doi :10.3969/j.issn.1671-1122.2015.01.004

基于微博的安全事件实时监测框架研究

李凌云[1]，敖吉[2]，乔治[3]，李剑[1]

（1. 北京邮电大学计算机学院，北京 100876；2. 中国科学院信息工程研究所，北京 100093；
3. 中国科学院计算技术研究所，北京 100190）

摘 要：文章根据微博事件发展规律和传播特点，在微博社会感知器网络基础上，提出了针对微博安全事件的实时监测框架，该框架包含若干项核心算法，如异常检测算法、地理位置定位算法、相关事件推荐算法和事件相关度分析算法。基于此框架，文章实现了微博事件实时监测系统。该系统采用混合网络爬虫和开放API接口方式采集微博数据，并实现了事件检索模块、事件实时监测模块和热点模块。同时该系统以多维度展示微博事件结果信息，且运行稳定、效果良好。总体上看，文章主要解决的问题是探索虚拟社交网络与物理世界时空相关性，监测特定事件，并在其爆发前发现并进行地理定位，从而提供预警。

关键词：微博事件；实时监测；异常检测；地理定位

中图分类号：TP309 **文献标识码：**A **文章编号：**1671-1122（2015）01-0016-08

中文引用格式：李凌云，敖吉，乔治，等. 基于微博的安全事件实时监测框架研究[J]. 信息网络安全，2015，(1)：16-23.

英文引用格式：LI L Y, AO J, QIAO Z, et al. Research on Security Event Real-time Monitoring Framework Based on Micro-blog[J]. Netinfo Security, 2015, (1):16-23.

Research on Security Event Real-time Monitoring Framework Based on Micro-blog

LI Ling-yun[1], AO Ji[2], QIAO Zhi[3], LI Jian[1]

(1. *School of Computer Science, Beijing University of Posts and Telecommunications, Beijing* 100876, *China;*
2. *Institute of Information Engineering, CAS, Beijing* 100093, *China*; 3. *Institute of Computing Technology, CAS, Beijing* 100190, *China*)

Abstract: According to the discipline of event's development and the social characteristic of event's propagation, this paper proposes a framework of real-time monitoring events which propagating on micro-blog, based on the theory of social sensor network, and this framework includes several key algorithms, such as abnormal detection algorithm, geography location positioning algorithm, related events recommendation algorithm, and event correlation analysis algorithm. Based on this framework, this paper develops and implements a real-time monitoring system about micro-blog events. This system applies hybrid web crawler and the way of open API interface to capture micro-blog data, and also implements the event retrieval module, real-time monitoring module and hot topic module. This system also displays the result information of micro-blog event in multiple dimensions, and operates stably. In conclusion, this paper is to explore the field of spatial-temporal correlation between the virtual network and the physical world, monitor the specific "event", and position its location before outbreak, and provide early warning.

Key words: micro-blog events; real-time monitoring; anomaly detection; location

收稿日期：2014-12-09

基金项目：国家自然科学基金[61472048，61402058]

作者简介：李凌云（1991-），男，河南，硕士研究生，主要研究方向：信息安全与数据挖掘；敖吉（1988-），女，内蒙古，硕士研究生，主要研究方向：面向社会媒体数据流的安全态势分析；乔治（1986-），男，山东，博士研究生，主要研究方向：数据挖掘与社会计算；李剑（1976-），男，陕西，副教授，博士，主要研究方向：信息安全和数据挖掘。

通讯作者：李凌云 lilingyun@bupt.edu.cn

0 引言

微博，即微博客（micro-blog），作为 Web 2.0 的产物，是一个基于用户社交关系的信息分享、传播以及获取的平台，用户可以通过 Web、WAP 等客户端组建个人社区，以最多 140 字左右的文字更新信息，并实现即时分享[1]。微博作为一种社交传播媒体，具有传播速度快、互动性强、信息更新方便等特点，对社会生活产生巨大影响，成为我国主要传播媒介之一。比起传统媒体，微博更可能占据信息发布的制高点，这点在突发事件中表现尤为突出。例如，2014 年 2 月，新疆和田地区发生 7.3 级地震，微博只用了不足 1 分钟的时间就对该事件做了报道，而国家官方网站第一次发布该信息是在 15 分钟之后。

由此可见，微博的出现拓宽了信息传播的渠道，对经济发展、社会进步和科技普及起到了积极的作用。

但是另一方面，进入微博平台的门槛极低，且用户不受过多限制，无论说得对还是错甚至谣言和辱骂等，都可以在微博平台上自由地传播。2011 年 3 月，“抢盐危机”直接导致盐价疯涨，严重影响百姓的日常生活；2011 年 6 月，“郭美美炫富”事件在 3 天内毁掉了中国红十字会的百年声誉，当年全国社会捐款降幅高达 73.6%。

反动、淫秽、迷信、暴力等有害信息在微博上传播，严重危害了国家和社会的稳定，侵蚀人民的思想。社会突发事件经微博快速传播后，造成网络上小道消息和流言广泛传播，容易引起公众的不理性判断和混乱行为，从而酿成严重后果，特别是经过实名大 V 账号转发后，后果更加严重。这就要求要加强对于微博的及时监测和有效的引导，对于维护社会安全发展，稳定民情民心有着重要的作用，对于促进国家的发展与进步更是有着重要的现实意义。

在此迫切需求下，研究基于微博的事件实时监测技术就变得非常有意义、有必要。此类技术可以帮助人们了解和分析在微博上传播的社会事件和自然事件的实时变化和趋势，以便对那些具有负面影响的话题或事件做出合理防范或提出解决方案[2-5]。正是在此背景下，我们认真调研和分析了微博平台上的事件发展和传播特点，并基于此提出了微博事件实时监测框架，其包含若干项主要技术，如实时异常事件检测技术、地理位置定位技术以及事件相关度分析技术。其中实时异常事件检测技术和地理位置定位技术是本文的核心技术和创新点。实时异常事件检测技术采用了改进的经典波峰识别算法，能够高效准确地挖掘出事件的异常点；地理位置定位技术则采用数据挖掘领域相关的知识和方法对微博文本内容进行处理并赋以权重筛选实现。本文介绍的基于该框架的微博事件实时监测系统如图 1 所示。该系统除使用了事件监测框架作为事件监测的核心模块，也使用了混合数据采集技术收集数据以及相关事件推荐技术增强用户体验。实验表明，该系统效果良好，且运行稳定。

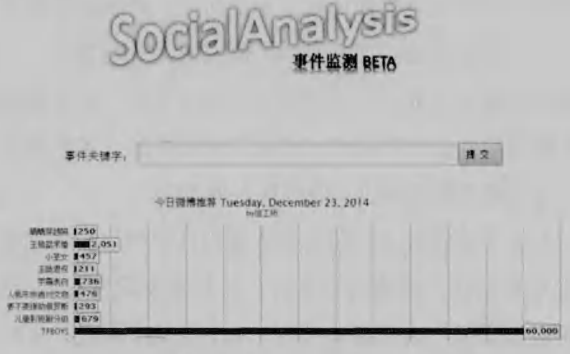

图1 事件监测系统首页

1 国内外研究现状

该类型研究工作始于 2010 年，国内的研究成果主要集中在微博事件搜索上，其本质上就是基于微博的搜索引擎进一步提升和发展，对收集到的数据进行过滤和聚合，以提供更优秀、细致的微博事件服务[6,7]。其研究成果多以网站的形式展示，比较有名的是北京大学网络和信息研究所自主研发的“天网搜索—中国事件检测和发现”和湖南蚁坊软件有限公司出品的“商用系统—鹰击舆情系统”。

而国外同类型研究成果相对丰硕，且模型多样化，这是因为国外社交网络诞生较早，普及较广，且被科研领域重视。Andrew 等人把微博作为一个混合形式的事件影响区域识别和定位的传感器，微博海量用户构造出分布式的社会传感器网络，并基于此网络进行数据挖掘和社会计算[8,9]。本文使用地震这一突发自然灾害事件为例，深入调研和挖掘社交媒体内容的使用潜力，并以实验验证这样一个观点，即微博用户作为感知器带给我们及时可比较的结果，能够补充其他来源的数据以提高态势感知的效率和帮助人们应对此类事件。本文比较完善地提出了

社会感知器等相关概念，并结合真实地震事件进行分析对比，对虚拟社交网络与物理世界的时空相关性研究工作起到了理论性建设作用，对基于微博的相关研究工作意义重大，这也是本文研究的主要理论基础。在系统研发方面，比较著名的系统有麻省理工学院的 TwitInfo，宾夕法尼亚州立大学的 Sense Place2 等[10-17]。

这些研究与系统均能对网络舆情发挥一定的监测、分析和预警作用，为社会和谐稳定提供了有效的技术和决策支持。但是从国内外研究现状可以看出，对微博平台上传播的事件进行实时监控还处于起步阶段，各项理论和技术分析还在摸索前进。尤其是国内研究成果相对简单，中文微博事件研究还是一个非常新型的研究方向，对在微博平台传播的事件进行实时监控还处于起步阶段，各项理论和技术分析还在摸索前进，值得深入地研究。

本文主要解决的问题是探索虚拟社交网络与物理世界的时空相关性，监测特定事件，在其爆发前发现并定位，从而提供预警。为实现该目的，本文提出了高效、实时的微博事件监测框架，用以监测微博平台上事件的传播和发展趋势。该框架旨在高效快速地检测微博事件爆发的异常时刻点，并展示给用户在该时刻该事件发生的地理位置信息。基于该框架，本文也开发了一个微博事件监测系统。

2 微博事件监测框架

从网络舆情和社会计算的原理和基本流程来看，框架至少应包含数据采集模块、数据处理模块和可视化展示模块等。根据微博平台的特点和具体功能需求可将基于微博的事件实时监测框架分为 5 个核心模块，分别为数据采集模块、事件检索模块、事件实时监测模块、热点模块和可视化展示模块，框架基本功能模块设计如图 2 所示。

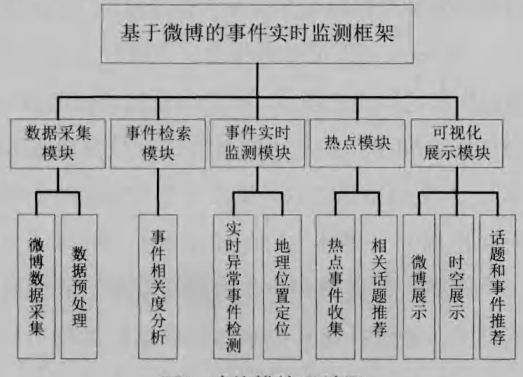

图2 功能模块设计图

其中数据采集模块又细分为微博数据采集和数据预处理模块。事件检索模块主要为事件相关度分析模块。事件实时监测模块是该系统不同于以往传统的微博事件监测系统的最大特点，其包含实时异常事件检测技术和地理位置定位技术模块，是本框架的核心技术要点，也是本文主要阐述的部分。实时异常事件检测技术采用了改进的波峰识别算法，能够在流数据上高效准确地挖掘出事件的异常点；地理位置定位技术则采用数据挖掘领域相关的知识对中文微博正文内容进行处理，并赋以权重筛选而实现。热点模块属于本框架的辅助功能模块，其中热点事件收集子模块的目的是在用户登录具体实现系统时提供微博平台的热门话题和事件，帮助快速了解最新微博火热动态。至于可视化展示模块，本框架除了对事件相关数据进行传统的时间趋势可视化分析之外，同时也对微博数据进行了地理空间可视化分析，在展示与事件最相关微博之外，也进行话题推荐以帮助用户修复事件关键词的偏差。

根据以上框架设计要求和功能需求，构建框架架构如图 3 所示。

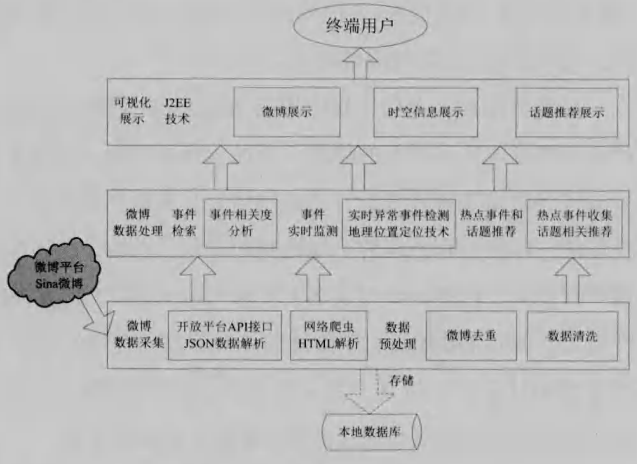

图3 基于微博的事件实时监测框架设计图

数据采集模块主要采用混合网络爬虫和微博开放 API 接口方式实现微博数据的采集，由于采集到的微博信息量比较大，故需要对原始微博数据进行数据预处理工作，主要包含微博去重和数据清洗，以便结构化存储在数据库中。微博事件检索模块主要负责从数据库中提取与事件关键词相关的微博文本信息，以缩小相关微博量而减轻系统运行压力。基于事件检索模块得到的微博文本集，系统进行事件监测，检测事件的异常点并分析异常点的地理位置信息，

同时根据已有的热点事件信息进行话题或事件推荐。热点事件模块在系统后台长时间运行，收集微博平台上的热点话题和事件。在得到相关结果数据后，通过页面层从不同的角度，如时间和空间来展示事件的结果信息，这就使得相关人员在对该事件进行决策时能直观地看到所需要的多角度辅助决策信息，以便对那些具有负面影响的话题或事件做出合理防范或提出解决方案。本文实现的监测系统主要基于新浪微博平台，采用混合网络爬虫和微博开放 API 接口方式实现微博数据采集。微博数据经过预处理后存储在本地数据库中，数据库采用非关系型的 MongoDB。系统接收到用户输入的事件关键词后，从数据库中检索事件，并在该检索结果微博集上进行异常检测和地理位置定位。业务层相关计算执行完毕后，将事件结果传递至前端页面以展示。其系统流程图如图 4 所示。

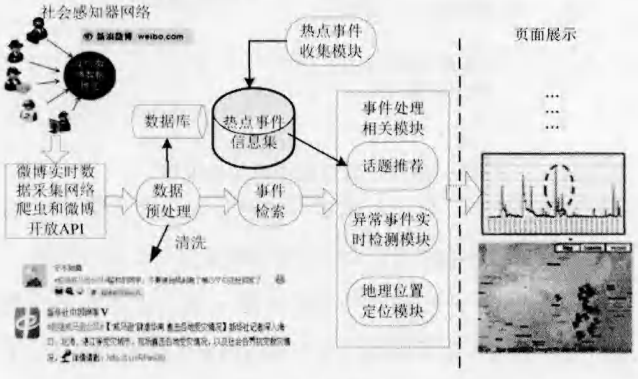

图4 框架流程图

该框架主要包含若干项重要算法要点，如事件相关度分析算法、异常事件检测算法[18]以及地理位置定位算法[19]。

2.1 事件相关度分析算法

事件相关度算法主要功能是检索与事件关键词相匹配或相近的微博文本数据集，以便在此微博集上进行高效的事件监测和分析，检索质量的好坏直接影响该框架的性能。

考虑到微博平台的传播特性，当突发事件信息在微博上传播时，人们更倾向于转发该事件相关的微博，而不是编写新的微博，尤其是大 V 知名或权威博主所发布的事件微博，且越早发布的事件微博，其转发量就越大。所以在设计事件相关度算法时，如果考虑到微博转发量的因素就会使微博事件检索的准确度提高。正是基于这样的想法，结合余弦相似算法和微博传播特点，本文提出了基于公式（1）的微博文本相似度算法。

$$L=\alpha\times\frac{retweet}{Maxretweet}+(1-\alpha)\times\frac{Sim}{MaxSim} \quad \cdots\cdots\cdots\cdots (1)$$

其中，*retweet* 表示该条微博的转发数；*Maxretweet* 表示中间结果集中微博最大转发量；*Sim* 表示该微博与事件关键字的余弦相似度；*MaxSim* 表示中间结果集中余弦相似度最大值；α 为因子参数。

算法的输入分为两部分，事件关键词 keyword 和原始微博数据集 $Ws=(w_1,w_2,...,w_i,...,w_n)(i\in1,2,3,...,n)$；输出结果为微博数据集 Wr（$Wr\subseteq Ws$）。根据以上分析和改进思路，改进的相似度算法 wbSim 的整体流程如下：

1）遍历原始微博集 *Ws*，分别计算 keyword 与微博 w_i 的余弦相似度 $Sim(keyword,w_i)$，把满足 $Sim\geqslant\beta$ 的微博加入中间微博集合 *Wm*，并更新 *Wm* 中微博最大转发量 *Maxretweet* 和 *MaxSim*。

2）遍历中间微博集 *Wm*，分别计算每条微博 w_k 的 $L(w_k)$ 值，把满足 $L(w_k)\geqslant\varepsilon$ 的微博加入结果集 *Wr* 中。

其中 β、ε 为相关标准阈值，根据相关实验结果分析，当参数满足 α=0.2，β=0.1，ε=0.2 时，效果显著。

该算法过滤了采集模块所得到的原始微博文本内容，减少了计算复杂度，进而优化了事件监测运行的数据集，对本框架的性能有着重要的影响。

2.2 实时异常事件检测算法

实时异常事件检测技术主要将该事件相关的微博数量以曲线图展示，采用波峰识别（peak-finding）方法实现，可以高效准确地挖掘事件的异常时间点。根据事件关键词检索出相关的微博数据集，以分钟为单位统计每个时段的微博数据量 c_i，形成微博数量的时间序列 C_n，然后在该序列上检测异常点。最原始的方法是寻找该序列的最大值，以最大值作为异常点，但是这种方式常常会误判异常点，导致准确性不高。根据微博事件的发展规律，事件在未发生之前其微博数量的时间序列是小数值平稳的，当事件发生后的很小时间段内会陡然波动剧烈上升，这个波动点就是异常点，也是事件异常检测算法想要检测的时间点。由此可见事件的异常点应该与之前的历史数据关联比较大，需要考虑到历史数据的情况，尤其是平均值和方差，具有指导意义。一个与之相似的问题就是 TCP 协议中的数据包重传的超时时间计算问题，超时时间设置长了，重发就会导致性能变差；但超时时间设置短了，重发就快会导致

网络拥塞。为了解决该问题，TCP 协议引入了 RTT（round trip time）算法（即 Jacobson/Karels 算法），以历史数据包往返时间修正超时间隔。借鉴该算法的思想，本文采用了加权移动平均值 *mean* 和方差 *dev*，以修正数据的偏差。

$$diff = |oldmean - value| \cdots\cdots (2)$$

$$newdev = \alpha \times diff + (1-\alpha) \times olddev \cdots\cdots (3)$$

$$newmean = \beta \times value + (1-\beta) \times oldmean \cdots\cdots (4)$$

其中，α、β 为参数，*value* 为新增值。实验发现 α=0.25，β=0.125 时，效果最好。

根据以上阐述和设计，事件异常检测算法的输入为基于事件检索后的相关微博数据而形成的微博数据量时间序列 C_n，输出为事件异常时间点序列 L_a。该算法的整体流程如下：

1）初始化平均值 $mean=c_0$，方差 $dev=\mathrm{var}(C_p), p \leqslant n$。平均值 *mean* 直接赋值为 c_0，处理 $c_i, i \geqslant 1$ 时会加权修正该值；方差 *dev* 只需要取前 p=3 个即可，处理 $c_i, i \geqslant 1$ 时同样会加权修正该值。

2）寻找满足 $\frac{|c_i - mean|}{dev} \geqslant \tau$ 且 $c_i > c_{i-1}$ 的下标 i 值，并记录 $start=i-1$，此点的意义在于在该点后的区间内存在潜在的异常点。τ 为参数，一般取值为 2。

3）寻找 *start* 后的区间内的极大值点下标，贪心思想使得该点可能不是异常点。在寻找区间极大值点过程中，同时使用公式（2）、公式（3）和公式（4）去加权修正平均值 *mean* 和方差 *dev*。

4）调整步骤 3）所得到的极大值点下标，基于修正后的平均值 *mean* 和方差 *dev*，若满足 $\frac{|c_i - mean|}{dev} \geqslant \tau$ 且 $c_i > c_{i-1}$，则极大值点下标 $i=i-1$，循环直至不满足条件，此时 i 即为异常点。

5）把步骤 4）得到的异常点的时间属性加入到结果序列 L_a（此前 $L_a=\varnothing$），并根据步骤 2）、步骤 3）、步骤 4）遍历 C_n 寻找所有异常点，并把异常点的时间属性加入到结果序列 L_a 中。

该算法能够高效准确地识别出异常点，而且基于滑动窗口的实现机制能够使其很容易地嵌入到数据流处理过程中，为实现实时监测技术提供了保证。根据奥卡姆剃刀定律“如无必要，勿增实体”，该算法并没有采用复杂的模型或理论，如时间序列分析模型，而是基于统计和规则，并借鉴经典的 TCP 协议中相关算法而实现。因为根据相关实验具体分析，即使采用一些复杂模型，在真实数据流上工作的效果并不见得比相对简单的统计和规则方法更优，甚至可能效果更差。

2.3 地理位置定位技术

地理位置定位技术主要基于大量与该事件相关的微博文本内容，从中抽取地理位置实体，去掉重复实体并附加权重，同时采用聚类的方式筛选出群体性地理位置，其实现过程如图 5 所示。

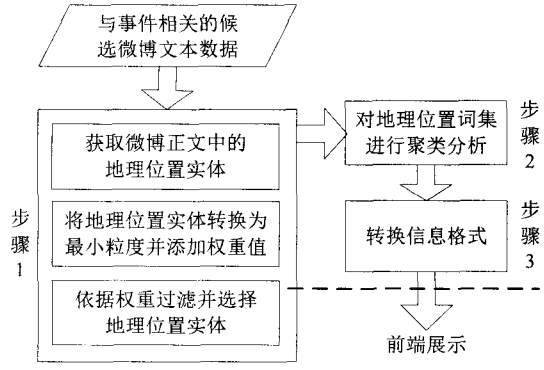

图5 地理位置定位技术数据流程

如图 5 所示，该项技术要点主要分为 3 个步骤：1）抽取微博集中每一条微博正文中的地理位置实体，并进行进一步分析和处理，形成地理位置实体集；2）对上述所得到的地理位置实体集采用聚类的方式筛选出群体性地理位置；3）将上述所筛选的地理位置实体转变为便于展示的信息格式。

步骤 1）抽取微博集中每一条微博正文中的地理位置实体，并进行进一步分析和处理，形成地理位置实体集。

（1）对每条候选微博正文进行中文分词处理以识别地理位置命名实体，分词采用中科院计算所 ICTCLAS 技术，并设置二级标注，地理命名实体词性标注为“/ns”。词性标注是唯一的且准确性非常高，以此标注可获得微博正文中的地理命名实体。分词后抽取该正文中的所有地理命名实体加入序列 *geo* 中，并记录其在微博正文中的具体位置索引 *geoIndex*[]，同时记录事件关键词和标点符号在正文中的具体位置索引，分别为 *keyIndex*[] 和 *punc*[]。

（2）基于位置索引序列 *geoIndex*[]、*keyIndex*[] 和 *punc*[]，分别计算地理位置实体与事件关键词之间的欧式距离 *length*[]，如公式（5）所示。

$$length_i = (keyIndex_i - geoIndex_i)^2 \cdots\cdots (5)$$

若在事件关键词和地理位置实体之间存在标点符号，则适当增大其距离值。本文采用比例系数方式增大距离值，句号、叹号等结束性标点符号对应的增大系数 α=1.125，其他情况 α=1.05。这是因为根据中文语言使用习惯，发布事件消息时，事件关键词都与其发生地理位置处于同一句段里，尤其是结束性标点符号意味着关联性降低。如微博“汶川地震了！我在成都，感觉非常明显”，叹号前面句段是发布事件消息，其后字段与事件的关联性并不大，但地理名称“成都”会产生干扰作用，设计算法时需要尽量消除此类影响。根据距离序列 *length*[]，选择其中距离事件关键词最近的地理位置实体 *geoNear*，并赋予权重值 *weight*=min(*length*[])。

（3）重复上述步骤，对微博集中每条候选微博正文做同样的处理，选择最近地理实体并赋予权重，形成地理位置实体集 *geoNear*[] 和其对应权重集 *weight*[]。若集合 *geoNear*[] 中元素个数大于 3，则丢弃处于权重集中最大的前 20% 的地理位置实体 *geoNear*，因为权重处于最大前 20% 的地理位置实体比较粗糙，会对后续部分产生误导。

（4）根据收集的地理位置实体词典，清洗从候选微博正文中抽取的最近地理位置实体集 *geoNear*[]，只保留最细粒度的地理位置实体（例如，四川和汶川，则保留较细粒度的汶川），但不改变其权重值。该词典如图 6 所示（部分）包含了中国所有县级以上的地理位置实体名，共 6225 行。

```
中华人民共和国 0 000000000000 116.380681 39.919126
中国 0 000000000000 116.380681 39.919126
陕西省 1 610000000000 109.503789 35.860026
陕西 1 610000000000 109.503789 35.860026
四川省 1 510000000000 107.194642 30.821702
四川 1 510000000000 107.194642 30.821702
...
宁波市 2 330200000000 121.579006 29.885259
宁波 2 330200000000 121.579006 29.885259
温州市 2 330300000000 120.690635 28.002838
温州 2 330300000000 120.690635 28.002838
...
梵皇县 3 130129000000 114.289553 37.628132
梵皇 3 130129000000 114.289553 37.628132
井陉县 3 130121000000 114.077952 38.000891
井陉 3 130121000000 114.077952 38.000891
...

桦南县 3 230822000000 130.637015 46.306672
桦南 3 230822000000 130.637015 46.306672
桦川县 3 230826000000 130.963018 46.989258
桦川 3 230826000000 130.963018 46.989258
```

图6 地理位置实体词典（部分）

词典每行的字段依次是：地理位置实体名，行政级别 0~3，唯一 id，经纬度信息。其中行政级别 0 代表国家级，1 代表省级，2 代表市区级，3 代表县级。唯一 id 字段是人工生成的数字，该数字中含有根据实际的地理位置关系产生的前缀信息，该前缀信息包含那些行政级别小于自身的地理位置实体的信息，根据此唯一 id 的前缀信息能够实现高效的保留最细粒度地理位置实体的功能。清洗后得到最细粒度地理位置集 *geoFine*[]，其对应权重为 *weightFine*[]。若集合 *geoFine*[] 中的元素个数大于 3，则选取对应权重值处于最小的前 80% 的地理位置实体，生成步骤 1）的结果集合 *geoList*[]，因为权重处于最小的前 80% 的地理位置实体是最优的，可以聚类出更好的结果。

步骤 2）对步骤 1）所得到的地理位置实体结果集 *geoList*[] 采用聚类的方式筛选出具有群体性的地理位置实体。

（1）统计集合 *geoList*[] 中各个地理位置元素出现的频度，频度小于 *threshold* 且小于 0.5*len*(*geoList*) 的地理位置实体，则认为其出现次数较少，不足以代表事件发生的地理位置，故舍弃。常设置 *threshold*=5，该部分伪代码不再给出。

（2）根据步骤 1）中收集的地理位置实体词典，获取上述结果集 *geoList*[] 中各个地理位置实体对应的经纬度信息 *latlongIngo*[]。若词典中不含某一地理位置实体，则可以通过百度地图提供的 *GeoCoding API* 接口获得经纬度信息。 *GeoCoding API* 是一类简单的 HTTP 接口，用以提供从地理位置信息到经纬度坐标或逆向解析的转换服务，支持 C#、C++、Java 等主流开发语言，以 JSON、XML 格式返回数据结果。

（3）基于上一步骤所生成的经纬度信息 *latlongIngo*[]，计算任意两个地理位置实体之间的空间欧式距离，进而形成 $n \times n$ 的距离矩阵。根据该矩阵，采用凝聚层次聚类的单链接算法（Single-Linkage）对地理位置实体 *geoList*[] 聚类，根据不同聚类簇的大小提取出概率较大的几组聚类簇 *geoRs*[]。选择单链接算法是因为其类间距离定义为两类数据之间的最小距离，可用于分裂式聚类以区分最近邻距离和最远的两组数据。事件发生的地理位置只有两种可能，事件只发生在某一地方或同时发生在多个地方，若发生在多个地方，两两之间距离会比较明显，区分度较大，满足单链接算法的使用特点，尤其是在清洗和选优后的地理位置数据集上执行，效果会较好。地理位置实体集 *geoRs*[] 即是基于全部候选微博正文内容，对事件进行地理位置定位的结果。

步骤 3）将步骤 2）所筛选出的地理位置实体集 *geoRs*[] 转变为便于页面展示的信息格式。

以上就是地理位置定位算法介绍，该算法基于与事件相关的微博正文内容，能够精准地定位出事件发生的地理位置，为实现事件监测技术提供了强有力的手段，这也是本文重点研究内容。

3 系统实现

基于上述微博事件实时监测框架，本文实现了完整的事件实时监测demo系统，其地址为http://202.43.148.168:8080/WeiBo/index.jsp。系统硬件环境为两台Linux主机，8G内存，其中一台主机主要负责相关业务层计算和提供Web服务；另外一台主要提供数据服务，硬盘为2T，其上部署MongoDB数据库并实时采集微博数据。

根据上述基于微博的实时事件监测框架和相关核心算法，本文实现了基于页面展示层、业务逻辑层、数据库层3层架构设计的事件实时监测系统，如图7所示，该架构保证了各模块间的松耦合性，利于程序开发和后期维护。

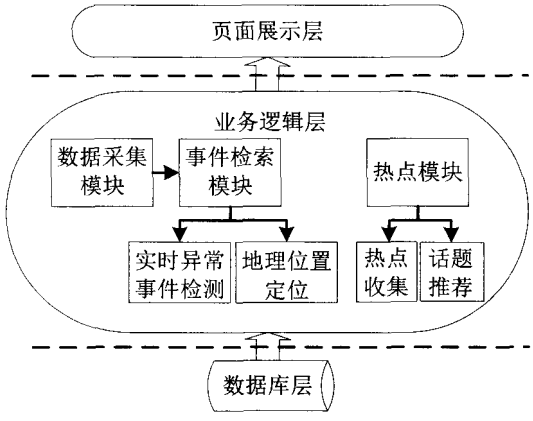

图7 事件时实监测系统

该系统主要使用Java编程语言，并基于J2EE标准的SSS框架而实现，其中页面展示层采用Struts2框架实现，业务逻辑层采用Spring框架进行管理，数据库层采用Spring-Data-MongoDB框架连接数据库。图8为SSS框架。

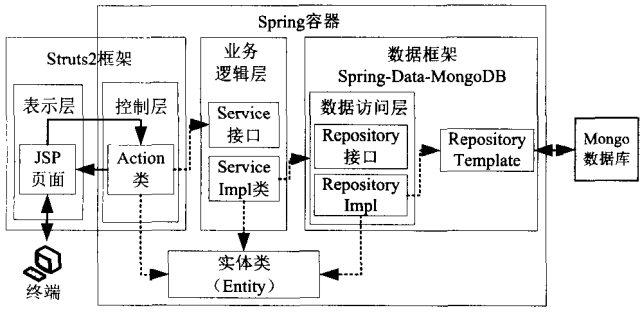

图8 系统J2EE SSS实现框架

3.1 数据采集

微博数据采集主要采用混合网络爬虫和微博开放API方式实现[4]。其中微博开放API接口可以简洁地获取相应的数据，为程序高效获取微博数据提供了保障。而且国内主流的微博平台均提供相应的开放API接口，通过OAUTH 2.0认证可以方便灵活地使用，如图9所示。

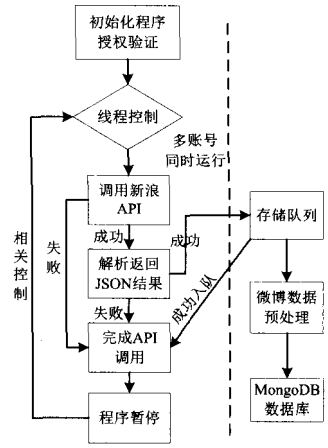

图9 使用新浪微博开放API接口采集数据流程图

虽然通过调用API接口可以实现微博数据的便捷抓取与解析，但所有微博服务商都不会无条件将完整API开放给普通用户，因此使用API的方式永远只可以解决微博数据获取中的一部分问题。例如，在新浪微博中，很多重要查询功能的API是不开放的，同时对于开放的API，一条查询的返回结果数目上限为5000，而往往那些拥有较大信息量的节点才是微博研究最关心的问题。于是在API之外，还需要引入网络爬虫与网页解析技术，来获取更多的微博数据。

基于微博开放API接口的数据抓取策略性能高，但因为服务器限制所以不能获得完整数据集，然而基于网页解析的数据获取方案可以获得最大的数据文本但效率低下。通常需要将两者结合起来，以实现最佳数据抓取效果。

考虑到demo系统微博数据抓取的效率和数量问题，本demo系统采用小时级别的实时监测。具体数据流程为：用户在前端页面输入感兴趣事件的相关关键词，提交后转接到Action中，服务器根据请求开始处理事件相关信息，并返回给前端此时刻之前事件的相关信息，包含相关推荐信息、事件发展趋势信息和异常信息；之后每隔1个小时，前端页面把监测事件的历史信息发回给服务端，服务端根据此请求信息把最新1个小时的事件信息返回给客户端以

展示给用户。在此过程中，服务器端使用 session 保存相关信息，过期时间稍微大于 1 小时。

3.2 事件推荐

为便于增强用户体验，本文增加了事件推荐功能模块。该模块采用相关事件推荐技术，基于用户查询的事件关键词和查询历史，分析用户可能感兴趣的相关事件并推荐给用户。

相关事件推荐技术主要根据用户输入的事件关键词，从热点事件数据库中抽取相关的 N 个热点事件，按照预设的推荐算法排序并返回给用户。

首先构造热点事件数据库，为确保事件平台的一致性，该数据库主要收集微博平台上的热点事件。其次基于事件关键词推荐热点事件。对数据库中的热点事件进行打分，其中打分函数 F 如公式（6）所示。

$$F(\text{事件关键词，热点事件}) = \delta \times S + (1-\delta) \times b \cdots\cdots (6)$$

其中，δ 为参数，s 是采用 Levenshtein 距离（编辑距离）得到的事件关键词和热点事件的相似度，b 取 0 或 1，若热点事件中包含事件关键词则为 1，否则为 0。根据打分函数得到的评分，选取前 N 个热点事件作为推荐事件。该项技术在事件平台一致的基础上向用户推荐感兴趣的热点事件，帮助用户修正事件关键词或发现感兴趣的事件是本 demo 系统中不可或缺的一部分。

4 结束语

本文基于真实事件在微博平台上的传播规律和发展趋势，提出了微博事件实时监测框架的若干项算法要点，分别是事件相关度分析算法、异常事件检测算法和地理位置定位算法，并基于此框架实现了事件实时监测系统。该系统除使用了事件监测框架作为事件监测的核心模块，也使用了新颖的混合数据采集技术来快速海量地收集数据，并采用相关事件推荐技术增强用户体验。本文主要解决的问题是探索虚拟社交网络与物理世界时空相关性，监测特定事件，并在其爆发前发现并进行地理位置定位，从而提供预警。（责编 潘海洋）

参考文献：

[1] 王先国，陈勇军．Web2.0 时代网站建设特征、内容及其应用研究 [J]. 软件导刊，2013，12（12）：9-11.

[2] 李婧，刘志明，崔朝国．基于微博的舆情监测与分析的研究 [J]. 智能计算机与应用，2013，3（2）：50-53.

[3] 林大云．基于 Hadoop 的微博信息挖掘 [J]. 计算机光盘软件与应用，2012，1（4）：7-8.

[4] 廉捷，周欣，曹伟，等．新浪微博数据挖掘方案 [J]. 清华大学学报（自然科学版），2011，51（10）：1300-1305.

[5] 田野．基于微博平台的事件趋势分析及预测研究 [D]. 武汉：武汉大学，2012.

[6] SHAN D, ZHAO W X, CHEN R, et al. EventSearch: A System for Event Discovery and Retrieval on Multi-Type Historical Data[C]// Proceeding of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, Beijing: ACM Press, 2012.

[7]LEE C H, YANG H C, CHIEN T F, et al. A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs[C]//International Conference on Advances in Social Networks Analysis and Mining, 2011, Songdo: IEEE Press, 2011.

[8]TAKESHI S, MAKOTO O, MATSUO Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors[C]//World Wide Web, 2010, Raleigh: ACM Press, 2010.

[9]ANDREW C, ARIE C, ANTHONY S, et al. Earthquake: Twitter as a Distributed Sensor System[J]. Transaction in CIS, 2013, 17(1): 124-147.

[10]LEE R, WAKAMIYA S, SUMIYA K. Discovery of Unusual Regional Social Activities Using Geo-tagged Microblogs[C]//World Wide Web, Hyderabad, 2011.

[11] SINGH V. From Multimedia Data to Situation Detection[C]//Proc. ACM Multimedia, 2011, Scottsdale: ACM Press, 2011.

[12]SINGH V, JAIN R. Structural Analysis of the Emerging Event-Web[C]//World Wide Web, 2010, Raleigh: ACM Press, 2010.

[13]CHOUDHURY S, BRESLIN J G. Extracting Semantic Entities and Events from Sports Tweets[C]//Workshop on Making Sense of Microposts, 2011: 22-32.

[14]ANTOS A, WIVES L, ALVARES O. Location-based Events Detection on Microblogs, Augusto Dias Pereira dos Santos[J]. arXiv preprint arVic, 2012, 3(3): 1-10.

[15]MARCUS A, BERNSTEIN M, BADAR O, et al. TwitInfo: Aggregating and Visualizing Microblogs for Event Exploration[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, Vancouver: ACM Press, 2011.

[16]MACEACHREN A M, ROBINSON A C, JAISWAL A, et al. Geo-Twitter Analytics Applications in Crisis Management[C]//Proceeding of the 25th International Cartographic Conference, 2010, Paris: ACMPress, 2010.

[17]ADOMAVICUYS G, TUZHILIN A. Toward the Next Generation of Recommender Systems: a Survey of the State-of-the Art and Possible Extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[18] 陈斌，陈松灿，潘志松，等．异常检测综述 [J]. 山东大学学报（工学报），2009，39（6）：13-22.

[19] 王波，甄峰，席广亮，等．基于微博用户关系的网络信息地理研究—以新浪微博为例 [J]. 地理研究，2013，32（3）：380-391.

[20] 周立柱，林玲．聚焦爬虫技术研究综述 [J]. 计算机应用，2005，25（9）：1965-1969.

[21] 田董涛．微博客数据的获取与分析方法研究 [D]. 北京：北京交通大学，2011.