基于短文本的食源性疾病事件探测技术

祝天刚[1,2]，郭旦怀[1]，王学志[1]，黎建辉[1]，周园春[1]
1. 中国科学院计算机网络信息中心，北京 100190；2. 中国科学院大学，北京 100049

摘要
微博数据是短文本事件探测的典型数据源，由于微博内容的多样性、稀疏性和碎片性，现有事件探测方法使用的数据源单一且噪声较大，在时空信息的发现上粒度过大，导致结果的准确性差。因此，在事件探测算法上提出动态上下文窗口算法，构建候选微博进行事件探测，提高了事件探测的效率和精度。并提出利用微博内容发现特定事件地理位置信息的算法，提高了事件时空信息的获取精度。最后应用于食源性疾病事件的自动探测中，相比以往的事件探测方法，扩大了数据来源，且时间和空间维度上的准确性得到显著提高。

关键词
短文本；事件探测；时空信息；微博；食源性疾病

中图分类号：TP399　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2016022

Foodborne diseases event detection based on short text

ZHU Tiangang[1,2], GUO Danhuai[1], WANG Xuezhi[1], LI Jianhui[1], ZHOU Yuanchun[1]
1. Computer Network Information Center, Chinese Academy of Science, Beijing 100190, China
2. University of the Chinese Academy of Sciences, Beijing 100049, China

Abstract
MicroBlog is a typical short text data source for event detection. Because of diversity, sparsity and debris in MicroBlog content, using existing event detection method is ineffective, and the event spatio-temporal information is inaccurate. To the end, a dynamic context window algorithm was proposed, improved the efficiency and precision of event detection of foodborne diseases based on MicroBlog. Moreover, an algorithm was developed which can get spatio-temporal information from MicroBlog more accurate. Finally, extensive experiments on event detection of foodborne diseases show the proposed method can help to expand the data source and improve the accuracy of time and space dimension.

Key words
short text, event detection, spatio-temporal information, MicroBlog, foodboorne disease

1 引言

随着互联网的全面普及，大量的数据随之产生[1,2]，经常为人提及的“信息爆炸”迅速具体化为“数据爆炸”。每一名互联网用户，不仅可以是互联网信息的浏览者，同时也是互联网信息的制造者[3]。大量数据的涌现可以充分丰富人们的信息来源，但是人们获得高质量信息的难度大大增加[4]。这一问题在短文本领域表现更为突出。从大量繁杂的短文本数据中找到有价值的信息，即基于短文本的事件探测，目前正成为文本领域最热的研究话题之一。

微博数据作为一种典型的短文本数据，成为了基于短文本事件探测的典型数据来源。每个人能随时随地、方便快捷地将发生在自身或身边的事通过微博共享给整个网络。因此微博数据除了拥有短文本数据的特点之外，还具有数据量大、内容丰富且转换快、群体性强和有时序性等典型特点[5,6]。微博的大数据量、广覆盖度和高活跃度等特性使其数据本身蕴含丰富的有价值的事件。微博数据也成了短文本分析的高质量数据[7,8]。

食源性疾病是指通过摄食而进入人体的有毒有害物质（包括生物性病原体）等致病因子所造成的疾病，目前已经成为我国食品安全的头等问题。随着人们生活水平的不断提高，对食源性疾病的关注程度也迅速增长。微博作为人们日常生活中网络社交的最主要手段之一，大量的数据中含有各种有关食源性疾病的信息[9]。从微博数据中，利用短文本数据事件探测的关键技术发现食源性疾病事件，不仅拥有很高的可行性，而且具有很大的价值。

本文以基于微博数据的食源性疾病事件探测为例，对基于短文本的事件探测关键技术进行了研究和实践。以微博作为数据源，利用其数据量大、内容丰富且实时性强等特点，对引起人们广泛关注的食源性疾病进行事件探测。其中包括两个重要部分：一是从大量的微博数据中，发现有关某特定食源性疾病事件，并抽取出合适的关键词来描述该事件；二是确定食源性疾病事件后，找到该事件的时空信息。

2 相关工作

在文本挖掘领域，基于短文本的事件发现占有极重要的位置。随着移动互联网时代的到来，“碎片化”已经成为现在互联网数据的一个最典型特点[10]，其中最具代表性的莫过于微博数据。由于微博数据本身的特点，事件发生的时间可以直接获取，事件的两个重要要素（即事件的关键词[11,12]和事件发生[13]的地点）则成为基于微博数据的事件发现这一问题的研究重点。

随着互联网时代的到来和广泛普及，微博的使用用户越来越多，微博的功能也越来越丰富。微博不仅可以方便用户在任何时候获取世界各地的信息，还使每一个微博用户成为一个信息提供者，甚至是新闻发布者。人们频繁地、实时地发布信息，使得微博成为发现事件的一个重要信息来源[14-17]，基于微博数据的事件探测，可以帮助人们解决越来越多的问题。例如众多体育赛事的战况，可以在微博中近乎实时地获得[18,19]；通过实时分析Twitter数据，在侦测地震事件中其响应速度甚至超过了任何一家传统媒体[20-22]。但是，利用微博数据对人们日益关心的食源性疾病事件进行探测、分析的相关研究还比较少。主要面临如下两方面的

挑战：一是微博内容比较短、信息噪音比较大[23,24]、主题变化快[25,26]，每条微博最多只能写140个字，而且用户比较随意，导致微博内容中会有大量短语、简略语等，与标准的语句略有不同，噪声较大[27-29]，而且微博有很高的实时性[30,31]，导致相邻微博可能完全表达不同主题的事件[32,33]；二是事件地理位置信息难以获取[34]，微博数据中食源性疾病事件的地理位置信息数据可以通过用户签到信息获取，但是这类数据十分稀疏[35]，而通过其他途径获取事件地理位置信息则十分困难[36]。

针对上述挑战，本文提出了动态上下文窗口算法，构建候选微博集合，扩充了较高质量数据源来进行事件探测，提高了事件探测的效率和精度；又提出利用微博内容发现特定事件地理位置信息的算法，提高了事件时空信息的获取精度。

3 微博数据的食源性疾病事件探测

基于微博数据的食源性疾病事件探测主要需要解决两个问题：一是从微博数据中发现食源性疾病事件；二是确定食源性疾病事件时空信息。针对这两个问题，本文提出了新的方法，即动态确定事件的上下文微博来发现食源性疾病事件；利用指定微博内容结合辅助数据，准确获取事件时空信息。本节将对这两个方法进行详细的介绍。

3.1 数据预处理

尽管微博数据量很大，很有价值，但是其中含有的噪音数据也很多。为了能够更好地进行后续实验、研究，需要对微博数据进行一定的过滤等预处理工作。抓取的是从2014年8-10月北京用户的新浪微博，并从中进行筛选。从北京市疾病防控中心获得食源性疾病的症状表现描述词，然后通过领域专家，即医院相关门诊医生对该词表中的词进行选择并口语化，最终得到一份描述食源性疾病的关键词词表，共32个词（吞咽困难、口干、虚脱、腹部不适、胀痛、胀气、血尿、昏迷等）。在筛选微博时，用最基本的字符串匹配方式，如果一条微博中含有食源性疾病关键词表中的一个或多个词，那么该微博用户的最近200条微博就被选取留下。按照这种规则，选取了共93万户左右的用户，他们的微博中至少含有一个食源性疾病关键词。这些用户的微博总共9 500多万条。这样筛选过的微博都是与食源性疾病相关的微博，筛选过的微博用户都是与食源性疾病有关的用户。

为了避免微博营销账户、“僵尸”账户等非真实账户的负面影响，利用SVM[37]算法训练分类器，根据微博账户分类特征（关注数、粉丝数、个人描述长度、微博总数、平均转发数、平均点赞数、平均评论数、微博平均长度、微博发布时间段、平均@数、微博平均链接数），对微博数据进行过滤，得到占总量31%左右的微博数据。对过滤后得到的微博数据进行分词，并建立词向量等，为进行食源性疾病事件探测做准备。

为了更准确地利用微博数据进行食源性疾病事件探测，本文还使用了其他外部数据作为辅助。通过对食源性疾病本身特点和发病人群的分析发现，大部分食源性疾病事件的发生通常都是人们在某饭店就餐后。而在微博中，人们通常不会准确写出饭店的地理位置，而是会直接提及饭店名。为了从微博数据中获取准确的时间、地理位置信息，利用大众点评网（www.dianping.com）上的饭店信息数据作为辅助数据。其中含有大量商户信息，包括商

户地理位置信息。另外，在用户的微博内容中还有可能提及街道名、地名等行政区划名称，虽然口语化严重，但仍然可以为事件地理位置的确定提供重要线索。利用博雅信息网上北京地区的行政区划信息，可以准确地确定微博内容中提及的地理位置信息，从而大大提高食源性疾病事件探测中事件地理位置发现的准确性。

3.2 发现食源性疾病事件关键词

由于微博本身具有数据量大、实时性高、碎片性强、话题转换快等缺点，单条微博所含信息量太小，很难完成食源性疾病事件探测。扩展数据来源就成为提高事件探测准确率的必要手段[38]。利用上下文窗口方法选定更多的微博作为事件候选微博。假设一名用户微博的时间序列为$S=\{T_1,T_2,\cdots,T_k,\cdots,T_{200}\}$。$T_k$为这名用户含有食源性疾病关键词的一条微博。如果简单地采取关键词字符串匹配的方法选择微博，只有T_k这条微博会被选出，但是其他相关微博也可能含有有关食源性疾病的其他重要信息，如地理位置等。为了避免这种状况发生，一般方法是设置上下文窗口选取多条微博，构成候选微博集，从而扩充数据源，可用式(1)表示。

$$C=\{T_{k-i+1},T_{k-i+2},\cdots,T_k,\cdots,T_{k+j}\} \quad (0<i<k,\ k<j<200) \tag{1}$$

C表示利用固定上下文窗口得到的候选微博集，上下文窗口为$[T_P,\ T_Q]$，即将T_k之前的P条微博到T_k之后的Q条微博，加入候选微博集中。具体过程如算法1所示。

算法1 固定上下文窗口算法。

输入：一名微博用户按时间排序的微博序列S；微博上下文窗口上界P；微博上下文窗口下界Q；食源性疾病所在微博T_k。

输出：候选微博序列C。

初始化参数C←空；

push T_k into C
for i=1 to P do
    push T_{k-i} into C
end for
push T_k into C
for j=1 to Q do
    push T_{k+j} into C
end for
return C

利用固定上下文窗口算法，虽然可以有效扩充微博数据来源，避免数据稀疏带来的事件探测不准确的问题，但所选微博之间没有任何语义关系，由于微博内容主题变换极快，导致上、下界之间的候选微博很可能并不是描述食源性疾病事件，形成明显的噪声数据，最后影响食源性疾病事件探测的结果。

为了解决以上问题，设计了动态设定上下文窗口算法。微博的上下文窗口依据微博之间的语义相似度，通过计算微博词向量之间的余弦值计算，两个向量之间的余弦值越大，相似度越高。具体过程如算法2所示。

算法2 动态上下文窗口算法。

输入：一名微博用户按时间排序的微博序列S；衰减率η；微博相似度阈值U；微博上下文窗口上界P；微博上下文窗口下界Q；食源性疾病所在微博T_k。

输出：候选微博序列C。

初始化参数T←T_k，C←空
push T_k into C
for i=1 to P do
    if Sim(T, T_{k-i}) > U
        push T_{k-i} into C
        T←T+T_{k-i}
        U←$U\times\eta$
    else
        break
    end if

end for
for j=1 to P do
if Sim(T,T_{k+j}) > U
push T_{k+j} into C
T←T+T_{k+j}
U←U×η
else
break
end if
end for
return C

动态下文窗口构建候选微博集的方法，在确定食源性疾病关键词所在微博T_k后，分别向前、向后利用微博间的文本相似性来确定上下文窗口。每一条微博与它本身到T_k之间所有的微博之和（即微博分词结果的并集）求相似度，如果相似度大于一定阈值U，这条微博就会被选入候选微博集。以此类推，最终会动态确定上下文窗口，并得到候选微博集。该算法在选定候选微博时，充分考虑了微博间的语义关系，保证所选出的微博与食源性疾病事件有很大的相关性。在有效避免了微博数据稀疏性缺点的同时，也避免了过多的噪音微博被选入，从而提高了食源性疾病事件探测的准确率。

从文本中提取关键词，最常见且简单易实现的方法是利用TF/IDF的方法，本文的实验选择该方法作为基准。但是，这种方法仅仅考虑了词语的统计性质，并没有考虑词与词之间的出现关系，会忽略掉低频词语的影响。这显然不适合微博文本短、词语多变性强这一特点。

TextRank是基于词图模型的关键词抽取算法，不需要提前对语料进行训练，保证了该算法的简洁、有效，可以广泛应用。TextRank的思想来源于信息检索中著名的PageRank算法，通过把文本分割成若干组成单元并建立图模型，利用投票机制对文本中的重要成分进行排序，即可获取按重要程度排序的关键词。TextRank算法仅利用单篇较短文本本身的信息即可实现关键词的抽取。

3.3 发现食源性疾病事件地理位置信息

通过微博数据获得食品安全事件后，还希望获得关于该事件多维度的更全面的信息。其中，事件发生的时间和地点是最关注的。由于每条微博都有其发出时间，所以获取事件的时间并不难，而获取事件发生的地点相对较难。微博数据中，用户的个人注册信息含有地理位置，但该地理位置信息通常是区县级别，这一级的地理位置信息粒度明显太大，精度太低。而移动端的微博还有签到信息，可以精确地反映用户发表微博的地点。但是签到信息数据过于稀疏，很难说明问题。

本文提出了一种通过微博内容来获取事件地理位置信息的方法。通过对人们日常行为和微博数据的分析不难发现，人们在微博上表达食品安全事件时，微博很大可能会包含食源地点信息，如饭店名或食物名。利用大众点评网中关于饭店名和位置的数据，结合食物名在百度地图API上返回的数据，通过设计的相近点算法，便可以最大程度上获取用户食品安全事件的地理位置信息。图1是获取食品安全事件地理位置的流程。

经过对食源性疾病候选微博的内容进行分析发现，微博中所含有关地理位置的信息主要包括：饭店名、食物名、直接地理位置信息和用户注册的地理位置信息。其中，饭店名可以通过大众点评数据找到地理位置信息，如“麦当劳”。而食物名可以通过百度地图API检索，找到所含该食物的饭店名及其地理位置信息，如食物名“水煮鱼”可以找到“沸腾鱼乡”这个饭

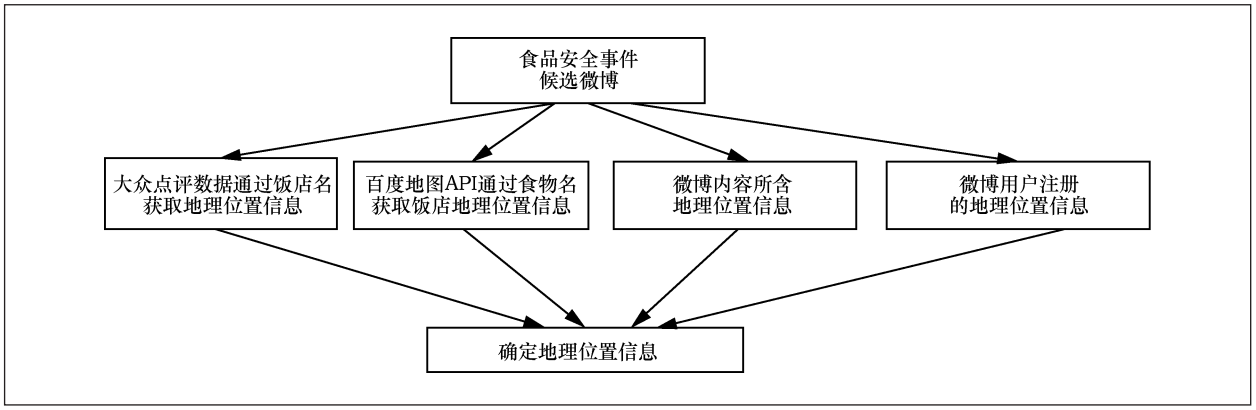


图1 事件地理位置信息发现流程

店。将这两种饭店地理位置信息统称为A，这是用户饮食发生的具体地理位置，A可能有多个地理位置。微博内容中直接含有的地理位置信息，如“中关村”，可以通过博雅地名网（http://www.tcmap.com.cn/beijing/）提供的北京市行政区划数据来找到该词的准确地理位置信息。将这类地理位置称为B，这是用户可能活动的具体地理位置，B可能有多个地理位置。将微博用户注册信息中的地理位置信息称为C，这是用户可能活动的地理位置信息，粒度比较大，通常为区县级别，如“海淀区”，C只有一个地理位置。一个用户的食源性疾病候选微博中，一定含有C，而A和B可能含有，也可能不含有。根据微博中含有的A、B、C这3类信息，可以相对准确地找到食源性疾病事件的地理位置信息。

当拿到一个用户的食源性疾病候选微博数据时，一定含有C类信息，而A、B类信息则不确定。本文根据不同情况，设计了不同的算法来发现事件的地理位置信息。如果微博数据中同时含有A类和B类信息，那么，利用计算相近距离的方法确定事件的最终地理位置信息。如算法3，计算A中地理位置与B中地理位置的距离，找到距离最近的两点A_i和B_j，A_i为最终的地理位置信息。

算法3 事件地理位置发现算法1。

输入：微博所含的3类地理位置信息A、B、C。

输出：地理位置信息Location。

初始化参数Location←空，D←Max

```
if A≠∅ && B≠∅
    for A_i in A do
      if DISTANCE(A_i,B_j) < D
            D = DISTANCE(A_i,B_j)
            Location = A_i
      end if
    end for
end for
return Location
```

当微博中只含有A而没有B时，寻找A中属于C的地理位置信息作为事件发生的地理位置信息。也就是说，微博中含有多个（或一个）饭店地理位置信息，而没有直接地理位置信息，但是该用户注册信息中含有地理位置信息C，这是该用户注册的行政区域。选择属于该行政区域内的饭店地理位置信息作为事件发生的地理位置，算法4描述了这一过程。

算法4 事件地理位置发现算法2。

输入：微博所含的3类地理位置信息A、B、C。

输出：地理位置信息Location。

初始化参数Location←空

if A≠∅ && B=∅
  if $A_i \in C$
    Location = A_i
  end if
end if
return Location

当微博中只含有B而没有A时，寻找B中属于C的地理位置信息作为事件发生的地理位置信息。也就是说，微博中含有多个（或一个）用户的直接地理位置信息，而没有饭店地理位信息，但是该用户注册信息中含有地理位置信息C，这是该用户注册的行政区域。选择属于该行政区域内的用户直接地理位置信息作为事件发生的地理位置信息。当微博中不含有A，也不含有B时，直接利用C来代表事件发生的地理位置信息。

利用这种方法，尽可能地利用了候选微博数据中的地理位置信息来确定事件发生的地理位置。

4 实验

4.1 数据预处理

本文所使用的微博数据为北京市在2014年8-10月产生的所有含有食源性疾病关键词的新浪微博数据。这个数据集中一共包含933 313个微博用户，每个用户200条新浪微博，共9 500万条微博，将近80 GB的数据。为过滤营销账号和“僵尸”账号，利用微博用户的关注、粉丝比、微博总数量等作为筛选条件，选出真正的个人微博账户及他们的微博，符合条件的微博占31%。对这些数据分词，并建立词向量。

抓取了外部数据进行辅助实验。抓取了大众点评网上北京地区餐饮数据，包括饭店名和饭店地理位置，共160 429条数据，还抓取了博雅地名网上北京地名及行政区划数据，共305个地名。

4.2 发现食源性疾病事件关键词

本文的实验共分为4组：固定上下文窗口获取候选微博，分别利用TF/IDF算法和TextRank算法抽取食品安全事件关键词；利用动态上文窗口法确定食品安全事件候选微博，分别利用TF/IDF算法和TextRank算法抽取食品安全事件关键词。对这4种方法的实验结果进行了比较。

主要的评价指标是事件发现的准确率。随机选择了一部分数据，通过人工标注，先找出其中的食品安全事件关键词，以此代表食品安全事件。如果算法找出的事件关键词有80%以上与人工标注的事件关键词相同，就代表事件关键词准确。对上述4种方法进行实验，根据不同方法得到的结果与人工标注结果的对比，得到不同方法的准确率，其比较结果见**表1**和**表2**。

表1展示了固定上下文窗口获取候选微博，分别利用TF/IDF算法和TextRank算法抽取食品安全事件关键词的实验结果。可以明显看出，TextRank算法在关键词抽取的准确率上，明显高出TF/IDF算法。

表2展示了动态上下文窗口获取候选

表1 固定上下文窗口，不同关键词抽取算法的结果

方法	总数/条	正确数/条	准确率
TF/IDF	3 000	639	21.3%
TextRank	3 000	960	32.0%

表 2 动态上下文窗口，不同关键词抽取算法的结果

方法	总数/条	正确数/条	准确率
TF/IDF	3 000	676	22.5%
TextRank	3 000	1 021	34.0%

微博，分别利用TF/IDF算法和TextRank算法抽取食品安全事件关键词的实验结果。同样，TextRank算法在关键词抽取的准确率上，明显高出TF/IDF算法。

从上述4种方法的实验结果可以看出，相比固定上下文窗口选取候选微博，动态上下文窗口法的准确率明显高出固定上下文窗口法。动态上下文窗口是利用微博间的语义关系来确定候选微博的，候选微博的主题更统一，噪声数据更少，故而实验结果表现更突出。而在关键词抽取上，TextRank考虑了词语之间的指向关系，比只考虑频率的TF/IDF在短文本的关键词抽取上更具优势。

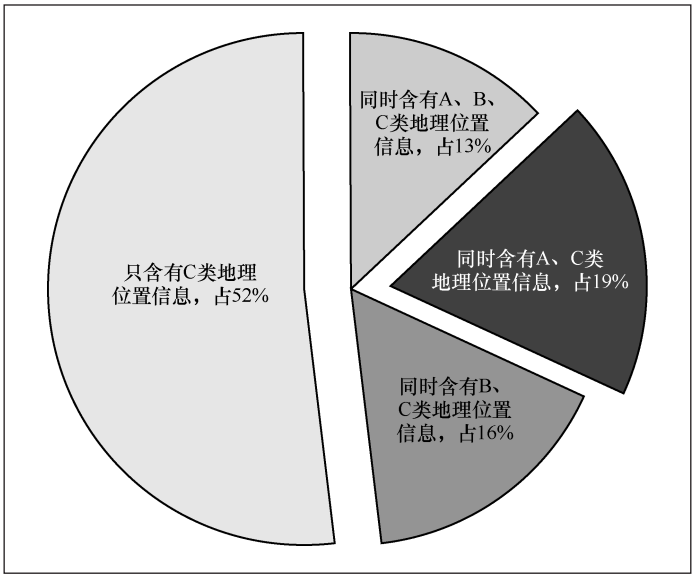


图 2 3 类地理位置信息分布

4.3 发现食源性疾病事件地理位置信息

本文还实现了利用相近点确定地理位置信息的方法，利用用户的微博内容，结合大众点评网、百度地图API、博雅地名网等数据，尽可能地确定食品安全事件发生的地理位置信息，从而对传统的事件地理位置信息的获取提供强有力的补充。本实验主要对同时包含上文中提到的A、B、C 3类信息的微博数据进行相近点算法验证。这样的数据大概占所有数据的13%，同时含有A、C类地理信息的数据占19%。同时含有B、C类地理信息的数据占16%，只含有C类地理信息的数据占52%，如**图2**所示。

主要评价指标是地理位置信息发现的准确率。选择了一部分同时包含上文提到的A、B、C 3类地理信息的微博数据，人工标注出事件的地理位置，然后对实验结果进行对比，得到算法的准确率，见**表3**。

表3是事件地理位置发现的实验结果，将人工标注数据分为500条、1 000条、1 500条、2 000条、2 500和3 000条6组数据，分别统计准确率。

本节全面介绍了基于微博数据的食源性疾病事件探测实验相关内容，包括数据预处理、实验过程和实验结果。通过分析得知，利用动态上下文窗口算法，可以更准确地选取微博，扩充数据来源。实验结果显示，在该算法基础上抽取事件关键词，准确率明显提高。而在充分利用了微博内容数据和外部辅助数据之后，也得到了更多的食源性疾病事件地理位置信息，其准确率在65%左右。

5 结束语

本文基于新浪微博数据对食源性疾病

表 3 事件地理位置信息实验结果

总数/条	正确数/条	准确率
500	332	66.4%
1 000	647	64.7%
1 500	1 009	67.3%
2 000	1 280	64.0%
2 500	1 603	64.1%
3 000	2 005	66.8%

事件探测进行了深入研究，提出了面向短文本数据挖掘的事件探测方法，并应用于食源性疾病事件的自动探测中，相比以往的事件探测方法，扩大了数据来源，时间和空间维度上的准确性得到显著提高。在下一步工作中，将融合微博用户好友关系、微博评论内容等多源数据进行食源性疾病事件探测的研究。

参考文献:

[1] 中国互联网络信息中心. 第32次中国互联网络发展状况统计报告[R]，北京：中国互联网络信息中心，2013.
CNNIC. The 32th Chinese Internet Development Report[R], Beijing: CNNIC, 2013.

[2] 祝华新，单学刚，胡江春，等. 2011年中国互联网舆情分析报告[R]. [出版地不详：出版者不详]，2011.
ZHU H X, SHAN X G, HU J C, et al. 2011 China Internet Public Opinion Analysis Report[R]. [S.l.:s.n.], 2011.

[3] LI R, LEI K H, KHADIWALA R, et al. Tedas: a twitter-based event detection and analysis system[C]// IEEE 28th International Conference on Data Engineering (ICDE), April 1-5, 2012, Arlington, Virginia, USA. New Jersey: IEEE Press, 2012: 1273-1276.

[4] GUPTA M, LI R, CHANG K C C. Towards a social media analytics platform: event detection and user profiling for twitter[C]// The 23rd International World Wide Web Conference, April 7-11, 2014, Seoul, Korea. [S.l: s.n.], 2014: 193-194.

[5] LI C, SUN A, DATTA A. Twevent: segment-based event detection from tweets[C]// The 21st ACM International Conference on Information and Knowledge Management, Oct 29-Nov 2, 2012, Maui, USA. New York: ACM Press, 2012: 155-164.

[6] LEE K, AGTAWAL A, CHOUDHARY A. Real-time disease surveillance using twitter data: demonstration on flu and cancer[C]// The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 11-14, 2013, Chicago, USA. New York: ACM Press, 2013: 1474-1477.

[7] 黄永光，刘挺，车万翔，等. 面向变异短文本的快速聚类算法[J]. 中文信息学报，2007，21(2)：63-68.
HUANG Y G, LIU T, CHE W X, et al. A fast clustering algorithm for abnormal and short texts[J]. Journal of Chinese Information Processing, 2007, 21(2): 63-68.

[8] 杨震，段立娟，赖英旭. 基于字符串相似性聚类的网络短文本舆情热点发现技术[J]. 北京工业大学学报，2010，36(5)：669-673.
YANG Z, DUAN L J, LAI Y X. Online public opinion hotspot detection and analysis based on short text clustering using string distance[J]. Journal of Beijing University of Technology, 2010, 36(5): 669-673.

[9] 徐君飞，张居作. 2001-2010 年中国食源性疾病暴发情况分析[J]. 中国农学通报，2012，28(27)：313-316.
XU J F, ZHANG J Z. Analysis of foodborne disease outbreaks in China between 2001 and 2010[J]. Chinese Agricultural Science

Bulletin, 2012, 28 (27):313-316.

[10] PARKER J, WEI Y, YATES A, et al. A framework for detecting public health trends with Twitter[C]// The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Aug 25-28, 2013, Niagara Falls, Canada. New Jersey: IEEE Press, 2013: 556-563.

[11] PETROVIĆ S, OSBORNE M, LAVRENKO V. Streaming first story detection with application to Twitter[C]// Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, June 2, 2010, Rochester, NY, USA. [S.l: s.n.], 2010: 181-189.

[12] MATHIOUDAKIS M, KOUDAS N. Twittermonitor: trend detection over the twitter stream[C]// The 2010 ACM SIGMOD International Conference on Management of Data, June 6-11, 2010, Indianapolis, USA. New York: ACM Press, 2010: 1155-1158.

[13] MARCHETTE D J, HOHMAN E. Tracking Disease Outbreaks Using Twitter[R]. [S.l: s.n.], 2014.

[14] CHENG Z, CAVERLEE J, LEE K. You are where you tweet: a content-based approach to geo-locating twitter users[C]// The 19th ACM International Conference on Information and Knowledge Management, October 26-30, 2010, Toronto, Canada. New York: ACM Press, 2010: 759-768.

[15] CULOTTA A. Towards detecting influenza epidemics by analyzing Twitter messages[C]// The 1st Workshop on Social Media Analytics, July 25, 2010, Washington DC, USA. [S.l: s.n.], 2010: 115-122.

[16] THOM D, BOSCH H, KRüGER R, et al. Using large scale aggregated knowledge for social media location discovery[C]// IEEE 47th Hawaii International Conference on System Sciences (HICSS), January 6-9, 2014,Washington DC, USA. New Jersey: IEEE Press, 2014: 1464-1473.

[17] MAHMUD J, NICHOLS J, DREWS C. Where is this tweet from? Inferring home locations of Twitter users[C]// The 6th International AAAI Conference on Weblogs and Social Media, June 4-8, 2012, Dublin, Ireland. Palo Alto: AAAI Press, 2012: 511-514.

[18] PAUL M J, DREDZE M. You are what you tweet: analyzing Twitter for public health[C]// The 6th International AAAI Conference on Weblogs and Social Media, June 4-7, 2011, Barcelona, Spain. Palo Alto: AAAI Press, 2011: 265-272.

[19] SIGNORINI A, SEGRE A M, POLGREEN P M. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic[J]. PLoS One, 2011, 6(5): e19467.

[20] HARRIS J K, MANSOUR R, CHOUCAIR B, et al. Health department use of social media to identify foodborne illness-Chicago, Illinois, 2013-2014[J]. MMWR Morb Mortal Wkly Rep, 2014, 63(32): 681-685.

[21] PAUL M, DREDZE M. A model for mining public health topics from Twitter[D]. Baltimore: The Johns Hopkins University, 2011.

[22] IMRAN M, CASTILLO C, DIAZ F, et al. Processing social media messages in mass emergency: a survey[J]. arXiv Preprint, 2014, arXiv:1407.7071.

[23] SAYYADI H, HURST M, MAYKOV A. Event detection and tracking in social streams[C]//The 3rd International AAAI Conference on Weblogs and Social Media,

May 17-20, 2009, San Jose, California, USA. Palo Alto: AAAI Press, 2009: 1-4.

[24] SCALLAN E, HOEKSTRA R M, ANGULO F J, et al. Foodborne illness acquired in the United States-major pathogens[J]. Emerging Infectious Diseases, 2011, 17(1): 1339-40.

[25] ALVANAKI F, SEBASTIAN M, RAMAMRITHAM K, et al. EnBlogue: emergent topic detection in web 2.0 streams[C]// The 2011 ACM SIGMOD International Conference on Management of Data, June 12-16, 2011, Athens, Greece. New York: ACM Press, 2011: 1271-1274.

[26] PAL A, COUNTS S. Identifying topical authorities in microblogs[C]// The 4th ACM International Conference on Web Search and Data Mining, February 9-12, 2011, Hong Kong, China. New York: ACM Press, 2011: 45-54.

[27] CHEW C, EYSENBACH G. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak[J]. PLoS One, 2010, 5(11): e14118.

[28] WENG J, LEE B S. Event detection in Twitter[C]// The 6th International AAAI Conference on Weblogs and Social Media, June 4-7, 2012, Barcelona, Spain. Palo Alto: AAAI Press, 2011: 401-408.

[29] YANG Y, PIERCE T, CARBONELL J. A study of retrospective and on-line event detection[C]//The 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA. New York: ACM Press, 1998: 28-36.

[30] HUTWAGNER L C, MALONEY E K, BEAN N H, et al. Using laboratory-based surveillance data for prevention: an algorithm for detecting Salmonella outbreaks[J]. Emerging Infectious Diseases, 1997, 3(3): 395.

[31] STERN L, LIGHTFOOT D. Automated outbreak detection: a quantitative retrospective analysis[J]. Epidemiology and Infection, 1999, 122(1): 103-110.

[32] CHUNARA R, ANDREWS J R, BROWNSTEIN J S. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak[J]. The American Journal of Tropical Medicine and Hygiene, 2012, 86(1): 39-45.

[33] POLGREEN P M, CHEN Y, PENNOCK D M, et al. Using internet searches for influenza surveillance[J]. Clinical Infectious Diseases, 2008, 47(11): 1443-1448.

[34] ARAMAKI E, MASKAWA S, MORITA M. Twitter catches the flu: detecting influenza epidemics using Twitter[C]// The Conference on Empirical Methods in Natural Language Processing, July 27-31, 2011, Edinburgh, UK. [S.l: s.n.], 2011: 1568-1576.

[35] BUSANI L, SCAVIA G, LUZZI I, et al. Laboratory surveillance for prevention and control of foodborne zoonoses[J]. Annali Dell'Istituto Superiore Di Sanità, 2005, 42(4): 401-404.

[36] COLLIER N, DOAN S, KAWAZOE A, et al. BioCaster: detecting public health rumors with a web-based text mining system[J]. Bioinformatics, 2008, 24(24): 2940-2941.

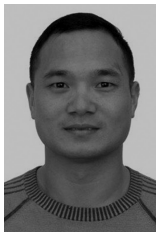
[37] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. arXiv Preprint, 2013, arXiv:1310.4546.

[38] 刘知远. 基于文档主题结构的关键词抽取方

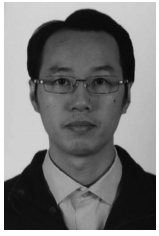
法研究[D]. 北京: 清华大学, 2011.
LIU Z Y. Research on keyword extraction using document topical structure[D]. Beijing: Tsinghua University, 2011.

作者简介

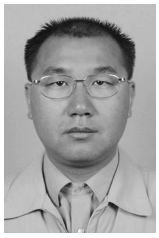
祝天刚(1988-), 男, 中国科学院大学硕士生, 主要研究方向为数据挖掘。



郭旦怀(1973-), 男, 博士, 中国科学院计算机网络信息中心副研究员、硕士生导师, 主要研究方向为海量时空数据挖掘、大数据可视分析。



王学志(1979-), 男, 中国科学院计算机网络信息中心副研究员, 主要研究方向为海量时空数据处理与分析。



黎建辉(1973-), 男, 博士, 中国科学院计算机网络信息中心研究员、博士生导师, 主要研究方向为大数据管理、大数据分析与处理。

周园春(1975-), 男, 博士, 中国科学院计算机网络信息中心研究员、博士生导师, 主要研究方向为数据挖掘、大数据分析与处理。

收稿日期: 2015-09-30

通信作者: 周园春, zyc@cnic.cn

基金项目: 国家自然科学基金资助项目(No.91224006); 国家“十二五”科技支撑计划资助项目(No.2013BAD15B02); 中国科学院战略性先导专项资助项目(No.XDA06010307); 国家卫生和计划生育委员会行业专项资助项目(No.201302005)

Foundation Items: The National Natural Science Foundation of China(No.91224006), The 12th Five-Year Plan for Science & Technology Support (No.2013BAD15B02), The Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDA06010307), Special Research Funding of National Health and Family Planning Commission of China (No.201302005)