基于速度增长的微博热点话题发现*

薛素芝[1,2]，鲁 燃[1,2]，任圆圆[1,2]
(1. 山东师范大学 信息科学与工程学院，济南 250014；2. 山东省分布式计算机软件新技术重点实验室，济南 250014)

摘 要： 在微博热点话题发现中，由于微博文本短、词量少、用词不规范等特征，使得传统的热点话题检测方法力不从心。针对这一问题，提出了基于速度增长的微博热点话题发现方法。首先把经过预处理的微博按等数量窗口划分，统计每个窗口内各词语的词频，并表示成时间二元组序列；然后通过计算每相邻两个窗口的个词语的增长斜率来发现增长速度快的词语；再通过计算与该词语有关的用户的增长速度和微博条数的增长速度来确定该词语是否是热点主题词；最后通过热点主题词聚类产生热点话题。通过实验验证了该方法的可行性。实验结果表明，该方法在一定程度上提高了检测效率，降低了漏检率和误检率，可以有效地及时发现微博热点话题。

关键词： 增长斜率；增长速度；时间二元组序列；热点发现
中图分类号： TP391.3　　**文献标志码：** A　　**文章编号：** 1001-3695(2013)09-2598-04
doi:10.3969/j.issn.1001-3695.2013.09.009

Hot topics found on micro-blog based on speed growth

XUE Su-zhi[1,2], LU Ran[1,2], REN Yuan-yuan[1,2]
(1. *School of Information Science & Engineering, Shandong Normal University, Jinan* 250014, *China*; 2. *Shandong Provincial Key Laboratory for Normal Distributed Computer Software Technology, Jinan* 250014, *China*)

Abstract: In hot topics found on micro-blog, because the text of micro-blog is short and less words, and the terms are not standard, so the traditional hot topic detection method can not find hot topics effectively. In order to solve this problem, this paper presented a method of hot topics found based on speed growth. Firstly, it divided the pretreated micro-blogs on the basis of the equal number of window, and added up the term frequency in each window, and expressed as feature trajectory of binary group sequence. Secondly, it calculated the growth slope of every adjacent two windows to find the words with growth speed. Thirdly, it calculated the growth speed of the word's relevant users and the growth speed of the word's relevant micro-blogs to ensure the word was hot subject or not. Finally, it found hot topics through the hot subject clustering. The experimental proves the feasibility of the algorithm, results show that the method improves the efficiency of the detection to a certain extent, and reduces the undetected rate and false detection rate, it can effectively discover hot topics on micro-blog in time.

Key words: growth slope; growth speed; feature trajectory of binary group sequence; hot topics found

0 引言

近年来微博服务蓬勃发展，再加上 Web 2.0 的使用，正改变着互联网用户的传统使用模式。据统计，自 2006 年 Twitter 发布使用，到目前为止信息总量已经超过 200 亿条，每天有 800 万条新微博信息发布[1]。由于微博不同于传统的新闻媒体，它的互动性强，用户不仅是被动的信息接收者，更是信息的发起者，这激发了用户参与互动的兴趣，使得信息在网络上得以迅速、及时地传播，一些突发事件往往在这时会表现出来。但由于微博信息表现形式复杂多样，包括文本、音乐、图片、视频等多种表现形式，加之微博信息量大，使得从微博上获取突发话题成为一个艰难的任务[2]。因此，如何充分利用微博的灵敏、即时性等优点，及时从海量的数据中获取热点话题显得越来越重要。

微博为广大用户提出了一个畅所欲言的开放服务平台，也正是由于它的开放性加之互动性，每天有大量的信息涌现在微博上，而且不同于传统的新闻形式，微博具有文本短、信息量少、用词不规范、不严谨等特点，使得利用传统的聚类或分类方法提取热点变得尤为困难。传统的话题检测任务一般集中在新闻语料上，因为新闻语料具有格式严谨、用词规范、文本长短适中等优点，利用聚类方法得到热点话题相对容易，但新闻具有一定的滞后性，如何从含有巨大噪声内容的微博中挖掘热点是微博研究者面临的共同问题。

本文提出一种基于速度增长的微博热点话题发现方法。该方法首先利用微博提供的微博标签如“@”“#”等格式进行微博筛选，过滤掉僵尸账户信息以及对话互动信息。微博文本虽然具有文本短、用词随意不规范、信息量大等缺点，但它有突发性、裂变性、即时性等优点，可以较传统的新闻媒体作出反应。因此，既要利用微博的优点，又要克服掉微博的缺点，对微博作了上述过滤的处理。处理采用如下方法：a）本文认为粉

收稿日期： 2013-01-03；**修回日期：** 2013-02-18　　**基金项目：** 国家自然科学基金资助项目(60873247)；山东省自然科学基金资助项目(ZR2009GZ007,ZR2011FM030)；国家社科基金资助项目(12BXW040)；公安部科技创新计划资助项目(2011YYCXSDST057)

作者简介： 薛素芝(1986-)，女，山东嘉祥人，硕士研究生，主要研究方向为话题检测与追踪、网络信息安全(lcuxuesuzhi@163.com)；鲁燃(1972-)，男，山东郓城人，副教授，博士，主要研究方向为网络信息安全、网络模型、计算机网络；任圆圆(1989-)，女，山东鱼台人，硕士研究生，主要研究方向为网络信息安全、信息过滤.

丝数小于一定阈值的用户所发的信息、“@”格式信息、“#”格式信息对词频统计影响大但作用小,过滤掉这些格式的信息[3],对处理后的微博不作传统的聚类;b)采用基于速度增长的热点主题词检测方法。本文不仅对某一词语计算其增长速度,还要对与该词语有关的用户数量计算其增长速度,综合两方面的速度来确定一个热点主题词。本文的主要内容有如下两个方面:

a)笔者仍认为时间信息在话题发现中有重大作用,时间间隔越短的两条微博相似性越大。本文对每一个词语用一个二元组的时间序列表示,然后在这个序列上计算每相邻两个窗口的增长斜率,利用斜率计算该词语的增长趋势,从而发现热点主题词。此方法能够根据增长趋势曲线及早预测并提取热点主题词。

b)对于热点主题词聚类抽取热点话题,本文从两方面进行词语的相似度测量,从基于词林的相似性和微博中两词语共现的条件概率两个方面来衡量词语之间的相似度,通过词语聚类产生主题词簇,进而推断出热点话题。该方法的有效性已经得到了验证。

1　相关研究

自从话题检测与追踪这个任务进入人们的视野以来,众多研究学者纷纷加入研究行列。但大多学者的研究都是针对新闻语料,因为新闻内容格式严谨、用词规范、文本长度足够,研究起来简单易行[4]。近年来随着微博的快速崛起,许多突发事件首先表现在微博上,而新闻却有一定的滞后性,于是学者们纷纷把研究重点转向微博。由于微博的文本短、用词随意等特点,传统的话题检测方法已力不从心。但微博有一个独特的特性——即时性,在同一时间段内相关微博的数量往往大于相关新闻的数量,并且由于微博的互动性,通过“关注”与“粉丝”两个功能可以对用户进行分类。目前为止在解决微博上的话题抽取任务主要从以下两个方面进行:一是从用户的角度,二是从微博内容的角度。首先,从用户角度的相关研究有文献[5]利用微博用户的关系,形成一个用户网或用户树,然后利用用户树进行热点检测、热点判断,进而发现热点话题。文献[2]对微博用户引入活跃指数进行用户活跃度排名。他们都认为在这种互动性的微博平台上,用户参与度比较大,对热点的传播起决定性作用,所以从用户的角度进行热点检测。其次,从微博内容方面相关的研究有同济大学的郑斐然利用热点主题词检测、热点主题词聚类等方法进行热点话题发现。目前对内容方面的分析大都采用聚类的方法,但干扰噪声大、聚类效果不好[6]。

虽然微博作为一个新生事物,国内外对其研究尚不成熟,尤其是国内对微博的研究尚处于起步阶段,但是微博作为 Web 2.0 时代的宠儿,其蓬勃发展的态势已引起了国内外众多研究学者和著名国际会议的注意。

2011、2012 年的 ACL 会议以及 2012 年第八届全国信息检索学术会议都涉及到社会媒体(论坛、博客、微博)的相关搜索与挖掘。

Mario 等人[7]提出了基于时序和社会关系评价的 Twitter 热点话题发现方法。在一段时间内,如果一个话题多次被检测到,而在之前很少被检测到,就认为此话题有可能成为热点话题。

Swit 等人[8]提出了一种对 Twitter 中爆炸性新闻进行检测的方法,利用采集、分组和排序等方法进行检测。

北京交通大学的孙胜平[9]提出基于空间向量模型的 SP&HA 算法用于热点话题检测。

华南理工大学的张邵捷[10]利用垂直搜索技术、文本分析和挖掘技术进行热点话题发现。

现在有的学者从用户角度和内容角度两个方面进行研究。本文采取基于时间序列的热点主题词发现,从用户角度和内容角度综合判断词语的增长速度,通过主题词的相似度测量聚类成一个个簇,进而发现热点话题。

2　基于时间序列的热点主题词检测

2.1　时间序列

在这里要得出时间序列,就要对微博进行预处理。预处理包括微博筛选和微博排序两个部分。由于所采集到的微博格式复杂多样,并且干扰噪声大,严重影响词频统计,所以先要对微博进行筛选。本文进行如下三种筛选:

a)去除“@用户名”格式的信息。这种格式的信息类似于人们平常的对话互动聊天,由于词量少,用词随意、不规范,上下文关系不明显等特点,并且它直接描述事件的可能性比较小,会严重影响聚类效果,所以去除这种格式的信息。

b)去除“#话题名”格式的信息。这种话题通常由微博平台发布,如“#国家公务员考试”,由于其人为因素较大,导致参与人数众多,会影响用户数量的统计。

c)去除粉丝数小于阈值 T 的信息。如今微博广告营销成为商家的一种营销模式,他们所发的信息量少,而且大多是与他们利益相关的广告信息,这种账户关注数量多,而粉丝数量少,所以删除这些广告账户所发的信息。

对微博筛选后就要对其排序,本文采用如下的排序方式:

a)提取微博的时间信息,并且按时间顺序排序。

b)按一定的数量间隔把微博分为若干组,每一组称为一个窗口(如每 300 条为一组)。

本文对每一个词形成一个时间序列,如对词 w 表示成

$$C_w = (C_w(1), C_w(2), C_w(3), \cdots, C_w(k))$$

其中 $C_w(i) = TF_w(i)$,即词语 w 在第 i 个窗口内的频率。

前面在微博排序时已提取了微博的时间信息,统计每个窗口的最早时间与最晚时间,并计算两者之差的一半,记为 td,因此每一词语的时间序列可用一个二元组序列表示为

$$V_W = [(C_w(1), td_1), (C_w(2), td_2), (C_w(3), td_3), \cdots, (C_w(k), td_k)]$$

即每一个窗口内词语 w 的词频和该窗口的中间时刻组成一个二元组。

2.2　斜率计算

当前有很多方法可以判断一个一元的时间序列是否含有潜在的热点词[11],其中典型的是 Kleinberg 提出的自动机模型,并且该模型得到了多方面的应用验证[12]。虽然这些典型的方法取得了良好的效果,但是实施起来较为困难,特别是在微博这种时间敏感性较强的平台上。微博热点的话题抽取应充分利用微博反应迅速、即时性强等优点,然而这一方法并没有在这一方面显示出很好的优越性。为了及时掌握事件的发展趋势,本文利用斜率对二元组的时间序列进行分析,及早发现疑似热点主题词的演变趋势,从而检测出热点话题。

本文采用的时间序列划分方法在前面已经提到,就是等数量窗口划分,这时数量窗口的确定是一个大难题,数量窗口如果选择不合适,将会平缓或加剧词语的演变趋势。因此,要采用合理大小的窗口,并且此数量窗口能够及早发现潜在的热点词语,同时又满足了即时性的要求。

给定一个时间序列:

$$V_W=[(C_w(1),td_1),(C_w(2),td_2),(C_w(3),td_3),\cdots,(C_w(k),td_k)]$$

利用斜率计算词语 w 的增长趋势:

$$SLO_w=\frac{C_w(i)-C_w(i-1)}{td_i-td_{i-1}}$$

其中 td_i 的单位为小时。

即计算某个词语的每相邻两个窗口的斜率,如果斜率越来越大或斜率大体上呈现增长的趋势,就认为该词语很可能成为热点主题词。这时只是预测了该词语的增长趋势,它将如何发展还要取决于传播该词语的用户数量。因此,计算该词语的增长速度,包括该词语的增长速度和该词语有关的不同用户数量的增长速度,分别记为 GR_w 和 GR_{user}。

$$GR_w=\frac{C_w(i)}{\sum_{i=1}^{l-1}C_w(i)/(i-1)}$$

$$GR_{\text{user}}=\frac{C_{w\ \text{user}}(i)}{\sum_{i=1}^{l-1}C_{w\ \text{user}}(i)/(i-1)}$$

即当前窗口的频率值除以之前窗口频率的平均值,其中 $C_{w\ \text{user}}(i)$ 是在窗口 i 内与词语 w 有关的用户数量,即出现词语 w 的不同用户数量。

这时为了兼顾词语自身的增长和传播词语的不同用户的增长,该词语的增长速度 GR 描述如下:

$$GR=a\times GR_w+b\times GR_{\text{user}}$$

其中:a、b 为两个参数,在实验中均被设置为0.5。

3 话题抽取

本文所做工作的一切都是为了抽取热点话题,前面的工作也是为话题抽取所做的准备。所以话题抽取包括以下五个过程:微博预处理、词语表示成二元组的时间序列、增长速度计算、提取热点主题词、主题词聚类发现热点话题。微博预处理包括删除僵尸用户新信息、删除“#话题名”格式的信息、删除“@用户名格式的信息”;然后对筛选后的微博进行切词,由于动词和名词(包括命名实体词)具有较高的区分能力,保存每一条微博切词后的动词、名词、用户名和时间信息,并利用一个二元组序列表示每一个词语;随后通过计算斜率和增长速度来提取热点主题词;最后通过热点主题词聚类来抽取热点话题。

3.1 主题词抽取

通过第2章的计算可以发现有些词语急速增长,有些词语增长缓慢,甚至增长斜率趋近于零,本文把各个词语的增长速度进行排序,选取增长速度大于阈值 T 的作为主题词。

3.2 主题词聚类

热点主题词抽取以后,如何将这些主题词进行聚类来得到热点话题是本节的重点。本文从基于词林的相似度和词语出现的上下文两个方面来衡量词语之间的相似度。之所以采用这两种相似度计算相结合的方法,因为词语的相似度不仅取决于词语意义上的相似度,还取决于在实际上下文中的相似性。在同义词林中,基于这样一个假设:词语的相似度取值符合词义相近的规律,一般情况下,默认为词义越相近词语的相似度越大,而词义越远就认为两词语无关。很多实际情况下,这种想法并不符合实际情况,比如“今年粮食产量普遍提高,其中小麦产量增产最大,比去年增长百分之十”这个话题,如果只用基于词林的相似度计算的话,就会有{粮食、小麦}这样的话题,而不是{粮食、小麦、产量、提高}这样的理想话题。所以本文还采用了基于上下文的相似度计算。

本文中还要比较两词的上下文通过计算条件概率来衡量两词的相似性。本文采用如下的方法计算相似度,对于主题词 t_1、t_2,统计每个窗口内同时出现 t_1、t_2 的条件概率,即

$$P(t_1/t_2)=\frac{P(t_1,t_2)}{P(t_2)}=\frac{C(t_1,t_2)}{C(t_2)}$$

即 t_1、t_2 同时出现的微博数除以 t_2 出现的微博数,则

$$\text{sim}(t_1,t_2)=\max P(t_1/t_2)$$

即 t_1 与 t_2 的相似度值为所有窗口内的最大值。

对于两种相似度计算方法,前者只考虑了词语固有的相似性,默认为两个词语词义越接近相似性越大,属于同一话题的可能性越大,它忽略了两个词出现的上下文语境,后者正弥补了这一缺陷。因此,本文采用基于词林的相似度和基于上下文的相似度进行主题词聚类。具体的聚类方法如下:

a)采用基于词林的相似度计算方法。对于主题词 t_1、t_2,查找基于词林的相似度,如果两词语基于词林的相似度判定为两词语相似,则继续步骤b)。

b)对于主题词 t_1、t_2,如果 $\text{sim}(t_1,t_2)$ 大于阈值 T_1,这时就认为 t_1、t_2 属于同一簇。

主题词聚类完成以后,可以得到若干个簇,每个簇可能包含一个或多个主题词,并且这些主题词构成一个热点话题。例如{中国、钓鱼岛、日本}这样的簇,表示一个描述“中日双方争夺钓鱼岛”的话题。

4 实验

为了验证本文所提出方法的准确性与高效性,笔者利用新浪微博的相关数据做了实验,该实验可以输出热点主题词和热点主题词簇。

4.1 实验数据

实验数据采用新浪微博十天的数据。新浪微博是目前国内使用人数最多的微博,由于其巨大的用户参与量,使得突发事件在网络上迅速传播,对突发事件反应非常灵敏。本文的实验数据是新浪微博上2012年5月15日~2012年5月24日之间的200万条微博数据,人工标注了该段时间内的主要热点话题,包括“朝鲜扣押中国渔船事件”“中菲南海对峙事件”和“大庆老人抗强拆事件”“公务员逼小吃店老板下跪事件”等几个事件。

4.2 实验结果说明

在话题检测与追踪中,一般使用漏检率和误检率来评价话题检测的效率,它是美国国家标准与技术研究院(NIST)发布的评测指南,漏检就是系统没有检测出新话题,误检则是系统将原来话题的后续相关报道错误地判断为新话题。一般情况下,漏检率和误检率用如下式表示,假设漏检率和误检率分别用 P_{miss} 和 P_{FA} 表示。

$$P_{\text{miss}}=\frac{c}{a+c},P_{\text{FA}}=\frac{b}{b+d}$$

其中:a 为检测到的相关微博文档数;b 为检测到的不相关微博文档数;c 为未检测到的相关微博文档数;d 为未检测到的不相关微博文档数。

为了评估本文所设置的参数对实验结果的影响,把实验数据分成 8 组进行热点检测,对于参数的确定,本文采用这 8 组数据所计算得出的平均值。为了确定本文所提出算法的有效性,把采用本文方法与未采用本文方法所得出的话题准确率进行比较。比较结果如图 4 所示。在这里,准确率如下定义:$P_{\text{accuracy}} = 0.5 \times P_{\text{miss}} + 0.5 \times P_{\text{FA}}$。考虑到词语的增长速度与相关用户的增长速度息息相关,取 $a = b = 0.5$,分别比较等数量窗口的数量和阈值 T、T_1 对实验结果的影响。例如图 1 所示的是等数量窗口为 2.5 万,一个窗口内的词频分布。从图中可以看出,只有少数词出现多次,很多词语只出现少数。

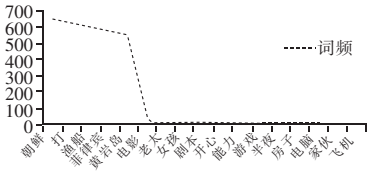

图 1　数量窗口为 2.5 万时的词频分布

从图 1 可以看出,只有少数词语频繁出现,而绝大多数词语只出现很少的次数。这是因为微博是一个畅所欲言的大平台,用户很少能聚集在一个话题下,除非是一个非常热点的话题,在大量的词语下直接进行聚类来寻找热点话题就非常困难。为了以后便于词语聚类,首先划定特定的等数量窗口,利用阈值 T 对词语进行筛选。图 2 和 3 所示的是不同等数量窗口和阈值 T 的漏检率和误检率(相对于人工标注)。

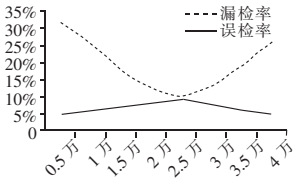

图 2　不同等数量窗口对漏检率和误检率的影响

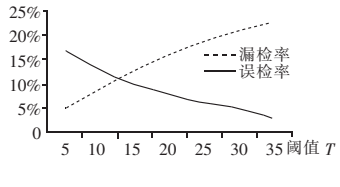

图 3　不同阈值 T 对漏检率和误检率的影响

由图 2 可以看出,等数量窗口较小时,噪声内容比例大,误检率比较高,由于窗口较小,大部分词语分散存在,把两篇相关报道误认为不同话题的可能性较大,这时检测出的话题数较多,所以漏检率相对低;等数量窗口较大时,由于大部分词语聚集在一个窗口内,所以漏检率较高,但误检率相对较低,因为窗口过宽,大部分报道包含在一个窗口内,把它们误认为同一个话题的可能性较大。由图 3 可以看出,当阈值 $T = 5$ 时,漏检率最低,但也会包含很多噪声数据。随着 T 值增大,漏检率会相应地升高,但误检率会降低。由此可见,阈值 T 对主题词抽取的影响甚大。综合图 2、3,选取等数量窗口为 2.5 万左右,阈值 T 为 15 左右抽取热点主题词。图 4 所示的是未使用本文方法和使用本文方法进行热点主题词抽取前后话题准确率的比较。

由图 4 可以看出,先使用本文方法抽取主题词再进行聚类,可以删除掉过多的冗余词,相对于对大量词语直接进行聚类有较好的效果。其中“未使用本文方法”使用的是传统的基于词频的统计方法,对文本进行切词后进行词频统计,词频高的认为是主题词。

对于主题词聚类,本文考察了 T_1 对聚类效果的影响(相对于人工标注)。图 5 表示了在不同值下的漏检率和误检率。

由图 4、5 可以看出,本文选取 T_1 为 0.6 左右为最佳值,所聚出的话题如表 1 所示。

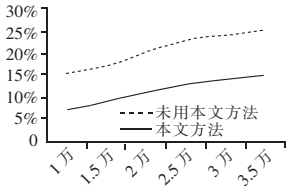

图 4　是否使用本文方法对话题抽取准确率的比较

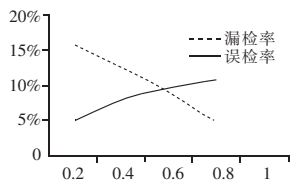

图 5　T_1 对聚类效果的影响

表 1　聚类得出的热点话题

主题词	对应的话题
中国/朝鲜/打/渔船/金	朝鲜扣押中国渔船
拆/死/大庆/老人/司机	大庆抗强拆事件
中国/黄岩岛/菲律宾/打/美国	中菲南海对峙
公务员/苍蝇/道歉/下跪/老板	公务员逼小吃店老板下跪
女孩/白血病/折磨/加油	血癌女孩鲁若晴

5　结束语

本文提出一种基于速度增长的微博热点话题发现方法。本文考虑到微博的即时性、对突发事件反应灵敏等特点,热点话题在微博平台上传播速度非常快。首先,对采集到的微博按时间顺序进行排序并划分成若干个窗口;然后针对每一个窗口计算每个词语的增长速度,提取增长速度大于阈值 T 的那些词语作为热点主题词,最后通过主题词聚类完成话题抽取。

本文对从新浪微博上采集到的微博做了大量实验,通过漏检率和误检率的比较,本文提出的方法有一定的效果,但是漏检率和误检率还有很大的改进空间。因此,笔者将进一步提高算法的精度,精简算法过程,使算法有更好的研究前途。

参考文献:

[1] GIGA T. Counting the number of Tweets[EB/OL]. (2010-08). http://popacular. com/gigatweet.

[2] 石磊,张聪,卫琳. 引入活跃指数的微博用户排名机制[J]. 小型微型计算机系统,2012,33(5):110-114.

[3] 郑斐然,苗夺谦,张志飞,等. 一种中文微博新闻话题检测方法[J]. 计算机科学,2012,39(1):138-141.

[4] ALLAN J,WADE C,BOLIVAR A. Retrieval and novelty detection at the sentence level[C]//Proc of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2003:314-321.

[5] 杨冠超. 微博客热点话题发现策略研究[D]. 杭州:浙江大学,2011.

[6] GIRIDHAR K,JAMES A. Text classification and named entities for new event detection[C]//Proc of SIGIR. 2004:297-304.

[7] MARIO C,LUIQI D C,CLAUDIO S. Emerging topic detection on twitter based on temporal and social terms evaluation[C]//MDMKDD 10 Proc of the 10th International Workshop on Multimedia Data Mining. 2010:1-10.

[8] SWIT P,TSUYOSHI M. Breaking news detection and tracking in twitter[C]//Proc of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2010:120-123.

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

[10] 张邵捷. 基于微博社交网络的舆情分析模型及实现[D]. 广州:华南理工大学,2012.

[11] QI He,CHANG Kui-yu,LIM E P. Analyzing feature Trajectories for event detection[C]//Proc of the 30th Annual International ACM SIGIR Conference. 2007:207-214.

[12] NISH P,NEEL S. Scalable and near real-time burst detection from ecommerce queries[C]//Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008:972-980.