitter 上的数据.提出通过话题模型[1]、事件检测[2]、热词感知[3]等方法获得突发性话题,以及通过反馈迭代[4]、时间概率[5]、稀疏矩阵[6]等方法解决话题漂移的问题.

在国内,一些学者针对中文微博的话题跟踪展开了研究.针对微博短文本处理问题,提出多种扩展特征向量的方法[7~9],对于微博文本相似度的计算也提出了基于热度[10]、语义结构[11]、特征词关联度[12]、相似距离[13]等多种方法.在特征提取和相似度计算等支撑算法基础上,利用 K-means 的话题模型的聚类[14],基于微博信息熵的动态话题跟踪[15]等微博话题跟踪方法也得到了研究与发展.

针对微博的话题跟踪,国内外的许多学者进行了相关的研究,但是现有方法仍存在诸多的问题,目前多数方法基于文本分类技术实现话题跟踪,该类方法依赖于初始样本训练,但通常在微博话题产生的初始阶段,没有足够的可用于训练的初始样本,过少的训练样本会造成分类器的泛化能力严重降低.同时,该类方法是通过逐条判断的方法实现话题的跟踪,并没有充分利用用户信息及其历史行为,也没有结合当前话题的背景语义,仅依赖于及其短小且口语化的文本信息,往往无法提取出特异性属性.针对微博的特点,本文提出了一种融合用户关系的自适应微博话题跟踪方法.该方法具有以下贡献:(1)引入用户属性协助推文的相关性判断,用户属性来自于其历史推文,从而借助其历史行为增加推文判断的稳定性.(2)对推文特征空间进行坐标变换,使其近似线性可分,从而使聚类问题简化为二元聚类,有效的提高了推文判断的效率与准确性.(3)采用迭代跟踪方式替代逐条分类,不需要样本进行初始训练;利用当前跟踪到的相关推文集合生成新一轮的话题跟踪模型,并能够密切跟踪话题的焦点演变.(4)话题跟踪在推文集合上进行,关注话题的整体走向,强调热度的变化与焦点的演变.话题跟踪效果不依赖于单条推文相关度的判别.

3 研究内容

3.1 话题跟踪模型概述

在本文的话题跟踪模型中,主要包括三个集合:分别是“网状”动态的用户集合、“带状”延展的推文集合以及“柱状”附有权值的话题特征集合.如图 1 所示,在话题跟踪过程中,用户集合会不断地被更新,包括将非相关用户标记为相关用户、更新已有用户的相关权值以及将相关用户标记为非相关用户.其次,推文集合类似于一条“传送带”,沿着时间线不断地增加新推文、淘汰过期推文.而话题特征集合是附有权值的话题焦点集合,在跟踪过程中,随着话题的发展,话题特征集合中焦点不断演变.本文主要是通过用户集合、推文集合以及话题特征集合三者之间的相互作用和影响来实现融合用户关系的自适应微博话题跟踪,基本思路如下:首先,整个话题跟踪系统由用户驱动,用户集合中的所有用户行为直接推动推文集合按时间线延展,其中强相关用户的推文被标记为高相关度候选推文,作为后续跟踪的重要依据.然后,在当前跟踪时间窗内,利用话题

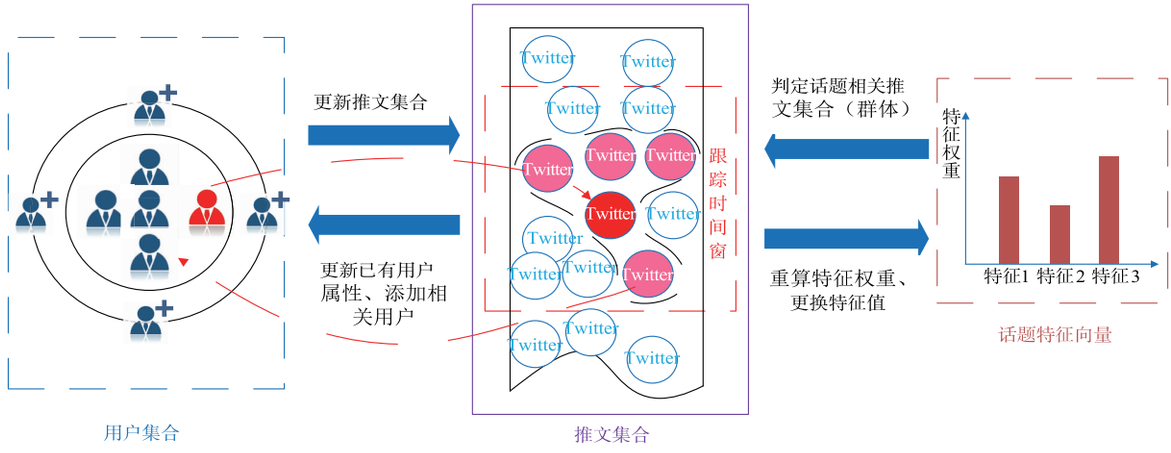


图1 融合用户关系的自适应话题跟踪模型

特征模型在候选推文集合中聚类相关话题子集,类中心即当前话题目标模型. 最后,将当前话题目标模型迭代更新至话题特征模型,用于下一轮话题跟踪. 同时,将强相关推文的关联的用户添加至用户集合.

3.2 特征向量的提取与权重计算

定义 $\boldsymbol{t}_i(w_i^1,w_i^2,w_i^3,\cdots,w_i^l,\cdots,w_i^L)$ 表示第 i 条推文的话题的特征向量,其中 w_i^l 为第 l 个特征词的权重. 在当前话题追踪时间窗内,将时间窗内的所有推文进行分词处理,由于推文中存在大量的高频但无意义的词语,首先使用中文停用词表过滤这些噪声干扰,只保留句子的核心词语. 然后,对所有核心词按其词频进行排序,取前 L 个词作为特征词. 从而组成维度为 L 的特征向量.

我们采用改进 TF-IDF 模型进行话题特征的提取和权重的计算. 特征权重的计算公式如式(1)所示:

$$w_j^i=\mathrm{tf}(i,j)\times\mathrm{idf}(i,j)\times k(i) \tag{1}$$

其中:$\mathrm{tf}(i,j)=n_{i,j}/\sum_k n_{k,j}$,$n_{i,j}$表示特征 i 在推文 j 中的个数,$\sum_k n_{k,j}$表示推文j 中所有词的个数. $\mathrm{idf}(i,j)=\log N_t/N_i$,$N_t$ 表示推文集中包含的推文的总数量,N_i 表示包含特征 i 的推文总数量. $k(i)=p_{ci}/p_{ti}$,p_{ci}表示在一个话题类 c 中包含特征 i 的推文数与类中推文数的比值,p_{ti}表示所有包含该特征的推文数与推文集大小的比值.

在一个文档中,当两个特征词的 tf 和 idf 相等时,其话题的分辨能力越大,则特征的权重越大. 显然,k 可以很好地反映一个特征的话题分辨能力,当某一特征在一个话题中出现次数较多,而在其他话题中出现次数较少时,该特征的 k 值就会较大.

在话题追踪的过程中,本文将上一轮聚类得到的相关话题的中心点映射到当前的话题向量空间中,从而得到下一轮的聚类初始中心点. 利用这种特征值更换的方法可以有效的解决话题漂移的问题.

由于话题跟踪的目的是划分出相关话题推文的子集,而不关心其他话题的子类别. 本文采用二元聚类的方式将整个推文集合聚类为相关推文集合和非相关推文集合,所以需要推文集合的分布为线性可分. 但是,推文在特征空间内按话题混杂分布,本身并不是线性可分. 所以针对话题跟踪的目标以及推文集合的分布特点,我们将对原始的推文特征空间进行坐标变换. 原始推文集合的分布特点如图 2 所示, 推文按照话题聚集为“话题簇”,“话题簇”是围绕某话题而聚集的高密度推文集合,其密度取决于话题的热度,体积取决于话题焦点的分化程度,话题热度越大、焦点分化程度越高,其相关推文集合越是呈现为密度高、体积大的“话题簇”. 同时,推文集合中还存在大量的非话题性推文,这些推文不规则的分布在话题特征向量空间中,形成了话题向量空间的随机噪声与背景.

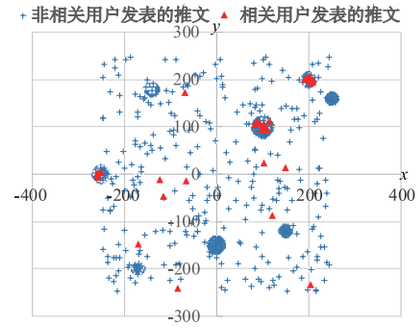


图2　原始推文集合分布图

3.3　极坐标变换及线性可分证明

首先,我们将话题向量空间的坐标原点进行平移,将原点平移到当前待跟踪的目标模型上. 得到了以话题目标模型为原点的新的坐标分布图. 其中,坐标原点平移过程如下:

(1)定义当前话题目标模型为 $\boldsymbol{t}_g(w_g^1,w_g^2,w_g^3,\cdots,w_g^l,\cdots,w_g^L,)$;定义原推文特征向量为 $\boldsymbol{t}_i(w_i^1,w_i^2,w_i^3,\cdots,w_i^l,\cdots,w_i^L)$.

(2)平移后目标模型变换为坐标原点,即:$\boldsymbol{t}_g(w_g^1,w_g^2,w_g^3,\cdots,w_g^l,\cdots,w_g^L,)\rightarrow\boldsymbol{O}(0,0,0,\cdots,0,\cdots,0)$;原推文特征向量 $\boldsymbol{t}_i(w_i^1,w_i^2,w_i^3,\cdots,w_i^l,\cdots,w_i^L)$ 变换为 $\boldsymbol{t}_i(w_i^1-w_g^1,w_i^2-w_g^2,w_i^3-w_g^3,\cdots,w_i^l-w_g^l,\cdots,w_i^L-w_g^L)$.

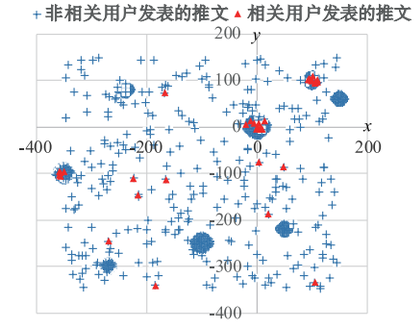


图3　坐标原点平移后推文集合分布图

经过坐标平移变换后得到的推文集合分布如图 3 所示,话题相关推文集合聚集在坐标原点周围形成“目标话题簇”,其他非目标话题推文聚集在各自的话题中心周围形成相应的“话题簇”,其密度和体积随着各自话题的热度及焦点的分化而改变. 同时在“话题簇”之间存在着零散分布的背景推文.

如图 3 所示,除了目标话题的推文集合外,其他非目标话题相关的推文集合的 “话题簇”不均匀地散列在目标“话题簇”周围. 在这种分布下,利用 K-means 算

法进行二元聚类的效果很差. 针对上述问题,我们将平移后的特征向量进行极坐标变换.

其中,极坐标变换过程如下:

(1)定义坐标平移后的推文特征向量为 $\boldsymbol{t}_i(w_i^{'1},w_i^{'2},w_i^{'3},\cdots,w_i^{'l},\cdots,w_i^{'L})$.

(2)根据极坐标变换公式如式(2)至式(4)所示

$$\rho=\sqrt{(w_i^{'1})^2+(w_i^{'2})^2+(w_i^{'3})^2+\cdots+(w_i^{'l})^2+\cdots+(w_i^{'L})^2} \tag{2}$$

$$\theta^1=\tan^{-1}(w_i^{'2}/w_i^{'1}),\theta^2=\tan^{-1}(w_i^{'3}/w_i^{'1}) \tag{3}$$

$$\theta^{L-1}=\tan^{-1}(w_i^{'L}/w_i^{'1}) \tag{4}$$

可得极坐标变换后的推文特征向量为:$\boldsymbol{t}_i=(\rho_i,\theta_i^1,\theta_i^2,\cdots,\theta_i^l,\cdots,\theta_i^{L-1})$.

(3)定义两条推文的相似度计算式(5):

$$\|\boldsymbol{t}_1-\boldsymbol{t}_2\|=|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2| \tag{5}$$

经过极坐标变换后得到的推文集合分布如图4所示,话题相关推文集合呈"带状"近似平行的分布在极径上,由图可知,相关话题推文集合以及非相关话题推文集合线性可分.

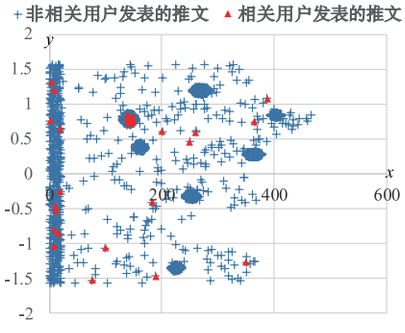


图4 极坐标变换后推文集合分布图

实际上,在原始的推文特征空间中,目标话题簇的外围存在一个包络面,该包络面为一个近似以话题目标模型为中心点的超球面,包络面使几乎所有相关的话题推文均在这个超球中,几乎所有不相关推文在这个超球面外面. 该超球面的方程如式(6)所示:

$$(x_1-w_g^1)^2+(x_2-w_g^2)^2+\cdots+(x_l-w_g^l)^2+\cdots+(x_L-w_g^L)^2=r^2 \tag{6}$$

其中,近似所有在超球内部的话题相关的推文 $\boldsymbol{t}_i(w_i^1,w_i^2,w_i^3,\cdots,w_i^l,\cdots,w_i^L)$ 满足:

$$(w_i^1-w_g^1)^2+(w_i^2-w_g^2)^2+\cdots+(w_i^l-w_g^l)^2+\cdots+(w_i^L-w_g^L)^2<r^2$$

近似所有在超球外面的话题非相关推文 $\boldsymbol{t}_j(w_j^1,w_j^2,w_j^3,\cdots,w_j^l,\cdots,w_j^L)$ 满足:

$$(w_j^1-w_g^1)^2+(w_j^2-w_g^2)^2+\cdots+(w_j^l-w_g^l)^2+\cdots+(w_j^L-w_g^L)^2>r^2$$

当坐标原点转移到当前的话题目标模型上时,此时的超球面公式转换为如式(7)所示:

$$x_1^2+x_2^2+\cdots+x_l^2+\cdots+x_L^2=r^2 \tag{7}$$

同时,近似所有在超球内部的话题相关的推文 $\boldsymbol{t}_i(w_i^{'1},w_i^{'2},w_i^{'3},\cdots,w_i^{'l},\cdots,w_i^{'L})$ 满足:

$$(w_i^{'1})^2+(w_i^{'2})^2+\cdots+(w_i^{'l})^2+\cdots+(w_i^{'L})^2<r^2$$

近似所有在超球外面的话题非相关推文 $\boldsymbol{t}_j(w_j^{'1},w_j^{'2},w_j^{'3},\cdots,w_j^{'l},\cdots,w_j^{'L})$ 满足:

$$(w_i^{'1})^2+(w_i^{'2})^2+\cdots+(w_i^{'l})^2+\cdots+(w_i^{'L})^2>r^2$$

最后,当向量空间进行极坐标变换后,该超球面在极坐标系下的图形为垂直于极轴的超平面,已知

$$\rho=\sqrt{x_1^2+x_2^2+\cdots+x_i^2+\cdots+x_L^2}$$

则超平面的方程为:$\rho=\sqrt{r^2}=r$.

此时,超平面将特征空间向量划分为两部分,使得绝大部分的话题相关推文分布在一侧,即这部分空间中的推文 $\boldsymbol{t}_i=(\rho_i,\theta_i^1,\theta_i^2,\cdots,\theta_i^l,\cdots,\theta_i^{L-1})$ 满足:$\rho_i<r$ 同时,绝大部分话题非相关推文分布在空间的另一侧,即这部分空间中的推文 $\boldsymbol{t}_i=(\rho_j,\theta_j^1,\theta_j^2,\cdots,\theta_j^l,\cdots,\theta_j^{L-1})$ 满足:$\rho_i>r$.

综上证明可知,将推文的特征向量空间进行极坐标转换后,存在一个超平面将特征向量空间划分为两部分,即此时的特征向量空间线性可分. 再此基础上,利用 K-means 算法对极坐标下的推文集合进行二元聚类分析.

3.4 K-means 聚类算法实现

对推文特征向量空间进行坐标原点平移与极坐标变换后,下一步是利用聚类算法找到相关话题的推文子集,其中由相关用户标注的点在更新聚类中心时会被加权.

定义当前的跟踪时间窗内的推文特征向量集合为:$T=\{\boldsymbol{t}_i\mid \text{time}(\boldsymbol{t}_i)\in[\text{date}_1,\text{date}_2]\wedge(\boldsymbol{u}_j\uparrow\boldsymbol{t}_i),\boldsymbol{u}_j\in U\}$;

定义当前用户集合为:$U=\{\boldsymbol{u}_0,\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k,\cdots,\boldsymbol{u}_K\}$,其中$[\text{date}_1,\text{date}_2]$表示当前跟踪时间窗的时间范围,$(\boldsymbol{u}_j\uparrow\boldsymbol{t}_i),\boldsymbol{u}_j\in U$ 表示 t_i 为用户 u_j 发表的推文.

利用 K-means 聚类算法找出相关推文集合的流程如下:

(1)初始化类中心

在跟踪时间窗内的推文样本集合 T 上,指定两个初始类别中心点,其中一个主类别中心点 $\boldsymbol{m}_g^{(1)}$ 为上一个跟踪周期的目标,另外指定一个非相关话题推文集合的中心点为 $\boldsymbol{m}_f^{(1)}(\rho_f,0,\cdots,0,\cdots,0)$,且指定 $\rho_f=C\times\rho_g$,其中 C 为常用系数. 同时设定迭代终止条件

$\max(\|\boldsymbol{m}^{(n+1)}-\boldsymbol{m}^n\|)<\Delta$,其中 $\max(\|\boldsymbol{m}^{(n+1)}-\boldsymbol{m}^n\|)$

表示第$(n+1)$次迭代得到的中心点与第 n 次迭代得到的中心点之间的相似度距离,Δ 为聚类中心收敛误差容限.

(2)样本归类

对 T 中的每条推文计算其到每个类中心的距离 $\|\boldsymbol{t}_1-\boldsymbol{t}_2\|=|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2|$,并把它归到最近的质心 $\boldsymbol{m}_{hit}^{(n)}$ 所代表的类别中 S,当且仅当$\|\boldsymbol{t}_i-\boldsymbol{m}_{\text{hit}}^{(n)}\|\leqslant\|\boldsymbol{t}_i-\boldsymbol{m}^{(n)}\|$.

(3)更新聚类中心

本文对 K-means 聚类算法更新聚类中心操作进行了相应改进,主要做法为:对相关话题推文类别的聚类中心进行更新,以此类的平均向量作为新的聚类中心 $\boldsymbol{m}_g^{(n+1)}$,其计算公式如式(8)所示:

$$\boldsymbol{m}_g^{(n+1)}=\sum_{i=0}^{\text{size}}(q_i\boldsymbol{t}_i)/\text{size},\boldsymbol{t}_i\in S \tag{8}$$

其中 q_i 为每个推文特征向量的权重,非相关话题推文类的聚类中心保持不变,即 $\boldsymbol{m}_f^{(n+1)}=\boldsymbol{m}_f^{(n)}=\boldsymbol{m}_f^{(1)}(\boldsymbol{\rho}_f,0,\cdots,0,\cdots,0)$.

(4)迭代(1)~(3)步直至满足上述收敛条件,得到 $\boldsymbol{m}_g^{(n+1)}$ 作为跟踪的话题目标,S 为跟踪到的话题目标集合. 如图 5 所示,为此轮二元聚类的结果.

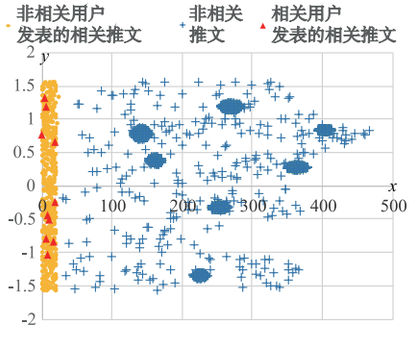


图5　聚类结果示意图

最后,利用上述聚类步骤得到的 $\boldsymbol{t}_a=\boldsymbol{m}_g^{(n+1)}$ 话题跟踪模型计算得出下一轮的话题跟踪模型,同时,根据本轮得到的相关推文集合,将与其关联的用户的相关度进行更新. 计算公式如(9)所示

$$\boldsymbol{t}_g^{'}=\delta\times\boldsymbol{t}_a+(1-\delta)\times\boldsymbol{t}_g \tag{9}$$

4　实验结果与分析

本文采用的分词程序是基于中科院的 ICTCLAS 中文分词算法. 为了验证本文方法的有效性,本文通过 Twitter 平台获取中文推文作为实验数据. 如表 1 所示:分别选取了在 2013 年、2014 年以及 2015 年发生的三个热点事件,通过清洗掉长度小于 4 字符或者只包含 url 和表情的数据,在这 3 个话题事件的时间周期内,本文总共选取 72993 条推文,涉及到的推特用户有 20134 名.

表 1　实验话题事件相关信息统计表

事件 ID	事件名称	起始时间	结束时间	推文总数
Topic1	复旦投毒	2013/4/14	2013/4/30	18390
Topic2	马航失事	2014/3/8	2014/4/1	37213
Topic3	长江沉船	2015/6/1	2015/6/16	17390

4.1　自适应性实验分析

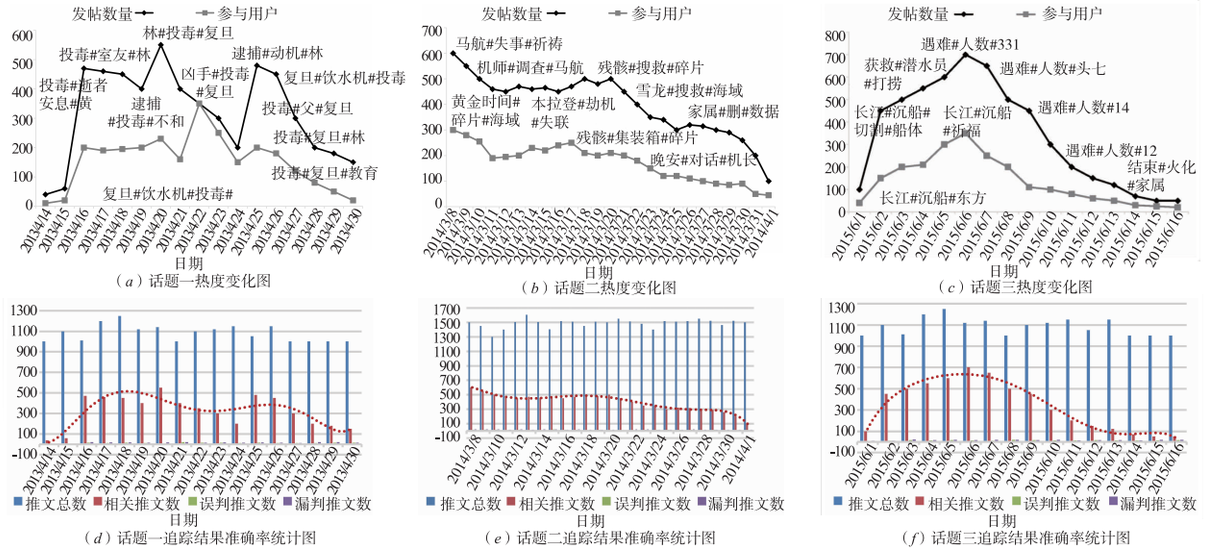
图 6(a)、(b)和(c)三个子图分别说明了三个热点话题追踪过程中推文集合、用户集合以及话题特征集合的变化和相互影响. 随着话题的不断发展,追踪到的相关话题数量和相关用户数量变化趋势大体一致,如当话题处于高峰期时,相关推文的数量和相关用户的数量达到最大,可见,本文的话题追踪模型中用户集合和推文集合的相互作用保证了追踪结果的准确性. 其次,在进行下一轮聚类之前,模型会根据上一轮得到的相关话题中心点计算下一轮的初始聚类中心点. 由三个子图可以看出,实验中话题追踪模型每次聚类得到的话题中心点随着三个热点事件的发展而改变,并且符合话题的发展趋势. 可见,推文集合和话题特征集合的相互作用有效的避免了话题漂移的问题.

图 6(d)、(e)和(f)三个子图分别为话题一、话题二和话题三的准确率统计图. 本文分别统计了时间窗内所有推文的总数量、话题模型跟踪到的相关话题推文数量、相关话题推文的误判数量以及相关话题推文的漏判数量. 其中,在三个图中,话题模型跟踪到的相关话题数量曲线图分别与图 6(a)、(b)和(c)中曲线形状相似. 可见,本文的话题跟踪模型可以有效地为用户提供话题的整体走向. 同时,从三组话题的跟踪结果的准确率统计图可以看出,本文的话题跟踪模型有较低的误判率以及漏判率.

4.2　对比实验结果分析

为了进一步验证验证本文中融合用户关系的自适应微博话题跟踪方法的性能,利用上述的实验数据,将本文方法与文献[5]和文献[6]中的方法进行对比实验. 其中,文献[5]中的方法旨在通过计算不同单词随时间而产生不同变化的概率来解决话题跟踪过程中话题漂移的问题. 文献[6]中的方法是利用给传统的查询扩展来解决数据稀疏的问题,并借助 Rocchio 的算法的思想解决话题漂移的问题.

本文实验采用 TDT 评测标准,利用召回率(R)、准确率(P) 以及两者综合性能指标 F 值这三个指标进行评价. 根据 TDT 评测标准,三种话题跟踪模型的实验结果如表 2 所示. 其中,针对三个热点事件,本文的话题跟踪模型跟踪结果的召回率、准确率以及 F 值均高于其他两种方法. 由此看出,本文提出的融合用户关系的自适应话题跟踪模型可以有效准确地对热点话题进行跟踪.



(a) 话题一热度变化图 (b) 话题二热度变化图 (c) 话题三热度变化图

(d) 话题一追踪结果准确率统计图 (e) 话题二追踪结果准确率统计图 (f) 话题三追踪结果准确率统计图

图6 话题追踪实验结果统计图

表 2 话题跟踪实验测评结果

	Topic1			Topic2			Topic3		
	本文	文献[5]	文献[6]	本文	文献[5]	文献[6]	本文	文献[5]	文献[6]
R(%)	90.37	87.20	86.39	92.98	89.13	86.69	90.60	85.1	90.0
P(%)	68.51	61.49	64.47	72.42	67.91	65.12	68.09	62.11	62.20
F(%)	77.94	72.12	73.84	81.43	77.09	74.38	77.45	71.74	71.40

本文的话题追踪模型对推文特征空间进行坐标变换,使其近似线性可分,从而使聚类问题简化为二元聚类,为了进一步验证坐标变换方法可以提高推文判断的效率,我们对以上三个热点事件进行话题追踪实验,并分别对坐标变换前后的追踪结果进行评估. 由于 K-means 聚类算法的复杂度为 $O(n\times k\times t)$,其中 n 是所有文档的数目,k 是类别簇的数目,t 是迭代的次数. 我们通过统计聚类过程中的簇的数目、迭代次数以及 F 值来进行验证,具体实现结果如表 3 所示. 可以看出,对推文特征空间进行极坐标变换后,话题追踪模型的准确性和效率性有了明显的提高.

表 3 极坐标变换前后话题跟踪结果对比统计

	Topic1			Topic2			Topic3		
	类数	次数	*F* 值	类数	次数	*F* 值	类数	次数	*F* 值
直坐标	7 个	78. 次	62.31%	6 个	74 次	65.73%	4 个	75 次	64.35%
极坐标	2 个	35 次	77.94%	2 个	37 次	81.43%	2 个	33 次	77.45%

5 结论

本文提出了一种融合用户关系的自适应微博话题跟踪方法. 主要包括:一、借助用户的历史信息协助推文的相关性判断,从实验结果可以看出该方法增加了推文判断的稳定性. 二、对推文特征空间进行坐标变换,使其近似线性可分,从而使聚类问题简化为二元聚类,有效的提高了推文判断的效率与准确性. 三、采用迭代跟踪方式替代逐条分类,不需要样本进行初始训练;利用当前跟踪到的相关推文集合生成新一轮的话题跟踪模型,有效的避免了在话题跟踪过程中出现话题漂移的现象;对推文特征空间进行坐标原点平移和极坐标转换,利用二元聚类有效地划分出相关推文集合,增加了推文判断的准确率. 四、话题跟踪在推文集合上进行,关注话题的整体走向. 实验结果表明,本文的话题跟踪模型可以成功划分出热点事件的相关话题推文集合,并可以提供相关话题的整体走向.

参考文献

[1] Lin J, Kolcz A. Large-scale machine learning at Twitter

[A]. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data[C]. USA: ACM, 2012. 793 – 804.

[2] Petrović S, Osborne M, Lavrenko V. Streaming first story detection with application to Twitter[A]. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics[C]. USA: ACL, 2010. 181 – 189.

[3] Phuvipadawat S, Murata T. Breaking news detection and tracking in Twitter[A]. IEEE/WIC/ACM Web Intelligence and Intelligent Agent Technology[C]. USA: ACM, 2010. 120 – 123.

[4] Duan Y, Wei F, Zhou M, et al. Graph-based collective classification for tweets[A]. Proceedings of the 21st ACM International Conference on Information and Knowledge Management[C]. USA: ACM, 2012. 2323 – 2326.

[5] Nishida K, Hoshide T, Fujimura K. Improving tweet stream classification by detecting changes in word probability [A]. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. USA: ACM, 2012. 971 – 980.

[6] Albakour M, Macdonald C, Ounis I. On sparsity and drift for effective real-time filtering in microblogs[A]. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management[C]. USA: ACM, 2013. 419 – 428.

[7] 崔争艳. 基于语义的微博短信息分类[J]. 现代计算机, 2010, (8): 18 – 20.
CUI Zheng-yan. Short message classification of microblogging based on semantic[J]. Modern Computer, 2010, (8): 18 – 20. (in Chinese)

[8] 路荣, 项亮, 刘明荣, 杨青. 基于隐主题分析和文本聚类的微博客中新闻话题的发现[J]. 模式识别与人工智能, 2012, 25(3): 382 – 387.
Lu R, Xiang L, Liu M R, Liu Q. Discovering news topics from microblogs based on hidden topics analysis and text clustering[J]. Pattern Recognition & Artificial Intelligence, 2012, 25(3): 382 – 387. (in Chinese)

[9] Tang J, Wang X, Gao H, et al. Enriching short text representation in microblog for clustering[J]. Frontiers of Computer Science in China, 2012, 6(1): 88 – 101.

[10] 孙胜平. 中文微博客热点话题检测与跟踪技术研究[D]. 北京: 北京交通大学, 2011.

[11] 洪宇, 张宇, 范基礼, 刘挺, 李生. 基于语义域语言模型的中文话题关联检测[J]. 软件学报, 2008, 19(9): 2265 – 2275.
Hong Y, Zhang Y, Fan JL, Liu T, Li S. Chinese topic link detection based on semantic domain language model[J]. Journal of Software, 2008, 19(9): 2265 – 2275. (in Chinese)

[12] 郝建波. 微博突发话题检测、跟踪与传播预测技术研究[D]. 哈尔滨: 哈尔滨工程大学, 2013.

[13] 刘彦伟. 微博话题追踪系统的研究与实现[D]. 北京: 北京交通大学, 2013.

[14] 王慧. 微博话题追踪方法研究与设计[D]. 北京: 北京交通大学, 2014.

[15] 史存会. 微博客话题追踪及实时检索的相关研究[D]. 大连: 大连理工大学, 2011.

作者简介

柏文言　女, 1990 年 5 月出生于河北省唐山市. 现为北京邮电大学计算机学院硕士研究生, 主要研究方向为社交网络、信息安全.
E-mail: baiwenyan@ bupt. edu. cn

张　闯(通信作者)　男, 1982 年 1 月出生在辽宁省鞍山市. 博士, 现为中国科学院信息工程研究所高级工程师, 硕士生导师. 主要研究方向为云计算、社交网络、信息内容安全等.
E-mail: zhangchuang@ iie. ac. cn

徐克付　男, 1977 年 8 月出生在湖北省随州市. 博士, 现为中国科学院信息工程研究所副研究员, 硕士生导师. 主要研究方向为分布式系统、网络与信息安全、智能信息处理等.
E-mail: xukefu@ iie. ac. cn

张志明　男, 1987 年 12 月出生在四川省广元市. 现就职于北京英孚泰克信息技术股份有限公司, 高级项目经理. 主要业务方向为大数据、软件项目集成、软件安全和逆向工程等.
E-mail: zhangzhiming@ itcchina. com