基于爆发词识别的微博突发事件监测方法研究

陈 国 兰

(南京邮电大学图书馆,南京邮电大学科技情报研究所 南京 210003)

摘 要 鉴于近年突发事件在微博传播的巨大影响力,提出基于爆发词识别的微博突发事件监测方法。把微博突发事件监测分为微博预处理、爆发词抽取、爆发词聚类三个模块。先提出微博噪声数据过滤规则从海量数据中过滤噪声微博,然后采用相对词频、词频增长率、爆发词权重三个计算指标来提取出爆发词特征,最后采用共词分析方法来实现爆发词聚类,从而提取突发事件。最后以新浪微博数据为例进行实验,验证了提出的算法对突发事件监测能取得不错的效果。

关键词 微博 爆发词 共现分析 突发事件

中图分类号 TP391.1 **文献标识码** A **文章编号** 1002-1965(2014)09-0123-06

DOI 10.3969/j.issn.1002-1965.2014.09.022

Micro-blog Emergencies Detection Approach Based on Burst Words Distinguishing

Chen Guolan

(Library of Nanjing University of Posts and Telecommunications, Libraries and Information Institute of Nanjing University of Posts and Telecommunications, Nanjing 210003)

Abstract Giventhehuge influenceof emergencies in micro-blog spread in recent years,this paper presentsa study of the emergency detection on Chinese micro-blog based on burst words distinguishing. In this paper, the burst event detection on micro-blog is divided intothree modules,namely micro-blog pretreatment, burst word extraction and burst word clustering. Firstly, it puts forwardmicro-blog filtering rules to filter micro-blog noise data from massive data, and then usesthree indexes of therelative word frequency, the frequency rate of growth and the weight of words to extract the features for burstwords. Finally it usesthe co-word analysis and "absolute cluster" to achieve the burst word clustering. Experiments on Sina Micro-blog afterwards verifies the effectiveness of the algorithm proposed in this paper on emergency extraction.

Key words micro-blog burst word co-occurrence analysis emergency

0 引 言

自2006年第一个微博平台twitter使用以来,由于其传播速度快、互动性强、信息更新方便等特点使其作为一种新型的媒体得以在国内外迅速推广,近几年更是随着手机、平板等移动终端的普及,越来越多的人可以随时随地的分享信息,进入了全民媒体时代。与传统媒体相比,事故现场的任何微博用户可以通过微博第一时间发布很多重大新闻,这在突发事件中表现更为突出。纵观近两年的国内大事记,微博已经成为众多突发事件的首发平台,因此通过对主流微博的监测可以第一时间捕获大部分的突发事件,及时进行疏导处理,降低突发事件的危害。

对微博事件监测的研究,近几年引起了广大学者的关注。国内外针对微博突发事件的研究主要分为两种:一种是传统的以文本为中心的事件检测[1-3],其思路是先抽取文本主题词,然后对主题词进行文本聚类,再在类中抽取突发特征识别突发事件。由于微博含有很多垃圾信息,先聚类再抽取突发特征,会引入很多噪声信息。同时在文本聚类过程中涉及到很多阀值的选取,阀值选取的好坏也会影响聚类结果。另一种是以突发特征为中心的事件检测[4-7],其思路是先抽取突发特征,然后根据突发特征聚类识别突发事件。由于微博文本简短、用词随意等特征,也使得传统的以文本为中心的事件检测方法力不从心。基于特征的事件检测算法受到越来越多的重视。以特征为中心的方法可

收稿日期:2014-05-09 修回日期:2014-08-07

作者简介:**陈国兰**(1981-),女,硕士研究生,馆员,研究方向:数据挖掘、信息处理。

以有效避免阈值的问题,但是垃圾信息的问题没有很好的解决。

与以上方法不同,本文结合微博突发事件的特征,提出利用爆发词的识别来监测突发事件。爆发词是指那种在一段时间大量出现的有意义的代表话题走向的词[8],因此爆发词的识别需要识别候选爆发特征,同时标记特征出现的时间并跟踪变化、最后统计并得到爆发词,正确识别爆发词对突发事件监测和追踪具有重要意义。

为了处理垃圾信息,本文结合微博突发事件的传播特征,提出了微博噪声数据过滤规则,实验证明取得了良好效果。本文把突发事件提取主要分为三个过程:微博预处理,爆发词提取,爆发词聚类发现事件。微博预处理主要过滤噪音微博,切词分词为爆发词提取做准备;爆发词提取主要是设定时间段获取爆发特征集合,怎么正确获得爆发词及计算爆发权重关系着事件提取准确率。本文提出相对词频、词频增长率、词语权重三个算法公式来提取爆发词。爆发词聚类本文利用共词分析理论来聚类相关事件簇,最后提取事件。

1 微博预处理

微博作为一种新的交流平台,发展迅猛,单单新浪微博用户已超过 2.5 亿,平均每天都有超过 2 500 万条新微博内容发布。要从这些海量数据中挖掘出突发事件,就要在检测前对微博进行合理的预处理,过滤噪声微博,可以提高后续检测环节的效率及准确性。

本文将微博预处理分为三部分,首先结合微博突发事件的特点,对微博进行噪声过滤;然后利用 NLPIR 汉语分词系统提供的用户自建词典的接口,接入用户自建的变体词典,使其分词适应微博口语化、变体词多的特点;最后对分词后的微博文本剔除掉对话题表征贡献度不大的词语,从而提高算法的准确性和效率。

1.1 噪声微博过滤 微博作为个人的生活发布平台,因此包含很多个人的日常生活描述、生活感触以及一些广告信息等,这些信息占有比例大,会对突发事件监测造成很大干扰。本文结合微博突发事件的传播特点,提出以下过滤原则对微博进行过滤:

a. 过滤微博文本字数少于 10 以下的微博。微博作为个人生活记录的平台,充斥着大量对事件提取没有实际意义的心情贴,像“今天好郁闷”之类的信息。据调查,用户个人琐事文本占据微博传播信息的 60% 以上;通常过短的微博内容无法较准确地描述一个事件,本文认为字数少于 10 的微博无法准确地描述一个事件予以过滤。

b. 过滤粉丝数少于一定阈值的用户发布的微博。通常粉丝数量接近于 0 的这些用户是机器平台产生的僵尸用户,为了商业目的定期发送大量重复的广告信息。这些广告重复内容多,会对我们的事件监测造成很大干扰。

c. 过滤以“#话题#”为格式的微博。这类微博通常是微博平台发起的微话题,虽然是热点话题,引起众多人数参与,但大多为对某一热门话题的讨论,引发突发事件的概率较低,且主题词多次出现,会强烈影响基于词频的统计算法。

d. 过滤以“@ 用户名”格式的微博。通常这类微博是用户之间的互动信息,而我们监测的突发事件通常面向公众话题,用户的报道或观点很少会指向另一个特定用户,所以带有@ 格式的消息直接描述公共话题的突发事件的可能性较小。

e. 过滤微博文本中带有 URL 链接的微博。利用文献[9]的实验结果,对随机下载的 1 000 条微博进行人工标注,发现噪声微博覆盖率高达 29.9%。,进一步分析发现这其中 85% 的噪声微博都是带有 URL 链接的广告微博。

f. 过滤标签标注是娱乐明星的原创微博,包括后面的转发、评论微博。娱乐明星由于其极高的人气可以作为意见领袖在突发事件的传播过程中起至关重要的作用,但其原创微博大多心情贴、活动宣传贴等生活记录事宜,虽然能在短时间内引起众多粉丝的转发评论,但其引发突发事件的概率极低。因此过滤此类微博,通常不会影响突发事件结果的检测。

1.2 未登陆词识别 未登录词(Out-of-Vocabulary,OOV) 是指通过某种途径产生的、具有基本词汇所没有的新形式、新意义或新用法的词语,并且是在分词系统中未被分词词典收录的词语,包括各类专名(人名、地名、企业字号和商标号等) 、某些术语、缩略语和新词等[10]。

微博作为大众发表个人观点和分享信息的平台,参与的人素质参差不齐,往往语言表达较为随意,经常会大量使用谐音词或者网络新词热词,并且新词层出不穷,不像媒体新闻类报道语言规范、准确。像“沙发”“板凳”等普通词语在微博里却有着特殊意义,再比如新词“高大上”“何弃疗”等分词工具根本无法识别出这些新词。正确分词是数据特征准确提取的基础。因此为了提高微博文本的分词正确性,本文从网络上搜集了常见的网络变异词语组成网络变异词典,利用 NLPIR 汉语分词系统提供的用户自建词典的接口,再对微博文本进行分词。实验证明,分词效果某种程度上得到改善,一些常见的变异词得到正确分词。当然此种方法未登陆词识别的好坏,基于变体词典构造的优劣。这个变体词典是一个不断完善的过程,后期可以基于机器学习的方法从维基百科或百度热词库

中不断优化添加。

1.3　**文本预处理**　词单元是爆发词抽取的最小语义单元,本文利用 NLPIR 汉语分词系统对微博进行分词。NLPIR 汉语分词系统是张华平博士多年工作研究的成果,其具有良好的分词效果,且支持词性标注,命名实体识别,新词识别等[11]。对一条微博进行分词及词性标注后,可以得到多个"词/词性标记"对。以"来自重庆南开的彭书涵同学,刚刚被比牛津哈佛更牛的深泉学院录取。"为例,经过分词及词性标注后,最后内容如下:"来自/v 重庆/ns 南/f 开/v 的/u 彭书涵/nr 同学/n,/w 刚刚/d 被/p 比/v 牛津/ns 哈佛/ns 更/d 牛/a 的/u 深/a 泉/n 学院/n 录取/v。"

为了提高算法的效率,本文只关注名词、动词、形容词、人名、地名、时间、机构团体名、其他专名这 8 种对事件表征贡献度大的词性,其他词性的词一律过滤。同时考虑对人名、地名、机构团体名对事件的描述性更强,应赋予更高的权重。像上述例子其中重庆、彭书涵、牛津、哈佛四词被赋予更高权重。

其次,对过滤后剩下的词汇统计其在微博集中出现的频率(包含该词的微博个数)。如果该词出现的频率太低,说明该词不太具有代表性,而如果该词频率太高,则说明该词不太具有区分度。因此设定一定的阀值,当该词的频率小于某个阈值或大于某个阈值时,从词汇集中去除该词。

基于词性和频率的双重过滤后,可以有效地剔除掉对话题表征贡献度不大的噪声词汇,从而提高算法的准确性和效率。最后再将剩余的词汇作为后面爆发词提取的处理对象。

2　爆发词识别

爆发词识别作为突发监测方法的基础性工作,正确识别爆发词对突发主题监测和话题追踪具有重要作用[12]。根据爆发词的描述可知,要正确识别爆发词,不仅需要考虑词语的频率,还要结合时间因素考虑其词频变化率。微博空间内存在一类词语,它们的词频增量较高,但却不具有明显的实际意义。例如到就餐时间的时候,微博上关于吃的讨论会剧增,到下班时间的话,关于下班或者交通的词汇也会有一定的起伏变化。然而,这类词语与突发事件并无显著关系,将其作为爆发词用于突发事件检测,往往会对检测结果造成一定的干扰。因此本文从相对词频、词频增长率、爆发权重三个维度来对爆发词进行筛选,使抽取的爆发词更能准确的描述突发事件。

2.1　**相对词频**　微博上同一事件主题词重复率较高,在经过文本预处理后,如果一段时间内,一个词汇明显高于该时间段内出现的其他词汇,在某种程度上说明这个词汇热度比较高,可能是当前关注度较高的热门事件主题词。

在同一个话题讨论时,人们可能会用相同或相近的主题词汇。因此统计词频时涉及到一个相似词汇的统计问题,应该把同义词或近义词统一化算作同一种词。目前较常用的计算词语相似度的方法主要是是基于语义知识词典(如同义词词林、知网中文词库等)的词语语义相似度计算,通过计算词语间的语义相似度,判断两个词是否相似。对于相似度较高的词,在词频统计时我们把它归一为同一个词。相对词频定义如下:

$$A_{ij} = \frac{F_{ij}}{F_{\max}} \tag{1}$$

式中:A_{ij} 是词汇 i 在 j 时间窗口的相对词频,F_{ij} 是词汇 i 包括其相似词在 j 时间窗口的出现的频率,$F_{\max}$ 是当前时间窗口出现的最高词频。

2.2　**词频增长率**　根据突发事件的发展规律,都有个潜伏、发展、爆发、消退的时期,如果一个词汇相对词频较高,但在相邻时间段内数量相当,变化不大,则其描述突发事件的概率较低。突发监测算法是 Kleinberg 在 2002 年提出的,它的基本原理是关注单位时间窗文档流中相对激增的词,当一个词汇在某时间段内频繁出现,且出现的频率要比上一个时间段内明显增加,则在一定程度上意味着与此相关的话题热度在提升,可能是某个潜在的突发事件主题词[13]。词频变化率定义如下:

$$B_{ij} = \frac{F_{ij} - F_{i(j-1)}}{1 + F_{i(j-1)}} \tag{2}$$

式中:表示词汇 i 在 j 时间窗口的增长斜率, 是词汇 i 包括其相似词在(j - 1) 时间窗口的出现频率。

根据相对词频和词频增长率可以识别出爆发新词,那些在时间窗内一出现即达到一定规模的新词,本文即认为其已经成为爆发词,而且还可能继续爆发,需要跟踪和关注。

2.3　**爆发词权重**　信息检索领域常用的权重计算方法是 TF-IDF 算法,词频 (Term Frequency, TF) 指的是某一个给定的词语在该文件中出现的次数。逆向文件频率 (Inverse Document Frequency, IDF) 是一个词语普遍重要性的度量。某一特定词语的 IDF,可以由总文件数目除以包含该词语之文件的数目,再将得到的商取对数得到。TF-IDF 是一种统计方法,用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降[14]。该方法倾向于使出现频率高且区分度大的词拥有较大的权重。然而,突发事件

中,爆发词往往以较高的频次出现在描述同一事件的不同微博中,使用 TF-IDF 算法会使这类词拥有较低的权重,容易被忽略。因此本文采用文献[15]提出的 TF-PDF 算法来计算爆发词的权重。计算方法如下:

$$C_i = \sum_{k=1}^{n} | F_{ik} | \exp\left(\frac{N_{ik}}{N_k}\right) \quad (3)$$

$$| F_{ik} | = \frac{F_{ik}}{\sqrt{\sum_{j=1}^{n} F_{jk}^2}} \quad (4)$$

式中:C_i为词 i 的权重,n 为取得时间窗的数量,F_{ik} 为词 i 在 k 时间窗的词频,N_{ik} 为 k 时间窗中包含词 i 的微博数量,N_k 为 k 时间窗的微博总数。

根据上述定义,文中对单位时间窗口内的词分别计算其 A_{ij}、值 &1、&2、&3。&1 代表词汇的热度,&2 代表词汇的突发度,&3 代表词汇的重要度。如果 $A_{ij} \geqslant \delta 1$、$B_{ij} \geqslant \delta 2$ 且 $C_i \geqslant \delta 3$,就把词汇 i 提取出来,放在爆发词集里。

3 爆发词聚类

爆发词抽取之后,如何采用合适的聚类算法将爆发词进行聚类来识别突发事件是本小节研究的重点。经过聚类,把识别的爆发词聚类到不同的类簇,一个类簇就代表一个相关突发话题。可见聚类效果的好坏会影响突发事件的检测准确率。

当前关于文本聚类的研究较多,传统的聚类算法对处理低维数据的聚类问题比较成功。但是在处理高维和大型数据的时候,通常效果很不理想。LDA 模型是最常见的话题建模模型,利用 LDA 模型对微博数据集进行隐主题建模,进而通过隐主题模型计算文本之间的相似度,某种程度上可以处理微博数据稀疏的特点[3]。它的缺点是计算量较大,速度较慢。

文本聚类方法很多,应用最广泛的是 K-means 聚类[16]算法。它的主要缺点是算法准确度有限,响应时间偏长,同时对用户介入的要求较高。文献[17]针对传统 K-means 算法的局限性,提出了一种基于文本平均相似度的改进 K-means 算法,实验证明,改进后的算法更适合微博文本的特点,微博舆情监测的效果大幅度提高。

与以上方法不同,本文采用基于共词分析的主题相似度距离计算和文献[6]的基于"绝对聚类"的微博文本聚类算法来对爆发词进行聚类。共词分析方法目前已经在人工智能、信息科学、信息管理系统和信息检索等多领域都得到了很好的应用。共词理论是基于这样的一个假设:如果在大规模语料中,两个词经常共现在文档的同一窗口单元(如一句话、一个自然段等),则认为这两个词在意义上是相互关联的,并且共现的概率越高,其相互关联越紧密[18]。在共词分析中,本文设计了一个共词矩阵以方便对爆发词两两共现频率的运算。我们统计微博集中爆发词两两之间共现在同一条微博的频率,便可以构成一个由这些爆发词组成的共词矩阵。如果有 N 个爆发词,经过共词分析便形成了 $N * N$ 的共词矩阵。本文根据生成的共词矩阵来度量爆发词间的距离,如果频率越高,则表示两词之间相似度越大,距离越小,属于同一个簇的概率越大。

文献[6]提出的绝对聚类的思想,是假定一个对象属于一个类,那它应该和这个类簇中的任何对象都相似,即"绝对"属于这个类,否则不属于这个类。本文借鉴这个思想,其算法思想主要如下:

a. 从爆发词集中任意选择两个爆发词,如果它们满足一定条件的相似(根据共词分析的结果),就归为一个类,否则归为两个类。并且将这两个爆发词从爆发词集中删除。

b. 对爆发词集中的任意爆发词,如果存在某个类中的所有文本与其都满足一定的相似条件,则将其归入此类中。否则新建一个类。并将该词从爆发词集中删除。

c. 循环执行步骤 2,直至爆发词集为空结束。

因此本文先用共词分析构造共词矩阵计算爆发词间的相似度,然后借助"绝对聚类"的思想对爆发词进行聚类,主要算法思路如下:

a. 对于共词矩阵以数组的形式进行输入,初始值均为 0。然后对单位时间段内的所有微博逐一进行匹配,如果共词矩阵里行列上的词共同出现在一条微博中,则将该值加 1。这个值代表他们的共现次数,也代表该时间段内两词共同出现的微博条数。

b. 对生成的共词矩阵,进一步处理。本文设定:

$$D_{ij} = \frac{1}{P_{ij} + 1} \quad (5)$$

其中 D_{ij} 表示两个爆发词之间的距离,P_{ij} 为共词矩阵中爆发词两两的共现频率。频率越大,距离越小,加 1 是为了防止频率为 0 的情况。

c. 根据计算的爆发词之间的距离相似度,采用绝对聚类的算法对爆发词进行聚类,形成突发事件话题簇。

4 实验结果与分析

为了验证本文所提出方法的准确性与高效性,本文以目前国内使用人数最多的新浪微博为实验平台,利用新浪微博开放 API 接口获取 2013 年 8 月 6 日至 2013 年 8 月 13 日约 60 万条的微博数据。首先对原始数据进行预处理,借助本文提出的过滤规则对微博文本进行去噪,最后剩余 38 万条微博数据。本文以天为

一个单位时间段,根据爆发词提取算法进行爆发词识别,然后根据共词分析理论进行爆发词聚类,最后根据聚类的爆发词提取最相关且热度最高的一条微博来代表抽取的突发事件,并与该段时间内新浪微博的热门话题榜进行对比。

4.1　**爆发词抽取结果**　在经过微博预处理之后,对剩余的词汇依据本文提出的相对词频、词频增长率、爆发词权重算法提取爆发词。为了较准确地抽取爆发词,需要设置合适的阀值 δ_1、δ_2 和 δ_3。本文采用动态阀值的方法,以适应不同时间段的爆发词提取。δ_1 选取单位时间窗口中相对词频排在第 100 位的词频值,δ_2 选取词频增长率排在第 50 位的值,δ_3 选取词语权重排在第 30 位的值,最后抽取的具有一定热度(排名前 100),突发性较强(排名前 50),同时又最重要(排名前 30) 的 30 个词汇构成的爆发词集如表 1 所示。

表 1　抽取的爆发词及权重值

爆发词	权重	爆发词	权重	爆发词	权重
光大	0.0832	梦鸽	0.1326	斯诺登	0.0872
李天一	0.2105	高温	0.0813	北京	0.0617
酒吧	0.1624	乌龙指	0.0585	杨女士	0.0616
证券	0.0586	棱镜门	0.0823	昌都	0.0739
台风	0.0724	尤特	0.0788	热带风暴	0.0845
律师	0.1530	中暑	0.0726	别墅主人	0.0627
楼顶	0.0528	控告	0.1538	登陆福建	0.0755
云计算	0.0791	地震	0.0826	中暑	0.0713
郑钧	0.0812	刘芸	0.0823	马尔代夫	0.0921
婚礼	0.0733	勒索	0.0628	西藏	0.0763

4.2　**爆发词共词分析**　对提取的爆发词集进行共词分析,构建共词矩阵来提取事件。根据已经提取出来的 30 个爆发词统计其在微博中两两出现,构成 30 * 30 的共词矩阵值如表 2 所示。

表 2　部分共词矩阵值

共词矩阵	光大	李天一	证券	梦鸽	律师	…
光大	0	0	2595	0	75	…
李天一	0	0	0	3796	892	…
证券	2595	0	0	0	52	…
梦鸽	0	3796	0	0	2678	…
律师	75	892	52	2678	0	…
北京	7	28	75	28	68	…
台风	0	0	0	0	0	…
…	…	…	…	…	…	…

共词矩阵里的值代表爆发词两两共现的频率,频率大小代表了两词共同出现的微博条数,也说明两词之间的紧密程度,利用公式(5)计算爆发词两两之间的相似度,然后依据“绝对聚类”的算法思想对爆发词进行聚类,最后构建的爆发词事件簇如表 3 所示。

表 3　爆发词事件簇

编号	绝对聚类后构成的事件簇
1	光大、证券、乌龙指、股市
2	李天一、梦鸽、酒吧、控告、律师、强奸、领班、受害人
3	台风、尤特、登陆 、热带风暴
4	斯诺登、棱镜门、美国
5	北京、别墅主人、楼顶
6	西藏、昌都、地震
7	郑钧、刘芸、马尔代夫、婚礼
8	全国、高温、中暑

根据已经抽取的事件簇,本文选取与事件簇的爆发词最相关且热度最大的一条微博作为当前突发事件微博的描述。微博的热度代表一个事件被关注的程度,在微博平台中主要靠微博的转发和评论使得一个事件快速传播。因此,用微博的转发数和评论数可以大致衡量一条微博的热度。计算方法如下:

微博热度 $=\lambda_1 \times$(转发数) $+\lambda_2 \times$(评论数),其中 $\lambda_1 + \lambda_2 = 1$

考虑到微博中转发对突发事件的传播影响更大一些,因此设置参数 $\lambda_1 = 0.6$,$\lambda_2 = 0.4$。

本文最后抽取的与事件簇里的爆发词最匹配且热度最大的 8 条微博如表 4 所示。

表 4　抽取的突发事件微博

编号	突发事件代表微博
1	日前,曝出李天一母亲梦鸽控告酒吧涉嫌介绍卖淫(pimping)和敲诈(extortion)。涉事酒吧回应,称梦鸽为救儿子丧心病狂…
2	【高温天气出游如何防中暑】今天全国大部地区高温橙色预警中,出游防暑防晒小窍门…
3	【今年第 11 号热带风暴尤特诞生在即日 或登陆福建】据日本气象厅,96W 可能在未来 2 天内加强为今年第 11 号热带风暴…
4	【斯诺登“坑”了云计算】美国独立智库近期研究显示,受棱镜门事件影响,美国的云计算服务出口潜力将受制约…
5	【北京一教授楼顶盖别墅假山大树俱全 称名人常来不怕被告】初看,这像是居民楼顶盖的假山花园,但住户称,其实是顶层住户张教授私自盖的别墅…
6	【光大证券发布公告承认乌龙指:成交 72 亿 涉 150 只股】传光大的自己核查报告已经交到会里,下单 230 亿,成交 72 亿,涉及 150 多只股票…
7	西藏昌都地震重灾区房屋倒塌严重 – 西藏昌都地区 8 月 12 日发生 6.1 级地震后,武警交通二支队全力投入抗震救灾…
8	郑钧刘芸马尔代夫补办婚礼,3 岁儿子捧钻戒,婚礼以“你,必须幸福”为主题,郑钧深情弹唱《爱的箴言》…

为了验证本文抽取的突发事件的效果,本文提取该时间段内新浪微博里的热门排行前十话题(见表 5)。

把本文实验抽取的突发事件微博与表 5 的热门话题榜进行比较,发现利用本文提出的突发事件检测方法能较好的提取新浪微博的热门事件。由于事件 4、6 和 10 是新浪微博平台发起的微话题讨论,本文实验已经当做噪声微博进行过滤,事件 9 由于首发贴是娱乐

明星,因此在微博过滤当中也已经过滤。虽然本文在微博预处理的时候把某些热门话题过滤了,但是不影响我们对突发事件的提取。

表5 新浪热门话题前10排行榜

编号	热门话题榜	编号	热门话题榜
1	#梦鸽控告酒吧#	6	#七夕,一起看英仙座流星雨#
2	#全国各地持续高温#	7	#北京楼顶盖别墅#
3	#光大乌龙指#	8	#台风尤特登陆#
4	#中国好声音#	9	#泰国白龙王病逝#
5	#斯诺登棱镜门#	10	#中国好外婆#

5 结 语

本文结合微博突发事件传播的特性,提出基于突发事件的微博过滤原则,对原始微博进行过滤,然后借助中科院的分词工具进行分词,为了使得分词效果更好,本文构建了变体词典用于识别未登录词,之后提出采用相对词频、词频增长率、词语权重三个算法公式来提取爆发词,最后对抽取的爆发词利用共词分析理论来建立话题簇抽取事件。实验证明,本文提出的方法能比较准确的检测微博的突发事件。如果需要检测更多的突发事件,可以通过改变阀值,增加爆发词集中爆发词的个数。本文的监测方法还有一些不足,也是后续研究的重点:

a. 爆发词权重算法的改进。使得抽取的爆发词可以更准确的描述突发事件特征,可以根据权重确定突发事件的核心词汇。

b. 更准确的突发事件的描述。能够根据聚类的爆发词比较自动、准确的进行事件描述。

c. 加入情感分析。结合突发事件的情感特征,忽略一般的热门事件,只抽取可能造成恶劣影响需要事件相关部门参与引导与介入的事件。

参 考 文 献

[1] Diao Q M,Jiang J,Zhu F D. Finding Bursty Topics from Microblogs[C]. In: Proceedings of ACL,2012:536-544.

[2] 陈莉萍,杜军平. 突发事件热点话题识别系统及关键问题研究[J]. 计算机工程与应用,2011,47(32):19-22.

[3] 路 荣,项 亮,刘明荣,等. 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.

[4] 郭跑秀,吕学强, 李卓基. 基于突发词聚类的微博突发事件检测方法[J]. 计算机应用,2014,34(2):486.

[5] 闫光辉,赵红运,任亚缙,等. 基于时间特性的微博热门话题检测算法研究[J]. 计算机应用研究,2013.

[6] 王 勇,肖诗斌, 郭跑秀. 中文微博突发事件检测研究[J]. 现代图书情报技术,2013(2):57-62.

[7] Du Y Y,Wu W,He Y X, et al. Microblog Bursty Feature Detection Based on Dynamics Model[C]. In: Proceedings of the International Conference on SystemsandInformatics (ICSAI), 2012:2304-2308.

[8] 逯万辉,马建霞,赵迎光. 爆发词识别与主题探测技术研究综述[J]. 情报理论与实践,2012,35(6):125-128.

[9] 王 琳,冯 时,徐伟丽. 一种面向微博客文本流的噪音判别与内容相似性双重检测的过滤方法[J]. 计算机应用与软件,2012,29(8):25-29.

[10] 魏莎莎. 一种中文未登录词识别及词典设计新方法[D]. 重庆: 西南大学,2011.

[11] http://ictclas.nlpir.org/.

[12] 逯万辉,马建霞. 基于CRFs的领域爆发词识别的研究与实现[J]. 情报科学,2014,32(1):89-93.

[13] Kleinberg J. Bursty and Hierarchical Structure in Streams[J]. Data Mining and Knowledge Discovery,2003,7(4):373-397.

[14] http://baike.baidu.com.

[15] Bun K K, Ishizuka M. Topic Extraction from News Archive UsingTF * PDF Algorithm[C]. In: Proceedings of the 3rd International Conference on Web Information Systems Engineering, 2002:73-82.

[16] Diao Q M,Jiang J,Zhu F D. Finding Bursty Topics from Microblogs[C]. In: Proceedings of ACL,2012:536-544.

[17] 朱晓峰,陈楚楚,尹婵娟. 基于微博舆情监测的K-means算法改进研究[J]. 情报理论与实践,2014,37(1):136-139.

[18] 王小华,徐 宁,谌志群. 基于共词分析的文本主题词聚类与主题发现[J]. 情报科学,2011,29(11):1621-1624.

(责编:刘武英)