The Analysis of Natural Disasters in China from 1998 to 2016 based on Text Mining

Xiao Liu[1], Haixiang Guo [1,2,3*], Yijing Li[1], Chunmiao Yang[1,2]

1 School of Economics and Management, China University of Geosciences, Wuhan 430074, China

2Research Center for Digital Business Management, China University of Geosciences, Wuhan 430074, China

3Mineral Resource Strategy and Policy Research Center of China University of Geosciences (WUHAN), Wuhan 430074, China

基于文本挖掘的历史自然灾害事件分析（1998-2016）

刘晓 [1,2]，郭海湘 [1,2,3*]，李诒靖 [1,2]，杨春苗 [1,2]

1. 中国地质大学经济管理学院，湖北武汉 430074；

2. 中国地质大学数字化商务管理研究中心，湖北武汉 430074；

3. 中国地质大学中国矿产资源战略与政策研究中心，湖北武汉 43007

Abstract

This paper takes the 248 disaster events happened in China from 1998 to 2016 as samples, uses text mining, descriptive statistics, association rules and other methods to implement event extraction for disaster events in history, and then analyzes characteristics of the disasters including type, location and time. The results show that Flood, earthquake, landslide, typhoon and rainstorm are found to be the most frequent natural disasters in China, according to the analysis of disasters time .most of disasters were occurred during May to August, specially flood and heavy rain. In the end, the paper analyzes the spatial distribution of disasters .The results show a significant difference of the geographical locations among different types of disasters. And Yunnan is the region where the most times of disasters happened. In addition, the spatial association of disaster events is also analyzed. The result of this study will offer insight into the identification of disaster risks and carrying out the disaster prevention and mitigation.

Keywords: Natural disasters, text mining, spatial association, disaster distribution

摘要

本文以我国 1998-2016 年发生的 248 次灾害事件为研究对象，运用文本挖掘、描述统计、关联规则等方法对我国历史的灾害事件进行了事件抽取，对自然灾害的类别、发生地点和发生时间等特征进行了分析。研究发现洪水、地震、滑坡、台风和暴雨是发生最频繁的五种自然灾害。灾害的发生时间集中在 5-8 月份，不同灾害发生的地理位置存

基金项目：国家自然科学基金资助 项目编号：71103163， 71573237；教育部新世纪优秀人才支持计划 项目编号：NCET-13-1012；教育部人文社会科学研究规划基金资助 项目编号：15YJA630019 中央高校基本科研业务费专项资金资助 项目编号：CUG120111，CUG110411，G2012002A，CUG140604，CUG160605ceyn

***通讯作者**：郭海湘（1978~），男（汉族），湖南湘乡市人，教授，博士，博士生导师，主要从事软计算、复杂系统模拟与决策研究， email: faterdumk0732@sina.com

在差异，云南是发生灾害次数最多的地区。另外，对灾害事件的空间关联关系进行了分析，分析发现灾害在福建、江西等临近地区表现了一定的地理相关性，比如福建、江西经常同时发生暴雨灾害等。研究结果对识别我国灾害风险、开展防灾减灾工作具有一定的参考价值。

关键字：自然灾害, 文本挖掘, 空间关联, 灾害分布

1. 引言

随着全球气候变化加剧以及人类社会活动的影响，灾害呈现出成灾条件复杂、频率高、强度大等特点[1]。灾害往往给人类及赖以生存的环境造成严重的破坏。据瑞士再保险报告，仅 2015 年全球发生灾害 353 起（自然灾害 198 起），经济损失高达 920 亿美元。而我国是一个自然灾害频发的国家，南涝北旱、地震、滑坡、泥石流等多种自然灾害使人们生活受到严重影响，社会经济损失惨重。2015 年由于各类自然灾害使得全国 18620.3 万人次受灾，直接经济损失 2704.1 亿元。因此，识别各种自然灾害风险，科学应对各种灾害及由灾害引发的一系列连锁反应并及时、有效地处理，减轻灾害带来的损失，已成为当前我国及国际社会面临的重大挑战问题之一。

自然灾害的发生与地理位置、天气状况等有紧密的联系，因此具有一定的时间、空间分布规律性。通过对历史灾情的分析，挖掘出各类自然灾害发生的特点，及早进行预防和应对，从而减少灾害造成的损失。然而，现有的研究和防灾减灾机构仅对每年的灾害数据进行了统计分析，并且往往是以国家为单位的灾害数据也基本上只包括对灾害事件的汇总，并没有对每个灾害事件的介绍。另外，传统的统计方式存在数据滞后性，只是简单的静态统计，无法反映历史灾害的变化趋势和分布特点。我国目前还没有比较完善的灾害数据库，获取灾害数据显得尤为困难。这些问题都使得动态分析我国历史灾害特征很难进行。随着大数据时代的到来，web 挖掘成为数据获取的潮流，通过对亚洲减灾中心网站（http://www.adrc.asia/）数据的获取，本文应用文本挖掘技术、统计分析及关联规则挖掘等方法，通过对我国 1998-2016 年 248 起自然灾害事件信息的提取、分析，挖掘各类自然灾害发生的频率、时间和空间分布规律。为预防和应对各类自然灾害提供参考。

2. 文献综述

随着互联网和社交媒体的迅速发展，网络和社交媒体逐渐成为一种获取信息的重要资源。企业、研究机构和政府都开始关注社交媒体的大数据，并尝试用数据挖掘、文本挖掘、机器学习等方法去挖掘重要的信息。比如在线用户评论[2]、网络舆情分析[3]、票房预测[4]等。

在突发事件领域，比如自然灾害和突发公共事件，由于消息的滞后性，以往的消息报道都需要通过现场记者或政府新闻发布会获取信息。现在借助社交媒体平台，可以及时的了解事件的进展、受灾人员情况、受灾影响等一些信息。国内外有一些学者已经针对突发事件发生时 Twitter 等社交媒体的信息进行获取、挖掘，获得有价值的信息。Middleton(2014)搭建了能够获取实时灾害地图的平台，该方法能够通过社交媒体数据对灾害的发生地点进行提取，从而绘制灾害危机地图[5]。Sakai（2015）结合贝叶斯分类、时空聚类和脉冲检测技术提出了一种能及时识别和分析地区突发事件主题的方法，并用该方法通过 Twitter 数据准确挖掘出了天气的主题[6]。Takahashi（2015）以菲律宾的台风为例，分析了不同的 Twitter 用户在灾害发生时和发生后的信息发布，发现不同类型的用户主要用社交媒体获取二手报道信息和协调救灾等[7]。国内有王昊（2012）对日本发生地震后，以新浪微博上的一周讨论为数据，对地震事件的主题词和用户情感进行了分析[8]。张宇（2015）以上海外滩事件为研究实例，通过对突发事件发生之后政务微博的发布形式、响应时间、响应速度、微博内容以及微博交互情况进行分析，对突发事件中政府信息发布和应对能力做出分析和评价，最后为突发事件中政府提高）信息发布的有效性提出具体的方法和建议[9]。

现有的基于社交媒体和网络的灾害信息挖掘基本都是对某一具体事件的分析。很少有对多种灾害的综合分析，为了全面分析历史灾害的特征和分布规律，需要对所有的历史灾害

事件进行信息提取，考虑到对多个灾害事件的进行提取时，微博等社交媒体的数据过大，信息从灾情、应灾、救灾到恢复建设都有涉及，内容重复性过高，数据处理费时。因此本文选取亚洲减灾中心网站的我国 1998-2016 年的灾害数据进行了挖掘和分析。

3. 数据来源及信息提取

3.1 数据来源

本文数据来源于亚洲减灾中心（Asian Disaster Reduction Center, ADRC）,ADRC 是 1998 年在日本成立的，该网站收集并提供了其 26 个成员国的自然灾害信息。本文对该网站的中国 1998-2016 年的灾害信息进行了检索（截止到 2016 年 5 月 8 日），并用软件八爪鱼采集器获取了所有的灾害信息，一共 248 起自然灾害。该网站每条灾害信息都是通过列表形式记录的，属于半结构化数据，主要有灾害发生时间、灾害名称、概要（outline）、人员伤亡情况等信息。由于概要和人员伤亡情况是文本形式，需要借助文本挖掘方法进行预处理，提取出诸如灾害地点、人员伤亡等灾害信息进行灾害分析。因此，本文的数据预处理主要是通过对文本进行自然语言处理，去掉一些噪声数据，提取出特征矩阵。并在文本预处理的基础上从文本中进行命名实体识别和基于事件框架的灾害事件抽取，从而获得灾害事件的完整结构化数据，便于进行历史灾害的分析。

3.2 文本预处理

文本预处理是通过去除标点、大小写字母转换、去除停用词等一系列操作将非结构化文本转化成可用于数据挖掘的结构化形式。一般用自然语言处理工具对文本进行预处理，目前国内外有很多成熟的针对中英文文本的自然语言处理工具。综合考虑各种工具的特点和适用性，本文主要用 R 语言的 tm 包和 Python 的 NLTK 包进行文本预处理。下面介绍了主要的文本预处理流程。

（1）去除停用词（Stop words）

主要是指在文本中高频率出现但对表达内容没有实际意义的介词、冠词和连词等，比如 the ,a , at 之类的。需要去除这些噪声词以免给关键词提取造成干扰。

（2）词根提取（Stemming）

与中文表达不同，英文中表示同一个名词的单词可能存在单复数情况，比如“flood”和“floods”表示的都是洪水。应该算一个关键词。因此需要对文本进行词根提取，即将单词缩减为词根形式，主要通过去除”s”,去除”ing”加“e”等固定算法进行。

（3）特征词抽取

文本的基本单位成为文本的特征或特征项，这里指的是文本中的单词，虽然对文本进行了以上操作，但是文本的特征词还是很多，如果把所有的特征词作为文本的中间表示，特征词的向量过大，从而导致计算量大，使得文本分类和聚类等技术都无法完成，因此需要在不破坏文本主要内容的情况下减少特征词，提取出关键的能代表文本内容的特征词。本文用的是 TF-IDF 算法[10]进行词频统计和特征词提取。操作层面用的是 R 语言的 tm 包。通过对 248 条灾害文本提取关键词。并对关键词进行了相关分析。

3.3 灾害事件地理信息提取

信息抽取（Information extract, IE）是文本挖掘中最为关键的技术，通过对非结构化文本的处理，从而获得可分析的结构化数据。这些数据一般是以表格形式存在的，便于数据分析师和学者进行研究。信息抽取将文本表示为实体和框架的组合。最简单常用的信息抽取为实体抽取，即命名实体识别，命名实体识别(Named Entities Recognition, NER)是自然语言处理(Natural Language Processing, NLP)的一个基础任务。其目的是识别文本中人名、地名、组织机构名等命名实体。命名实体是命名实体识别的研究主体，一般包括 3 大类(实体类、时间类和数字类)和 7 小类(人名、地名、机构名、时间、日期、货币和百分比)命名实体。

命名实体识别方法主要分为 3 类：基于规则的方法、基于统计的方法、规则与统计相结合的方法[11]。统计方法中常用的有最大熵模型、隐马尔可夫模型、条件随机场等。命名实体识别有很多比较成熟的工具，比如英文的有 Gate、Yooname 等软件，本文用斯坦福大学的基于条件随机场的 NER 软件包在 Python 软件环境下

对自然灾害事件发生的地点进行了抽取。主要抽取流程如下：(1) 将预处理后的文本进行词性标注，即确定单词是名词、动词、形容词还是副词，方便后面进行信息抽取，这部分可以在 python 的 nltk 模块下完成；(2) 命名实体识别，目前比较流行的命名实体识别是利用统计分类器来进行，即分类器分析句子中的每个单词，确定该词处于命名实体的哪个部分，通过训练和预测，就可以使用该分类器识别命名实体。本文用斯坦福大学的命名实体识别软件包 NER 在 python 环境下进行地点实体识别。

4. 结果分析

通过文本挖掘技术提取出 248 条灾害信息的时间、名称、地点、伤亡人数等信息，转换成结构化数据保存，并用统计分析和关联规则对灾害信息进行分析和挖掘。

4.1 主要灾害种类与灾害发生的时间分布

由于有些灾害不是单一发生的，比如台风和暴雨、洪涝和滑坡。灾害信息中如果是同时发生的灾害，标题直接是同时发生的灾害名称，因此本文通过标题统计各类灾害出现的次数，进而计算了各类灾害的发生频率，如图 1 所示，其中的其他灾害包括火山、寒潮、热浪等一些不常见的灾害。另外，为了分析各类灾害造成的影响，本文以灾害造成的死亡人数为指标，进行了分析。在 248 条灾害信息中，有死亡人数的信息有 158 条，对这 158 条信息中同类灾害的死亡人数进行求和（见表 1）。

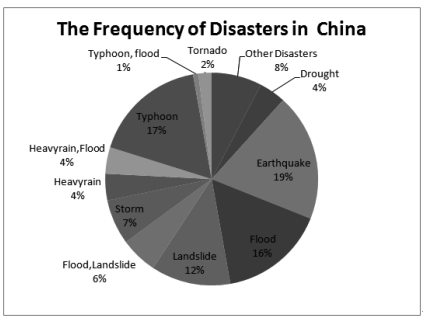

图 1　1998 年-2016 年我国各类自然灾害发生频次图

我国在 1998-2016 年主要发生的自然灾害是洪水、地震、台风、滑坡和暴雨。大概占到灾害总数的 80%。为了进一步分析各类自然灾害的发生频次，对各类自然灾害进行了统计（如图 1 所示），从图中可以看出，单独发生的灾害事件中，地震灾害的发生次数最多，占灾害总数的 19%。其次是台风、洪水和滑坡。另外，灾害的发生往往并不是独立的，一种灾害的发生可能引起其他灾害的发生，这样一种灾害的发生引发一系列次生灾害和衍生灾害，呈现多米诺骨牌效应，称之为灾害的多级联动[12]。比如暴雨灾害的发生，引起了洪水灾害，洪水灾害可能引起滑坡、泥石流灾害。台风灾害也可能引起洪水灾害。这种灾害多级联动的情况占到了 11%。可见灾害的联动应该引起政府和防灾减灾机构的重视，及时采取应对措施，预防灾害的多级联动，从而最大程度的减少人员伤亡和经济损失。

表 1 不同灾害的死亡人数总计

Disaster	Number of dead	Frequency
Earthquake	77040	19%
Flood	1952	16%
Landslide	1207	12%
Typhoon	550	17%
Tonado	475	2%
Storm	304	7%
Flood,landslide	293	6%
HeavyRain,Flood	115	4%
HeavyRain	108	4%
other disasters	237	13%

从表 1 中可以看出，以死亡人数作为衡量灾害影响程度的指标，灾害的影响和发生频率基本上是正相关的，地震的发生频率最高，死亡人数也是最多的。其次是洪水、滑坡和台风。龙卷风虽然发生频率很低，但是其灾害影响程度不低。另外，还有风暴及一些并发灾害，从表中可以看出，灾害影响程度较大的是地震、洪水、滑坡、台风、龙卷风。因此这些灾害预警或防范时，应注意及时疏散人群，平时应加强这些灾害频发区的人员紧急疏散演练、宣传灾害防御知识，以最大限度地减少灾害发生时的人员伤亡。

一般而言，灾害的发生与天气有着密不可分的关系，而天气是随时间变化的，那么灾害事件的发生是否呈现一定的时间分布规律呢？

为了探讨这个问题，我们对灾害事件的发生时间进行了月份抽取，即分析不同灾害的时间分布特征。如图 2 所示，5-8 月是灾害发生的高峰期，尤其是洪水、台风、滑坡、暴雨都呈现“两头低，中间高”的分布趋势，因为台风、暴雨等灾害与天气有着紧密的联系，从图 2 中可以看出，暴雨灾害发生频率最高是在 5 月，滑坡发生次数最多的是 6 月，洪水灾害发生次数最多的时间是 7 月，而台风灾害发生次数最多的是在 9 月。这四种灾害的发生的时间集中，因此这段时间应该加强灾害监测预警，通过防灾减灾宣传、加强基础设施建设等方案提高洪水、滑坡等灾害高发区的防灾减灾能力，减少不必要的损失。另外，相比较而言，地震的时间分布趋势比较平稳，可以得出与天气的变化不大。但我国是地震的频发地区，2008 年的“汶川地震”，2010 年的“玉树地震”，2013 年的“雅安地震”造成了巨大的社会影响和经济损失，因此，提高地震监测预警能力，加强城镇的防震减灾能力，尤其是地震频发区，还要关注相关机构应对灾害的应急管理能力。

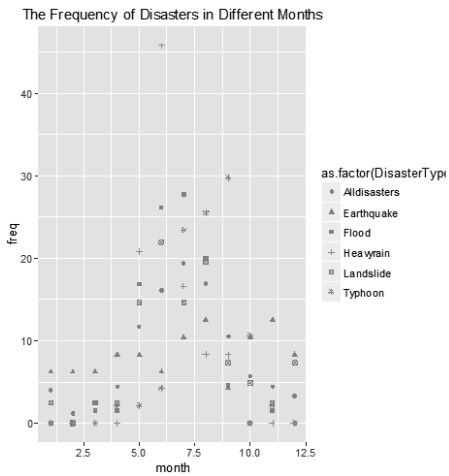

图 2 不同灾害类型的时间分布

4.2 不同灾害的地理分布特征

自然灾害具有一定的区域特点，由于天气、地理位置的差异，南方降雨较多，一般易发生暴雨、洪水灾害，北方则易发生干旱灾害等。但所说的“南方”和“北方”具体界定，一般都是气象专家根据历史天气变化得出的结论，一般人们都只是有个模糊的概念。因此，本文通过对我国近 18 年的灾害发生的地理位置信息抽取，分析不同灾害的频发区，为不同地区制定防灾减灾方案提供参考依据。

1998 年-2016 年的灾害主要分布在云南、四川、福建、广东、贵州、湖南等地区，云南发生灾害的次数最多，约占灾害总数的 13%。四川的灾害发生次数也超过了 10%。政府和防灾减灾机构应高度重视这些地区防灾减灾设施的建设、提高这些地区的承灾能力，最大限度地减少灾害风险。

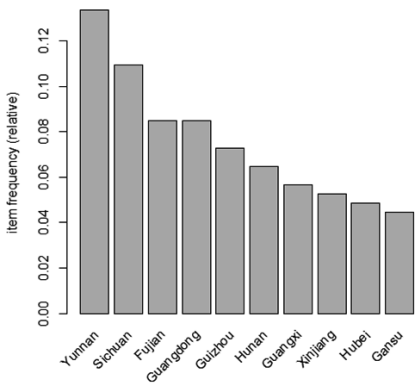

图 3 灾害的发生频率最高的十大地区

为了进一步分析不同灾害的空间分布，把频发的地震、洪水等五种灾害的地理位置进行了统计，根据每个地区的发生次数进行排序，取前八个地区的数据，如表 2 所示。可以分析得出，地震主要集中在云南、新疆、四川、西藏和甘肃等地，主要是我国西部地区，特别是西南地区。而洪水集中在四川、湖南、广东等地，暴雨主要集中在贵州、湖南、广西等，台风频发区是福建、广东和海南。从表中可以看出，有些地方不止是一种灾害的频发区，比如云南地震和滑坡灾害发生次数都是最多的，四川的洪水和滑坡灾害频次也很大。而地震和滑坡、洪水和滑坡都有很大的关联性，有可能是灾害的多级联动产生的。因此这些地区应特别关注灾害的多级联动性，一旦发生地震或洪水灾害，应及时监测预警、做好滑坡等次生灾害发生的防护工作，减少灾害联动造成的损失。

表 2 灾害易发区的灾害次数统计表

省份	地震	省份	洪水	省份	滑坡
云南	21	山西	3	台湾	4
新疆	10	湖北	3	广西	4
四川	8	云南	9	甘肃	1
西藏	5	四川	8	江苏	1
甘肃	4	贵州	5	贵州	4
内蒙古	2	福建	4	湖南	4
广西	1	广西	4	广西	3
湖北	1	湖南	3	湖北	2
四川	11	广东	3	江西	2
湖南	8	湖北	3	四川	2
广东	6	福建	10	北京	1
江西	5	广东	10	重庆	1
贵州	5	海南	7		
重庆	3	浙江	6		

4.3 灾害事件的空间关系挖掘

由于地理的相近等原因，不同灾害在空间上具有一定的相关性，空间上一个地区发生了自然灾害，临近的地区可能也会发生相同的灾害。为了探讨灾害发生地点的关联性，本文采用关联规则挖掘算法对 248 条灾害事件的地理位置进行了关联分析。

关联规则是 Agrawal 在 1993 年提出的[13]，是数据挖掘中的根本任务之一，该算法的目的是在数据项目中找出所有的并发关系，即关联关系。比较经典的关联规则案例就是购物篮。目的是找出顾客所选购商品之间的关联关系。关联规则挖掘算法的问题描述：设 $I=\{i_1,i_2,i_3\ldots,i_m\}$ 是一个项目集合，$T=(t_1,t_2,t_3,...t_n)$ 是一个事务（数据库）集合，其中每个事务 t_i 是一个项目集合，并满足 $t_i \subseteq I$。本文中每个事务代表一条灾害事件。

一条关联规则的表达形式如下：

$X \to Y$, 其中 $X \subset I, Y \subset I$, 且

$$X \cap Y = \varnothing \qquad (1)$$

X（或 Y）是一个项目的集合，称为项集，并称 X 为前件，Y 为后件。本文中 X 和 Y 代表的是灾害发生地点的集合。

支持度（Support）和置信度（Confidence）是两个常用的衡量关联规则强度的指标。

（1）支持度：规则 $X \to Y$ 的支持度指的是事务 T 中包含 $X \cup Y$ 的百分比，表示的是规则在事务集合 T 中使用的频繁程度，支持度的计算如下：

$$Support = \frac{(X \cup Y).count}{n} \quad X \cup Y \qquad (2)$$

其中，n 为事务的数目，在本文中指的是灾害事件的条数。

（2）置信度：既包含了 X 又包含了 Y 的事务的数量占所有包含了 X 的事务百分比，决定了规则的可预测程度，计算如下：

$$Confidence = \frac{(X \cup Y).count}{X.count} \qquad (3)$$

类似于购物篮，本文中把每条灾害事件看作一条购物清单（即事务），每条灾害事件发生的地点看作购物清单上的商品，从而找出灾害地点之间的关联关系。目前关联规则算法有很多，Apriori 算法[14]、FP-grown 算法[15]等等，Apriori 和 FP-growth 的主要区别在于 Apriori 需要多次重复扫描数据库，产生大量候选频繁子集，从而使得计算时间和空间复杂度较大，而 FP-growth 算法通过生成频繁树结构生成频繁集，无需产生候选子集，减少了扫描数据库的次数，从而提高计算性能。由于本文中的数据量较小，因此两个算法的计算效率不会相差太大。本文用经典的 Apriori 算法对灾害地点进行关联规则挖掘。Apriori 算法是基于频繁项集的关联规则挖掘算法，先是生成所有的频繁项目集，然后对频繁项目集进行合并和剪枝，最后生成可信的关联规则。操作上用 R 语言的 arules 包进行。由于数据样本比较少，而且不同灾害的发生地点是有差异的，因此地点出现的频次不是很高。所以本文设置的支持度偏小，设为 $Support=0.01$，$Confidence=0.8$ 挖掘的关联规则如下表。

表 3 灾害事件的空间关系挖掘结果

Rules	Support	Confidence	Lift
{Guangxi,Jiangxi} => {Guizhou}	0.016194	1	13.72222
{Guangxi,Guizhou} => {Jiangxi}	0.016194	1	22.45455
{Guizhou,Jiangxi} => {Guangxi}	0.016194	1	17.64286
{Guangxi,Jiangxi} => {Fujian}	0.016194	1	11.7619
{Fujian,Guangxi} => {Jiangxi}	0.016194	1	22.45455
{Fujian,Jiangxi} => {Guangxi}	0.016194	0.8	14.11429
{Guangxi,Guizhou} => {Fujian}	0.016194	1	11.7619
{Fujian,Guangxi} => {Guizhou}	0.016194	1	13.72222
{Fujian,Guizhou} => {Guangxi}	0.016194	1	17.64286
{Hunan,Jiangxi} => {Fujian}	0.016194	0.8	9.409524
{Fujian,Hunan} => {Jiangxi}	0.016194	1	22.45455
{Fujian,Jiangxi} => {Hunan}	0.016194	0.8	12.35
{Guizhou,Jiangxi} => {Fujian}	0.016194	1	11.7619
{Fujian,Jiangxi} => {Guizhou}	0.016194	0.8	10.97778
{Fujian,Guizhou} => {Jiangxi}	0.016194	1	22.45455
{Guangxi,Guizhou,Jiangxi} => {Fujian}	0.016194	1	11.7619
{Fujian,Guangxi,Jiangxi} => {Guizhou}	0.016194	1	13.72222
{Fujian,Guangxi,Guizhou} => {Jiangxi}	0.016194	1	22.45455
{Fujian,Guizhou,Jiangxi} => {Guangxi}	0.016194	1	17.64286

注：lift 提升度是衡量关联规则可靠性的一个指标，一般而言，lift>3 表示关联规则是有价值的

从表 3 中可以看到，符合设置条件的规则有 19 条，以第 1 条规则为例，{Guangxi,Jiangxi} => {Guizhou} 表示在 248 条灾害事件中，广西、江西和贵州同时出现的概率为 1.16%，在广西、江西发生灾害的前提下，贵州发生灾害的概率为 100%。即这三个地区总是同时发生灾害的。第 2、3 条规则和第 1 条规则类似，由于地点不如商品的特色明显，而且缺少相关知识对不同地点的灾害发生时间进行排序，即很难确定哪个灾害地点是频发的，该地点发生灾害引起其他地点发生灾害的可能性较大。在本文中，很难筛选出第 1、2、3 条规则哪条是最有价值的。因此，本文中不过分关注规则的前件和后件，而是重点分析地点的共现性，发现同时发生灾害的地点集合。比如，我们可以根据前三条规则得出，广西、江西和贵州同时发生灾害的可能性较大，一旦其中一个地区发生灾害，其他地区应该注意防范灾害。这样将表 2 进行简化，发现灾害地点的共现集合，如表 4 所示。可以看出仅仅是广西、江西、贵州、福建、湖南这几个省份之间的关联，从地理位置可以看出广西和贵州、江西和福建、湖南与广西和江西是相邻的，结合表 2 中不同地区的灾害发生次数分析得出，广西、江西和贵州同时发生暴雨的概率比较大；广西、江西和福建同时发生台风和滑坡的概率比较大；江西和湖南是暴雨、洪水和滑坡易发的地区，因此福建、江西和湖南应该是同时发生这三种灾害的可能性较大。广西、贵州、江西和福建应该是暴雨发生的频率较高。

表 4 灾害地点的共现集合

Guangxi	Guizhou	Jiangxi	
Guangxi	Jiangxi	Fujian	
Guangxi	Guizhou	Fujian	
Hunan	Jiangxi	Fujian	
Guizhou	Jiangxi	Fujian	
Guangxi	Guizhou	Jiangxi	Fujian

5 结论

本文以我国 1998 年-2016 年的灾害事件为研究对象，通过网站获取灾害事件信息，并用文本挖掘、统计分析、关联规则挖掘等方法对 248 条灾害信息的进行提取、分析主要的灾害、灾害的时间分布和空间分布，以

及灾害的空间关联性。结果表明我国过去 18 年的灾害主要集中在地震、洪水、暴雨、台风和滑坡灾害中， 并且灾害的频发时间为每年的 5-8 月。另外不同灾害的空间分布是不同的，云南、四川是灾害的高发区，不止一种灾害频发。最后，对灾害的空间关联性进行了挖掘。由于数据量较少， 仅发现广西、贵州、江西和福建呈现一定的灾害空间关联性。另外，灾害事件抽取出的经济损失情况过少，本文并未进行讨论，未来需要通过扩大研究时间跨度、多源数据（新闻、微博等）获取足够多的信息进行分析挖掘，深入挖掘不同地区的灾害损失情况，继而评价我国自然灾害风险。

参考文献

[1]Galindo G, Batta R. Review of recent developments in OR/MS research in disasteroperations management. European Journal of Operational Research 2013, 230(2): 201–211.

[2]李实, 叶强, 李一军. 中文网络客户评论的产品特征挖掘方法研究. 管理科学学报, 2009, 12(2): 142-152.

[3]陈卓群. 基于共词网络的社交媒体话题演化分析. 情报科学, 2015, 33(1): 120-125.

[4]王炼, 贾建民. 基于网络搜索的票房预测模型——来自中国电影市场的证据. 系统工程理论与实践, 2014, 34(12): 3079-3090.

[5]Middleton S E. Social media analytics are evolving: from twitter-based crisis mapping to large-scale real-time situation assessment with trust and credibility analysis. 2014.

[6]Sakai T, Tamura K. Real-time analysis application for identifying bursty local areas related to emergency topics. SpringerPlus, 2015, 4(1): 1-17.

[7]Takahashi B, Tandoc E C, Carmichael C. Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines. Computers in Human Behavior, 2015, 50: 392-398.

[8]王昊, 杨亮, 林鸿飞. 日本地震的微博热点事件分析. 中文信息学报, 2012, 26(5): 7-14.

[9]张宇, 王建成. 突发事件中政府信息发布机制存在的问题及对策研究——基于 2015 年“上海外滩踩踏事件”的案例研究. 情报杂志, 2015, 34(5): 111-117.

[10]Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information processing & management, 1988, 24(5): 513-523.

[11]Rizzo G, Troncy R. NERD: a framework for unifying named entity recognition and disambiguation extraction tools. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012: 73-76.

[12] 郭海湘, 李亚楠, 黎金玲, 等. 基于灾害多级联动模型的城市综合承灾能力研究. 系统管理学报, 2014, 1: 015.

[13] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record, 1993, 22(2): 207-216.

[14] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Proc. 20th int. conf. very large data bases, VLDB. 1994, 1215: 487-499.

[15] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. ACM Sigmod Record. ACM, 2000, 29(2): 1-12.

[16]He Z, Zhai G. Spatial effect on public risk perception of natural disaster: a comparative study in East Asia. Journal of Risk Analysis and Crisis Response, 2015, 5(3): 161-168.

[17]Tian Y, Xu C, Chen J. Regional risk assessment of earthquake-triggered landslides. Journal of Risk Analysis and Crisis Response, 2015, 5(4): 234 – 245.