

危害公共安全事件的关联关系挖掘及预测

陈夏明,强思维,王海洋,孙莹,石开元

OmniEye团队

上海交通大学

摘要: 公共安全是社会尺度下公民得到的外部环境和秩序的保障,其管理水平在一定程度上反映了一个国家或地区的公共服务水平。近年来,由于国内不同地区收入差距的加大、以及周边政治环境的动荡,危害公共安全的事件时有发生,给公民个人生命和财产带来了严重损害;同时互联网技术的普及使得事件消息的传播不再受空间限制,传播行为也更为复杂,给传统的公共安全管理模式带来了巨大挑战。针对这一需求,本文提出一种基于多维(时间、空间、语义)数据分析的公共安全事件管理方法,包括同类、异类事件的相关性分析、以及预测未来一段时间内同地区发生类似事件的可能性。研究首先基于公开的新闻和微博报道数据,结合其他多种数据源(如地区人口分布数据、GDP数据等),对公交车爆炸、暴力恐怖、以及校园砍杀三类事件进行识别和提取;然后通过相关性分析与数据可视化的方法,对已提取事件的媒体传播规律、事件发生的时空共性进行分析研究;最后通过特征工程方法对时间、空间、语义特征进行提取,并采用Gradient-Boosting算法对未来一段时期内某地区公共安全事件是否发生进行预测,同时利用回归树(Regression Trees)算法对该地区发生的频次进行预测。交叉验证的实验结果表明,我们提出的方法能够揭示在不同时空尺度下事件发生的内在联系,对多类事件在未来1~3个月内发生次数的预测准确度达到65%~82%,充分展示了该方法在以预防为主的新型公共安全事件管理中的重要意义。

关键词: 公共事件;开放数据;Gradient-Boosting算法;决策树回归算法;交叉验证

Intelligent Public Security: Mining Correlation and Prediction

Xiaming Chen, Siwei Qiang, Haiyang Wang, Ying Sun, Kaiyuan Shi

OmniEye, Shanghai Jiao Tong University

Abstract: Public security indicating the management level of a specific region is one of the most vital aspects in modern human civilization. Recently, public security incidents endanger the private properties and lives of citizens because of the enlarging gap of domestic incomes and a turbulent political situation around China. Meanwhile, the transmitting behaviors of incident messages indicate some complicated patterns regarding the prevalence of Internet technologies and social networks, which brings about challenges in traditional management of public security mechanisms. In this paper, we propose a novel, multi-dimensional data analyzing techniques for public security management, including intra- and inter-incidents correlation analysis and the prediction of similar events in the same region in near future. Our method is based upon the public, easily-accessed news and weibo data, in association with multiple data sources such as population and GDP distributions. We first identify and extract different incidents from raw media data, and then the transmission patterns of intra- and inter-incidents are analyzed with correlation analyses and visualization techniques. Afterwards, incident features over temporal, spatial and semantic dimensions are engineered for identified events while both Boolean prediction of incident occurrence and the prediction of occurrence times in near futures are performed. The cross-validation experiments on real datasets show that a prediction accuracy of 65%~82% is obtained in future 1~3 months, which indicates potential applications our approach in the intelligent management of public securities.

Key Words: Public Security, Open Data, Gradient-Boosting Algorithm, Regression Tree, Cross Validation

1 引言

公共安全指社会和公民个人进行正常的生活所需要的稳定的外部环境和秩序,其管理方式受到各国家地区的广泛关注、管理水平也逐渐成为除经济、政治、文化以外的另一个彰显地区公共服务实力和先进程度的重要标志。近年来由于改革开放在经济成果上的成果凸显,地区收入和经济水平的差距逐步增大,以及局部和周边政治局势的多变,危害公共安全的事件时有发生。各类危害公共安全的事件给广大人民群众的生命和财产带来了严重损害,极大地影响着社会稳定和民族团结,诸如公交车爆炸事件、极端民族主义导致的系列暴恐事件、幼儿园砍伤事件等,影响极其恶劣。例如2014年3月1日21时昆明火车广场发生暴力恐怖事件,事件造成29人死亡130余人受伤,严重危害到社会稳定和公民人身安全。

互联网媒体的兴起使得事件消息的传播不再受空间地理的限制,传播速度和传播模式都较传统媒体有很大不同。了解突发事件的新型传播模式,能够帮助很好地理解系列事件发生的内在规律,进而实现从被动处置到主动预防的公共安全管理模式的转变。

从公共安全事件发生的诱因来看,系列事件的发生并非偶然,有些是有组织有预谋的群体性破坏行动(近期越来越呈现离散化发展趋势),有些可能是经由某些社会因素影响(诸如媒体大规模报导、网民舆论传播带来的启发和情绪影响等)发酵形成的个体行为。个体行为导致的事件一旦形成模式,危险性不亚于群体性事件。了解这些危害公共安全事件在互联网上的触发、传播机理,找到相关事件的影响关系和共性,具有重要的研究意义。

针对这一需求,本文提出一种基于多维(时间、空间、语义)数据分析的公共安全事件管理方法,包括同

类、异类事件的相关性分析、以及预测未来一段时间内同地区发生类似事件的可能性。该方法基于可公开获取的互联网媒体数据（包括传统新闻媒体和微博），主要包括四个步骤：数据清洗和补充、独立事件识别和事件类型标记、事件相关性分析、以及事件未来发生的概率预测。由于互联网数据的非结构化特性以及危害公共安全事件的离散性，该方法需要克服来自以下三个方面的挑战：

- 1) 异构的互联网数据对识别独立事件带来的挑战。
互联网媒体数据种类多样，没有统一的结构化标识。一个事件发生后，可能会被多种类型、多家媒体报导，不同语境下采用的用词方式（即语义）不尽相同；同时事件报道的时间跨度不一，短则几天，长则数月，这些因素对独立事件的识别（即将多条不同但针对统一事件的报道进行分类）以及同类事件的标识（即将同性质的不同事件进行分类，如公交爆炸事件）带来了极大的影响。
- 2) 离散的独立事件对分析事件传播规律带来的挑战。事件的发生原因多种多样，有的事件可能是事件发起人受到媒体传播的类似事件的影响而引发，有的事件可能预谋已久。如何发现事件之间的关联，找到影响这些事件产生的因素，进而对今后未发生的事件进行预测，具有很大的挑战。
- 3) 高维度的事件特征对预测事件发生概率带来的挑战。由于公共安全事件背后的诱因复杂，有些事件是有组织有预谋的，有些事件是潜在因素诱发的随机事件；此外，同类事件在不同地区的诱发因素可能完全不同。针对这些不同的因素，需要通过不同的特征来量化表示，这样形成的高维度特征向量给预测算法的设计带来了很大的挑战。

针对挑战一，本文采用基于语义的事件提取和基于开放数据(Open Data)的事件标注方法。首先对原始新闻和微博数据进行预处理，包括数据清洗、修正、融合，然后把基于语义的事件提取算法与基于开放数据的事件标注相结合，完成对独立事件的识别。在基于语义的事件提取算法中，通过TF-IDF (Term Frequency-Inverse Document Frequency) 对单条新闻报道或微博的关键字进行提取，然后利用关键字序列的（余弦）相似度对同类事件进行聚类。另一方面，我们搭建了基于CKAN[1]的开放数据平台，以微信公众号的形式向公众开放，通过众包的方式对新闻进行人工标注，并利用人工标注的结果对事件提取算法进行修正。之

针对挑战二，我们从三个维度（时间、空间、语义）出发，采用可视化和相关性分析方法，对同类事件的相互触发关系、异类事件间的相似性进行量化研究。给定地区事件、给定事件类型发生数目的时间序列，通过计算相关性度量最大信息量相关系数(MIC)[3]，我们找到了同类事件相互触发的显著时间间隔，如在95%的显著水平上，公交车爆炸事件复发的时间间隔是15天左右，而校园砍杀事件复发的时间间隔是5天左右。

针对挑战三，我们通过特征降维的方法，提取对不同类事件的重要特征子集，然后通过局部时间、空间、语义分析与全局时空分析相结合对未来发生的公共事件进行了预测。交叉验证的实验结果表明，我们提出的方法能够揭示在不同时空尺度下事件发生的内

在联系，对多类事件在未来1~3个月内发生次数的预测准确度达到65%~82%，充分展示了该方法在以预防为主的新型公共安全事件管理中的重要意义。

在下文当中，我们首先介绍本研究的数据集和事件提取算法；然后介绍同类事件和异类事件内的相关性分析和可视化结果；最后介绍事件预测算法和交叉实验结果。

2 数据集与预处理

2.1 数据集介绍

针对本课题，我们有3个核心数据集合（如表2.1所示）。分别是新闻和微博数据集、新闻传播信息数据集、微博用户资料数据集。

- 1) 新闻和微博数据集：主要包含每条新闻的信息。包括新闻的唯一标识ID、新闻发布时间、新闻标题、新闻导语、新闻正文、发布媒体等。
- 2) 新闻传播信息数据集：数据集(1)中某条新闻在互联网上的传播情况。包括新闻的来源、评论数、转发数、点赞数。
- 3) 微博用户资料数据集：数据集(1)中微博用户的个人资料。包括用户所在地、出生日期、注册时间、关注数、粉丝数、状态数、活跃天数、等级数等。

表 1: 数据集概览

数据集	记录数
新闻和微博数据集	540205(54万)
新闻传播信息	257550(25万)
微博用户资料	243365(24万)

2.2 数据预处理

原始数据存在一些问题，主要有三点：存在重复记录、信息不完整、事件类型标注不准确。这些重复数据、杂质数据对后续的事件提取、特征提取、关联分析、预测、等工作造成很大影响，因此数据预处理工作十分重要，直接关系到后续处理，对结果的好坏具有很大影响。预处理工作分为数据去重、信息完整化、事件类型标签修正三步。

(1)数据去重

在新闻传播信息、微博用户资料数据集中，存在多条重复数据。我们对这两个数据集进行了去重处理，对于多条完全相同的记录，只保留一条；而对于多条差别很小的记录，仅保留最后更新的一条记录，删除其他记录。

(2)信息完整化

公共事件的发生有很多影响因素，比如某公共事件发生于哪个节日，发生地的经济发展状况、人口数等。这些因素对事件的传播触发、规则关联分析，以及事件是否发生及发生频次预测具有重要的意义。然而，原始数据集中并没有这些特征信息，这回对事件关联分析、事件预测的准确度产生很大的影响。因此，我们采用多源数据融合的方法，把已有数据完整化。我们通过互联网、公共数据库以及一些开放数据平台采集了相关信息，中国各省市GDP数据、日期与节日对

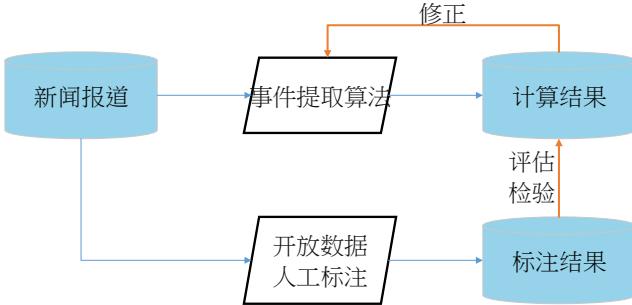


图 1: 事件提取的总体架构图

应表等，之后把这些数据与已有数据源融合，为完善数据特征做好了准备工作。

(3) 事件类型标签修正

新闻和微博数据集中，每一条新闻纪录已通过标签对每一条新闻进行了标注，共3类：公交车爆炸事件、暴恐事件、校园砍伤事件。然而经过对数据进行采样分析，我们发现有许多错标、漏标的情况，甚至有很多与公共事件无关的新闻也误被标注为三类事件之一，标注准确率很低，对我们后续的事件提取工作造成了很大的干扰。为了更准确地进行事件提取，我们对每一条新闻重新进行了事件类型标注。首先针对三个系列事件，人工建立含有少量相关词汇的语义库，之后采用相邻相关词语扩充算法，对每一类事件已有的语义库进行扩充。当针对每一系列公共事件建立好相应的语义库之后，对每一条新闻的标题进行分词，与所建立的词库进行词语匹配，计算匹配度，新闻与哪一类事件的匹配度最高，则把这一条新闻标注为哪一类事件，如果某一条新闻与三类事件的匹配度都很低(低于设定阈值)，则定义这条新闻定义为无关数据(杂质数据)。经抽样统计，修正后的数据事件类型标签准确度达95%以上。

3 事件提取

3.1 整体思想

经过数据去重、信息完整化、事件类型标签修正的预处理之后，与公共事件无关的杂质数据已被过滤掉，得到公交车爆炸、暴恐、校园砍杀三个系列事件的所有新闻。我们的事件提取工作正是基于针对这三个系列事件的所有新闻，分为三个系列分别进行处理，三个系列事件的处理流程相同。

事件提取工作分为两部分，采用TF-IDF相似度匹配算法进行事件提取；采用开放数据微信平台，以众包的方法对事件进行人工标注，对TF-IDF相似度匹配算法的计算结果进行评估与检验，进而对算法进行修正。整体架构如图1所示。

3.2 TF-IDF相似度匹配算法

3.2.1 概述与准备工作

一起公共事件会有许多相关媒体报导，包括新闻报导以及微博报导。即使是针对同一个事件的报导，也会有很大的差别，包括时间上的差别与内容上的差别。比如针对同一事件的不同报导在时间上可能相隔

数周，不同媒体对同一事件报导风格相差迥异。如何在这些海量的报导中识别出哪些属于对相同事件的报导，进而把事件提取出来具有很大挑战。TF-IDF相似度匹配算法通过设定时间阈值，定义并计算新闻间的相似度，根据相似度对不同的新闻分为不同的簇，每一簇中的所有新闻认为是对同一事件的报导，进而实现事件提取。

算法的输入为每一条新闻的主键、报导时间(精确到天)、标题、正文内容、新闻所属事件类别，输出为每一条新闻对应的事件主键，其中事件主键自动生成。在使用TF-IDF相似度匹配算法进行事件提取之前，我们做如下准备工作：报导事件类型分系列处理、时间地点人物信息提取、同系列事件时间排序。

(1) 报导事件类型分系列处理

数据预处理中我们把属于同一系列事件的新闻打好标签(公交车爆炸事件、暴恐事件、校园砍杀事件)，且具有很高的准确度。同一系列事件的媒体报导具有很高的相似性，因此我们把所有媒体报导按照系列类别分类(公交车爆炸、暴恐、校园砍杀)，对每一类分别进行事件提取，这会对分类准确度有很大的提高。

(2) 时间地点人物信息提取

为了精确地对一条媒体报导进行描述，我们首先做假设1：假设1 某事件可以由时间、地点、人物三个要素唯一确定。

基于假设1，我们对每一条媒体新闻所描述事件的时间、地点、人物进行提取。我们对每一条媒体报道的标题、内容通过基于词典与基于规则相结合的方法进行分词。并对其中的地点词语、任务词语自动识别，进行提取，作为这一媒体报导的地点、人物特征。

(3) 同系列事件时间排序

提取过程中我们发现，绝大多数的新闻中对事件的报道的描述很模糊(例如“昨日下午”、“傍晚”等很不精确的时间描述)，而根本无法从新闻的描述中得出具体的事件发生时间，我们有的只有这一条媒体的报导时间。我们做出假设2：

假设2 事件发生时间与第一条新闻发表时间相隔很近，即为时间发生时间就是第一条新闻的报导时间。

根据假设2，我们对每一个系列事件中的所有媒体报导按照时间从前到后的顺序进行排序。这样在把媒体报导分为n个事件后，每个事件最早的一个新闻报导即为事件发生的时间。

3.2.2 TF-IDF计算

在准备工作中，我们针对每一个媒体报道提出了一个人名地名词库。我们对每个系列的媒体报道中，结合每一条媒体报道的词库，整合成一个所有人物、地名词库。把每一条新闻用一个向量表示，向量的维度即为同系列事件中所有媒体报道的词语。之后，每一条媒体报道的向量计算每个词语的TF-IDF值，这是对每条媒体报道特征的数值化表示。计算方法如下：

(1) 计算词频

词频(Term Frequency, TF)指的是某一个给定的词语 t_i 在该文件中出现的频率。对于在某一特定文件里

的词语来说，它的重要性可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

其中 $n_{i,j}$ 表示词语 t_i 在文件 d_j 中的出现次数； $\sum_k n_{k,j}$ 表示文件 d_j 中所有词语出现次数之和。

(2) 计算逆向文件频率

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：

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \quad (2)$$

其中 $|D|$ 代表语料库中的文件总数； $\{j : t_i \in d_j\}$ 表示包含词语 t_i 的文件数目（即 $n_{i,j} \neq 0$ 的文件数目）。TF-IDF 值即为两者的乘积：

$$tfidf_{i,j} = tf_{i,j} \times idf_i.$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

3.2.3 余弦相似度计算

计算这一条新闻和已有事件的余弦相似度[5]，得出一系列的余弦相似度值：

$$\cos_1, \cos_2, \dots, \cos_n$$

这里我们取最大的值 \cos_i 。如果 \cos_i 大于阈值 η （本文设 $\eta = 0.15$ ）则把该条新闻标注为事件，否则把该条新闻归为新的事件。

余弦相似度计算方法如下：在计算两文件的相似度前，需将文件表达成向量的形式，亦即将文件中所有的重要词语都视为一个个的向量维度，以该词语的 TF-IDF 值为该维度的值，组合而成一向量，代表该文件，例如文件 i 表达成文件向量 $D_i = (W_{i1}, W_{i2}, \dots, W_{in})$ ，文件 j 表达成 $D_j = (W_{j1}, W_{j2}, \dots, W_{jn})$ ，则此两文件的余弦相似度计算公式为：

$$\text{COS}(D_i, D_j) = \frac{\sum_{k=1}^n W_{ik} W_{jk}}{\sqrt{\sum_{k=1}^n W_{ik}^2} \sqrt{\sum_{k=1}^n W_{jk}^2}} \quad (3)$$

3.3 基于开放数据的事件标注

上节中我们设计了事件提取算法，然而在自然语言理解的问题上，机器无法完全取代人类。对于一些具体词汇在不同语义语境下的理解，人与机器多多少少会出现一些区别。比如按照算法，会识别“街边咖啡厅爆炸，导致公交车站牌被炸毁”这一条新闻为公交车爆炸事件，而实际上该条新闻应该属于暴恐事件。因此算法会有一定误差，如何发现这些识别错误的数据记录对修正我们的算法是一项很重要的工作，最有效

的方法就是人工对各条数据集进行事件类别和独立事件标注。

然而数据集记录多达 50 余万条，如此庞大的数据量，为人工标注带来了巨大的挑战。我们利用上海交通大学开放数据共享平台 (<http://data.sjtu.edu.cn>, 如图 2 所示) 这一数据平台，借助微信公众账号服务，将这一简单却量大的任务众包给公众。



图 2: 开放数据共享平台

普通众包的方法存在两个不足：第一是便捷性，如此庞大的数据集一般会存储在数据库中，而数据库操作复杂，界面不够友好，操作难度大；第二是没有标注动力，如此多的数据量，让人望而却步，很难带动起人们的积极性。而开放数据为众包提供了可能，上海交通大学采用 CKAN 开源软件作为开放数据平台，提供实时修改数据的功能接口，并具有很好的数据共享隔离机制。

我们正是利用这一开放数据平台结合微信公众账号（如图 3 所示），实现方便用户标注操作的众包入口。为了安全起见。许多互联网入口登陆需要校验码，而人们需要消耗许多时间、精力，也会消耗很多网络与计算资源。为此我们把标注服务与校验码相结合，既可以解决人工标注的问题，又可以节省一大笔开销，把有效的资源用在最合适的地方。



图 3: 开放数据平台微信应用

最后，我们通过事件类别筛选和独立事件人工标注，利用事件标注算法的反例，对事件提取算法进行了修正。在反复人工标注、算法修正的迭代下，我们最终收获了很好的事件提取效果。

4 事件关联预测

4.1 事件关联分析

第 3 节中，我们针对不同系列的事件，分别从新闻

媒体、社交网络的数据中进行了事件提取。在这些提取出的危害公共安全事件中，我们尝试发现事件间的关联规律。比如，同系列事件中，事件之间在时间、空间的传播会有一定规律；而不同系列事件之间的发生则会有一些共性规律，这些规律对于我们进一步了解危害公共安全事件发生规律、对危害公共安全事件进行预测具有巨大的意义。

在本节中，我们首先从时间、空间、语义三个方面进行事件的特征提取；然后通过数据可视化的方法对同系列事件间触发关系、不同系列事件间共性规律进行定性发现[6, 7]，提出一些假设猜想；最后通过相关性度量最大信息量相关系数(MIC)，对事件之间的关联度进行定量分析，做出总结。

4.1.1 特征提取

危害公共安全事件的特征涵盖时间、空间、新闻媒体、社交网络等多维度，如事件发生的时间、事件发生的地点、事件发生的媒体报道情况。为了更详细地表征每一个事件，我们从时间、空间、语义三个方面对事件进行了特征提取，共提取近40个特征。具体特征条目如附录一所示。

4.1.2 同系列事件触发关系分析

表征一个事件有很多特征，比如事件发生的时间、事件发生的地点、事件发生的媒体报道情况。为了更详细地表征一个事件，我们从时间、空间、语义三个方面进行特征提取，共提取近40个特征。

(1) 时间触发关系研究

同系列事件在时间上存在一定的触发关系。在一定时间范围内，一起系列危害公共安全事件的发生很可能对另一起事情的发生产生触发作用。我们使用最大信息量相关系数(MIC)对公交车爆炸事件、暴力恐怖事件、校园砍杀事件3类事件进行了时间维度相关性分析，结果如图4所示。

从图4中我们可以看出：每起公交车爆炸事件时间分布特征和15天前的分布特征相似，每起暴力恐怖事件时间分布特征和5天前的分布特征相似；每起校园砍杀事件时间分布特征与4天前至18天前的事件分布特征均有一定的相似度。此外我们发现，三种系列事件在以月为时间粒度的规律分布上并无相关性可循。

每起公交车爆炸事件时间分布特征和15天前的分布特征相似，每起暴力恐怖事件时间分布特征和5天前的分布特征相似；每起校园砍杀事件时间分布特征与4天前至18天前的事件分布特征均有一定的相似度。此外我们发现，三种系列事件在以月为时间粒度的规律分布上并无相关性可循。

(2) 空间触发关系研究

我们对3类事件在空间触发关系进行分析。首先我们以省级单位为空间划分单位对各个省危害公共安全事件发生频次做相关性分析，但并没有发现明显的相关性特征。然而当我们把地理分区作为空间划分单位，每个地理分区包含若干个省，对各个地区事件发生频次做相关性分析，发现各地区事件发生频次之间具有较为明显的相关性特征。如图5~7所示。

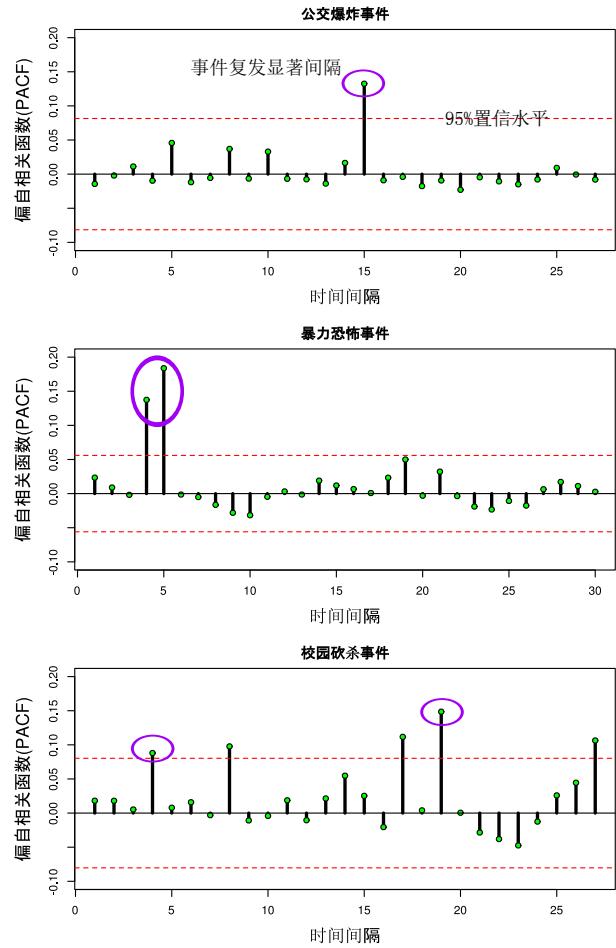


图 4: 同系列事件时间相关性分析

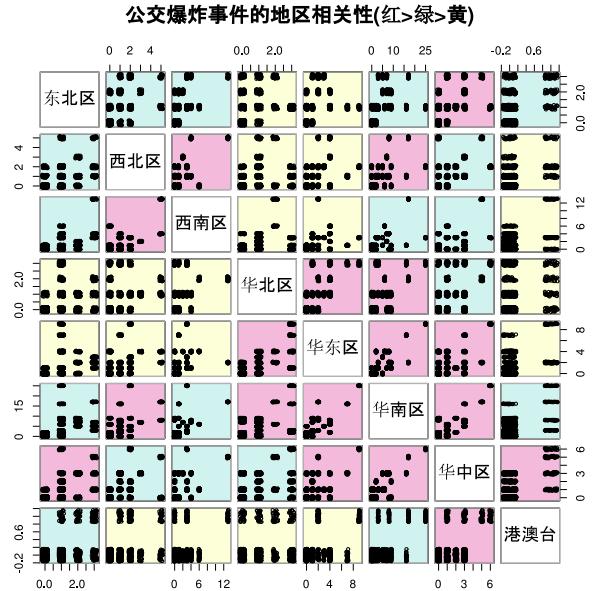


图 5: 公共交通事件空间传播变化规律

我们以校园砍杀事件为例进行分析，校园砍杀事件中，东北区、西北地区、港澳台地区和其他地区的相关性较小。东北地区和西北地区均属于中国的边境地

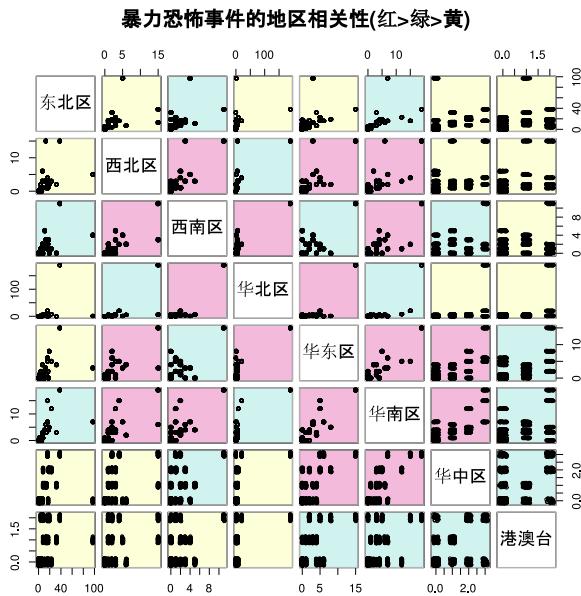


图 6: 暴恐事件空间传播变化规律

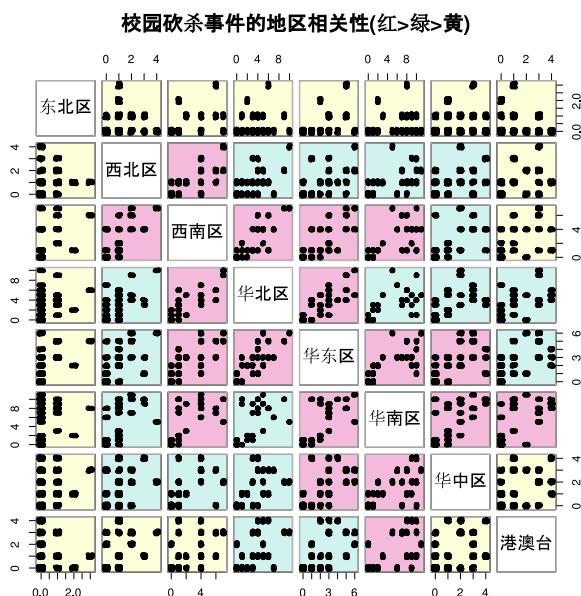


图 7: 校园砍杀事件空间传播变化规律

区，因此相关性与其他地区相关性较小，而港澳台地区由于经济体制与大陆不同，因此与大陆地区关联并不大。然而我们发现，西北地区与西南地区相关性较高，港澳台地区与华南地区相关性较高，这是因为地理位置较为临近所致，类似的有华中区与华南区、华东区相关性较高。

此外，西南区、华北区、华东区、华南区四省之间的相关性较高，而我们发现这四个地区有一些共同特征：每个地区都包含一个特大型城市，直辖市或者是经济高度发达城市，而且这些地区普遍经济水平在全国处于前列。

由以上分析我们得出如下结论：(1)边境地区危害

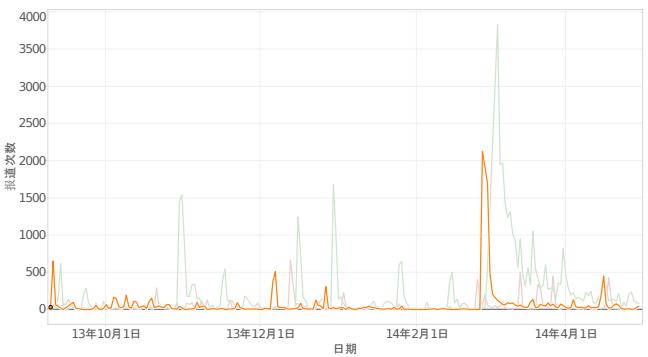


图 8: 公交车爆炸事件媒体传播随时间变化规律

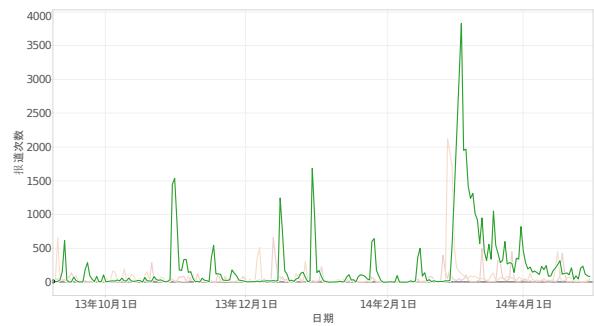


图 9: 暴恐事件媒体传播随时间变化规律

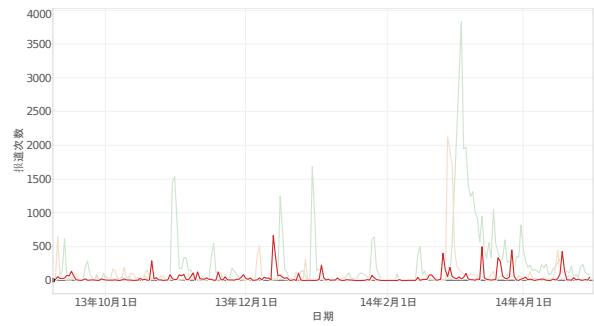


图 10: 校园砍杀事件媒体传播随时间变化规律

公共安全事件的发生与其他地区相关性较小；(3)地理空间上相邻地区相关性较高；(3)政治中心、经济中心危害公共安全事件的发生特点具有较高的相关性。

(3)新闻媒体传播：报导数、报导篇幅、媒体数、评论数。

首先我们做如下定义：当日媒体报导量超过200的危害公共安全事件称为大事件，而日媒体报导量小于200的事件称为小事件。

由图8~10中三类危害公共安全时间新闻媒体传播量随时间变化规律图，我们可以发现，在没有大事件发生时，往往在全国范围内很少有危害公共安全事件的发生，即使有也是程度很小的事件(日媒体报道量小于10)。然而当发生一起大事件时，新闻媒体会把这件事件以很快的速度传播到全国各地，而这种媒体的传播会带动同系列事件的发生，甚至会触发另一起大事件的发生。可见新闻媒体的传播对同系列事件的发生

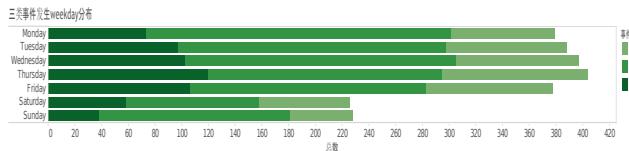


图 11: 三类事件weekday时间分布

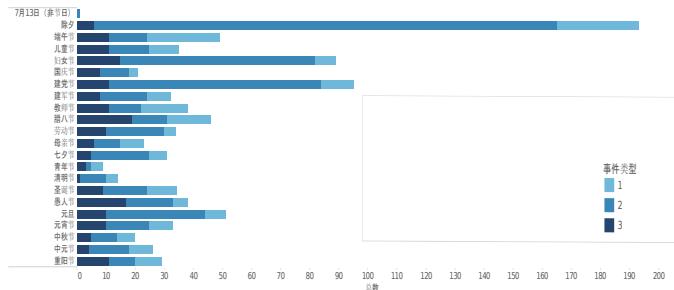


图 12: 三类事件节日分布

具有较大影响。

4.1.3 不同类事件共性分析

在上一节中，我们分析了同系列事件在时间、空间、新闻媒体传播上的触发关系，找到了一些规律。我们发现其实不仅是同系列存在这些触发关系，在不同系列事件的分布规律中，也会存在一些共性。这一节中，我们仍然从时间、空间、新闻媒体三个角度对三类危害公共安全事件进行分析，发现三类事件之间的分布规律共性，进而找到事件发生的影响因素。

(1)时间特征共性分析

从图11中我们可以看出三类危害公共安全事件均在工作日发生次数的较多，而在双休日发生次数的较少。而从节日分布的角度来看(图12)，元旦、除夕、建党节等均是三类危害公共安全事件的多发时段。

(2) 空间特征共性分析

从三类事件的发生次数空间分布图13~15中可以看出：公交车爆炸事件多发生于华东地区，包括山东、江苏、浙江以及福建、广东等省；校园砍杀事件多发生于西部边境省份，包括新疆自治区、云南省；校园砍杀事件则多分布于中国南方地区，河南、广东、江苏等省。

(3) 媒体特征共性分析

如图16所示，当某一系列一起特大事件(日报导量超过1000条的事件)发生时，在一周时间内往往伴随着不同系列事件的重大事件(日报导量超过500条的事件)发生。

4.2 公共安全事件可视化

为了更好地一体化展示事件发生的时间、地点、严重程度、事件类型，我们以动态网页的形式对三类公共事件从进行了可视化展示，如图17所示。

可视化展示的数据为2013.1~2014.4的15个月的公共事件数据。以周为单位进行数据展示，每三秒钟动态更新一次。可视化以颜色表示三类公共事件，绿色

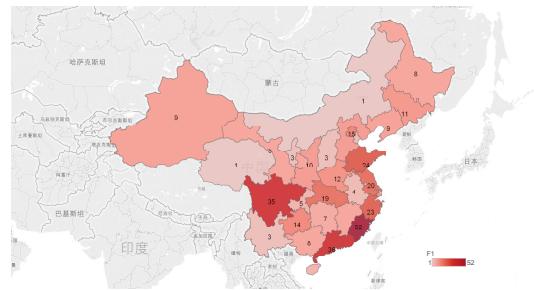


图 13: 公交车爆炸事件发生次数空间分布

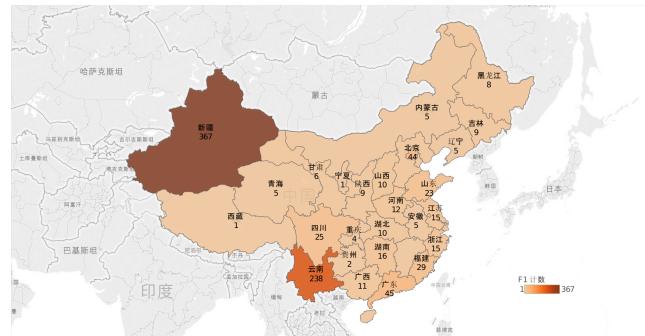


图 14: 暴力恐怖事件发生次数空间分布

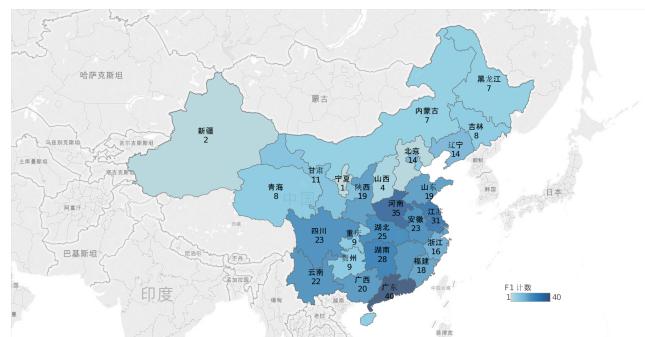


图 15: 校园砍杀事件发生次数空间分布

为公交车爆炸事件、黄色为暴恐事件、红色为校园砍杀事件；以圈大小表示事件的严重程度(用每个事件的总报道量表示)；每个圈出现时间表示媒体报导时间。

4.3 事件预测

了解公共危害事件的触发以及传播机理，找到事件间的影响关系和共性，最终的目的是为了抑制事件的发生，通过对事件可能发生的时间和地点进行准确预测，能够提前做好相应的预防措施（加强管制）和管控方案。本节提供一种事件预测方法，主要针对各区域（省）在未来一段时间，事件是否发生以及发生的次数进行预测。

事件预测的框架和数据流如图18所示，主要包括：特征提取，特征降维，分类与预测，结合事件提取的结果进行检验4个阶段。

4.3.1 事件预测整体思想

上节中我们通过关联分析的方法发现了同系列事

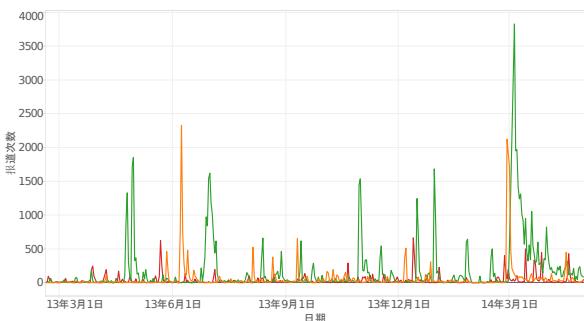


图 16: 三类事件的媒体报道量随时间的变化趋势



图 17: 公共安全事件的数据可视化

件在时间、空间、媒体三个维度的触发关系，以及不同系列事件之间的共性。得到了一些结论与规律。总结起来主要有以下规律：

(1) 某区域(省)内某类事件在当月是否发生以及发生的次数与同类事件在该区域(省)或全国在前期(近一个月内)是否发生与发生次数有较高的相关性，通常相近的事件段，事件的发生具有一定关联性。

(2) 某区域(省)内某类事件在当月是否发生以及发生次数与前期(近一个月内)，同类事件在全国发生的区域的分布相关，通常相近的区域事件的发生具有相同的趋势。

(3) 某区域(省)内某类事件在当月是否发生以及发生次数与前期(近一个月内)媒体对该类事件的报道以及社会舆论有一定关系(诸如媒体大规模报道、网民舆论传播带来的启发和情绪影响等)。

(4) 某区域(省)内某类事件在当月是否发生以及发生次数与当月的时间特征相关，通常事件的发生按年可能具有周期性，此外也可能与当月所涉及的重大节日相关联，通常节日前后也是事件的高发期。

(5) 某区域(省)内某类事件在当月是否发生以及发生次数与该区域(省)的空间地点特征相关，空间地点特征包括：该类事件在本区域内过去发生的频率，本区域的经济发展情况(通过GDP衡量)、人口数量、民族组成等。

因此，根据前文的分析，可以将某类事件发生的可能影响因素归为5大类：前期(δt 时间段内)时间(发生频率，距离上一次发生的时长)因素，对应规律(1)；

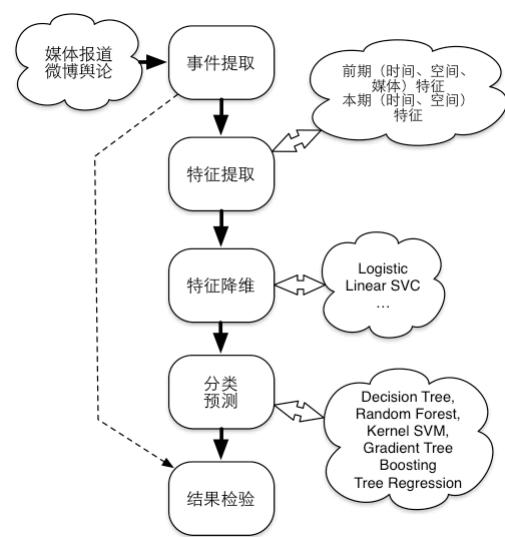


图 18: 预测整体框架

前期空间(事件发生点的空间分布)因素，对应规律(2)；前期媒体因素(媒体报道量，社会舆论情绪)，对应规律(3)；本期时间(月份、季节、是否包含重大节日)因素，对应规律(4)；本期空间(本地过往该事件发生的频率，经济水平，人口数量和民族组成)因素，对应规律(5)。上述5个因素也可分为两大类(如图19所示)，即前期时间(T_{t-1})、空间(S_{t-1})、媒体(M_{t-1})因素、以及本期时间(T_G)和空间(S_G)因素。

4.3.2 特征提取

(1) 前期特征

对前一个月本区域的信息，从时间、空间、语义(媒体)三个维度进行特征提取。

T_{t-1} : 前期时间特征，包括：前 Δt 时间内，本区域发生的事件数，全国发生的事件数，本区域前N个事件距离本期起始点的时距，全国前N个事件距离本期起始点的时距。

S_{t-1} : 前期空间特征，包括：前 Δt 时间内，事件发生在全国(省/地区)的分布，分布包括绝对次数和占比。

M_{t-1} : 前期语义(媒体)特征，包括：前 Δt 时间内，本地发生的各事件的媒体报道数、报道篇幅、报道媒体的总数，微博中对该事件评论的总数，总人数，回复数，转发数，点赞数，以及内容中包含正、负情感的词数的占比等。

(2) 本期特征

T_G : 本期时间特征，包括：预测期所对应的月份、季节、是否包含重大节日，包含重大节日的数量等；

S_G : 全局空间特征，历史上本区域发生该事件发生的频率，本区域的经济水平(GDP指数，人均GDP指数，GDP增幅)，人口数量，及民族组成(少数民族占比)等。

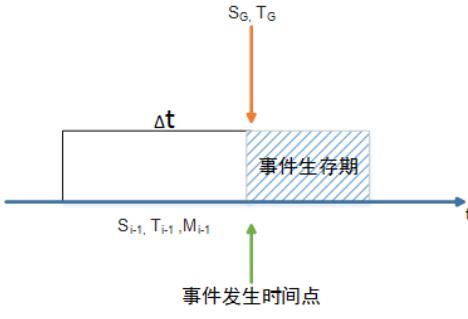


图 19: 事件预测的特征维度

4.3.3 机器学习算法

由于事件的发生具有离散性的特点，因此，针对某区域、某时间段内事件的发生，我们主要对事件是否发生、事件发生频次，这两个指标进行预测。

(1) 事件是否发生预测

事件是否发生的预测问题，即：类别1（事件发生）、类别2（事件不发生），属于标准的二类分类问题。我们主要采用了包括Decision Tree（决策树）、Random Forest（随机森林）、SVM（支持向量机）、Gradient Boosting（梯度提升决策树）等多种方法，并对其预测的准确率进行比较。

决策树是以实例为基础的归纳学法，它着眼于从一组无次序、无规则的实例中推理出以决策树表示的分类规则。构造决策树的目的是找出属性和类别之间的关系，用来预测将来未知类别记录的类别。它采用自顶向下递归的方式，在决策树内部节点进行属性比较，并根据不同属性值判断从该节点向下的分支，在决策树的叶节点得到结论。主要的决策树算法有ID3、C4.5等。

随机森林是包含多个决策树的分类器，输出的类别由某些树输出的类别的众数而定。它利用自助重抽样的方法从原始样本中抽取多个样本，对每个自助重抽样样本进行决策树建模，然后对多棵树的预测结果进行组合，通过投票得出最终预测结果。大量的理论和实证研究都证明了随机森林具有很高的预测准确率，对异常值和噪声具有很好的容忍度，且不容易出现过拟合。

支持向量机通过寻求结构化风险最小的方案来提高学习机泛化能力，实现经验风险和置信范围最小化，其原本是针对线性可分情况进行分析，对于线性不可分的情况，可以通过使用非线性映射算法将在低维特征空间线性不可分的样本转化到高维特征空间使其线性可分，从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能。

梯度提升决策树是一种提升算法，它主要的思想是，每一次模型都建立在之前建立模型损失函数的梯度下降方向。损失函数描述模型的不靠谱程度，损失函数越大说明模型越容易出错。如果模型能够让损失

函数持续下降，说明模型在不停改进，而最好的方式就是让损失函数在其梯度的方向上下降。

实验证明，Gradient Boosting（梯度提升决策树）效果最好，因此将其作为我们的最终预测算法。

Gradient Boosting（梯度提升决策树）算法的伪代码如下，其中， $F_k(x)$ 是函数估计值，迭代次数是 M ，每次迭代根据预测准度计算损失函数的梯度，进而更新估计函数。

```

Algorithm
 $F_{k0}(\mathbf{x}) = 0, \quad k = 1, K$ 
For  $m = 1$  to  $M$  do:
   $p_k(\mathbf{x}) = \exp(F_k(\mathbf{x})) / \sum_{l=1}^K \exp(F_l(\mathbf{x})), \quad k = 1, K$ 
  For  $k = 1$  to  $K$  do:
     $\tilde{y}_{ik} = y_{ik} - p_k(\mathbf{x}_i), \quad i = 1, N$ 
     $\{R_{jkm}\}_{j=1}^J = J$  terminal node tree( $\{\tilde{y}_{ik}, \mathbf{x}_i\}_1^N$ )
     $\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{\mathbf{x}_i \in R_{jkm}} \tilde{y}_{ik}}{\sum_{\mathbf{x}_i \in R_{jkm}} |\tilde{y}_{ik}|(1-|\tilde{y}_{ik}|)}, \quad j = 1, J$ 
     $F_{km}(\mathbf{x}) = F_{k,m-1}(\mathbf{x}) + \sum_{j=1}^J \gamma_{jkm} \mathbf{1}(\mathbf{x} \in R_{jkm})$ 
  endFor
endFor
end Algorithm

```

图 20: Gradient-Boosting 算法伪代码

(2) 事件发生频次

对于事件发生次数的预测问题，即连续值的预测问题，我们主要采用了回归树的方法对其进行预测。

分类回归树类似于决策树，在计算过程中充分利用二叉树的结构，即根节点包含所有样本，在一定的分割规则下根节点被分割为两个子节点，这个过程又在子节点上重复进行，成为一个回归过程，直至不可再分为叶节点为止。预测阶段类似与分类过程，利用对应类别的样本回归计算得到预测值。

按照要求需要给出三类事件（任选一类）的发生数量和发生地区的概率。发生数量已有上述方法给出，发生地区的概率可通过该地区的预测发生数在全国各省预测总数的占比计算。

5 实验对比与分析

5.1 事件提取算法评估

抽样选取个事件($n = 1000$)，包含事件ID、新闻ID、新闻标题、导语、内容等信息。对每一个事件，实际包含该事件的所有新闻集合为 U_i ，已标注的为事件*i*的新闻集合为 E_i ，则 $E_i \cap U_i$ 为已标注为事件*i*的新闻中标注正确的集合， $U_i - E_i$ 漏标注的新闻集合。 $(|A|$ 表示几个A所含元素的数目)。我们定义误报率和漏报率为：

误报率(Error Report Rate, ERR):

$$ERR = \frac{\sum_i |E_i - E_i \cap U_i| / |E_i|}{n} \quad (4)$$

漏报率(Missing Report Rate, MRR):

$$MRR = \frac{\sum_i |U_i - E_i| / |U_i|}{n} \quad (5)$$

在我们的实验中，所得三类事件的误报率和漏报率如表2所示。

表 2: 事件提取算法针对三类事件的误报率和漏报率

事件类型	误报率	漏报率
公交车爆炸事件	14.28%	12.09%
暴恐事件	12.39%	14.05%
校园砍杀事件	14.10%	11.54%

表 3: 三类事件的误报率和漏报率

评估算法	评估算法	预测频次误差
常规算法	64.50	0.8956
Leave-one-out算法	82.34	0.5250
K-Fold算法	82.34	0.5234
滑动窗口	75.27	0.5525

5.2 预测算法评估

针对事件是否发生和发生次数两类预测值，分别采用准确率和平均绝对误差评估。

正确率 $Accuracy = \frac{N_{right}/N_{total}}{\times} 100\%$ ，其中， N_{right} 为预测正确的次数， N_{total} 表示预测总次数。

平均绝对误差 $AbsErr = \frac{1}{N} \sum |n_{pred} - n_{obs}|$ ，其中， n_{pred} 表示预测发生的次数， n_{obs} 表示实际发生次数。

我们使用了4种数据评估算法：

(1) 常规算法通过前期历史数据预测最后三个月每月各省时间是否发生以及发生次数。

(2) Leave-one-out 算法

由于本文研究的是相邻时间段（即 τ_n 和 τ_{n-1} ）内事件发生数与前期时间、空间、媒体因素以及当期时间、空间因素的关联性，并不考虑预测期的绝对时间 τ_n 的影响，因此可采用Leave-one-out算法将其他数据均做为训练数据，依次预测各省各时间段内的事件发生和数量，结果取均值。

(3) K-Fold 算法

类似于Leave-one-out算法，将数据分为 $K (=10)$ 份，每次将其中 $K - 1$ 份作为训练集，剩余的1份作为测试数据，依次预测 K 次，整体循环 N 次，结果取均值。

(4) 滑动窗口

类似与常规算法，由于本文所考察的关联性可能随时间变化，因此仅采用预测期（3个月）的前一小段时间即窗口期（6个月）进行训练，窗口不断滑动，依次预测出结果，结果取均值。

本文预测系统可针对3类不同事件分别进行预测，按照要求任选一类（校园砍杀事件）提供预测自评结果如下。

四种评估方法差异较大，其中，常规算法预测准确率较低的原因在于，相邻时间段的相关性可能随时间变化，常规算法采用除预测期其余全部数据训练，但仅预测了最后3个月，然而最后3个月规律可能和之前差异较大，因此，预测准确率较低，Leave-one-out算法和K-Fold算法预测采用了全部数据进行评估，因而能够更加准确全面的反应出预测方法的效果。

6 结束语

本文首先针对原始数据的不足及进行了数据预处理工作，包括数据的清洗、修正、融合。之后基于TF-IDF结合余弦相似算法的事件提，并通过开放数据人工标注的方法修正算法。创新点有如下几点：

(1) 采用ckan开放数据平台，用众包的方式进行人工标注事件。网页提交表单时需要输入验证码，如果能把人工标注事件以验证码的方式众包，将会使这部分资源有效利用，而开放数据平台为众包提供了基础，使之成为可能。

(2) 采用多源数据融合的方法进行数据分析。未来我们会在数据可视化方面做进一步的工作。

参考文献

- [1] <http://ckan.org/>.
- [2] <http://zh.wikipedia.org/wiki/TF-IDF>.
- [3] Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; Sabeti, P. C. (2011). "Detecting Novel Associations in Large Data Sets". *Science* 334 (6062): 1518 – 1524.
- [4] 孔秋强,贺前华. 基于TFIDF与分类树的工程文本信息分类法[J].计算机应用与软件.2014.
- [5] 李巍,孙涛,陈建孝,罗梓恒,李雄飞.基于加权余弦相似度的XML文档聚类研究[J].吉林大学学报.2010.
- [6] 洪娜,钱庆,范炜,方安,王军辉.关联数据中关系发现的可视化实践[J].现代图书情报技术.2013.
- [7] 张宁.统计数据的可视化关联分析[J].统计与决策.2012.

附录

A

该研究的数据集、源代码、及可视化链接:

- 原始数据存储CKAN: <http://202.121.178.242/dataset/ccfbdb>
- 代码管理github: git@github.com:happyjane/CCFBGDG_2014
- 可视化链接: <http://data.sjtu.edu.cn:1111/>

B

表 4: 时间特征

特征名称	特征意义
day_offset	事件报道时长
duration	事件报道时长
holiday	事件发生所在节日
weekday	事件发生于周几

表 5: 空间特征

特征名称	特征意义
heppen_city	事件发生城市
happen_province	事件发生省份
happen_area	事件发生地区(如华东区)
city_gdp_ranking	所在城市gdp排名
2013_gdp	城市2013年gdp总量
gdp_increase_from_2012	城市gdp增长率(2013)
province_gdp_ranking_2013	省份gdp排名
province_gdp_2013	省份gdp总量
avg_province_gdp_ranking_2013	省份人均gdp排名
province_pnum_2013	省份人口总数
avg_province_gdp_2013	省份人均gdp量
province_hanzu_ratio	省份汉族人口比例

表 6: 语义特征

特征名称	特征意义
wb_cnt_num	某事件在微博中传播到的人数
wb_person_num	事件发生省份
wb_repeat_num	某事件在微博中被重复关注的次
wb_loc_num	某事件的关注人中的地点分布数
total_news	某事件新闻报道总数
total_weibo	某事件发微博总数
media_total	某事件新闻报道(独立)媒体总数
media_origin	某事件新闻报道(来源)媒体总数
comments_news	某事件新闻评论数
comments_weibo	某事件微博评论数
quotes_news	某事件新闻转发数
quotes_weibo	某事件微博转发数
attitudes_news	某事件新闻点赞数
attitudes_weibo	某事件微博点赞数
words_med_news	某事件新闻报道长度中值
words_med_weibo	某事件微博长度中值
words_mean_news	某事件新闻报道长度均值
words_mean_weibo	某事件微博长度均值
pos_eval	新闻正情感词频数
pos_emo	正情感词频
pos_ntusd	正情感词频数
neg_eval	负情感词频数
neg_emo	负情感词频数
neg_ntusd	负情感词频数