面向信息处理的突发事件语料库分类体系研究

杨丽英　雷勇

山西大学商务学院信息学院　山西　030031

摘要：本文分析了突发事件新闻的特点、建设突发事件新闻语料库的目的和意义，研究了突发事件新闻语料的分类体系和编码。研究结果对突发事件新闻语料库的建设、突发事件新闻信息检索、国家相关部门对突发事件应急处理方案制定以及对流行病学的医学研究等领域具有一定的理论意义和应用价值。

关键词：突发事件新闻；新闻语料库；新闻分类体系；突发事件新闻编号

0　引言

随着互联网的高速发展，Web 已经成为最重要的新闻媒体之一。通过浏览 Web 新闻，人们能够在很短的时间内了解来自不同国家和地区近期所发生的各类事件。而在众多新闻当中，那些难以准确预测而突然爆发的，对国家和社会产生重大影响的突发事件新闻是人们普遍关心的焦点。为了满足各级政府和社会的需求，有必要对突发事件新闻做系统的研究分析和信息处理。而这项研究首要任务是建立突发事件新闻语料库，并对其进行信息加工，如突发事件新闻分类和编号等。

1　突发事件新闻语料库的建设

1.1　建设突发事件新闻语料库的意义

突发事件新闻语料库的建立是适应信息化建设的需求。第一，通过该语料库可以尽早地、准确地、全面地掌握国内外各种突发事件的发生情况和发展趋势，为国家和各级地方政府有关部门及时采取应急措施和制定防范计划等提供科学决策依据；第二，对医学研究疾病及其分布规律和影响因素提供实例资源；第三，为语言学关于突发事件新闻的语言研究提供语料资源。

1.2　语料来源

Web 新闻是利用万维网技术，采用网页的方式进行新闻发布的网络新闻业务，是传统新闻业务的一种延伸，但它比传统的新闻发布方式有着更强的时间观，更能体现出新闻的纪实性。突发事件的突发性、偶然性和不可预料性，使得新闻网页比其他媒体有着更快反应的优势。所以，互联网是收集突发事件新闻的最好来源。

1.3　语料库的加工

语料库的加工主要包括四部分：语料的人工分类，文本格式处理，语料编号，以及分词与词性标注。本文主要介绍人工分类和语料编号两部分。

2　现有的突发事件分类体系

目前，对突发事件的分类主要有以下两种。

2.1　《国家突发公共事件总体应急预案》中关于突发事件的分类

国务院颁布预案的目的是提高政府保障公共安全和处置突发公共事件的能力。根据突发公共事件的发生过程、性质和机理，突发公共事件主要分为以下四类：

(1) 自然灾害。主要包括水旱灾害，气象灾害，地震灾害，地质灾害，海洋灾害，生物灾害和森林草原火灾等。

(2) 事故灾难。主要包括工矿商贸等企业的各类安全事故，交通运输事故，公共设施和设备事故，环境污染和生态破坏事件等。

(3) 公共卫生事件。主要包括传染病疫情，群体性不明原因疾病，食品安全和职业危害，动物疫情，以及其他严重

作者简介：杨丽英(1983-)，女，硕士生，研究方向：中文信息处理，人工智能。

影响公众健康和生命安全的事件。

(4) 社会安全事件。主要包括恐怖袭击事件，经济安全事件和涉外突发事件等。

各类突发公共事件按照其性质、严重程度、可控性和影响范围等因素，一般分为四级：Ⅰ级(特别重大)、Ⅱ级(重大)、Ⅲ级(较大)和Ⅳ级(一般)。

2.2 《突发流行病学》中对突发事件的分类

突发事件流行病学是研究突发事件的原因、发生、发展及其后果和应对方法的一门学科。突发事件可以有多种分类方法，目前最常用的是按照原因和性质分类，将其分成自然灾害、人为事故和疾病爆发三大类。

(1) 自然灾害(natural disaster)。主要包括气象灾害，海洋灾害，洪水灾害，地质灾害，地震灾害，农业生物灾害，森林灾害，宇宙灾害。

(2) 人为事故(accident)。主要包括战争和暴力，恐怖活动，重大交通事故，严重火灾，意外爆炸，群体中毒，急性化学事故，放射事故，其他事故。

(3) 疾病爆发(outbreak)。主要包括肠道传染病，呼吸道传染病，虫媒传染病，自然疫源性疾病，性传播疾病等。

突发事件还可按其规模大小和严重程度分为：一般性突发事件、重大突发事件和特大突发事件。此外，也可以按发生地点、发生时间和事件的后果对突发事件进行分类。

3 突发事件新闻语料分类体系研究

根据新闻报道的特点和突发事件新闻语料库的基本功能，本文提出以下一些分类原则。

3.1 分类原则

(1) 主题分类与实际情况相结合

由于新闻报道的特点和实际新闻工作的需要，按主题进行分类也能满足用户希望在一个主题或专题下查全相关信息的需求。因而，本文二级类目按照主题或专题内容确定类目，这不仅直观与实用，也更能保证这些主要类目在较长时期内的稳定性。考虑到有些突发事件新闻信息量大，且是人们的关注热点，可以对其类目进行提升。

(2) 求大同存小异

分类层次在三级以上的类目应最大程度的统一，这有利于自动标引的标准化与网络分类浏览检索的资源共享。三级以下类目，特别是专业性太强的小类，有些确实是难在最大层面上作到统一的，只要大类得到统一，采用主题词就可以较好地解决这一问题。这种求大同(1－2 级类目高度统一)存小异的原则，可以保证突发事件新闻分类法的实用性与可推广性。

(3) 用语规范性与灵活性相结合

在力求达到准确性与通用性的前提下，充分考虑到突发事件新闻信息的特点。对一级和二级类目的命名，基本上参考了国务院的类名，力求科学规范。但由于新闻报道的特点是综合性强、时效性强，不断有新事物、新名称出现，变化性大，所以在类名的命名时也采取了相应的灵活处理原则。例如，三级及三级以下关键词语有些则采用了自然语言或习惯用语(即新闻语言)作为类名，但力求选用能够被广泛认可和语意明了的词。

(4) 具有层次性和可扩展性

突发事件的突发性和偶然性本身就需要我们在基本大类保持不变的前提下，可对相应类目进行扩充。分类体系拟采用三层结构，其中的二级类目、三级类目以及主题词都具有可扩充性。

3.2 突发事件新闻语料分类体系

根据以上分类原则提出的突发事件新闻语料分类体系包括 3 个层次，其中一级 4 类，二级 33 类，三级 94 类。下面给出了一、二级类别及其编码，三级编码种类多，不列出：

(1) 自然灾害类 N(Natural disaster)：

01 水旱灾害	02 气象灾害
03 地震灾害	04 地质灾害
05 海洋灾害	06 生物灾害
07 森林草原火灾	08 宇宙灾害

(2) 事故灾难类 A(Accident)：

01 战争和暴力
02 工矿商贸安全事故
03 交通运输安全事故
04 城市生命线事故
05 通讯安全事故
06 环境污染和生态破坏
07 严重火灾
08 中毒事件
09 急性化学事故

(3) 公共卫生事件 P(public health)：

01 传染病疫情
02 群体性不明原因疾病
03 食品安全和职业危害
04 动物疫情

05 其他严重影响公众健康和生命安全的事件

(4) 社会安全事件 S(social safety)：

01 恐怖袭击事件

02 重大刑事案件 00

03 经济安全事件 00

04 涉外突发事件 00

05 规模较大的群体性事件

06 民族宗教

07 反政府和反社会主义

3.3 突发事件新闻语料编号

编码是对新闻信息进行分类标引和检索的工具，是分类法的表现形式。一般是将同一主题从大类到小类，按照逻辑系统逐级展开。突发事件新闻语料的编码从新闻和语料库两方面入手，参考了《中文新闻信息分类及代码》和人民日报语料库编码规则，根据实际检索进行了编码。

(1) 类目代码

一级类目代码用每类事件英文首字母表示。二级和三级类目采用十进分类法，每一级类目用两位阿拉伯数字表示(01-99)。无三级目录的类目用“00”表示。

(2) 语料编码

采用突发事件发生日期+文档编号。日期用 8 位表示，年用 4 位，月和日都用 2 位表示。文档编号为 3 位(000—999)。

一篇新闻的完整编码为：类目编码+新闻编码， 全部代码共 16 位，具体如下：

一级分类号(1 位字母)+二级分类号(2 位数字)+三级分类号(2 位数字)+日期编号(8 位数字)+文档编号(3 位数字)。

(3) 实例说明

例如：编号为 A070120101115000 的新闻语料的解析如表 1 所示。它表示事故灾难类中的第 7 类严重火灾中住宅区火灾，报道时间是 2010 年 11 月 15 日，文档编号 000 表示在此类中对此事件的第一篇新闻报道。

表 1 语料编码方案解析示例

一级分类号 (1位字母)	二级分类号 (2位数字)	三级分类号 (2位数字)	日期编号 (8位数字)	文档编号 (3位数字)
A	07	01	20101115	000

4 结束语

本文在建立突发事件语料库的过程中，对语料进行了初级加工，对突发事件新闻分类体系进行了详细的研究。由于突发事件本身的不确定性，使得对突发事件的分类存在一定的困难，需要根据实际情况不断地扩充和完善。

参考文献

[1]http://www.gov.cn.

[2]李立明. 流行病学(第 4 版).北京:人民卫生出版社.1999.

[3]周文.基于应对视角的突发公共事件分类[J].商场现代化.2011.

[4]张玲玲,李鼎鑫.重大突发事件新闻报道的分类及特点[J].华北科技学院学报.2009.

[5]http://news.xinhuanet.com.

[6]俞士汶,段慧明,朱学峰等.规范[J].中文信息学报.北京大学现代汉语语料库基本加工. 2002.

[7]袁辛奋,胡子林.浅析突发事件的特征.分类及意义[J].科技与管理.2005.

[8]孙香勤.国内外重大突发事件管理模式分析[J].交通企业管理.2005.

The Research on Classification System of Accidental News Corpusfor Information processing

Yang Liying,Lei Yong

Information faculty of Business College of Shanxi University,Shanxi,030031China

Abstract: This paper not only analyzes the characteristic of accidental news,the purpose and significance of accidental news corpus,but also researches the classification system and code of accidental news corpus.The result has the certain theories meaning and the application values in the field of establishment of the accidental news corpus,information retrieval of accidental news,the establishment to the emergency plan of accident by relative departments of government and the medical science research of epidemiology and so on.

Keywords: Accidental news;News corpus;News classification system;Accidental news code