第 42 卷 第 6A 期
2015 年 6 月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

基于移动群智数据的城市热点事件感知方法

张佳凡 郭 斌 路新江 於志文 周兴社

(西北工业大学计算机学院 西安 710129)

摘 要 以新浪微博为研究对象,研究了基于移动群智数据的城市热点事件感知方法,对热点事件进行发现与分类。面向不同的应用需求,可将发现的热点事件分为物理事件与虚拟事件两大类。采用的方法首先根据热词的词频变化特征对新浪微博中的热词进行有效挖掘,然后根据热词的上下文语境进行层次聚类以得到热点事件描述。此外,通过分析信息量特征、时序特征及原创微博数目特征,采用不同方法进行事件分类。实验结果表明,不同的分类方法均可达到较高的准确率。

关键词 微博,热点事件发现,微博事件分类,移动群智感知

中图法分类号 TP391 文献标识码 A

Approach for Urban Popular Event Detection Using Mobile Crowdsourced Data

ZHANG Jia-fan GUO Bin LU Xin-jiang YU Zhi-wen ZHOU Xing-she

(School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract This paper proposed an urban popular event detection and classification approach using the crowdsourced data from Sina Weibo. The detected events can be categorized into physical events or virtual events, which can be used for different applications. Our approach firstly extracts the hot words from crowd posts according to the characteristic of word frequency. With the context of hot words, hierarchical clustering is then used to obtain the description of popular events. By analyzing the three proposed features, including lexical entropy, temporal dynamics, and content originality, we applied various methods to do event classification. The experiment results indicate that all different classification methods can achieve a higher precision under our approach.

Keywords Microblogging, Popular event detection, Microblogging event classification, Mobile crowd sensing

1 引言

移动群智感知[1,2]是指用大量普通用户使用的移动设备作为基本感知单元,通过物联网/移动互联网协作,实现感知任务分发与数据收集利用,最终完成大规模、复杂的社会与城市感知任务。换句话说,就是通过群智贡献的数据进行大规模感知。移动群智感知的数据包括两种产生方式:移动感知和移动社交网络数据[2]。作为一种新的移动社交网络形式,微博在过去的几年中获得了快速的发展,已成为人们交流、分享信息的主要网络平台之一。大量用户通过微博发布的信息十分丰富,这为人类社会的热点事件发现带来了新的思路。即可以将大量微博用户视为群体感知源,利用群智的思想通过众多微博用户发布的内容来进行热点事件的发现和感知。

一般来说,通过微博数据发现的热点事件可分为虚拟事件和物理事件两大类。虚拟事件包括一段视频、一部电影及对社会现象的讨论等;物理事件包括体育盛事、自然灾害、基于位置的事件等。不同种类的事件往往具有不同的应用范围,例如,物理事件对于建设智能城市、公共安全预警、进行地点及活动推荐等具有重要意义,而对虚拟事件进行分析则可及时了解网络舆情和进行话题推荐。因此,对热点事件进行分类,则可对事件进行细粒度分析,区分物理事件与虚拟事件,以满足不同用户及应用的需求。本文的主要研究内容就是通过微博提供的丰富的海量数据及时有效地发现热点事件,并且对事件的类型(虚拟事件、物理事件)进行分类。

目前,以微博作为平台进行事件发现的研究方法基本都基于文本分析[3-8],而微博的文本内容往往具有如下几个特点:第一,一条微博属于内容不超过 140 个字的短文本,由于字数限制,其提供的信息量是很少且不全面的。第二,除了微博认证的媒体用户外,大部分用户在撰写微博时使用语言都很不规范,并且存在中英文混用情况,这为文本分析带来了干扰。第三,微博用户经常会通过微博分享自己的心情与感悟,这类微博占有较大的比重,但从文本上来看,这些微博往往都不能反映为事件。第四,随着微博逐渐深入到每个人的生活中,广告商也开始将各类广告以微博的形式投放,从文本分析

本文受国家重点基础研究发展计划(973 计划)(2015CB352400),国家自然科学基金(61332005,61373119,61222209)资助。

张佳凡(1992—),女,硕士生,主要研究方向为移动社交网络、大数据处理,E-mail:zhangjiafanworld@gmail.com;郭 斌(1980—),男,博士,教授,CCF 高级会员,主要研究方向为普适计算、移动群智感知,E-mail:guob@nwpu.edu.cn;路新江(1984—),男,博士生,主要研究方向为普适计算、社会感知计算,E-mail:ramber1836@gmail.com;於志文(1977—),男,博士,教授,CCF 高级会员,主要研究方向为普适计算、社会感知计算,E-mail:zhiwenyu@nwpu.edu.cn;周兴社(1955—),男,教授,主要研究方向为嵌入式计算、普适计算,E-mail:zhouxs@nwpu.edu.cn。

的角度发现事件时，这类微博属于噪声数据。这些特点为热点事件发现的研究带来了挑战。

如何区分物理事件与虚拟事件是本工作的另一大挑战。对于事件分类，一个传统的方法就是以文本分析作为入口，提取不同类别事件文本内容的特征。由于物理事件涉及物理世界发生的事情，其一大特征就是文本内容应当包含时间、地点及人物信息。然而，不同于传统媒体，由于微博内容的不规范性，物理事件所应包含的时间、地点、人物信息往往在一些用户发布的微博中缺省。因此，传统的文本分析方法并不是一个有效区别物理事件及虚拟事件的方法。对于事件分类，应该提取那些独立于文本的特征作为分类的依据。除文本特征外，微博还包含其他维度的信息，如其传播模式、用户交互特性、信息发布的行为特征等。这些客观信息为事件分类提供了新的途径。

基于上述考虑，本文提出了一种基于热词挖掘和层次聚类的热点事件发现方法。对于发现的热点事件，选择信息量特征、时序特征及原创微博数目特征作为事件分类的依据。通过有监督的学习算法：朴素贝叶斯、支持向量机及 C4.5 决策树训练了二元分类器，将发现的事件分为物理事件或虚拟事件。本文基于新浪微博数据集[9]进行实验，结果表明本文提出的热点事件发现方法可以实时有效地发现一天中的热点事件，微博事件分类方法可达到较高的准确率。

2 相关工作

近年来，面向微博的事件发现已经成为了研究人员关注的问题之一。文献[3,4]提出了基于话题标签的事件发现方法，不同于国外的 Twitter，使用新浪微博的用户可以任意撰写话题标签，这就导致了在一些微博中其话题标签与内容并不相符，在这种情况下，基于话题标签的方法就会失效。文献[5]通过基于排队论的方法对 Twitter 中的热词进行实时发现，然后将处于相似上下文语境中的热词聚为一类，形成热点事件。与文献[5]类似，文献[6]是通过基于 TF-IDF 的方法发现某一时段的热词，然后将相关的热词聚集在一起，形成该时段的热点事件。国内文献[7,8]与文献[5,6]的基本思想一致，只是用于检测热词及热词聚类的算法有所不同。本文基于文献[5,6]的工作，提出了不同的热词发现方法及候选事件聚类方法。

前面已经提到，微博热点事件一般分为虚拟事件和物理事件两大类。与虚拟事件相比，物理事件的发现往往具有更重要的意义。物理事件反映人们在物理世界中的现实生活，因此，在物理事件中包含着许多实用的信息，比如哪个地方发生了自然灾害，哪个地方正在进行一场活动，这些信息可以为人们提供各种各样的生活服务。文献[10]利用 Twitter 实时地对橄榄球比赛(NFL)进行了检测；文献[11,12]利用带有地理标签的微博数据对社会事件发现方法进行了研究；文献[13]采用一种基于密度的聚类方法发现滑动窗口内的潜在事件，然后利用潜在事件的关键词及地理位置信息提取特征向量，利用有监督的学习算法训练一个分类器，将潜在事件分为本地事件或非本地事件。已有的事件分类方法大多数都要利用地理位置信息，但是，出于隐私、终端限制等多方面的原因，大量的微博是缺失地理位置信息的。为了解决这个问题，本文在特征选取中选择了信息量特征、时序特征以及原创微博数目特征，这些特征都独立于地理位置信息，因此，在缺失地理位置信息的条件下也可对发现的事件进行分类。

3 方法框架

本文方法框架如图 1 所示，微博数据库模块为整个系统提供了输入，主要功能模块为热点事件发现模块及热点事件分类模块。热点事件发现模块以微博数据库的内容作为输入，输出当前时间窗口内的热点事件集，热点事件分类模块以热点事件集及微博数据库作为输入，将发现的热点事件分为物理事件与虚拟事件。

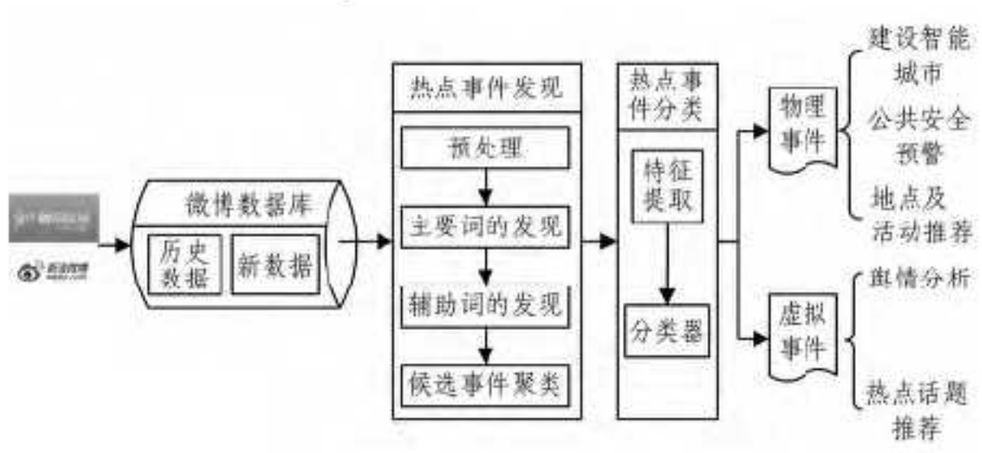

图 1 方法框架

3.1 热点事件发现

图 1 中的热点事件发现模块描述了热点事件发现的基本框架。对于微博数据库中的每一条微博，首先对其进行预处理。然后，利用主要词发现模块发现当前时间窗口内的主要词，即可以代表热点事件的关键词语。之后，利用辅助词发现模块发现构成某一主要词语境的辅助词，即可以代表该主要词上下文的词语。此时，发现的候选事件由一个主要词及若干个辅助词构成。存在这样的情况，一个事件有多个主要词而不仅仅有一个主要词，比如伦敦奥运会这一事件，“伦敦”及“奥运会”两个词都与该事件相关。因此，利用候选事件聚类模块，将涉及同一事件的候选事件聚集在一起。最后，输出一个当前时间段内的热点事件集合，集合中的热点事件由主要词及辅助词描述。

3.1.1 预处理

预处理分为 3 步：分词，去除停顿词，保留名词。

中文分词有多种不同的算法和工具，本文采用 NLPIR 分词系统对微博进行分词处理。它是中文文本处理中经常使用的一个工具，其分词效果较好且支持人名识别、地名识别、组织机构名识别等特殊词类。并且在本文所使用的 2013 版还支持新词发现与自适应分词功能，最重要的是它在分词的同时支持词性标出，所标出的词性用于后面的词语过滤。

经过分词处理后，再将停顿词去除。最后，考虑到大多数事件由名词就可表述清楚，根据第一步分词时得到的词性标注，本文只保留了微博中的名词。

3.1.2 主要词的发现

主要词也即热词，本文将那些在规定的时间窗口内频繁出现的词作为主要词，它是发现热点事件的入口。主要词是事件的主要组成部分。

通常来说，在一天中出现频率高的词语往往暗示着热点事件，然而，这些具有高词频的词语中还包含着一部分无意义的词语，这类词语并不代表事件，但却经常出现在微博中，比

如“话”、“图”、“时”、“称”等，本文称这一类词语为常用词。常用词给主要词的发现带来了很大的干扰，通过去除停顿词及词性过滤只可以消除一部分常用词。与文献[7,8]不同，本文采用了一种较简单的热词检测方法，该方法同时可以减少常用词带来的干扰。通过分析发现，常用词在多个时间段内出现的频率都是较高的，而真正暗示某一热点事件的词语只在某一时间段内出现频率较高。基于这个观察结果，提出了发现主要词的算法。图2示出了一个常用词的词频变化。图3示出了一个热点事件词语的词频变化。

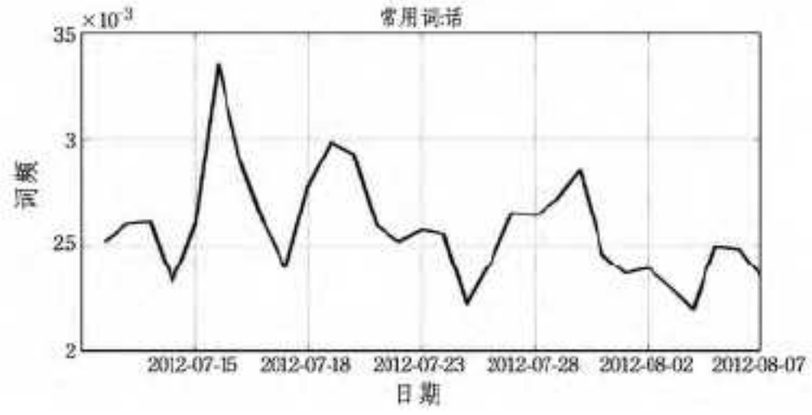

图2 常用词“话”的词频变化

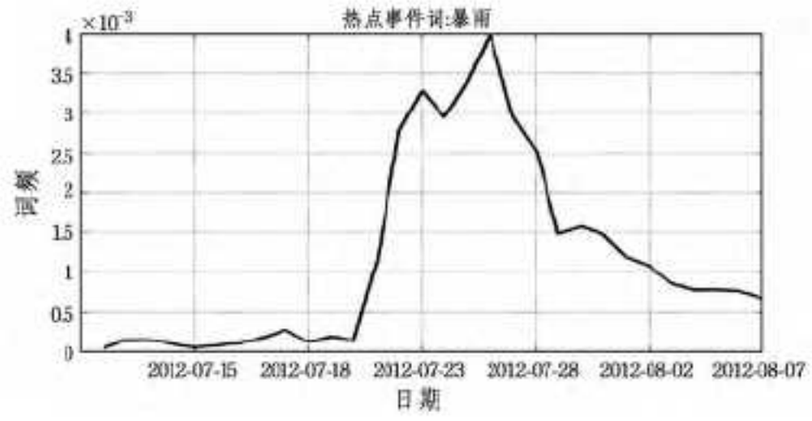

图3 热点事件词“暴雨”的词频变化

发现主要词的算法如算法1所示。

算法1 主要词的发现

输入：经过预处理的词语集S

输出：主要词集合Sc

1. for 每一个S中的词语w
2. w′←WS(w)
3. 把w′放入集合$\hat{S}$中
4. end
5. α←maxima($\hat{S}$)
6. CS←{w∈S|WS(w)←α}
7. for 每一个CS中的词语W
8. 计算DF(w)
9. end
10. 根据DF(w)对CS中的词语排序(降序)
11. 把排在前2%的词语放入集合Sc中

在发现主要词的算法中，

$$WS(w)=\frac{F_c(w)}{F_b(w)} \tag{1}$$

$$maxima(S)=2\times[Q_3(S)+1.5\times IQR(S)] \tag{2}$$

$$DF(w)=\frac{|\{t\in D:w\in t\}|}{N} \tag{3}$$

在式(1)中，$F_c(w)$表示在当前时间窗口内词语w出现的频率，$F_b(w)$表示词语w在微博中出现的常规频率，本文将词语w在微博中出现的常规频率称为词语w的基础频率。若词语w为热词，观察结果表明$F_c(w)$将远大于$F_b(w)$。因此，若其比值$WS(w)$大于阈值α，则将词语w视为潜在的热词。在文献[6]中，词语w的基础频率被视为是固定不变的，但实际上$F_b(w)$应该是动态变化的，由此引入$F_{t,b}(w)$，$F_{t,b}(w)$代表词语w在时间段t内的基础频率，计算公式如式(4)所示。

$$F_{t,b}(w)=\lambda\cdot F_{t-1,b}(w)+(1-\lambda)\cdot F_c(w) \tag{4}$$

其中，λ为衰减因子，为了及时跟踪词频的动态变化，添加了项$(1-\lambda)\cdot F_c(w)$。式(4)不仅考虑了历史信息，还考虑了当前时间段内新的信息，用以更新基础频率，及时地反映基础频率随时间的变化，这是符合实际情况的。

通常来说，在社交媒体中词频的变化呈指数型上升、幂律型下降。由此，衰减因子λ的计算公式如式(5)所示，根据文献[14,15]，本文将β的取值范围限定在$1<\beta\leqslant 1.5$。

$$\lambda=(1+|\varepsilon|)^{-\beta},1<\beta\leqslant 1.5 \tag{5}$$

式(2)是基于一个常用的用于发现数据中异常点的方法导出的，Q_3代表取第三四分位数，IQR代表第三四分位数与第一四分位数之差。

在式(3)中，D是当前时间窗口内出现的微博的集合，t代表一条微博，N是集合D的大小，即当前时间窗口内出现的微博的总数。高的$DF(w)$值表明词语w出现在当前时间窗口内的多条微博中，因此，热词应当具有高的$DF(w)$值。

3.1.3 辅助词的发现

这一步的主要目的是根据与主要词同时出现的频率来发现构成主要词上下文语境的辅助词。根据$DF(w)$来衡量词语w与主要词c同时出现的频率，此时，在式(3)中，D是当前时间窗口内包含主要词c的微博的集合。

之后，根据$DF(w)$对词语进行排序，将排在最前面的词语作为主要词c的辅助词。这里，选择合适数量的辅助词是一个难题，并且辅助词的选择会影响候选事件的聚类效果。为了解决这个问题，对于辅助词，本文定义了一个辅助词候选集合SS，与选择主要词候选词的方法相同，对于每一个词语w，计算其$WS(w)$，定义一个阈值ω，ω由$\frac{maxima}{2}$决定，$maxima$的定义见式(2)。对于词语w，若$WS(w)>\omega$，则保留词语w，否则去除词语w。

最后，选择辅助词$sw\in\{CC\cap SS\}$，$CC=\{w|w\in t,t\in D\}$，D是当前时间窗口内包含主要词c的微博的集合。这样选择辅助词是为了提供关于主要词c的较为丰富的上下文信息以便聚类。由$DF(w)$决定每一个辅助词对于其主要词的重要性。

3.1.4 候选事件聚类

通过主要词发现模块以及辅助词发现模块得到了一个候选事件集CE，该集合中的每一个候选事件ce都由一个主要词及若干个辅助词构成。这一模块的功能就是将涉及相同事件的候选事件聚集在一起。

图4为候选事件聚类的基本流程。将每一个候选事件ce视为一个文档，该文档由该候选事件的主要词及辅助词集合构成，该文档中每一个词语出现的频率由该词语的权重决定，而每一个词语的权重由$DF(w)$决定，见3.1.3小节。根据本文的分析，涉及相同事件的候选事件文档的主题概率分布是具有相似性的，因此，本文通过LDA(Latent Dirichlet Allocation)计算候选事件文档的主题概率分布，并将该概率分布作为每一个候选事件的特征向量。最后，本文再用JS散度(Jensen-Shannon Divergence)作为距离衡量标准，利用SL(Single-Linkage)层次聚类的方法对候选事件进行聚类。

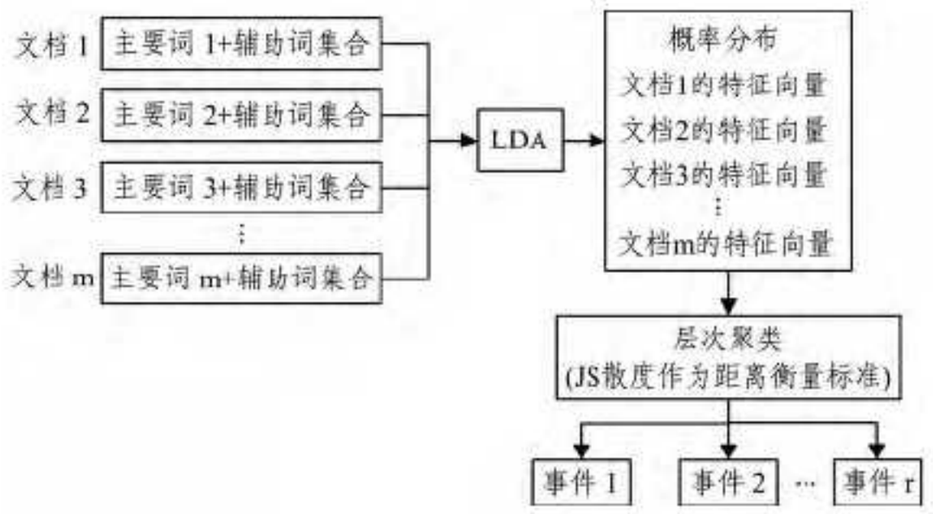

图 4　候选事件聚类基本流程

3.2　热点事件分类

图 5 描述了热点事件分类模块的基本框架。该模块以热点事件发现模块输出的热点事件集作为输入，在经过特征提取及分类器两个小模块后，将输入的热点事件集分为两类，一类是物理事件集，一类是虚拟事件集。

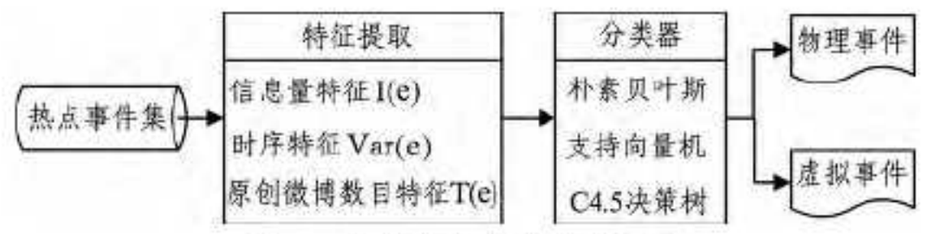

图 5　热点事件分类基本框架

3.2.1　特征提取

对于用于分类的特征，本文从以下 3 点进行了分析。

第一，映射到物理世界的物理事件除了事件的主要方面会被人们广泛讨论外，由其衍生而出的许多侧面也会被人们讨论，事件侧面的多样性导致事件信息量变得丰富。对于虚拟事件，该类事件往往不会衍生出侧面，人们只关注虚拟事件本身。因此，与物理事件相比，虚拟事件带来的信息量是较少的。本文利用热点事件主要词的信息熵作为热点事件的信息量特征 $I(e)$。计算公式见式(6)。其中，R 代表描述事件 e 的主要词的个数。$H(X)$ 代表主要词的信息熵，n 为在当前时间窗口内与主要词 c 同时出现的不重复的词的个数，$p(x_i)$ 取决于 $DF(x_i)$，见式(3)，D 为当前时间段内包含主要词 c 的微博的集合。

$$I(e)=\frac{\sum_{j=1}^{R} H_j(X)}{R} \tag{6}$$

$$H(X)=-\sum_{i=1}^{n} p(x_i)\log_b p(x_i) \tag{7}$$

第二，由于微博用户会对物理事件进行更广泛的讨论，有关该事件的信息熵会急剧升高，与物理事件相比，虚拟事件的信息熵变化不剧烈。图 6、图 7 分别展示了一个物理事件主要词和一个虚拟事件主要词的信息熵变化。根据以上分析，本文采用信息熵的方差来衡量热点事件的时序特征 $Var(e)$，计算公式见式(8)。其中，E 代表求期望值，μ 代表信息熵的平均值。

$$Var(e)=E[(I(e)-\mu)^2] \tag{8}$$

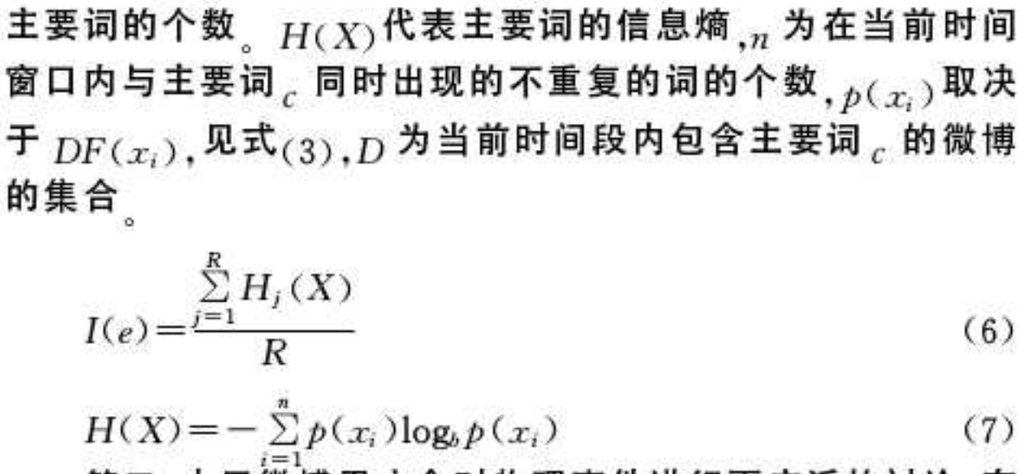

图 6　物理事件主要词的信息熵变化

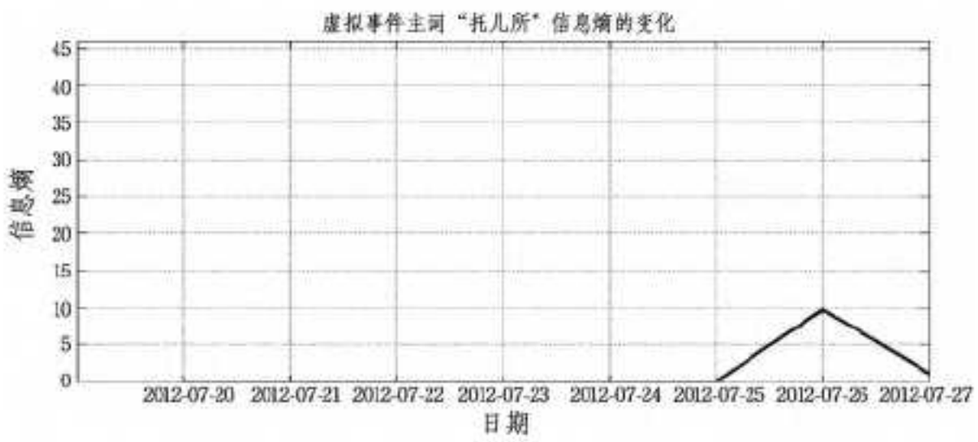

图 7　虚拟事件主要词的信息熵变化

第三，对于物理事件，由于其受众面广，关注人群多，因此在一段时间内，在微博中会出现多条原创微博。但对于虚拟事件，往往是由于大量的转发使其变成了热点事件，因此在一段时间内，微博中虽有大量的与其相关，但是原创微博数目却很少。通过以上的分析，本文将时间窗口内与热点事件相关的原创微博的数目 $T(e)$ 作为第三个区分物理事件与虚拟事件的特征。

3.2.2　分类器

本文主要采用了 3 种经典的分类算法，包括朴素贝叶斯、支持向量机以及 C4.5 决策树。利用基于 JAVA 环境下的开源数据挖掘软件训练二元分类器，将发现的事件分为物理事件或虚拟事件。

4　实验验证

4.1　实验数据集

本文使用了文献[9]发布的数据集，该数据集采集了从 2009 年 8 月到 2012 年 12 月发布的微博的数据。表 1 展示了该数据集的基本统计信息。图 8 展示了数据集中原创微博数据的分布。图 9 展示了加入转发微博后的数据分布。本文主要采用 2012 年 7 月及 2012 年 8 月这两个月的数据进行实验，对热点事件分类方法的有效性及热点事件分类方法的准确性进行验证。

表 1　微博数据集

数据集	用户数量	原创微博数量	转发微博数量
新浪微博	1787443	300000	37372573

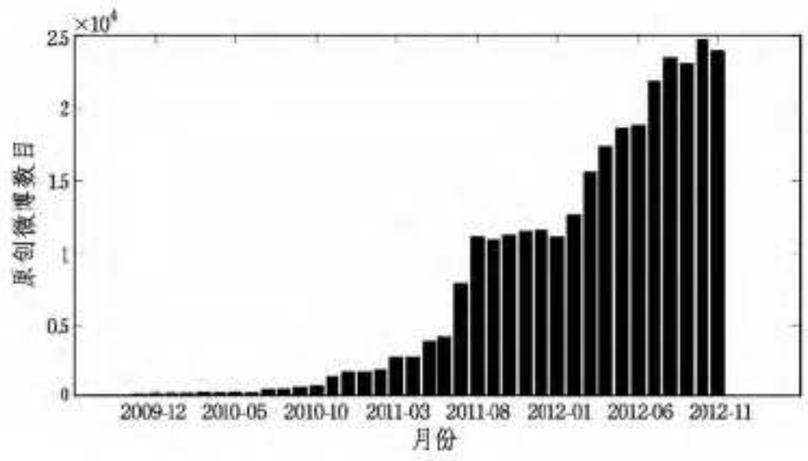

图 8　原创微博的数据分布

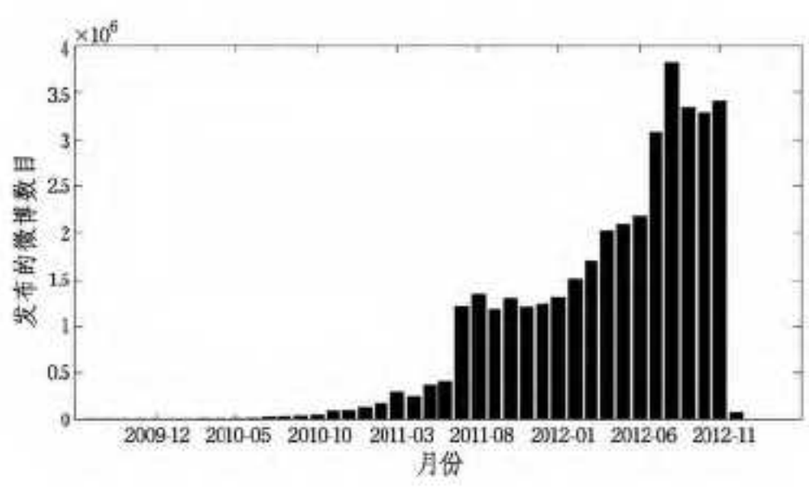

图 9　包含转发微博的数据分布

本文利用人工标注的方法，选择了 73 个样本用于热点事件分类器的训练，其中有 51 个正样例（物理事件），22 个负样例（虚拟事件）。利用十折交叉验证的方法对分类结果的准确率进行验证。

4.2 实验结果

图 10、图 11 展示了两个不同的实验结果示例。实验结果表明本文采取的方法可以有效地检测到热点事件并区别物理事件和虚拟事件，但仅依据主要词和辅助词较难获得事件的完整描述。图 12 展示了利用主要词及辅助词关联微博的准确率及召回率，召回率衡量的是本文算法的查全率，即检测出的相关微博数和数据集中所有的相关微博数之比。由图 12 可知，本文提出的方法可以较准确地将事件与微博关联在一起，但是对于侧面丰富的真实事件（事件 1 以及事件 4），召回率却较低。图 13 展示了选取不同单一特征及分类算法的分类结果准确率，图 14 展示了选择不同组合特征及分类算法的分类结果准确率。图 13、图 14 表明，使用信息量特征 $I(e)$ 分类效果较好，朴素贝叶斯及 C4.5 决策树比支持向量机更适合解决本文的分类问题。

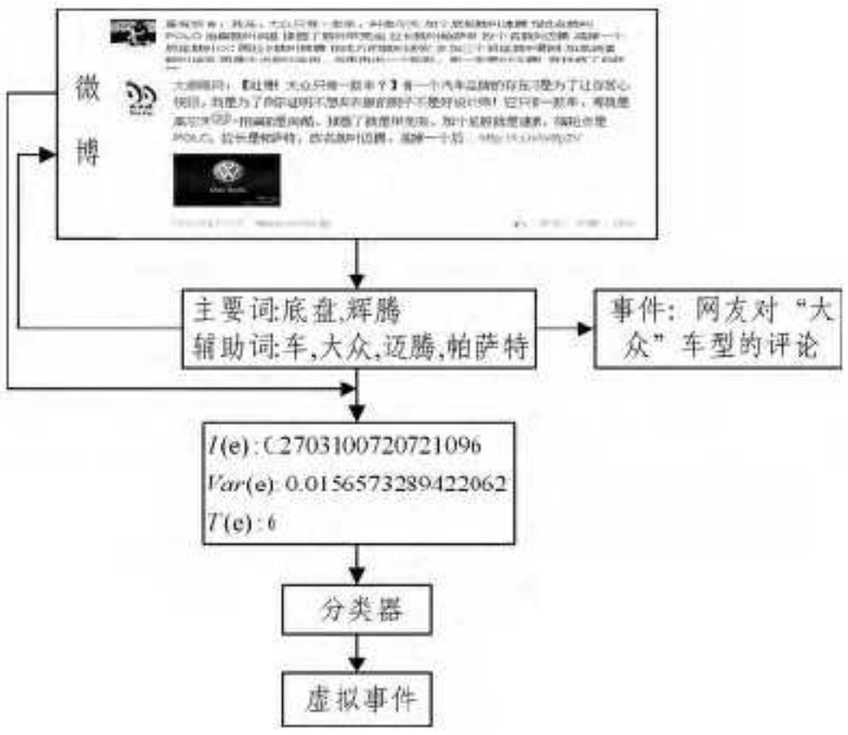

图 10 虚拟事件结果示例

图 11 物理事件结果示例

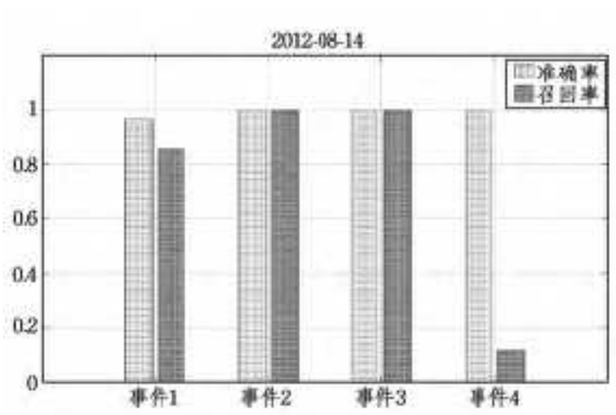

图 12 关联微博的准确率及召回率

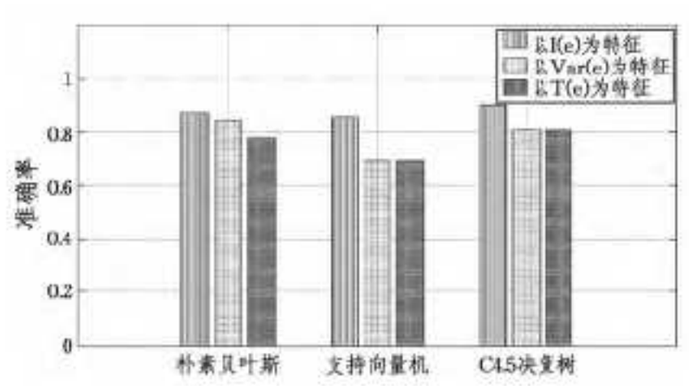

图 13 选取不同单一特征及算法的分类结果准确率

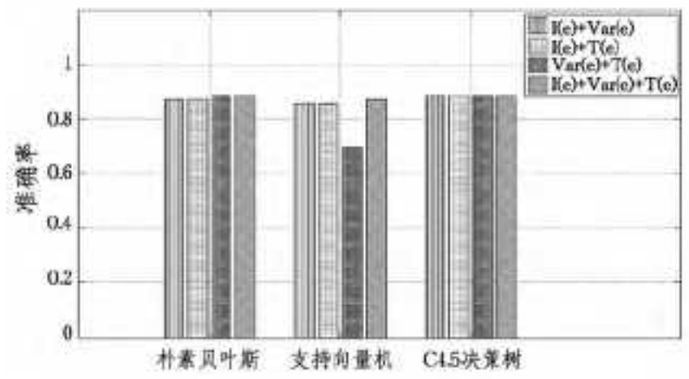

图 14 选取不同组合特征及算法的分类结果准确率

实验结果表明，本文提出的热点事件发现方法可以实时有效地发现当天的热点事件。对于热点事件分类方法，选取不同的特征及特征组合，朴素贝叶斯分类平均准确率为 86.3%，支持向量机为 79.5%，C4.5 决策树为 86.9%。

结束语 本文提出的基于移动群智数据的城市热点事件感知方法可以实时有效地发现一天中的热点事件并对热点事件进行分类。在对热点事件进行分类时，为了解决地理位置信息缺失及文本内容不规范的问题，本文提出了 3 个独立于地理位置信息及文本内容的特征用于微博事件分类，比较了选择不同单一特征或组合特征时，应用 3 种不同的分类算法得到的分类结果的准确率。

然而，通过实验结果也可以看到，本文所提出的方法还存在不足。对于热点事件发现，仅利用主要词以及辅助词并不能形成事件的完整描述，对于侧面丰富的真实事件往往不能发现所有与其相关的微博。这都是基于文本的事件发现方法所带来的不灵活性。在未来的工作中，可以考虑移动社交网络中信息的传播模式以改善上述不足。对于热点事件分类，并没有考虑到数据分布以及不同分类算法的优缺点，只是简单地采用了 3 种典型的分类算法进行了实验。因此，在未来的工作中应当根据数据分布选择合适的算法再进行实验，并且还需要探索其他分类特征，比如用户属性特征等。另外，基于本方法可以开发终端应用，提供事件推荐或城市管理服务等。

参 考 文 献

[1] 刘云浩. 群智感知计算[J]. 中国计算机学会通讯，2012，8(10)：38-41

[2] Guo B，Yu Z，Zhou X，et al. From Participatory Sensing to Mobile Crowd Sensing[C] // 2014 IEEE International Conference on Pervasive Computing and Communications Workshops(PERCOM Workshops). IEEE，2014：593-598

[3] Cui A，Zhang M，Liu Y，et al. Discover breaking events with popular hashtags in twitter[C] // Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM，2012：1794-1798

[4] Ozdikis O，Senkul P，Oguztuzun H. Semantic expansion of hashtags for enhanced event detection in Twitter[C] // Proceedings of the 1st International Workshop on Online Social Systems. 2012

（下转第 37 页）

Sets and Systems, 1995, 69: 125-139

[6] Bazan J, Nguyen H S, Nguyen S H, et al. Rough set algorithms in classification problem[M] // Rough Set Methods and Applications. Physica-Verlag, 2000: 49-88

[7] Alvarez J L, Mata J, Riquelme J C. OBLIC: classification system using evolutionary algorithm[C] // 6th International Work-Conference on Artificial and Natural Neural Networks. 2001

[8] Bandyopadhyay S, Pal S K. Classification and learning using genetic algorithms[M]. Springer Verlag, 2007

[9] Winkler S M, Affenzeller M, Wagner S. Advances in applying genetic programming to machine learning focussing on classification problems [C] // Parallel and Distributed Processing Symposium. 2006

[10] Zhou Chi. Gene expression programming and rule induction for domain knowledge discovery and management [D]. Chicago: University of Illinois at Chicago, 2003

[11] Zhou C, Xiao W, Tirpak T M, et al. Evolving Accurate and Compact Classification Rules with Gene Expression Programming [J]. IEEE Transactions on Evokutionary Computation, 2003, 7 (6): 519-513

[12] 张建伟, 吴志健, 黄樟灿. 基于多表达式编程的分类算法研究[J]. 小型微型计算机系统, 2010, 131(7)

[13] Pham D T, Ghanbarzadeh A, Koc E, et al. The Bees Algorithm [Z]. Technical Note, Manufacturing Engineering Centre, Cardiff University, UK, 2005

[14] Pham D T, Ghanbarzadeh A, Koc E, et al. The Bees Algorithm, A Novel Tool for Complex Optimisation Problems[C] // Proc 2nd Virtual International Conference on Intelligent Production Machines and Systems. Elsevier(Oxford), 2006: 454-459

[15] Ryan C, Collins J J, O'Neill M. Grammatical evolution: evolving programs for an arbitrary language[C] // First EuropeanWorkshop on Genetic Programming, 1998. Paris, France, April 1998: 83-96

[16] O'Neill M, Ryan C. Grammatical Evolution[J]. IEEE Trans. on Evolutionary Computation, 2001, 5(4): 349-358

[17] O'Neill M, Ryan C. Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language[M]. Kluwer Academic Publishers

[18] 刘坤起, 康立山, 赵致琢. 关于认知演化计算分支领域的研究简报(Ⅰ)[J]. 计算机科学, 2009, 36(7): 26-31

[19] 刘坤起, 康立山, 赵致琢. 关于认知演化计算分支领域的研究简报(Ⅱ)[J]. 计算机科学, 2009, 36(8): 35-39

[20] Witten I H, Frank E. Data mining: practical machine learning tools and techniques[M]. San Francisco, CA: Morgan Kaufmann, 2005

[21] 王卫红, 杜燕烨, 李曲. 基于克隆选择和量子进化的 GEP 分类算法[J]. 计算机科学, 2011, 38(10): 236-239

[22] 柳益君, 朱明放, 习海旭, 等. 基于最大隶属度原则的基因表达式编程分类[J]. 计算机工程与应用, 2012, 48(26): 48-52

[23] Sugiura H, Mizuno T, Kita E. Santa Fe Trail Problem Solution-Using Grammatical Evolution[J]. 2012 International Conference onIndustrial and Intelligent Information, Singapore, 2012, 12: 36-40

[24] Ahmad S A. A Study of Search Neighbourhood in the Bees Algorithm[D]. Cardiff University, 2012

[25] Alfonsecaa M, Gil F J S. Evolving an ecology of mathematical expressions with grammatical evolution[J]. BioSystems, 2013, 111(2): 111-119

[26] 王璞. 基于遗传规划的分类算法研究[D]. 安徽: 中国科技大学, 2013

[27] 陈剑, 马光志. 一种基于文法演化自动拟合非线性数据的蜂群算法[J]. 计算机应用研究, 2013, 30(10): 3257-3260

[28] Ganesh Kumar P. Hybrid Ant Bee Algorithm for Fuzzy Expert System Based Sample Classification[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014, 11 (2): 347-360

[29] UCI 数据集[OL]. http://archive.ics.uci.edu/ml/

(上接第 9 页)

[5] Mathioudakis M, Koudas N. Twittermonitor: trend detection over the twitter stream[C] // Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010: 1155-1158

[6] Gupta M, Gao J, Zhai C X, et al. Predicting future popularity trend of events in microblogging platforms[J]. Proceedings of the American Society for Information Science and Technology, 2012, 49(1): 1-10

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

[8] 郭跑秀, 吕学强, 李卓. 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): 486-490

[9] Zhang J, Liu B, Tang J, et al. Social influence locality for modeling retweeting behaviors[C] // Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. AAAI Press, 2013: 2761-2767

[10] Zhao S, Zhong L, Wickramasuriya J, et al. Human as real-time sensors of social and physical events: A case study of twitter and sports games[J]. arXiv preprint arXiv: 1106. 4300, 2011

[11] Lee R, Wakamiya S, Sumiya K. Discovery of unusual regional social activities using geo-tagged microblogs[J]. World Wide Web, 2011, 14(4): 321-349

[12] Weiler A, Scholl M H, Wanner F, et al. Event identification for local areas using social media streaming data[C] // Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks. ACM, 2013: 1-6

[13] Boettcher A, Lee D. EventRadar: A real-time local event detection scheme using twitter stream[C] // 2012 IEEE International Conference on Green Computing and Communications (GreenCom). IEEE, 2012: 358-367

[14] Barabasi A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435(7039): 207-211

[15] Leskovec J, McGlohon M, Faloutsos C, et al. Patterns of Cascading behavior in large blog graphs[C] // SDM. 2007, 7: 551-556