

分类号_____

学号_____

学校代码

密级_____

硕士学位论文

面向应急决策的突发事件案例相似度研究

学位申请人:

学 科 专 业: 公共管理

指 导 教 师:

答 辩 日 期:

**A Thesis Submitted in Partial Fulfillment of the Requirements for
the Degree of Master of Public Administration**

**Research On Case Similarity for
Decision-making of Sudden Incidents**

Candidate :

Major : Public Administration (MPA)

Supervisor :

May, 2014

独创性声明

学位论文版权使用授权书

摘 要

这些年来，我国各类公共突发事件频发。突发事件的突然性、灾难性与演变的复杂性交织，对突发公共事件管理者提出了巨大挑战。借助信息科学，很多学者对突发事件的应急管理从案例推理方面进行了研究。而且，一些学术机构已经建立了一些突发事件案例库。依据已有的案例和知识为决策者提供决策支持，成为了一种重要的辅助手段。

本文围绕不确定信息的知识表示及其推理进行研究，通过对国内外案例相似度的深入研究分析，提出用中文分词来分析历史案例。并通过对词汇词频、分布和词性的研究找出文本相似的内在关联。文章中详细分析了 123 个案例，307127 个词汇，638731 字，得到了大量的基础数据，并通过这些数据分析出案例之间的关系。文章设计了两个对比实证检验（一个是相似案例的相似度对比，一个是同一案例在不同阶段的对比），来验证得出的相关结论。通过这些实验和数据表明，本文提出的分析案例相似度的方法具有有效性，为处理历史突发事件案例提供了一种新的思路，为案例推理和案例检索提供必要的技术支撑，并为后期课题的研究打下理论和技术的基础。

最后，通过研究案例相似度在政府部门运用中的交互方式，应用路径和约束条件等，得出案例相似度的应用的方向和条件。

关键词：应急决策 突发事件 案例相似度 中文分词

Abstract

Recent years, a lot of public emergencies occurred in our country. All categories of sudden, catastrophic intertwined with the complex evolution of emergencies, presented a huge challenge to the public emergency managers. Information science, many scholars began to study the population problem Emergency Management Case-Based Reasoning (CBR) approach academic institutions have begun to build a database of emergency cases. Based on existing knowledge of the case and provide decision support for policy-makers, has become an important adjunct.

This paper focuses on the knowledge representation and logic reasoning with uncertainty information. Based on the similarity of both domestic and foreign related cases in-depth research and analysis, put forward by Chinese word segmentation to analyze historical cases. And through word frequency, distribution and research to identify parts of speech text similar to inter-relate. Article detailed analysis of 123 cases, 307 127 words, 638731 Chinese characters, collect a large number of basic data and through these relationships and data analysis between the cases. Article designed two comparative experiments to verify the conclusions drawn (one is comparison of similarity in the the similar cases, one is the same case in the comparison of different stages of). Through these experiments and the data show that the proposed method has a validity case similarity analysis for handling emergencies case history provides a new way of thinking, provide necessary technical support for the case based reasoning and case retrieval, and lay the theoretical and technical foundation for later research topic.

Finally, by studying the case similarity in government departments in the use of interactive mode, application of the path and the constraint condition, direction and conditions of application of the case similarity.

Key words: emergency decisions-making; sudden incidents; case similarity; the Chinese word segmentation

目 录

摘 要.....	I
Abstract.....	II
1 绪论.....	1
1.1 选题背景及意义	1
1.2 国内外研究现状	3
1.3 研究思路与方法	7
2 相关理论研究	9
2.1 突发事件理论	9
2.2 应急决策理论	11
2.3 基于案例推理的应急决策	13
3 案例相似度研究环境分析	16
3.1 案例库建设的机制	16
3.2 案例比较的条件	17
3.3 案例文本处理	18
4 案例分析处理模型	25
4.1 案例词汇词频分析	25
4.2 案例词汇分布分析	27
4.3 案例词汇词性分析	32
4.4 案例文本相似度分析小结	36
5 实证研究	37
5.1 相似案例相似度比较研究	37

5.2 同案例不同阶段相似度比较研究	43
6 案例相似度研究的应用	46
6.1 案例推理中决策者与案例库的交互	46
6.2 案例相似度应用路径分析	47
6.3 政府使用案例相似度应用的启示和要求	48
7 结束语.....	50
7.1 研究结论.....	50
7.2 研究不足.....	51
7.3 研究展望.....	51
注 释.....	52
致 谢.....	55
参考文献.....	56
附 录.....	61

1 绪论

1.1 选题背景及意义

1.1.1 选题背景

中国因为幅员辽阔，人口众多，不仅是世界上自然灾害最多的国家，而且各类紧急突发事件频发。加上我国正处在社会转型的矛盾激化期，某种程度上加重了这种趋势。中国已经进入乌尔里希·贝克所说的“风险社会”。

所谓“风险社会”中的紧急突发事件，包括常规性突发事件和非常规性突发事件。常规事件因其结构简单、特征明显，决策者较容易依靠紧急预案，甚至个人经验做出合理决策，并被认为是一种领导艺术；然而非常规性突发事件却完全不同。非常规性事件产生发展过程中，往往伴随着不可预知性、高度衍生性、快速扩散性、高破坏性和信息匮乏等特点，决策者缺乏相关处置经验，甚至对此类事件完全不了解，加之决策者往往会面对来自多方面的压力，导致应急决策迟缓、盲目甚至出现重大失误，不仅无济于事，还可能浪费了宝贵的资源，妨碍了应急救援顺利实施，最后导致事件影响扩大，造成二次损失。

例如 2011 年的日本福岛核泄漏事故，决策者需要在时间十分紧迫、信息极度缺乏、舆论压力巨大和决策环境多变等诸多不利情况下，利用有限的资源，来做出关乎全国人民生命安全甚至世界环境安全的重要决定。难度之大，压力之大可想而知。

国内外近年来突发事件应急管理形势严峻，各类自然灾害和社会安全事件频发。如 2001 年美国 911 恐怖袭击、2004 年印度洋地震海啸、2005 年美国新奥尔良飓风、2008 年南方雪灾和汶川大地震、2009 年禽流感事件、2010 年松花江水体污染、2011 年日本核泄漏事故、2013 年的余姚水灾和青岛爆炸事故等等，这些突发事件频频发生。其中，2003 年 12 月重庆开县特大井喷事故爆发后最初的 18 小时内应急响应失误，说明应急决策人员在突发事件的响应和处置过程中需要发挥聪明才智和坚强意志之外，还需要丰富的知识储备和现代科技的支持与协助。

总之，非常规应急决策活动由于事件本身的动态性和复杂性，不仅要求应急决

策人员具备较高的素质和快速反应能力，同时还需要现代科学技术手段为应急决策科学化提供必要的辅助和支撑。其中，信息科技的快速发展，为危机管理实现高效管理提供可以用于实践的技术保障。依据历史案例或案例知识库为应急管理过程提供决策支持，这越来越成为了一种提高危机管理能力的必要手段，其中以案例推理（Case-Based Reasoning）最为理论界和实践界广泛采用。案例辅助决策的核心是案例特征的表达、案例相似度分析和匹配，通过信息科学和技术对旧案例进行一系列的处理。以往研究更多从概念化定性角度或者从计算机科学角度进行，较少见综合管理学、信息科学及心理学等多学科交叉研究的成果。本文的选题依托国家自然科学基金重大项目“面向应急决策支持的非常规突发事件案例推理的理论与方法”，集中关注于对历史案例相似度研究，拟通过案例分词的方法，对案例相似度在词频、分布和词性等方面进行深入分析，并通过人机交互体系构建关联机制，为实践中提高应急决策能力提供借鉴。

案例推理辅助决策，它是一种重要的基于已有知识库的问题求解方式，利用过去积累的案例来解决新问题，即检索案例库中过去已解决的类似问题及其解决方案；或者检索看似毫无关联，但是特征表达类似，影响过程类似，灾害损失类似的案例及其解决方案，比较新旧案子时空差异，对旧案例的解决方案进行合适的调整和修改，快速形成新的解决方案的一种推理模式。目前相关研究主要集中在计算机科学技术领域开展的检索技术和算法研究实现，而对旧案例的分析、案例内部结构和相似度的研究，尤其是面向应急决策的案例相似度的研究是没有的。

1.1.2 研究意义

现实意义方面，由于非常规突发事件应急管理是动态全过程的，只有提前分析处理旧的案例，把握旧案例的特征、结构和联系，才能实现高效检索和匹配。本文为国家自然科学基金重大项目“面向应急决策支持的非常规突发事件案例推理的理论与方法”后续的案例检索与匹配工作提供前期的理论支撑和技术铺垫。

理论价值方面，以历史案例提示机理为研究对象，排除作者主观因素的影响作用，通过科学的分析方法剖析整个案例，探究案例内部关联，阐释案例相似度的关

系，实现人机交互下的案例提示—顿悟机理，从而为完善基于案例推理的非常规突发事件应急决策辅助支持的智能感知技术和方法提供重要的理论参考。

1.2 国内外研究现状

1.2.1 应急决策研究现状

突发事件应急决策的研究是近几年兴起的一项热门研究方向。不少学者尝试在传统经典决策理论上加上应用数学的相关理论形成新的决策方法，如概率分析等；或者将经典决策理论同风险管理理论相结合，如效用分析等；或者运用运筹学等理论来处理突发事件中的应急疏散、应急物资调用、应急力量调度等方面，但是都没有取得较为良好的效果。因为 Salvatore Belardo 与 Harold L. Pazer 等人（1995）在对突发事件决策进行研究时认为，应急决策具有：紧急性、时间有限性、信息有限性、决策量大、决策责任大、分散性等特性，导致人们很难用一个固化的模型或者模式去处理，加之一些不断来自于政治、法律、伦理道德、社会舆论方面的限制，目前还没有较为成熟的应急决策方法模型。^[1]

当然，这并不代表国内外在相关领域没有取得一些突破。国外学者在利用敏感性分析、概率分析、效用分析、数理分析和运筹学等方法，对突发事件的发生、过程、决策和结果进行了定量研究。如 Noel Pauwels 等人（2000）运用敏感性和效用分析方法对核泄漏事件发生后的撤退决策进行了研究。^[2]Hiroyuki Tamura 等人（2000）运用决策树分析方法对灾害风险进行了分析。^[3]L. Jenkins（2000）建立了如何选取特定应急场景使预案最具代表性的整数规划模型。^[4]此外，Donald（2004）对受灾者的决策能力进行了分析，从心理学的角度向政府提出灾后管理的若干建议。^[5]而国内的研究方面，袁辉（1997）认为应急决策是群体决策，需要综合考虑决策主体的专业知识结构、能力结构、年龄结构等。由于应急活动的特殊性，选用项目组织结构形式能够有效发挥决策者能力。^[6]姜卉和黄钧（2009）针对罕见重大突发事件，指出传统的“预测-应对”的应急决策范式已不适用，必须向“情景-应对”的应急决策范式转化。^[7]

1.2.2 案例推理的研究现状

目前已经有一些文章研究了基于案例推理（Case-Based Reasoning, CBR）的应急决策支持系统（CBR Emergency Decision Support System, CEDSS）的理论框架，既发挥计算机处理信息即时性优势，又依据人工智能领域中的案例推理方法对已经存储的应急案例进行检索，查找到与当前问题相类似的案例，提高决策的科学性与有效性。

国外学者多侧重于将 CBR 应用于故障诊断、计算机科学、企业管理、医疗领域、规划设计等多领域，对其在应急决策中的研究较少。相较而言，国内学者相关研究更多，如张荣梅等（2002）提出在交通事故处理当中，CBR 与多智能体和多库协同的智能决策支持系统^[8]；张建华、刘仲英（2002）提出了基于 CBR 的火灾应急响应决策支持系统^[9]；柳炳祥等（2002）提出了基于案例推理的企业危机预警系统的组成、框架结构、功能及工作原理并论述了基于案例推理的企业危机预警系统中的一些关键技术^[10]；郭泳亨（2006）运用了数据仓库、OLAP 和数据挖掘新技术在应急决策支持系统中增加案例库，并设计了案例推演的推理机制^[11]；周云海（2007）等设计了基于案例推理的打听点恢复系统；^[12]陈铭（2009）开发了一个航空事故相似度决策支持系统^[13]；贺清（2012）针对铁路枢纽站在应急决策方面存在的不足，提出一种新的基于案例推理方法的应急预案管理模型^[14]。

1.2.3 中文分词技术的研究现状

案例库中的案例主要由三部分构成：普通文本、图片和视频。针对图片和视频处理目前科技界难以通过技术手段来进行语义的表达，而且这两类案例数量相对较少，不是本文分析的重点。

分词技术是进行案例文本分析处理的前提，要实现诸多文本分析技术，如文本挖掘、语义表达、文献检索、文本相似度比较等等，分词都是十分关键的一步。可以说，分词的好坏直接影响到后续研究的最终实现。一般来说，分词系统处理的都是非常大量的语料对象，（如国家行政学院案例库案例数量超过 30 万，上海交大的案例库案例也有 3 万多）因此分词的处理速度十分重要。

分词对象包括英文分词和中文分词。应该说英文里并不存在分词的问题，英文里每个词根据语法规则，都是自然的用空格隔开，而且本文也不讨论英文案例，不再赘述。

汉语文本分词主要是指中文自动分词技术。现有中文分词方法主要包括：字典匹配法、预设标志位法、词频统计法、规则分词法、语义语用分词法、链接表分词法和神经网络的分词法等等近二十种分词方法。这些分词方法大致分为两大类，即机械分词和理解性分词。其中技术相对成熟的是机械分词技术，中文机械分词主要可以分为两种：一是基于概率统计的机械分词，另一种是基于分词字典的机械分词法。

基于字典的分词技术是基于字符串的机械匹配，即依据某种策略将待分词的中文字符串与一个词典中的词条依次匹配，若匹配成功则认为是一个词，可进行切分。匹配方法按照匹配方向差异分为正向匹配、逆向匹配和双向匹配，根据长度优先不同又可分为最大匹配和最小匹配等。而不同的方向与长度的匹配方式又可以进行组合，很明显总共可以有六种以上的组合方式。不过基于字典的分词技术总体上来看，精度是不高的，有统计表明无论用何种组合方式，错分率都在 5% 左右。因此，在实际应用中机械分词往往被用作初分方法，还需采用其他方式进一步提高分词效果。

基于统计的机械分词是基于这样的思路：“词应该是字的一种稳定组合，在文本中相邻的字共现次数越多成词的可能性就越大。”^[15]因此可统计出文本中相邻共现的各个字的组合的频率，并计算其互现信息。互现信息说明了字组合间关系的紧密度，当其超过某一个阈值时即可大致认定该字组合为一个词。

目前，国内关于中文分词的研究已经持续了 30 年了，也取得了较大的成就，建立了很多独立完善的中文分词系统。例如：北京师范大学现代教育研究所研制的书面汉语自动分词系统、微软中国研究院（Microsoft Research）开发的汉语句法分析器中的自动分词、北大计算语言所研制的分词系统、国家语委文字所开发的汉语自动分词、基于 lucene 的庖丁中文分词系统（Paoding Analysis）等；另外，国内的清华大学、复旦大学、浙江大学、北京航空航天大学、哈尔滨工业大学和山西大学等很多高校都研制出了自己的分词系统，且分词的性能各有所长，分词结果都很不错。

本文采用的是中科院计算机所的 ICTCLASS(Institute of Computing Tehnology,Chinese Lxieal Analysis system)汉语词法分析系统。^[16]

该系统可以根据用户需求定义个性化词典，可以实现中文文档的分词、词性标注和新词识别等操作。该系统原始版的 ICTCLAS 是完全基于 C++语言进行开发的，本文实验所采用的是基于 C/C#语言的免费开源版 ICTCLAS2013 版。

1.2.4 语义及文本相似度的研究

Dekang 于 1998 年提出了一组具有广泛意义的相似度定义：直觉告诉我们，对象 A 和 B 之间的相似度与它们之间共性和差别相关，两个对象所拥有的共性越多，则相似度越大，而两个对象之间的差异越多，则相似度越小。当两个对象 A 和 B 是同一个对象时，相似度达到最大。当 A 和 B 无关或独立时，相似度最小。^[17]

语义相似度的研究非常复杂，采用的研究方法也非常的丰富。心理学领域最初对相似度的研究可以追溯到 Osgood（1952）提出的语义微分方法（Semantic Differential）^[18]。可以用几何模型进行多维度、多层次语义分析。

而 Tversky 认为对象的一些属性无法用数字量化，这些属性更适合用定性的方式描述。Tversky 模型通过特征/属性集合描述对象，相似度被定义为关于特征共同性和差异性的函数^[19]。

Lee、Lipika、Ong 和 Blaz 等人认为，主要使用计算机进行学习，这样可以根据聚类 and 模糊模型去构建新的文本模型，其实就是运用本体论的方法研究文本相似度。而 Navigli、Sugumaran 等人更是指出就是要对文本进行处理^[20]。

国内学者潘谦红等提出利用属性论来计算相似度，晋耀红提出了基于语境框架的相似度计算方法^[21]，金博等也提出了利用《中国知网》已有的知识结构来描述文本语言的语法相似度^[22]。

1.2.5 研究现状评述

通过文献及研究现状分析可以看出，突发事件因为其自身特点，决策者因为种种原因难以在短时间内，利用有限的资源对复杂的事件进行高效的、科学的决策。而且大多数学者认为应该更多的利用现代的科学技术用以辅助决策，提高决策效率

和准确度，使用计算机和案例辅助决策目前来看是一条可行的路。而计算机和案例辅助决策的核心在于案例的分析，包括案例特征的表达，案例相似度的分析，案例的检索匹配等等。利用中文分词技术对旧的案例进行处理，实现文本语义的表达，找出文本之间在结构、分布、联系等方面的相似性，实现案例文本标示、案例文本预处理，为后续的检索研究提供支撑。

1.3 研究思路与方法

1.3.1 研究思路

本文的选题依托国家自然科学基金重大项目“面向应急决策支持的非常规突发事件案例推理的理论与方法”，根据本论文的研究内容和导师安排，论文将按照下图（图 1 所示）所示的技术路线开展研究。

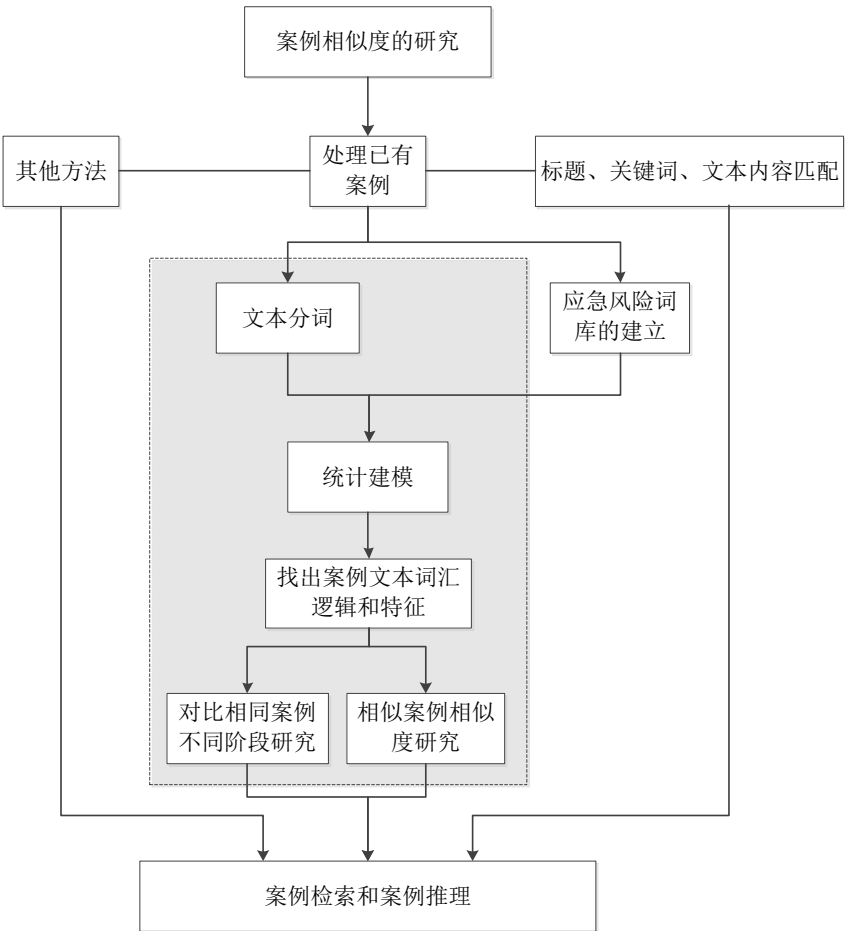


图 1 案例相似度研究思路

1.3.2 研究方法和工具

1.内容分析法：通过对历史案例文献进行研究，理清突发事件案例推理、基于案例推理的应急决策和案例相似度的内涵、方法等基本理论问题。

2.中文分词法：通过拆分案例文本的词和词组，分析词汇之间的内在逻辑和联系。

3.模型方法：通过多种数学模型来分析文本特征。

4.工具：C/C#、mySQL、SPSS、EXCEL。

2 相关理论研究

2.1 突发事件理论

2.1.1 突发事件定义和分类

突发事件 (emergency)，顾名思义可被理解为突然发生的事情。我认为突发事件有两层含义：首先是事件的发生、发展的速度很快，事先没有任何准备，事后发展速度也超过想象；其次是事件难以使用常规方法应对，可能需要采取非常规方法来处理。

根据中国 2007 年 11 月 1 日起施行的《中华人民共和国突发事件应对法》的相关规定：“本法所称突发事件，是指突然发生，造成或者可能造成严重社会危害，需要采取应急处置措施予以应对的自然灾害、事故灾难、公共卫生事件和社会安全事件。”^[23]

根据 2005 年 1 月 26 日，国务院发布的《国家突发公共事件总体应急预案》，突发事件被描述为：“突发公共事件是指突然发生，造成或者可能造成重大人员伤亡、财产损失、生态环境破坏和严重社会危害，危及公共安全的紧急事件。”

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）。^[24]

2.1.2 突发事件特征分析

突发事件虽然种类多样，但是都具有以下的共同特征：

（1）突发性。突发事件所发生的时间、地点、性质、形态及可能造成的后果都难以预测。如大地塌陷，有可能是由于自然灾害地震引起，也有可能是由于人们无序开采地下水造成的，也有可能是因为建筑引起的。而且，事件的发生的时间无法预知，有时候并没有明显的征兆，事件的开端、发展、衍生和次生影响都无法确定。

（2）破坏性。突发事件不仅会造成物质上的破坏损失，还会造成人们精神和心理上的创伤，严重的甚至可能给社会、国家稳定带来威胁。比如地震造成的破坏长达几十年，不仅一次性的造成了巨大的社会损失，对致残人士和失去亲人的受害者的影响会非常长久。而且，地震引起的次生地质灾害和次生卫生灾害可能的破坏性比地震本身还要严重。更为糟糕的是如果政府在地震救援中表现不力，可能会影响到政权更迭和社会稳定。

（3）扩散性。或者我更加愿意将其称作爆散性。突发事件传播速度极快，内部和外部在人们的关注下发生联动耦合，内部扩张，外部连带，短时间内就能造成非常强烈的社会影响。尤其是在这个移动互联网的时代，传播速度达到了前所未有的高度。如 7·23 甬温线特别重大铁路交通事故在发生后，微博上的信息传播速度甚至超过了政府救援响应速度。在整个救援过程中，政府的救援和信息的扩散是在同步进行的，信息的快速扩散将政府的行为放大到了整个社交网络之上，如果稍微处理不慎，造成的后果难以想象。

（4）衍生性。突发事件本身不可怕，次生灾害可能比事件还要可怕。为阻止突发事件引发更严重的负面后果，在事件发生后采取人为手段干预，处置不当反而放大了事件的灾难性。“杞县钴 60 案”就是一个典型的例子，所谓的“核泄漏”和“核污染”并没有发生，而处置过程的信息不公开、信息不透明和信息不对称衍生成为了居民的恐慌性逃难，给社会稳定造成了巨大的影响。

2.1.3 突发事件演化分析

我们讲突发事件的演化，往往会用到下面的词汇：“多米诺效应”、“跨界危机”、

“共振效应”和“黑天鹅式危机”。这些词汇无一不是在讲述突发事件演化的特点。

一方面，非常规突发事件具有极强的破坏性和诱发性，若不能及时进行有效地安全控制和管理，不仅会造成巨大的破坏，还极易诱发社会潜在风险，进而导致公共危机发生。另一方面，非常规突发事件集突发性、破坏性、扩散性、衍生性、不确定性、跨界性、突显性和耦合性等等特性于一体，与传统事件相比，实践中缺少有效的紧急应对手段。

综合文献，我认为突发事件整个演化的过程可以分为：爆发、扩散、衍生、变异和耦合。爆发是指突发事件发生的一瞬间，以及发生一瞬间所造成的事故和破坏，是整个事件的开端；扩散是指事件发生以后，事件的进程按照特定的规律不断发展，包括时间上的延续和空间上的扩张，是事件的初期发展阶段；衍生是指某一突发事件在处置过程中因为各种不确定的原因新的诱发了新的突发事件发生，两事件之间并无必然的因果逻辑关系；变异是指某一突发事件在发展过程中引发了另一突发事件的发生，两者之间具有一定的逻辑和因果关系；耦合则是指在事件发展过程中几个因素共同作用，产生“共振”，导致该突发事件程度加深和破坏加剧。后三个过程一般认为事件演化到了深入阶段。

其中衍生和变异是突发事件演化过程中最难控制的阶段，尤其是衍生。因为前后缺乏必然的联系，决策起来缺乏必然的逻辑联系，往往导致决策者无所适从，或者延误最优决策时机。

2.2 应急决策理论

决策(Decision Making)是指为了达到某特定目标，从不同角度制定多个备选方案，并从备选方案中选择一个最优方案的过程^[25]。决策理论的研究按照发展阶段，大致可以分为经典决策理论、有限理性决策理论、行为决策理论和自然决策理论四个阶段^[26]。

经典决策理论主要是从经济学和博弈论的角度来建立精确的决策公式，寻求规范的决策分析方法，属于传统决策理论。但是，由于突发事件发生后，决策者很难在有限的时间内掌握与事件相关的足够信息，无法快速的建立精确的决策公式，所

以经典决策理论不适用于突发事件的应急决策。

有限理性决策理论也属于传统决策理论，是在备选的方案中选择最优的一个方案。可是，在突发事件发生后，信息有限的情况下很难做出有效的方案，更不可能做出所谓的最优选择。所以有限理性决策理论也不适用于突发事件的应急决策。

行为决策理论是一种采用启发式决策理论。对比备选决策方案的偏差来分析新发生的突发事件，需要决策者在备选方案中选择一个各方面都较优的决策方案。但是，往往因为受到时间和信息不完全的约束，难以构建规范化、标准化的决策模型，就更谈不上把握各个方案之间的偏差，所以，行为决策理论也不适用于突发事件应急决策。

由于以上三个决策理论没有综合考虑决策过程中时间的有限性和信息的不完全性，更是完全没有考虑到决策者作为一个自然人在知识、认知和经验等方面的不足。导致不能在目标不明确、信息不对称且时间有限的复杂情况下的进行决策，这些种种决策中的实际情况导致了自然决策理论的产生。

自然决策理论(Naturalistic Decision-Making, NDM) 还原了在不确定环境，时间紧迫，信息不完全的情形下的真实决策过程，完全摒弃了对标准化解解决方案的追求。Gary Klein 于 1993 年提出了认知主导决策 RPD(Recognition-Primed Decision-making) 模式^[27]，指出决策环境是动态的，决策者往往需要根据经验，对新问题的环境做出评估，依据案例和经验找到类似的解决方案的做法。

在本项目中提出的“提示—顿悟”决策模式就是对自然决策理论的一种有效探索 and 延伸。试图在复杂多变的决策环境中，快速查找相似性较高的历史案例，进行快速的匹配，使用较低计算量地做出决策。当然这些典型的自然决策模式，已经在海军决策与军队决策训练系统中得到应用^[28]。案例推理核心思想是依据决策者的经验进行决策，适合处理真实的决策环境，特别是突发事件的应急决策，所以案例推理最适合突发事件应急决策。

2.3 基于案例推理的应急决策

2.3.1 案例推理理论的发展

基于案例的推理以下几个重要特点：（1）与案例相关的知识相对容易获取；（2）案例问题中自带解决方案，容易求解问题；（3）案例库的自动更新和案例学校；（4）可以借鉴历史案例，避免犯历史错误等，尤其是避免重复历史错误非常重要。因为针对人文、社科类案例，尤其是公共突发应急事件案例很难进行试验和模拟，历史案例的研究可以为解决新问题提供必要的对比。

案例推理理论的发展大致经历了四个阶段：萌芽阶段（1983 年以前）、早期发展阶段（1984-1993）、快速扩散阶段（1994-2003）和理论融合阶段（2004 年至今）。

萌芽阶段：美国耶鲁大学的 Roger C. Schank 和 Robert P. Abelson 在 1977 年提出了尝试用脚本的方法来表示文本知识，这被认为是案例推理理论（CBR）研究的萌芽。Roger C. Schank 在 1982 年首次提出了基于案例推理理论的认知模型及框架，这被视为案例推理理论的基础。

早期发展阶段：上世纪 80、90 年代，在美国出现了一些简单的，基于案例推理的应用系统，当然涉及的面并不是很广泛，其主要特点有：（1）利用简单的 K-NN 算法来进行案例的匹配检索；（2）简单的案例推理机制；（3）计算机在 CBR 系统中起辅助的作用，因为除了案例检索，其它功能只能靠人工完成。

快速扩散阶段：这个时期 CBR 研究有四个主要特点：第一，CBR 科研人员和 CBR 研究机构不限于美国，全世界都开始从事 CBR 研究活动；第二，CBR 不再是简单的应用，其领域迅速扩大，很多数据分析的技术和数据挖掘领域的思想被引入到在案例推理应用中来；第三，CBR 的理论方面研究依然集中在案例的表示和模型的构建上；第四，不确定性理论被引用到案例推理应用当中，为了解决案例推理中存在的 uncertainty 问题，Kaoru Hirota 等人尝试使用模糊理论^[29]来表示案例特征；用粗糙集理论来实现案例属性的简化^[30]

理论融合阶段：这一时期的主要特点是：（1）多技术的融合。数据挖掘、模糊集理论、粗糙集理论、人工神经网络、基于规则的推理等等理论与案例推理理论融

合。(2) 重视案例库的维护。案例库的维护是确保案例推理具有自增量学习（案例库的自我更新）的重要前提，案例维护包括案例的增加、移除和修改。(3) 案例的逻辑框架和数学模型。

2.3.2 案例推理的理论研究

综合国内外提出的案例推理模型，现在一般认为案例推理包括四个方面的内容：案例表示、案例检索、案例修正、案例学习。

案例表示：案例表示是案例推理理论中最为基础的内容。案例表示是针对某个案例的特征、结构、属性、细节和解决方案等进行有效的描述。案例表示基本上就决定了案例匹配和检索的效果，从而影响到整个案例推理的效能^[31]。常用的案例表示方法主要有两类：逻辑的非逻辑的。前者主要指谓词逻辑，后者则包含有框架^[32]和语义网等。无论谓词逻辑、框架还是语义网，虽然各自有着自身的优缺点，但是他们有着共同的基础表现元——词汇。通过对词汇的整理、归纳、提炼，构建有效的表达模型，实现案例的表示。

案例检索：现在，在案例推理中人们关注的焦点是案例的检索与匹配，其目的是能够快速地从大量的历史案例库中找到与当前案例问题最为匹配的案例。案例检索的前提是案例表示，并且能够用计算机来量化两个案例之间的“相似度”值。

案例修正：案例推理中很难找到与新问题完全一样的问题，尤其是针对非常规突发事件。一方面这类事件非常少发生，缺乏相应的案例；一方面人们又很难通过数学或者计算机进行模拟。这就需要对备用的案例解决方案进行适当的修正。目前尚无特别好的解决办法，很难实现大规模的修正。这也是“提示—顿悟”应急决策模式提出的意义所在。

案例学习：案例学习我们也可以称之为案例的自我修正，目的是把新发生的案例及其解决方案都汇总存储起来，实现案例的自我更新，是案例库不断的完善，使得案例中包含的知识、信息更加全面，更加符合实际效用。

在本文当中，我关注的重点是在案例表示和案例检索之间的——案例相似度表达。案例相似度表达，现在虽然有这样或者那样的一些算法，大都是基于计算机仿

真方面的研究，而对于案例本身关注很少，缺少了对于案例本身属性相似的思考。

我们注意到几乎所有的案例表示，无论是框架、语义网还是谓词逻辑，最终都落脚到了词汇方面。而一个案例，从语义学和词汇学的角度来看，其实就是一个普通的中文文本。而中文文本的构成基础是句子，是词组，是词汇。通过对文本的分解，对句子的分解，对词组的分解，我们可以得到文本构成的基础——词汇。这些词汇通过案例写作人的思考和某种逻辑融合到一起，组成了一篇包含丰富知识、信息的文献。我们掌握了这些词汇，通过对词汇进行整理、归纳、总结，从语义学的角度来审慎这篇案例。找出案例文本语义表达中的规律性、内在解释、不同语言在语义表达方面的个性以及共性，来帮助我们提高案例相似度匹配的成功率。

3 案例相似度研究环境分析

案例相似度研究环境是研究案例相似度的前提，这里面主要包括案例库的建设、案例文本的处理和案例文本的数据化表达。

3.1 案例库建设的机制

案例库的建设，其实就是案例学习。众所周知，案例库的建设周期很长，往往是很多专家学者多年智慧和积累的结晶。有必要采取一种机制来不断完善这样一个案例库建设。

为了达到案例库自身完善，具备“自我学习”的功能，可以将案例库建设设为以下五个步骤：（1）检索匹配度最高，和新案例最为相似的案例或者案例集；（2）将检索案例和新案例进行匹配，尤其是调用案例库中的信息和方案来解决问题；（3）调用案例和新案例不断匹配，不断优化，不断修正，尤其是经过实战检验的解决方案得到评估后，案例库更新推荐解决方案或者并提示新解决方案可能造成的后果；（4）决策人选择方案进行执行、评估，得出新案例的一个解决方案；（5）将这个案例进行整理、收纳进案例库当中为案例库提供新的经验和知识源泉。五个过程环环紧扣，可以用以下任务模型（图 2 所示）来进行描述。

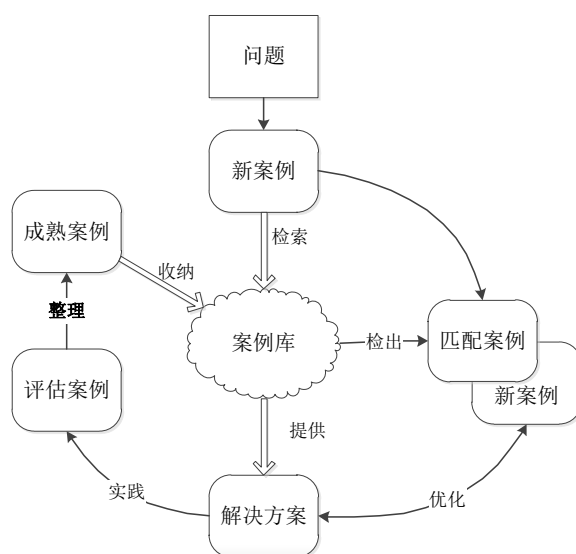


图 2 案例学习机制

3.2 案例比较的条件

公共突发事件的案例库与一般意义上的传统数据库相比，主要区别在于其核心数据总是以文本的形式进行展现，而非传统意义上的结构化的数据。进行案例比较研究，其实就是进行文本比较研究，需要对案例库中的历史案例文本进行结构化的处理，使得案例库中的案例以一种数据化的形式进行表达，而且案例文本的结构化、数据化的改造必须要能够准确体现原始文本的属性和特征。当然，对于不同领域的文本处理，因为要求和目标的不同，手段及方法也会有不同。一般针对文本数据挖掘和案例相似度分析来说，我认为必须要包含以下几个步骤。（图 3 所示）

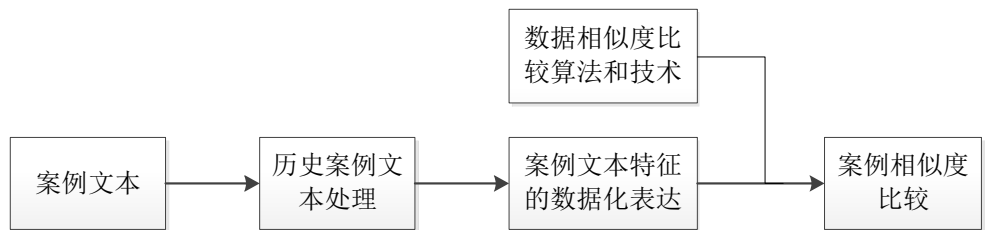


图 3 案例相似度研究条件

从上图可以看出，进行案例相似度的比较研究，必须要具备以下两个条件：（1）案例文本特征的数据化表达；（2）数据相似度比较的算法和技术。案例文本特征的数据化表达的前提是对案例文本的处理，将案例文本类的展现形式转变成数据库能够识别，或者能够构建数学模型的结构化数据形式。而数据相似度比较算法和技术，已经有很多人提出了各种各样的算法，比如求距离的相似度，求矢量相似度，我们可以从中得到一些启发，用来构建我们自己的数学模型。并且，我们必须要注意到，案例文本经过结构化、数据化后数据形式的特点，结合数据特点设计相应的数学模型。这些数据的特点包括，数量、分布、词性结构等等。只有结合到数据的特点才能对数据进行深入的挖掘，才能找出词汇或蕴藏的真正意义，最终达到找出文本真正意义的目的。

在本章中，我们需要做的首先是对案例文本进行处理，再在下一章深入讨论如何通过结构化的文本数据构建数学模型，寻找词汇之间的内在关联。

3.3 案例文本处理

在英文中单词与单词之间有通过自然的空格作为明确的分割符号，计算机能够通过识别空格来准确区分词汇。但是就中文字词而言，不仅没有这样的书写方式用以区分词汇，甚至一个文字同前后不同的长短的文字都能组成不同的词汇。凡是涉及到语义的研究，又都需要将词汇作为最基本的研究对象，将案例文本处理为分割词汇的过程，是要做的第一步。

当然，在进行文本处理之前，也就是在进行词汇分解之前，我这里有两个前提需要进行说明。

前提一：案例库案例的完整性和科学性是有保障的。因为本论文是基于国家自然科学基金重大项——“面向应急决策支持的非常规突发事件案例推理的理论与方法”，并且为后续的案例检索与匹配工作提供前期的理论支撑和技术铺垫，所以默认该案例库是完整且科学的。而且事实上该案例库也是在不断的更新、维护、完善当中。

前提二：本文分词所需要的紧急预警词库是完善的。因为中文分词想要达到较高的准确率，必须有相应的词库进行匹配。但是因为本文主要工作是做案例相似度的研究，紧急预警词库并不是本文的研究方向，所以默认是存在着这样的词库的。

3.3.1 中文机械分词法

本文主要采取的中文分词法是中文机械分词法。中文机械分词法又称为基于字符串的分词方法，即按照一定的逻辑或策略将待分的案例文本与一个词典中的词条进行逐一匹配。如果文本中的某个词汇与词典中词汇匹配成功，就作为优先分词逻辑。随后再对没有词典词条匹配的剩余文本进行正常逻辑匹配。当然，根据字符串长度的不同，匹配的情况也可以分为最长匹配法、最短匹配法、顺序匹配法、特殊词汇匹配法，正向匹配法和逆向匹配等等。当然这些具体的逻辑方法只是一种算法逻辑，词典的引入也是为了增加匹配成功率。因为这个项目是团队协作，本文默认存在应急词汇词典，这里只是使用这个词典。

当然，本文主要使用的是机械分词方法，但并不代表机械分词没有自身的缺陷，

比如穷举的词典并不能真正做到无一遗漏，比如涉及到前后文的逻辑关系词汇的理解，等等。所以在主要使用机械分词方法的同时，还会采用其他方法对机械分词进行优化。比如，基于理解的分词方法，这个方法是通过人工对句法进行定义，用标点作为词汇分隔符，从而判断句子类型，模拟出人对句子的理解，达到识别语义的目的。再比如，基于统计分词法，这个方法是统计词汇频率，来分辨词汇构成。当然，这些方法在本文中只是辅助方法，在这里只做简要的介绍。引进更多的方法的主要目的还是为了提高分词准确性。

3.3.2 文本相似度方法算法研究

没有现成的可以匹配文字、词汇和文本相似度的算法，我们只能从其他的一些匹配相似度的算法当中寻找灵感。当然，针对任何事物，使用不同的处理方法当然有可能会得到不同的结果。同样的，对文本相似度的处理，如果算法不一样，结果也应该会有很大的差别。例如，如果用 $s(x,y)$ 来表示文本样本 x, y 之间的相似度。一般可以通过某种算法得到 s 的值。可能当 s 的值比较小的时候， x 与 y 的相似度较低，反之亦然。当然，也有可能当 s 的值比较大的时候， x 与 y 的相似度较高。

如果用样本距离来类比描述相似度，可以理解为，某两点之间的距离 $d(x,y)$ 作为衡量标准。距离越大，相似度越低，距离越小，相似度越高。

设样本 $x = (x_1, x_2, x_3, \dots, x_n)$ ， $y = (y_1, y_2, y_3, \dots, y_n)$ ，我们可以通过不同的数学模型求出距离。

如，欧几里得距离为：

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

曼哈顿距离公式为：

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$$

切氏 (Chebyshev) 距离为：

$$d(x,y) = \max_i |x_i - y_i|$$

从上面的各种距离公式可以设想，要想得到数据相似度的模型，有两个必要条件，一是穷举的 x_n ，即必须要知道每一个待比较参数的值；二是这些待比较参数 x_n 总是和某一个参考值进行比较运算。虽然运算方式和最终得出结果的方式可能不一样，但是这两点却是不变的。

如果根据向量来描述相似度，可以理解为相似度是两个矢量的长度和方向一致的时候视为同一，而长度和方向相比，两个矢量的方向是考虑问题的基础，矢量的长度相对不那么重要。设两矢量样本为 $A=(x_1, y_1, z_1)$, $B=(x_2, y_2, z_2)$ ，则矢量相似系数为矢量余弦夹角。

$$\cos \angle(A, B) = \frac{AB}{|A||B|}$$

$$\cos \angle(A, B) = \frac{X_1X_2 + Y_1Y_2 + Z_1Z_2}{\sqrt{X_1^2 + Y_1^2 + Z_1^2}\sqrt{X_2^2 + Y_2^2 + Z_2^2}}$$

根据上式得出指数相似系数：

$$e(x, y) = \frac{1}{n} \sum_{i=1}^n \exp \left[-\frac{3(x_i - y_i)^2}{4\sigma_i^2} \right]$$

其中， σ_i^2 为 AB 矢量分量的方差，n 为矢量维数。

通过对于矢量相似度算法的研究，我们发现矢量的方向性才是决定矢量是否相似的关键，而不是数量。

基于距离公式和矢量公式相似度算法的比较研究，我们大胆猜测，文本相似度比较要具备三个要素：一、需要对词汇进行穷举分析；二、对分析词汇都要有一个参照值进行比较；三、可能决定文本相似度的并非词汇的多少，而是词汇某种特性的比例是否相似。

那么，针对这三个假设，我们采取下面三种方法来进行实验验证：一、对文本词汇词频进行分析；二、求出词汇词频匹配的期望，作为参考值来分析词汇的分布；三、研究词汇词性构成结构情况。

3.3.3 中文分词系统的演示使用

中文分词 (Chinese Word Segmentation)是文本挖掘的基础，输入任意的一段中文，成功的进行中文分词，可以达到电脑自动识别语句含义的效果，极大的减少工作量。比如以《奥地利卡普伦滑雪胜地列车隧道火灾》案例为例进行分词。

待分词文本：

表 1 《奥地利卡普伦滑雪胜地列车隧道火灾》案例结构表

案例描述	案例标题	奥地利卡普伦滑雪胜地列车隧道火灾
	案例时间	2000 年 11 月 11 日 9 时许
	案例地点	通往奥地利卡普伦滑雪场的列车上
	背景介绍	列车和隧道概况 卡普伦滑雪胜地位于奥地利萨尔茨堡州基茨施坦霍恩山，山的海拔 3202m。为了便于滑雪者上下山，开辟了一条长 3300m 的铁路隧道。山脚起点站海拔 911m，隧道上端终点站海拔 3029m。列车由一条钢缆牵引，沿着 45°倾斜角的铁轨行驶，通往滑雪训练基地。1994 年，该列车经过改造，车厢由一节增至两节，定员 180 人，每小时可运送 1500 名乘客。初学者的滑雪场海拔比较低，坐列车到达中转站再改乘小型缆车到达各个滑雪场。
	事件过程	<p>起火经过 2000 年 11 月 11 日 9 时许，列车从山脚起点站发车，沿轨道徐徐上升，不多时进入隧道继续上行。忽然，后面车厢有人闻到焦糊味，并发现有浓烟冒出。人们呼喊救命，有人隔着车厢玻璃给前节车厢的乘客打手势，叫司机采取措施，司机没有反应。9 时 30 分，司机终于察觉列车失火，便通过电话向本部报告，但很快通讯中断。列车在进入隧道 600m 处停下来。照明电源也被切断了。列车上的乘客去拉门，但因电源中断无法开启。</p> <p>逃生情景 列车缓缓停了下来，车厢里一片惊慌。这时有人用雪橇打碎车门玻璃，一位幸存者把女儿举着送到车外，自己和另外一些人也爬了出去。他们正准备向隧道顶部的出口走的时候，有人喊：“千万不能往上跑！火往上烧，烟往上冒，大家往下走啊！”十几个人听了他的话安全地逃生。而有些人弄错了方向，上逃中被浓烟熏死在隧道中。</p>
	事件原因	火灾原因 当时火灾原因还没有肯定的说法。据媒体报道有 5 种传闻：电线短路；乘客违章吸烟；有人携带易燃易爆物品；列车违反规定，人货混运，装载了本应夜间专车运送的液化石油气、汽油、柴油等；列车上的润滑油或制动液压装置漏油受车轮摩擦产生高温致灾。
	事件损失	火灾损失 火灾造成 155 人死亡，18 人受伤。遇难者中有奥地利人 92 名、德国人 37 名、日本人 10 名、美国人 8 名、斯洛文尼亚人 4 名、荷兰人 2 名、英国人和捷克人各 1 名。

将案例文本导入分词软件，可以迅速得到分词结果，下图（图 4）是软件的分词选项截图，表 2 时分词后统计的排名前 100 名的词汇和词频。

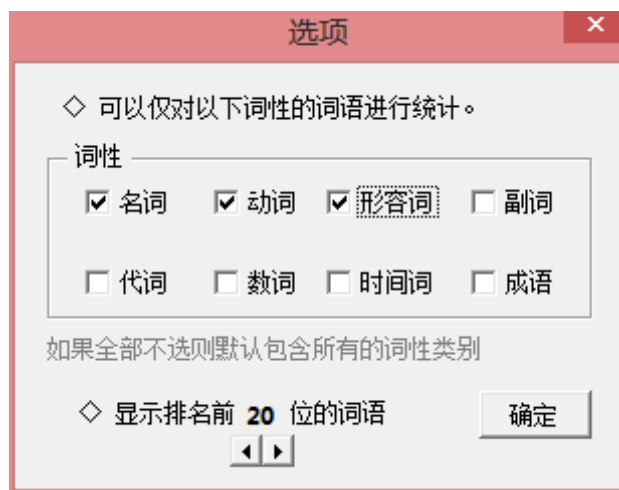


图 4 分词系统选择词性

分词过程：

奥地利\卡普伦\滑雪\胜地\列车\隧道\火灾\列车\和\隧道\概况\卡普伦\滑雪\胜地\位于\奥地利\萨尔茨堡州\基茨施坦霍恩山\山\的\海拔\3202m\为了\便于\滑雪者\上下山\开辟\了\一条\长\3300m\的\铁路\隧道\山脚\起点站\海拔\911m\隧道\上端\终点站\海拔\3029m\列车\由\一条\钢缆\牵引\沿着\45°\倾斜角\的\铁轨\行驶\通往\滑雪\训练\基地\1994 年\该\列车\经过\改造\车厢\由\一节\增至\两节\定员\180 人\每小时\可\运送\1500 名\乘客\初学者\的\滑雪场\海拔\比较\低\坐\列车\到达\中转站\再\改乘\小型\缆车\到达\各个\滑雪场\起火\经过\2000 年\11 月\11 日\9 时\许\列车\从\山脚\起点站\发车\沿\轨道\徐徐\上升\不多时\进入\隧道\继续\上行\忽然\后面\车厢\有人\闻到\焦糊味\并\发现\有\浓烟\冒出\人们\呼喊\救命\有人\隔着\车厢\玻璃\给\前节\车厢\的\乘客\打\手势\叫\司机\采取\措施\司机\没有\反应\9 时\30 分\司机\终于\察觉\列车\失火\便\通过\电话\向\本部\报告\但\很快\通讯\中断\列车\在\进入\隧道\600m\处\停下来\照明\电源\也\被\切断\了\列车\上\的\乘客\去\拉门\但\因\电源\中断\无法\开启\逃生\情景\列车\缓缓\停了\下来\车厢\里\一片\惊慌\这时\有人\用\雪橇\打碎\车门\玻璃\一位\幸存者\把\女儿\举着\送到\车外\自己\和\另外\一些人\也\爬了\出去\他们\正\准备\向\隧道\顶部\的\出口\走\的\时候\有人\喊\千万\不能\往上\跑\火\往上\烧\烟\往上\冒\大家\往下\走\啊\十几\个人\听了\他\的\话\安全\地\逃生\而\有些人\弄错了\方向\上逃\中\被\浓烟\熏死\在\隧道\中\火灾\原因\当时\火灾\原因\还\没有\肯定\的\说法\据\媒体\报道\有\5 种\传闻\电线\短路\乘客\违章\吸烟\有人\携带\易燃\易爆\物品\列车\违反\规定\人货\混运\装载\了\本应\夜间\专车\运送\的\液化\石油气\汽油\柴油\等\列车\上\的\润滑油\或\制动\液压\装置\漏油\受\车轮\摩擦\产生\高温\致灾\火灾\损失\火灾\造成\155 人\死亡\18 人\受伤\遇难者\中\有\奥地利人\92 名\德国人\37 名\日本人\10 名\美国人\8 名\斯洛文尼亚人\4 名\荷兰人\2 名\英国人\和\捷克人\各\1 名\

分词统计（统计前 100 个词汇）

表2 《奥地利卡普伦滑雪胜地列车隧道火灾》分词统计表

排名	词语	词性	词频	排名	词语	词性	词频	排名	词语	词性	词频
1	列车	名词	12	35	荷兰人	名词	1	69	行驶	动词	1
2	隧道	名词	8	36	雪橇	名词	1	70	察觉	动词	1
3	火灾	名词	5	37	不多时	时间词	1	71	电线	名词	1
4	车厢	名词	5	38	缆车	名词	1	72	吸烟	动词	1
5	海拔	名词	4	39	幸存者	名词	1	73	轨道	名词	1
6	有人	代词	5	40	专车	名词	1	74	液压	名词	1
7	乘客	名词	4	41	弄错	动词	1	75	传闻	名词	1
8	起点站	名词	2	42	铁轨	名词	1	76	媒体报道	名词	1
9	浓烟	名词	2	43	短路	动词	1	77	便于	动词	1
10	滑雪场	名词	2	44	增至	动词	1	78	携带	动词	1
11	司机	名词	3	45	起火	动词	1	79	高温	名词	1
12	逃生	动词	2	46	发车	动词	1	80	开启	动词	1
13	山脚	名词	2	47	终点站	名词	1	81	缓缓	副词	1
14	胜地	名词	2	48	运送	动词	1	82	装置	名词	1
15	斯洛文尼亚人	名词	1	49	本部	名词	1	83	概况	名词	1
16	中断	动词	2	50	徐徐	名词	1	84	送到	动词	1
17	倾斜角	名词	1	51	德国人	名词	1	85	顶部	名词	1
18	滑雪者	名词	1	52	打碎	动词	1	86	情景	名词	1
19	钢缆	名词	1	53	英国人	名词	1	87	停下	动词	1
20	奥地利人	名词	1	54	切断	动词	1	88	照明	名词	1
21	到达	动词	2	55	并发	动词	1	89	忽然	副词	1

22	经过	动词	2	56	制动	动词	1	90	这时	代词	1
23	改乘	动词	1	57	车轮	名词	1	91	违反	动词	1
24	电源	名词	2	58	车门	名词	1	92	各个	代词	1
25	原因	名词	2	59	润滑油	名词	1	93	现有	动词	1
26	玻璃	名词	2	60	呼喊	动词	1	94	说法	名词	1
27	进入	动词	2	61	下山	动词	1	95	出去	动词	1
28	没有	动词	2	62	汽油	名词	1	96	很快	名词	1
29	遇难者	名词	1	63	柴油	名词	1	97	损失	名词	1
30	拉门	名词	1	64	开辟	动词	1	98	上升	动词	1
31	定员	动词	1	65	通往	动词	1	99	下来	动词	1
32	上端	名词	1	66	夜间	时间词	1	100	人们	名词	1
33	石油气	名词	1	67	闻到	动词	1				
34	中转站	名词	1	68	初学者	名词	1				

3.3.4 文本分词情况小结

在上一节中，我们把《奥地利卡普伦滑雪胜地列车隧道火灾》这篇案例进行了详细的拆解，得到了排名前 100 的词汇。这篇案例只是一个示范，在实际的分析当中，我像这样一共拆解了 123 篇各式案例，分析的词汇总量是 307127 词，字数为 638731 字。当然，这只是进行文本分析前的一项必要工作，是进行文本分析的前提。

4 案例分析处理模型

4.1 案例词汇词频分析

4.1.1 词汇词频分析概念

在中文分词的基础上，对词汇的词频进行分析(Word Frequency Analysis)是指对文本的中词汇出现的次数进行整理、统计，是文本数据挖掘的重要手段。在开始词频分析之前，先介绍几个概念。

词频：指的是在一定范围内的语言文字材料中某个词汇出现的总次数；词频占比：指的是在一定范围内的语言文字材料中，词汇的使用频率，即某个词汇在一定文字语料中出现的总次数和该词汇在这段文字语料中所占百分比；词条：指的是词表中收录的词汇；语量（词汇量）：指的是该段文字语料中所包含的所有的词汇的数目。假设某段案例或者某段文字材料中包含有 X 个词汇，词语 A 出现了 N 次，则该词汇的词频占比是 $N/X100\%$ 。

当然，进行案例库相似度分析，语料规模，或者说抽样样本不能过于狭小。如果过于狭小，词频统计的结论可靠性将会降低，错误率将会上升。当然，也不能过于庞大，甚至穷举所有案例也是不现实的。本文最终完成的案例分析是 123 个案例，词汇总量是 307127 词，字数为 638731 字。案例平均字数为 5193 字，最短为《奥地利卡普伦滑雪胜地列车隧道火灾》，736 字，最长为《杞县钴 60 案》，10953 字。文章尤其针对 2013 年 10 月 7 日至 2013 年 10 月 12 日的，余姚水灾进行了案例跟踪分析，从刚刚开始台风预警到内涝结束。通过对这一案例的跟踪分析，试图找出一个那里在不同阶段，特征相似度发展的变化。

本文针对这 123 个案例进行了详细的分词、并统计了词频排名前五的词汇，与文章标题和文章关键词做了比对。因为现在一般的检索方法是对文章进行标题、关键词和摘要扫描，作为检索匹配依据。或者从某种意义上讲，标题和关键词的匹配就决定了文章是否匹配。但是我们通过统计发现，文章中出现的高频词汇和文章标题以及关键词差距很大。换句话说，因为案例写作人的主观意志以及概括能力的参

差不齐必定会导致现有的匹配方法是无法做到精确匹配的。

4.1.2 历史案例词频分析展示

通过对这 123 个案例的分词分析以及对案例的标题和关键词的统计汇总，我们得到了《历史案例分词情况汇总表》（表 25），详情见附录。

4.1.3 历史案例词频处理结果分析

经过对案例分词分析表的统计，我们可以得到：

表 3 历史案例分词情况统计表

灾害类别	案例数量	平均匹配度
公共卫生	12	51.67%
社会安全	19	29.47%
事故灾难	45	31.11%
自然灾害	47	32.34%
总计	123	33.33%

从上表我们看到，就算是最高的“公共卫生”类，高频词汇与标题、关键词的匹配度也只有 51.67%，案例总样本的匹配度只有 33.33%，这其实是非常低的。也就是说，标题和关键词的匹配不仅不能代表案例，甚至差距很大。

既然如此，我们是否可以设想，案例的标题和关键词因为作者的能力和主观判断的不同，并不能真实的反应案例的主旨，并不能将标题或者关键词作为判断案例相似的依据。相反的，案例的中文分词，没有人的主观因素的影响，可能更加客观一些。或者说，更能反映出案例的特征。

当然，单纯通过中文分词得出的词汇，准确率是比较低的。需要我们用合适的方法进行修正。就像上面的距离公式给我们的启示：不仅需要对全部文本进行处理，还需要把握其文本内在联系。接下来我们对案例词汇的分布情况和词性构成情况进行分析。

4.2 案例词汇分布分析

4.2.1 数据分布模型研究

我们对案例分词词频排前 30 的词汇，进行进一步的分析他们的分布情况。需要解释一下为什么不选择更多地词汇来分析，是因为：做过三个全案例词汇的分布情况分析对比，发现在排名超过 30 的词汇进行分布分析因为词汇量太小，不能代表文章的特征和重点，处理这些词汇没有意义，所以只选择对词频分析排前 30 的词汇进行分析。同样的，我们来看看一些处理数据分布的一些数学模型。

一般来说，对于数据（随机变量）进行分析，假设随机变量服从一个位置参数为 μ 尺度参数为 σ 的概率分布，且其概率密度函数为：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

也可以记作 $x \sim N(\mu, \sigma^2)$ 。当 $\mu = 0, \sigma = 1$ 时，为标准正态分布：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2}\right)$$

当然，我们还可以用求方差的方法来计算数据的波动和分布情况。

一般来说，当数据分布比较分散（即数据在平均数附近波动较大）时，各个数据与平均数的差的平方和较大，方差就较大；当数据分布比较集中时，各个数据与平均数的差的平方和较小。因此方差越大，数据的波动越大；方差越小，数据的波动就越小。

设 x 是一个随机变量，若 $E\{[x - E(x)]^2\}$ 存在，则称 $E\{[x - E(x)]^2\}$ 为 x 的方差，记为 $D(x)$

$$D(x) = \frac{1}{n} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

当然我们也可以用样本方差来估算方差。

$$D(x)' = \frac{1}{n-1} \sum_{xi} X^2 = \frac{1}{n-1} [(x_1 - x)^2 + (x_2 - x)^2 + \dots + (x_n - x)^2]$$

4.2.2 文本数据分布模型设计

我们利用上节的各类基本公式建立用以计算案例词汇分布的数学模型。首先，我们将案例均分为 m 段，设为样本空间 Ω 。针对不同的词汇 $A_1, A_2 \dots A_n$ ，他们是相互独立的，设 $P(A)$ 为词汇 A_n 在样本空间 Ω 的词频比例。 $E(X)$ 为该离散性随机变量的期望值（或数学期望、或均值，亦简称期望）。

在样本空间 Ω 里词频 $P(A)$ 为：

$$P(A_n) = \frac{A_n}{\Omega} \times 100\%$$

平均概率为：

$$\overline{P(A_n)} = \frac{P(A_1) + P(A_2) + \dots + P(A_n)}{n}$$

数学期望为：

$$E(X) = A_1 \times P(A_1) + A_2 \times P(A_2) + \dots + A_n \times P(A_n)$$

该案例文本的词汇分布方差为：

$$D(x) = \frac{1}{n} [(A_1 - E(X))^2 + (A_2 - E(X))^2 + \dots + (A_n - E(X))^2]$$

根据方差的概念，我们可以知道方差越大表示该词汇分布越不平均；方差越小，该词汇分布更加均匀一些。

4.2.3 历史案例词汇分布分析演示

下面，我们以《河南省杞县钴 60 案》为例来演示说明计算过程。

先通过分词得出《河南省杞县钴 60 案》前 30 位的词汇词频情况。

表 4 《河南省杞县钴 60 案》词汇词频情况统计表

排名	词语	词性	词频	排名	词语	词性	词频
1	的	副词	414	16	公开	动词	31
2	信息	名词	91	17	对	副词	28
3	政府	名词	91	18	公共	形容词	25
4	在	介词	85	19	公众	名词	23
5	了	副词	73	20	钴 60	名词	22
6	事件	名词	63	21	及时	形容词	22
7	和	连词	45	22	不	介词	21
8	杞县	名词	45	23	都	形容词	21
9	没有	动词	42	24	7 月	名词	20
10	是	动词	42	25	被	介词	20
11	危机	名词	42	26	并	连词	20
12	中	连词	40	27	等	副词	20
13	辐照	动词	33	28	民众	名词	20
14	突发	动词	32	29	人	名词	20
15	发生	动词	31	30	也	连词	20

经过与紧急预警词库比对和人工筛选，我们筛选出以下 11 个相对比较重要的高频词汇，作为进一步分布分析的词汇。

表 5 《河南省杞县钴 60 案》词汇词频情况筛选表

排名	词语	词性	词频	排名	词语	词性	词频
1	信息	名词	91	7	突发	动词	32
2	政府	名词	91	8	公开	动词	31
3	事件	名词	63	9	公共	形容词	25
4	杞县	名词	45	10	公众	名词	23
5	危机	名词	42	11	钴 60	名词	22
6	辐照	动词	33				

再比如以“钴 60”为例，我们得到“钴 60”的 22 个词在文本中的位置

表 6 《河南省杞县钴 60 案》“钴 60”词汇分布情况表

词汇	词汇分布位置	词汇	词汇分布位置
钴 60	1	钴 60	4078
钴 60	478	钴 60	4467
钴 60	1174	钴 60	4497
钴 60	1533	钴 60	4730
钴 60	1557	钴 60	4763
钴 60	1607	钴 60	5622
钴 60	1828	钴 60	6032
钴 60	1901	钴 60	6351
钴 60	3230	钴 60	6878
钴 60	3555	钴 60	7610
钴 60	3985	钴 60	10031

如果样本空间 Ω 为 1000，那么我们可以汇总如下：

表 7 《河南省杞县钴 60 案》“钴 60”概率分布表

$\Omega(n)$	0-1000	1001-2000	2001-3000	3001-4000	4001-5000	5001-6000
A(x)	2	6	2	1	5	1
P(A)	0.60%	1.80%	0.60%	0.30%	1.50%	0.30%
$\Omega(n)$	6001-7000	7001-8000	8001-9000	9001-10000	10001-11000	
A(x)	3	1	0	0	1	
P(A)	0.90%	0.30%	0.00%	0.00%	0.30%	

求出平均概率 $\overline{P(A_n)}$ 为：

$$\overline{P(A_n)} = 0.60\%$$

数学期望 E(X) 为：

$$E(X) = 2$$

得出“钴 60” 词汇分布方差为 $D(x) = 3.4545$

根据上面的计算方法，我们得出《河南省杞县钴 60 案》有效分布词汇的方差表，如下：

表 8 《河南省杞县钴 60 案》高频词汇分布重排表

词频排名	词语	词性	词频	方差	根据方差排名
4	杞县	名词	45	2.82	1
11	钴 60	名词	22	3.46	2
2	政府	名词	91	3.56	3
6	辐照	动词	33	3.77	4
7	突发	动词	32	4.56	5
9	公共	形容词	25	5.67	6
1	信息	名词	91	6.43	7
3	事件	名词	63	7.42	8
8	公开	动词	31	8.23	9
10	公众	名词	23	10.89	10
5	危机	名词	42	19.01	11

我们看到根据方差排名的顺序相比通过词频排名的顺序发生了明显的变化。也就是说，包括“杞县”、“钴 60”等词汇在文章的分布中更加平均。再考虑到这些词汇本身也是高频词汇，我们可以说，这些本身词频较高且分布方差较小的词汇是贯穿全文的，更能反应出案例的中心思想。事实上，通过对词汇分布情况对高频词汇匹配的修正，我们发现前五个词汇与标题和关键词的匹配度从 40%上升到了 60%。通过对词频排名和方差排名的对比，我们可以看出根据方差排名的词汇更能反应出文章的主旨和中心思想，在检索匹配的时候可能更加适合做为待检索匹配词汇。

4.2.4 历史案例词汇分布结果分析

通过使用上一节的方法，我们对所有 123 个案例的词汇分布情况进行了分析。通过分析，我们发现基本上所有种类的“分布修正后平均匹配度”都有了一定幅度

的提升，详细情况见下表。

表 9 历史案例词汇分布情况统计表

灾害类别	案例数量	分布修正后平均匹配度
公共卫生	12	56.47%
社会安全	19	41.41%
事故灾难	45	38.13%
自然灾害	47	42.74%
总计	123	42.19%

当然，也并非所有的案例都会在修正后匹配度上升，包括《中央电视台火灾事故报告》和《广东中山沙溪镇群体性聚集事件》在内的 16 个案例的匹配度出现了下降。不过总体上来看，通过分布算法修正后的词频匹配度有了明显的提升。

4.3 案例词汇词性分析

词性，一般是指根据的词的特点对词汇进行分类，用以表示词汇的属性和性质。现代汉语的词可以分为两类 12 种词性。包括实词：名词、动词、形容词、数词、量词和代词，以及虚词：副词、介词、连词、助词、拟声词和叹词等。我们分词系统设置的时候主要区分了包括名词、形容词在内的七种词性，再加上成语，一共八种。

在对案例词汇做词性分析的时候，我有两个思路，一是根据匹配案例中的词性构成占比来统计分析；另一个是根据现代汉语的句式构成来调整统计词汇词性构成。猜想，这两种词汇构成都应该能够有效的提升案例词汇匹配程度。

4.3.1 高匹配度词词性结构样式修正

根据个人的书写习惯的不同，每一篇案例肯定有着不同的词性构成，分词后的五个高频词汇也有着不同的词性构成比例。那么什么样的词性构成比例能够代表文章，更准确的匹配文章标题和关键词呢？一般认为有着更高匹配度的案例具有指导意义，我们把所有在高频词匹配中超过 60% 的案例拿出来做分析，如下图所示：

表 10 高匹配度词汇词性结构统计表

匹配度高案例的高频词汇	词 1 词性	词 2 词性	词 3 词性	词 4 词性	词 5 词性
疫区、疫情、禽流感、防控、事件	名词	名词	名词	名词	动词
口蹄疫、疫情、事件、疫区、防控	名词	名词	名词	名词	动词
火灾、燃放、烟花、央视、事故	名词	名词	名词	名词	动词
风雹、灾害、洪涝、暴雨、冰雹	名词	名词	名词	名词	名词
营养餐、乳业、事件、学生、乙方	名词	名词	名词	名词	名词
学生、大学、出现、腹泻、事件	名词	名词	名词	动词	动词
症状、医院、群体性、瘰病、患者	名词	名词	名词	名词	形 容 词
馒头、食品、超市、调查组、问题	名词	名词	名词	名词	名词
事件、双黄连、患者、注射液、救治	名词	名词	名词	名词	动词
征地、村民、事件、拆迁、原因	名词	名词	名词	动词	动词
暴徒、恐怖、暴力、团伙、工作人员	名词	名词	名词	名词	名词
和田、势力、事件、机场、机组	名词	名词	名词	名词	名词
塌陷、抢险、地面、管线、交通	名词	名词	名词	动词	动词
供热、事故、单位、抢修、处置	名词	名词	名词	动词	动词
隧道、油罐车、火灾、货车、堵截	名词	名词	名词	名词	动词
列车、隧道、火灾、车厢、海拔	名词	名词	名词	名词	名词
盐酸、泄漏、事故、处置、货车	名词	名词	名词	名词	动词
搜救、船员、遇难者、强台风、货船	名词	名词	名词	名词	动词
塌陷、岩溶、人员、顶板、发生	名词	名词	名词	动词	动词
事故、抢救、坍塌、事件、组织	名词	名词	动词	动词	动词
洪水、暴雨、事件、围困、多名	名词	名词	名词	动词	数 量 词
冰雹、灾害、大风、受损、绝收	名词	名词	名词	动词	动词

经过统计发现匹配度高的案例中，其中名词有 85 个，动词 23 个，形容词和数量词各 1 个，合计 110 个词。那么，名词的占比达到了 77.27%，动词 20.91%，形容词和数量词的占比分别为 0.91%。同样的，我们统计了匹配度较低的案例高频词汇词性占比。这个数字分别是名词占比 61.12%，动词占比 34.67%，形容词 2.23%，其他词 1.98%。如果我们将低匹配案例中的词汇做出调整，将多余的非名词词汇调整出匹

配词汇，顺位将后面的高频词纳入到匹配词汇。我这里以《山东威海海域沉没天津籍货船已致 7 人遇难》、《东北地区发生严重的洪涝灾害》和《甘肃定西市岷县、漳县暴洪泥石流灾害》三个案例来进行演示说明。

表 11 示范案例高频样式修正表

案例名称	原来匹配词汇	修正匹配词汇
甘肃定西市岷县、漳县暴洪泥石流灾害	灾害、抢修、群众、乡镇、抢险	灾害、抢修、群众、乡镇、泥石流
山东威海海域沉没天津籍货船已致 7 人遇难	救助、搜救、海上搜救、事件、航次	救助、事件、航次、威海、货船
东北地区发生严重的洪涝灾害	灾区、降水、高压、热带、暖湿气流	灾区、降水、高压、热带、洪涝

很明显，三个案例经过词性比例的修正，匹配度都有了一定的提高，分析原因还不一样。《甘肃定西市岷县、漳县暴洪泥石流灾害》中，动词“抢险”“抢修”显然是累赘重复的，用一个更加能够表明文章属性的“泥石流”来取代其中一个，必定能够增加匹配度。《山东威海海域沉没天津籍货船已致 7 人遇难》的问题和前者一样，不再赘述。而《东北地区发生严重的洪涝灾害》中全是名词，没有表明行为的动词，引入动词“洪涝”后，改善了词性构成，增加了匹配程度。

当然，也并不是所有的修正都能产生好的效果，包括《青海省乌兰县突发山洪 21 人遇难》的 23 个案例通过修正，匹配度并没有提高。通过对所有 123 个案例的进行高频词样式修正后结果如下：

表 12 历史案例高频样式修正统计表

灾害类别	案例数量	高频样式修正平均匹配度
公共卫生	12	51.63%
社会安全	19	48.14%
事故灾难	45	37.35%
自然灾害	47	39.78%
总计	123	46.22%

4.3.2 现代汉语句式修正

我们知道，在现代汉语中，一个句子必须按照一定的模式来组织，这个模式称

为句式。比如排比句、命令句、判断句、被动句，等等。我们不难发现几乎所有的标题都是采用陈述句式，而且陈述句式的结构往往是一定的。陈述句式一般是主谓宾的机构。即

陈述句=主语+谓语+宾语

陈述句=定语+主语+谓语+定语+宾语

陈述句=形容词+名词（代词）+动词+形容词（副词）+名词

也就是，如果高频词汇中如果按照名词（代词）2个，形容词（副词）2个，动词1个的结构配比，最有可能同案例标题和案例关键词匹配。我们还是找《西藏昌都地区左贡县、芒康县交界发生 6.1 级地震》、《内蒙古包头市应急处理盐酸泄漏事故》、《贵州仁怀茅台园区发生群体性事件》三篇案例来说明修正方式。

表 13 示范案例句式修正表

案例名称	原来匹配词汇	修正匹配词汇
西藏昌都地区左贡县、芒康县交界发生 6.1 级地震	地震、震级、构造、震源、边界	地震、震级、构造、剧烈、突然
内蒙古包头市应急处理盐酸泄漏事故	盐酸、泄漏、事故、处置、货车	盐酸、泄漏、事故、紧急、应急
贵州仁怀茅台园区发生群体性事件	群众、安置、处置、社会、聚集	群众、安置、社会、群体性、激动

在这三个案例当中，有两个案例匹配度有了一定的提升。但是《西藏昌都地区左贡县、芒康县交界发生 6.1 级地震》案例的匹配度并没有什么变化。而，最后的全部 123 个案例根据现代现代汉语句式修正的统计结果也印证了这一点。

表 14 历史案例句式修正统计表

灾害类别	案例数量	汉语句式修正平均匹配度
公共卫生	12	46.87%
社会安全	19	37.42%
事故灾难	45	37.14%
自然灾害	47	33.94%
总计	123	41.27%

根据现代现代汉语句式修正的匹配度提高是最少的，甚至在公共卫生这一类当中还出现了 4.8% 的匹配度下降。不过整体上来看，相较于单纯的高频词匹配还是提高了不少。分析原因，案例书写不同于一般意义上的文章，有着自己的行文规范和要求，可能是造成这种情况的原因。

4.4 案例文本相似度分析小结

针对文本案例相似度分析，首先提出了建立一个可以不断完善的案例库的机制，并在这个机制下，对于案例文本进行文本预处理和文本中文分词。通过类似的数学模型找到研究文本相似度的方向，即对文本全部词汇进行词频分析；并计算词汇分布期望，以此为参考来分析词汇的分布情况；最后根据词汇的两种词性结构构成情况来进行修正。通过大量的文本和数据验证。我们可以得出以下几个结论：

- 一、对案例文本的匹配不等于对于案例标题和关键词的匹配；
- 二、通过统计词汇词频来进行匹配可能是一种新的寻找文本案例相似度的思路；
- 三、在词汇词频进行匹配的基础上，通过对高频词汇分布状况的修正可以提高匹配成功率；
- 四、在词汇词频进行匹配的基础上，通过依照高频词汇词性结构的修正可以提高匹配成功率；
- 五、在词汇词频进行匹配的基础上，通过依照汉语句式词性结构的修正可能会适度提高匹配成功率，但是效果并不是很明显。

我这篇论文是选取了文本前五的高频词，当然也可以选取更多的高频词来分析（比如十个）肯定是可以提高匹配成功率的。但是，那样做也会造成检索匹配结果数量庞大，并不具备指导意义和实用性。

5 实证研究

在上一章中，分析过程基本上都是分词得到的词汇同案例文本本身进行对比，主要是同文本的标题和关键词进行对比。不难发现，文本标题和关键词并不能完全代表案例文本本身。这就需要通过交叉对比来判别上面的结论是否继续有效。那么，根据上一章的分析和模型构建，本章主要通过两组（六例）公共事件案例的对比分析，来对上一章的结论进行验证。当然，判别是否相似的方法，我们会采用到《知网》判别文本语义相似度的算法来进行验证。

在《知网》中是这样定义词汇之间相似度的，对于两个汉语词汇 w_1 和 w_2 ，如果 w_1 有 n 个意向： s_1, s_2, \dots, s_n ； w_2 也有 m 个意向： s'_1, s'_2, \dots, s'_m ，就规定 w_1 和 w_2 之间的相似度是各个的最大值，也就是：^[33]

$$\text{Sim}(w_1, w_2) = \max_{i=1 \dots n, j=1 \dots m} \text{Sim}(s_i, s'_j)$$

当然，这是《知网》关于语义的判断，我这里只是拿来用以验证第二章的模型和结论，具体过程，不再详述。

5.1 相似案例相似度比较研究

在本小节，我要对比的是这样三个案例，《北京市西城区西西工程坍塌案》、《北京市东三环京广桥路面坍塌案》和《湖北襄阳市南漳建设工地脚手架坍塌案》进行详细分析来验证。之所以选择这样三个案例，主要因为三个案例都是建筑物坍塌案，而时间、地点、特征等却不尽相同。

经过中文分词，我们得到了三个案例不同的分词结果，根据顺序分别记作 A、B 和 C。

《北京市西城区西西工程坍塌案》的分词结果：

A1={事故、公司、坍塌、建筑、处置}

《北京市西城区西西工程坍塌案》的分布修正结果：

A2={事故、公司、坍塌、建筑、工程}

《北京市西城区西西工程坍塌案》的高匹配度词词性结构样式修正结果

A3={事故、公司、坍塌、工程、单位}

《北京市西城区西西工程坍塌案》的现代汉语句式结构修正结果

A4={事故、公司、坍塌、处置、工程}

那么我们可以根据上面四个分词和修正结果得出 A 的构成。

A= {事故、公司、坍塌、处置、工程、单位、建筑}

《北京市东三环京广桥路面坍塌案》的分词结果

B1= {抢险、管线、事故、污水、恢复}

《北京市东三环京广桥路面坍塌案》的分布修正结果

B2= {抢险、管线、事故、恢复、交通}

《北京市东三环京广桥路面坍塌案》的高匹配度词词性结构样式修正结果

B3= {抢险、管线、事故、交通、隧道}

《北京市东三环京广桥路面坍塌案》的现代汉语句式结构修正结果

B4= {抢险、管线、事故、恢复、交通}

那么我们可以根据上面四个分词和修正结果得出 B 的构成。

B= {抢险、管线、事故、污水、恢复、交通、隧道}

《湖北襄阳市南漳建设工地脚手架坍塌案》的分词结果

C1= {事故、建设、脚手架、事件、工地}

《湖北襄阳市南漳建设工地脚手架坍塌案》的分布修正结果

C2= {事故、建设、脚手架、工地、倒塌}

《湖北襄阳市南漳建设工地脚手架坍塌案》高匹配度词词性结构样式修正结果

C3= {事故、建设、脚手架、工地、公司}

《湖北襄阳市南漳建设工地脚手架坍塌案》的现代汉语句式结构修正结果

C4= {事故、建设、脚手架、工地、倒塌}

那么我们可以根据上面四个分词和修正结果得出 C 的构成。

C= {事故、建设、脚手架、事件、工地、倒塌、公司}

我们将 A、B、C 的不同类别进行归类，将 A= {事故、公司、坍塌、处置、工程、单位、建筑} 和 B= {抢险、管线、事故、污水、恢复、交通、隧道} 可以写成矩阵 AB。如下：

表 15 AB 矩阵相似度计算表

	事故	公司	坍塌	处置	建筑	工程	单位
抢险	0.044444	0.044444	0.117647	0.186047	0.186047	0.074074	0.044444
管线	0.111628	0.111628	0.074074	0.074074	0.074074	0.369231	0.096533
事故	1.000000	0.188889	0.044444	0.044444	0.044444	0.132554	0.160269
污水	0.122997	0.134367	0.044444	0.044444	0.044444	0.160269	0.12037
恢复	0.044444	0.044444	0.166667	0.242424	0.210526	0.242424	0.044444
交通	0.266667	0.171429	0.074074	0.074074	0.126316	0.615385	0.140412
隧道	0.122997	0.134367	0.044444	0.044444	0.406838	0.12037	0.12037

也就是

$$AB = \begin{bmatrix} 0.044444 & 0.044444 & 0.117647 & 0.186047 & 0.186047 & 0.074074 & 0.044444 \\ 0.111628 & 0.111628 & 0.074074 & 0.074074 & 0.074074 & 0.166667 & 0.096533 \\ 1.000000 & 0.188889 & 0.044444 & 0.044444 & 0.044444 & 0.229952 & 0.160269 \\ 0.122997 & 0.134367 & 0.044444 & 0.044444 & 0.044444 & 0.12037 & 0.12037 \\ 0.044444 & 0.044444 & 0.166667 & 0.242424 & 0.210526 & 0.074074 & 0.044444 \\ 0.266667 & 0.171429 & 0.074074 & 0.074074 & 0.126316 & 0.615385 & 0.140412 \\ 0.122997 & 0.134367 & 0.044444 & 0.044444 & 0.406838 & 0.12037 & 0.12037 \end{bmatrix}$$

使用公式

$$\text{Sim}(A, B) = \frac{1}{m \times n} \sum_n \text{sim}(A) \sum_m \text{sim}(B)$$

计算出两个案例 AB 的相似度为 $\text{Sim}(A, B) = 0.150048$

再来计算 B= {抢险、管线、事故、污水、恢复、交通、隧道} 和 C= {事故、建设、脚手架、事件、工地、倒塌、公司} 构成的矩阵 BC。如下：

表 16 BC 矩阵相似度计算表

	抢险	管线	事故	污水	恢复	交通	隧道
事故	0.044444	0.111628	1.000000	0.122997	0.044444	0.266667	0.122997
建设	0.186047	0.074074	0.044444	0.044444	0.242424	0.074074	0.044444
脚手架	0.074074	0.896000	0.111628	0.126316	0.074074	0.165947	0.600000
事件	0.074074	0.186047	0.266667	0.111628	0.074074	0.165947	0.111628
工地	0.074074	0.188632	0.111628	0.171429	0.074074	0.166698	0.126316
倒塌	0.117647	0.074074	0.044444	0.044444	0.166667	0.074074	0.044444
公司	0.044444	0.111628	0.188889	0.134367	0.044444	0.171429	0.134367

也就是

$$BC = \begin{bmatrix} 0.044444 & 0.111628 & 1.000000 & 0.122997 & 0.044444 & 0.266667 & 0.122997 \\ 0.186047 & 0.074074 & 0.044444 & 0.044444 & 0.242424 & 0.074074 & 0.044444 \\ 0.074074 & 0.896000 & 0.111628 & 0.126316 & 0.074074 & 0.165947 & 0.600000 \\ 0.074074 & 0.186047 & 0.266667 & 0.111628 & 0.074074 & 0.165947 & 0.111628 \\ 0.074074 & 0.188632 & 0.111628 & 0.171429 & 0.074074 & 0.166698 & 0.126316 \\ 0.117647 & 0.074074 & 0.044444 & 0.044444 & 0.166667 & 0.074074 & 0.044444 \\ 0.044444 & 0.111628 & 0.111628 & 0.188889 & 0.044444 & 0.171429 & 0.134367 \end{bmatrix}$$

$$\text{Sim}(B, C) = 0.158561$$

最后计算 $A = \{\text{事故、公司、坍塌、处置、工程、单位、建筑}\}$ 和 $C = \{\text{事故、建设、脚手架、事件、工地、倒塌、公司}\}$ 构成的矩阵 AC 。如下

表 17 矩阵相似度计算表

	事故	公司	坍塌	处置	工程	单位	建筑
事故	1.000000	0.188889	0.044444	0.044444	0.229952	0.160269	0.132554
建设	0.044444	0.044444	0.126984	0.210526	0.074074	0.044444	1.000000
脚手架	0.111628	0.111628	0.074074	0.074074	0.166667	0.096533	0.369231
事件	0.266667	0.171429	0.074074	0.074074	0.615385	0.145455	0.126316
工地	0.111628	0.111628	0.074074	0.074074	0.166667	0.096533	0.145455
倒塌	0.044444	0.044444	1.000000	0.137931	0.074074	0.044444	0.126984
公司	0.188889	1.000000	0.044444	0.044444	0.175084	0.722222	0.139181

也就是说

$$AC = \begin{bmatrix} 1.000000 & 0.188889 & 0.044444 & 0.044444 & 0.229952 & 0.160269 & 0.132554 \\ 0.044444 & 0.044444 & 0.126984 & 0.210526 & 0.074074 & 0.044444 & 1.000000 \\ 0.111628 & 0.111628 & 0.074074 & 0.074074 & 0.166667 & 0.096533 & 0.369231 \\ 0.266667 & 0.171429 & 0.074074 & 0.074074 & 0.615385 & 0.145455 & 0.126316 \\ 0.111628 & 0.111628 & 0.074074 & 0.074074 & 0.166667 & 0.096533 & 0.145455 \\ 0.044444 & 0.044444 & 1.000000 & 0.137931 & 0.074074 & 0.044444 & 0.126316 \\ 0.188889 & 1.000000 & 0.044444 & 0.044444 & 0.175084 & 0.722222 & 0.139181 \end{bmatrix}$$

$$\text{Sim}(A, C) = 0.212028$$

我们不难发现， $\text{Sim}(A, B) \approx \text{Sim}(B, C) \ll \text{Sim}(A, C)$ ，也就是说，A 案例和 C 案例的相似度要远高于 A 案例和 B 案例的相似度，也要高于 B 案例和 C 案例的相似度。通过阅读文章我们也发现，A 案例和 C 案例虽然发生的地点不一样，但是都是建筑物在修建过程中发生的垮塌；而 B 案例，却是道路塌陷，并且造成了严重的管线受损和交通堵塞。案例本身的差别是造成案例文本相似度计算上差别的主要原因。

那么我们现在设 A 案例和 C 案例相似，那么 C 案例在分词到分布修正到词性修成中是否是越来越准确了。下面继续计算验证：

A1={事故、公司、坍塌、单位、处置}

A2={事故、公司、坍塌、建筑、单位}

A3={事故、公司、坍塌、工程、建筑}

A4={事故、公司、坍塌、处置、工程}

C= {事故、建设、脚手架、事件、工地、倒塌、公司}

那么 A1C 矩阵为

表 18 A1C 矩阵相似度计算表

	事故	公司	坍塌	建筑	处置
事故	1.000000	0.188889	0.044444	0.132554	0.044444
建设	0.044444	0.044444	0.126984	1.000000	0.210526
脚手架	0.111628	0.111628	0.074074	0.369231	0.074074
事件	0.266667	0.171429	0.074074	0.126316	0.074074
工地	0.111628	0.111628	0.074074	0.145455	0.074074
倒塌	0.044444	0.044444	1.000000	0.126984	0.137931
公司	0.188889	1.000000	0.044444	0.139181	0.044444

$$\text{Sim}(A1,C) = 0.216501$$

A2C 矩阵为

表 19 A2C 矩阵相似度计算表

	事故	公司	坍塌	建筑	单位
事故	1.000000	0.188889	0.044444	0.132554	0.160269
建设	0.044444	0.044444	0.126984	1.000000	0.044444
脚手架	0.111628	0.111628	0.074074	0.369231	0.096533
事件	0.266667	0.171429	0.074074	0.126316	0.145455
工地	0.111628	0.111628	0.074074	0.145455	0.096533
倒塌	0.044444	0.044444	1.000000	0.126984	0.044444
公司	0.188889	1.000000	0.044444	0.139181	0.722222

$$\text{Sim}(A2,C) = 0.235082$$

A3C 矩阵为

表 20 A3C 矩阵相似度计算表

	事故	公司	坍塌	建筑	工程
事故	1.000000	0.188889	0.044444	0.132554	0.229952
建设	0.044444	0.044444	0.126984	1.000000	0.074074
脚手架	0.111628	0.111628	0.074074	0.369231	0.166667
事件	0.266667	0.171429	0.074074	0.126316	0.615385
工地	0.111628	0.111628	0.074074	0.145455	0.166667
倒塌	0.044444	0.044444	1.000000	0.126984	0.074074
公司	0.188889	1.000000	0.044444	0.139181	0.175084

$$\text{Sim}(A3,C) = 0.240568$$

A4C 矩阵为

表 21 A4C 矩阵相似度计算表

	事故	公司	坍塌	处置	工程
事故	1.000000	0.188889	0.044444	0.044444	0.229952
建设	0.044444	0.044444	0.126984	0.210526	0.074074
脚手架	0.111628	0.111628	0.074074	0.074074	0.166667
事件	0.266667	0.171429	0.074074	0.074074	0.615385
工地	0.111628	0.111628	0.074074	0.074074	0.166667
倒塌	0.044444	0.044444	1.000000	0.137931	0.074074
公司	0.188889	1.000000	0.044444	0.044444	0.175084

$$\text{Sim}(A4, C) = 0.201135$$

我们通过实验案例可以看到，通过分布修正和高匹配度词词性结构样式修正后，词汇的匹配度相比单纯的分词均有了一定的增加，而通过现代汉语句式结构修正后的匹配度有一定的小幅下降。这也很符合在第二章得出的结论。

5.2 同案例不同阶段相似度比较研究

在这一节，我选择的是余姚水灾专题来进行对比说明。余姚水灾从开始台风预警，到后来变成城市内涝，再到后来演化为以救援为主和最后需要强力维持稳定，在不同阶段表现除了不同的应急需求和特点。这里找到了七篇当时的新闻稿，整理成了案例，用来验证一个案例在不同阶段的变化情况。

还是先进行中文分词（按照事件发生的时间排序），我们把七个案例分别命名为 S1, S2, S3, S4, S5, S6, S7

《台风“菲特”登陆与“丹娜丝”形成双台风效应》

S1= { 台风、紫菜、强台风、船只、船员 }

《浙江宁波余姚堤防决口》

S2= { 堤防、险情、决口、抢险、排涝 }

《宁波母亲河水位创新高 市民广场被淹不见底》

S3= { 水位、抢险、大闸、水闸、江水 }

《余姚遭菲特重击 大水围城乡镇被淹》

S4= { 水库、停运、泄洪、水位、乡镇 }

《浙江余姚市 7 成被淹 四个安置点电话均接不通》

S5= { 积水、云系、内涝、台风、海葵 }

《余姚 300 个安置点安顿民众 救援物资储备充足》

S6= { 安置点、村民、矿泉水、物资、镇政府 }

《宁波电视台关于余姚卫星车被围堵事件的说明》

S7= { 警察、记者、卫星、群众、特别报道 }

其实，不需要计算，从分词结果已经可以看出来整个事件发展的变化，从开始

的预防台风，并没有引起足够高的重视，案例中甚至还在重点关注一些生产的情况（紫菜）；到中间姚江堤防决口，江河倒灌，导致城市内涝；而后出现的安置问题和物资紧张问题凸显，最后发展到群众围堵记着。我们来验算几个案例，看看相似度的情况。

我们构建 S1S3，S3S4，S5S6，这样三个矩阵，

S1S3 矩阵如下：

表 22 S1S3 矩阵相似度计算表

	台风	紫菜	强台风	船只	船员
水位	0.044444	0.044444	0.044444	0.039781	0.074074
抢险	0.074074	0.074074	0.074074	0.044444	0.074074
大闸	0.210526	0.210526	0.210526	0.145455	0.166529
水闸	0.210526	0.285714	0.210526	0.145455	0.166529
江水	0.285714	0.210526	0.285714	0.111628	0.166529

$$\text{Sim}(S1, S3) = 0.144414$$

S3S4 矩阵如下：

表 23 S3S4 矩阵相似度计算表

	水库	停运	泄洪	水位	乡镇
水位	0.039210	0.044444	0.044444	1.000000	0.042112
抢险	0.044444	0.104828	0.096508	0.044444	0.044444
大闸	0.583043	0.074074	0.074074	0.042904	0.126316
水闸	0.583043	0.074074	0.074074	0.042904	0.126316
江水	0.122746	0.074074	0.074074	0.042904	0.171429

$$\text{Sim}(S1, S3) = 0.151637$$

S5S6 矩阵如下：

表 24 S5S6 矩阵相似度计算表

	积水	云系	内涝	台风	海葵
安置点	0.044444	0.074074	0.044444	0.074074	0.074074
村民	0.126316	0.186047	0.121707	0.186047	0.126316
物资	0.153439	0.155331	0.152047	0.145455	0.042112
矿泉水	0.145455	0.210526	0.111628	0.210526	0.044444
镇政府	0.113060	0.134893	0.193615	0.111628	0.126316

$$\text{Sim}(S1, S3) = 0.124321$$

通过计算，我们发现 $\text{Sim}(S1, S3)$ $\text{Sim}(S1, S3)$ $\text{Sim}(S1, S3)$ 三个矩阵的计算值都很低，也就是说，三个案例通过分词得到的词汇的相似度匹配度很低。即，这几个案例之间的相似度是比较低的，而且这种相似度匹配程度低很难通过修正来实现好转。

我们通过这样的一个对比实验发现，案例在发展过程当中，案例相似度的特征在发生着较大的变化，甚至从词汇的角度上来看变得毫无关联。这种变化之大可能是超过我们的想象的。正是突发事件在发展过程中的这种突变性，为我们处理解决这类事件提出了极大的挑战。

6 案例相似度研究的应用

6.1 案例推理中决策者与案例库的交互

传统的案例推理工作，就是收集目标问题中的信息，并与案例库中历史案例的全部信息或属性进行匹配。并随着信息量的逐步增多，决策者根据这些信息不断的进行调整，并得到一个完整的解决方案。但是当一次突发事件发生，因为信息的不完全和受限于决策者自身的知识结构等客观因素，导致应急决策者很难找出合适的案例用以参考辅助决策。在实际的决策过程中，案例的启发交互至少存在下面三种方式。

（1）直接借鉴式。此类案例主要属于常规突发事件类型。即，突发事件的类型、演变方式都与案例库中某历史案例发展过程高度一致。如火灾事件、煤矿渗水事故、煤矿瓦斯事故等。在这种情况下，决策者就能直接利用历史案例进行应急处置策略，快速得出新问题的解决方案。

（2）触类旁通式。事实上，世界上找不到完全相同的两片树叶，对于突发事件也是一样的。常规突发事件虽然事件本质属性相似甚至相同，但是这些事件发生的时间、地点等不可能完全一样。决策者如果能把握某一类事件或特定事件发展变化趋势和规律，那么目标事件的发展过程就可以通过类比推断，从历史案例中触类旁通，借鉴到相应的解决框架等，再根据具体特点进行微调。如玉树地震和汶川地震，虽然两者都是地震灾害，都发生在西部贫穷地区，但是玉树因为地处高原，救援方案必须要考虑到这一点。但是，汶川地震的救援方案仍然可以给玉树地震的决策提供重要借鉴意义。

（3）人机互助式。在应急决策实践中，最为困难的事处置非常规非典型性突发事件。因为非常规非典型性的突发事件的表现形式可能与既往经验和知识体系不一致。往往只有极少相同或者几乎没有相同的属性和信息。完全不能用简单的类比或者借鉴的方式来生成完整的解决方案。但是，如果对此类事件进行谨慎的，深入细致的分析，还是可以得到诸如事件情境、阶段特征、演化过程等方面部分相同的情

报信息。不仅如此，根据事件演化发展，对于一些从来没有发生过的事情，只能通过计算机模拟，来评估事件发展动态。换句话说，只有当决策人和计算机之间高效互动，彼此做出一些合理的推理联想，通过与案例库中的历史案例相互学习、不断修正，才能根据目标问题的特征进行合理的研判，直至找出最佳决策方案。

6.2 案例相似度应用路径分析

在整个应急处置过程中，按照突发事件发生的时间顺序，大致可以分为事前、事中、事后三个基本阶段。很显然，应急响应阶段是包括事前与事中两个阶段，而业务恢复则在事后。当然根据核心业务的不同，各阶段应急处置侧重点也是不同的。一般认为，事前属于防控阶段；事中属于响应阶段，而事后属于重建阶段。可以看出，事后重建与本文没有直接关系，不再赘述。

事前防控阶段。俗话说，防患于未然。只有当我们了解了运作过程，方便鉴别业务运用中可能产生的威胁，进行合理的风险评估，分析各种危险和问题可能对组织造成的冲击，探索各种可选择的策略，同时组织相关培训，才能在面对突发事件时冷静、灵活应对。

事中响应阶段。突发事件发生时，应该因势利导的按照优先级和各种约束要求做出救援计划，并加以贯彻实施。以确保支持关键业务活动的救援活动能够在合理的时间内有效持续进行。

而我们强调的是，在整个应急决策过程中可能都会使用到案例相似度的应用。也就是说，可能在整个突发事件发生的前期和中期，凡是涉及应急决策的领域，都有可能使用到案例相似度应用。这并不只是由突发事件本身的属性所决定的，还受到空间、时间、环境和决策者知识结构的约束。缺乏决策依据和决策经验的领域，都可以引入案例相似度应用来辅助决策。一般的，我们可以将该辅助推理过程概括为五个步骤——“采集—提示—启发—顿悟—验证”。

(1) 信息采集阶段。事件开始发生后，来自各方面的信息会汇总到各个信息采集中心，并进一步汇总到决策者。决策者通过将外界传递信息和记忆中的知识进行比对，并加以处理。

（2）提示阶段。当决策者发现储备的知识和经验无法解决当前问题，可以采取向外界寻求帮助的方式来辅助决策。一种，可以通过会议来辅助决策；另一种就是计算机辅助决策。在该阶段，需要向计算机输入客观，真实的信息，并可能会持续不断的更新这些信息。

（3）启发阶段。随着信息的输入，案例相似度应用会给出相应匹配的案例和解决方案，偶发的信息会触发决策者的思维联想。对比不同案例件的差异，可以更加明晰的比较不同决策方案的差异。随着信息的不断输入，匹配度越高，会使得提示的解决方案更加接近该突发事件。

（4）顿悟阶段。由于启发过程中产生的联想不断发展，使得决策者的认知发生飞跃。能够在更早的时候和信息量更少的时候认识到问题的本质。我这里所说的顿悟，并不是指空穴来风，而是在案例相似度应用辅助的情况下，利用较少的信息的提示，在更早的时间就可以做出更加符合当前突发事件的决策方案。

（5）验证阶段。这属于案例学习的一部分，是案例库的自我更新，自我完善的过程。

6.3 政府使用案例相似度应用的启示和要求

政府在面对突发事件过程中，即便是集体决策也难免会出现疏漏以及瑕疵，更不用说在决策实践中更多地是以领导意志为中心。领导转化为决策人，当然决策人所面临的问题和困扰也并不会因为领导的身份发生变化。信息不完全、环境多变、约束条件多、事件发展快速，以及领导的经验和知识储备并不能够完全控制整个决策过程等等。这些情况下就可以引入计算机辅助决策，即案例相似度的应用。

对于政府来说，使用案例相似度的应用，即使用计算机辅助决策，因为政府特有的权威性，有着自身特殊的约束条件，又有着同实验不同的使用要求。

（1）决策的责任主体在于决策者，而非机器。无论哪种计算机辅助决策，都只能起到辅助决策的作用，并不能代替决策者做出决策。无论从政治上还是法律上，责任的主体都是决策者，而不是计算机。计算机辅助决策只是一种科学辅助决策，通过不断的优化更新可以提供更优的决策方案，但是并不能替代决策者做出决策。

（2）历史案例库建立的科学性和完整性。政府需要面对的问题包罗万象，并不等同于实验中的决策可以选择一个方面作为研究对象。尤其是需要应用与实践的计算机辅助决策的领域，更加需要通过政府意志建立最为广范的案例库系统和数据库系统，以一种和多种算法来进行评估优化，并不断更新完善。

（3）信息采集的科学性。现有的决策模式更多的是采用人和数据共同报告相关信息，甚至在决策过程中更加偏重下属的报告情况。在案例相似度应用中我们可以看到，该算法本身就是为了避免人的主观观念的影响。而且，从论文标题关键词与论文中心思想的不一致也反应出了人的主观观念确实会影响到对于客观案例的判断。

（4）避免信息孤岛，实现信息联动。在决策实践中，信息来源往往很多，而信息的缺乏作为早期影响决策最关键的因素之一是我们关注的焦点。这里的信息孤岛包含两层意思：一、灾害引起的信息阻断，造成信息孤岛；二、沟通协调不畅，本位思想引起的信息孤岛。对于第一种情况，只能要求通信部门加快恢复通信，并采取一些受影响较小的通信方式进行沟通；对于第二种情况，要求在管理中打破本位思想，实现更大范围内的信息共享，至少在应急决策过程中必须如此。

（5）决策过程不断的更新。案例相似度匹配中，总是根据最新的信息给出相应的匹配案例和解决方案，而事件发生也一定会产生新的信息。随着信息的不断输入，案例的匹配和解决方案也会随之发生变化。实证案例二也很好的说明了这一点，随着事件的发生，会表现出不同的中心信息，自然也会得出不同的解决方案。这就要求决策者不断根据信息的变化对已有方案做出适当调整，以满足最新的决策需求。

（6）事件之后的验证与评估。人文事件，尤其是灾害事件，因为成本高昂，很难通过实验室模拟来实验。灾害本身就是验证解决方案最好的实验场。把握每一次灾害，对事件做出详尽的、全面的评估，更新历史案例库，方能使得每次灾害后决策水平有所提升。

7 结束语

7.1 研究结论

我的研究对象是突发事件案例的相似度。众所周知，突发事件有着极强的不确定性、突变型、迅速扩张性和极强的破坏性，客观要求政府部门在灾后第一时间响应，快速做出决策。然而决策者在短时间难以充分掌握全面的信息，更难以对信息进行快速分析处理，并做出最优决定。这一方面是因为客观的信息不完全造成的，另一方面也受限于决策者自身的能力和知识结构。甚至可以说，让每一个决策者在极短的时限内做出可以承担责任的决策，这本身就是一种不负责任。而，如果无法采取最优的处置措施，往往会使得突发事件造成更大的负面影响，甚至引发次生灾害。整个国家自然科学基金重大项目试图从案例相似度比较的角度入手，从计算机和大数据的角度来处理旧有的历史案例。通过对历史案例的分析以及案例的快速检索匹配，来为决策者提供做出最优决策的参考建议。本文的尝试用中文分词的方法来捕捉案例的特征，用词频分布分析案例词汇分布特点，用词性构成来优化匹配方案。

论文得出如下结论：

- 1、对案例标题和关键词的匹配不等于对案例文本的匹配。
- 2、通过统计文本词汇词频来进行匹配可能是一种新的寻找文本案例特征相似度的有效办法。这样做一方面可以避免主观偏差的影响，一方面可以进行量化计算。
- 3、在词汇词频进行匹配的基础上，可以通过对高频词汇分布状况来修正匹配词汇的构成，从而达到提高匹配成功率的目的；
- 4、在词汇词频进行匹配的基础上，通过分析高频词汇词性结构的，参照这种结构构成优化匹配词汇构成，可以提高匹配成功率；
- 5、在词汇词频进行匹配的基础上，依照现代汉语句式词性结构的修正可能会适度提高匹配成功率，但是效果并不是很明显。
- 6、通过对词频分析、词汇分布优化，词性构成优化后的匹配词汇组可以有效的

保证相同案例之间的匹配。

7、随着案子的发展，案例的特征也在发生变化，而这种变化往往是很激烈的，甚至是一种突变。案子发展的前后从词汇词频、分布和词性构成的角度上看不出必然的联系。相同案子，不同阶段的案例文本必定也是难以匹配的。

7.2 研究不足

1、在本文撰写的时候文章开头的两个假设（案例库和紧急预警词库），都还不够完善，尤其是紧急预警词库的缺乏会导致分词中误差增加。

2、文章没有能够进一步挖掘和比较多种算法之间的优缺点，文章得出的结论只能是初步的案例相似度方面的研究成果，距离真正的实际应用还有一定的距离。

3、虽然本文分析的数据量不小，但是真实的案例库数据量更加庞大，本文没有考虑到这两者之间在处理和管理上的差别。

7.3 研究展望

随着公共应急事件案例库（知识库）的发展和完善，结合案例推理，案例检索等理论和实际应用，公共应急案例库一定能够为决策者提供更好的决策支持。当然，这还需要做大量的工作：

- 1、结合更多地案例相似度分析方法继续优化案例相似度的研究。
- 2、对公共应急事件进行分类研究可能能够取得更好的效果。
- 3、走出案例本身，通过大数据，同更多的社会反应、舆论反应建立联系，优化研究。

注 释

- [1] Salvatore Belardo , Harold L. Pazer . A Framework for Analyzing the Information Monitoring and Decision Support Investment Tradeoff Dilemma :an Application to Crisis Management [C]. IEEE TRANSACTIONSONENGINEER ON GMANAGEMENT, 1995,42 (4)
- [2] Noel Pauwels, Bartel Van De Walle, Frank Hardeman, Karel Soudan, The implications of irreversibility in emergency response decisions ,Theory and Decision;Aug2000,49,1; ABI/INFORM Global Pg.25
- [3] Hiroyuki Tamura ,Kouji Yamamoto, Shinji Tomiyama, Itsuo Hatono, Modeling and Analysis of decision making problem for mitigating natural disaster risks[J], European Journal of Operational Research,2000,12:461-468
- [4] L. Jenkins. Selecting scenarios for environmental disaster planning[J]. European Journal Operational Research ,2000,121(2): 275-286
- [5] Donald L. Rosenstein, Decision-Making Capacity and Disaster Research, Journal of Traturatie Stress, Vol.17, No.5,October 2004, PP.373-381
- [6] 袁辉.应急决策群体的组织[J].安全,1997,(1):11-13
- [7] 姜卉,黄钧.罕见重大突发事件应急实时决策中的情景演变[J].华中科技大学学报,2009(1):104-108
- [8] 张荣梅,涂序彦.基于 CBR 的交通事故处理智能决策支持系统[J].计算机工程与应用.2002,(2):247-249
- [9] 张建华,刘仲英.案例推理和规则推理结合的紧急预案信息系统[J].同济大学学报.2002,30(7):890-894
- [10]柳炳祥,盛昭翰.基于案例推理的企业危机预警系统设计[J].中国软科学, 2003 (3):67-70
- [11]郭泳亨,卢兴华,刘云.应急决策的模糊综合评判研究[J].科学技术与工程,2006(1):588-592
- [12]周云海(2007)设计了一个应用于大停电事故恢复的基于案例推理系统[75];

-
-
- [13]周云海,胡翔勇,罗斌.基于案例推理的大停电恢复系统设计[J].电力系统自动化, 2007(18):87-90
- [14]陈铭.航空事故案例库设计及检索方法研究[D].南京航空航天大学,2009
- [15]贺清.基于案例推理的铁路枢纽站应急预案管理[J].铁路运营技术, 2012, 10(4): 18-24
- [16]刘挺,吴岩,等.串频统计和词形匹配相结合的汉语自动分词系统.中文信息学报, 1998, 12(1):17-25
- [17]张华平.汉语词法分析系统 ICTCLAS. <http://ictclas.nlpir.org/downloads>
- [18]Lin D. An Information — Theoretic Definition of Similarity[C]. In: Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. 296-304
- [19]Osgood C E. The nature and measurement of meaning[J]. Psychological Bulletin. 1952, 49(3):197-237
- [20]Tversky A. Features of Similarity[J]. Psychological Review. 1977, 84(4):327 — 352
- [21]Sugumaran, V. and V. C. Storey, Ontologies for conceptual modeling: their creation, use, and management. Data & Knowledge Engineering, 2002. 42(3): 251-271.
- [22]晋耀红. 基于语义的文本过滤系统的设计与实现. 计算机工程与应用 [J], 2003(17):22-25
- [23]金博,史彦军,滕弘飞. 基于语义理解的文本相似度算法. 大连理工大学学报[J], 2005 (02): 291-297
- [24]《中华人民共和国突发事件应对法》2007 版. 第三条
- [25]《国家突发公共事件总体应急预案》2005 版. 1.3 条
- [26]刘心报. 决策分析与决策支持系统[M], 北京: 清华大学出版社, 2009
- [27]郎淳刚, 刘树林. 国外自然决策理论研究述评[J]. 技术经济与管理研究, 2009, 4: 63-66
- [28]Klein G A. A Recognition-Primed Decision (RPD) Model of Rapid Decision Making[M]. US: New Jersey, Ablex Publishing, 1993
- [29]Thomas H. Killian. Decision Making and the Levels of War[J]. Military Review, 2000, 11-12: 66-70
-

-
-
- [30]Kaoru Hirota, Hajime Yoshino, Xu MingQiang, et al. An Application of Fuzzy Theory to the Case-based Reasoning of the CISG[J].Journal of Advanced Computational Intelligence,1997,1(2): 86-93
- [31]Wang Zhenyu, Hang Xiaoshu, Xiong Fanlun. Rough Fuzzy Case-Based Reasoning and Its Applications in Forecasting Insect Pests[C]. Second Asian conference for information technology in agriculture(AFITA 2000),Korea,2000:
- [32]Branting L K. Stratified Case-based Reasoning in Non-Refinable Abstraction Hierarchies[C]. Proceedings of the Second International Conference on Case-based Reasoning.Springer, 1997:519 -530.
- [33]]Chi R T,Whinston A B , Kiang M Y. Case-based Reasoning to Model Building[C].Proceedings of the 26th Hawaii International Conference on System Science, 1993:324-332.
- [34]刘群;李素建. 基于《知网》的词汇语义相似度计算[J].计算机语言学及中文信息处理,2002,7(2):59-67.

致 谢

参考文献

- [1] 刘双印.电子商务智能推荐系统中基于领域本体的案例检索算法[J].计算机应用, 2010(05):1304-1308
- [2] 王波,宋东,姜华男.基于粗集的 CBR 故障诊断案例的检索方法研究[J].计算机测量与控制, 2007(11):1430-1433
- [3] 尤军东.基于 CBR 面向企业诊断的流程案例检索与复用[J],2006,哈尔滨工业大学.
- [4] Salvatore Belardo , Harold L. Pazer . A Framework for Analyzing the Information Monitoring and Decision Support Investment Tradeoff Dilemma: an Application to Crisis Management [C]. IEEE TRANSACTIONSONENGINEER ON GMANAGEMENT, 1995,42(4)
- [5] Noel Pauwels, Bartel Van De Walle, Frank Hardeman, Karel Soudan, The implications of irreversibility in emergency response decisions ,Theory and Decision; Aug2000,49,1; ABI/INFORM Global Pg.25
- [6] Hiroyuki Tamura ,Kouji Yamamoto, Shinji Tomiyama, Itsuo Hatono, Modeling and Analysis of decision making problem for mitigating natural disaster risks[J],European Journal of Operational Research,2000,12:461-468
- [7] L. Jenkins. Selecting scenarios for environmental disaster planning[J]. European Journal Operational Research ,2000,121(2): 275-286
- [8] Donald L. Rosenstein, Decision-Making Capacity and Disaster Research, Journal of Traturatie Stress, Vol.17, No.5,October 2004, PP.373-381
- [9] 袁辉.应急决策群体的组织[J].安全,1997,(1):11-13
- [10]冯凯,徐志胜,冯春莹,王冬松.小城镇突发公共事件应急决策系统的研究[J].灾害学,2005,20(2):6-10
- [11]秦军昌,王刊良.应急管理中的决策支持系统研究[J].情报杂志,2008(12):37-39
- [12]姜卉,黄钧.罕见重大突发事件应急实时决策中的情景演变[J].华中科技大学学报, 2009(1):104-108
- [13]刘霞,严晓.突发事件应急决策生成机理:环节、要素及序列加工[J].上海行政学院学

-
-
- 报,2011,12(4):37-43
- [14]Suleyman Tufekci. An integrated emergency management decision support system for hurricane emergencies[J].Safety Science,1995,(20):39-48
- [15]John Cosgrave. Decision making in emergencies[J]. Disaster Prevention and Management. Bradford,1996,155(4):28
- [16]Alvin E. Roth. The economist as engineer: game theory, experimentation, and computation as tools for design economics[J]. Econometrical,2002, 70(4):1341-1378
- [17]周正平.社会安全事件的特征比较[J].北京人民警察学院学报,2008,(2):16-19
- [18]杨文国,贾屹峰,黄钧.非常规突发事件应急决策中的情景分析方法和决策准则研究[A].第四届国际应急管理论坛暨中国（双法）应急管理专业委员会第五届年会[C].2009,(12):421-424
- [19]韩传峰,王兴广等.非常规突发事件应急决策系统动态作用机理[J].软科学,2009(8):50-53
- [20]刘霞.非常规突发事件动态应急群决策：“情景-权变”范式[J].云南社会科学,2010(5):21-25
- [21]吴广谋,赵伟川,江艺平.城市重特大事故情景再现与态势推演决策模型研究[J].东南大学学报(哲学社会科学版),2011,13(1):18-23
- [22]舒其林.非常规突发事件的情景演变及“情景-应对”决策方案生成[J].中国科学技术大学学报,2012,11(11):936-941
- [23]L. Jenkins. Selecting scenarios for environmental disaster planning[J]. European Journal Operational Research ,2000,121(2): 275-286
- [24]]Hiroyuki Tamura ,Kouji Yamamoto, Shinji Tomiyama, Itsuo Hatono, Modeling and Analysis of decision making problem for mitigating natural disaster risks[J], European Journal of Operational Research,2000,12:461-468
- [25]王革.管理学中案例研究方法的科学化探讨[J].中国行政管理,2011(3):116-120
- [26]王继明;杨国林.基于 Lucene 的中文文本分词[J].内蒙古工业大学学报, 2007(3): 185-188
- [27]张贤坤.基于案例推理的应急决策方法研究[D].天津大学,2012
- [28]蒋鹏.基于本体的应急案例相似度算法研究[J].南昌高专学报,2009(3):159-161
-

-
-
- [29]王家钺.信息检索中的“相关性”概念的研究[J].现代外语,2001(2):181-191
- [30]史忠植.高级人工智能[M].北京:科学出版社,1998
- [31]李景.本体理论及在农业文献检索系统中的应用研究——以花卉学本体建模为例[D].中国科学院文献情报中心博士学位论文,2004
- [32]陶朱军.基于万维网环境下基于领域知识的信息资源管理模式研究[D].中国农业大学博士论文,2004
- [33]何爱元.基于词典和概率统计的中文分词算法研究[D].辽宁大学,2011
- [34]张守川.基于改进模糊案例推理算法的分类问题研究[D].华东理工大学,2013
- [35]刘恒文.基于网络语义挖掘的舆情监测预警研究[D].武汉理工大学,2010
- [36]张洪彬.基于语义网的信息内容检索[D].四川大学,2004
- [37]赵煜;蔡晓东;樊娜;李慧贤.利用词汇分布相似度的中文词汇语义倾向性计算[J].西安交通大学学报,2009(6):33-37
- [38]Aamodt, Plaza. Case-based Reasoning: Foundational Issue, Methodological Variation, and System Approaches[J]. AI Com-Artificial Intelligence Communications, IOS Press,1994,7(1):39-59
- [39]Gilboa. Case-based Decision Theory[J].Quarterly Journal of Economics,1995, 110:605-639
- [40]刘庆贤;肖洪钧.案例研究方法严谨性测度研究[J].管理评论,2010(5):112-119
- [41]杨志国.相似度计算在基于本体的自动问答系统中的应用[J].2010
- [42]Sugumaran, V. and V.C. Storey, Ontologies for conceptual modeling:their creation, use, and management. Data & Knowledge Engineering, 2002. 42(3): 251-271
- [43]赵俊杰,胡学钢.基于文本分类的文档相似度计算[J].微型电脑应用,2008(12):46-56.
- [44]Jiawei Han, Micheline Kamber,范明,孟小峰.数据挖掘概念与技术[M],北京:机械工业出版社,2011’ 254-256
- [45]金希茜.基于语义相似度的中文文本相似度算法研究[D],2009,浙江工业大学
- [46]Kolodner J L, Simpson R L. The Mediator: Analysis of an Early Case-based Problem Solver [J]. Cognitive Science, 1989, 13:507-549
- [47]Sycara K. Using Case-based Reasoning for Plan Adaptation and Repair[C].
-

-
-
- Proceedings of the DARPA Case-Based Reasoning Workshop. Morgan Kaufmann, San Francisco, 1988:425-434
- [48] Bareiss R. Exemplar-Based Knowledge Acquisition: A Unified Approach to Concept Representation [C]. Classification, and Learning, Boston: Academic Press 1989: 159-166
- [49] Hammond K J. Case-Based Planning: Viewing Planning as a Memory Task [M]. San Diego: Academic Press, 1989
- [50] Smsalamo M, Golobardes E. Rough Sets Reduction Techniques for Case-based reasoning [C]. Anon International Conference on Case-based Reasoning. Springer, 2005:467-482
- [51] 陆建江, 张亚非, 苗壮, 等. 语义网原理与技术[M]. 北京: 科学出版社, 2007
- [52] 王玉, 邢渊, 阮雪榆. 基于案例的推理循环中人工神经网络和遗传算法的四种模型[J]. 上海交通大学学报, 2000, 37(3): 202-204
- [53] 季赛, 沈星, 沈超. 基于粗糙集和相似度量的 CBR 检索方法[J]. 计算机工程与应用, 2006, 13: 172-174
- [54] Dubois D, Esteva F, Garcia P, et al. Fuzzy Modeling of Case-based Reasoning and Decision [C]. Proceedings of the Second International Conference case-based reasoning, LNAI 1266, Springer-Verlag, 1997: 599-610
- [55] [54] Watson I, Marir F. Case-based Reasoning: A Review[J]. The Knowledge Engineering Review, 1994, 9(4): 327-354
- [56] Vong C M, Leung T P, Wong P K. Case-based Reasoning and Adaptation in Hydraulic Production Machine Design[J]. Engineering Applications of Artificial Intelligence, 2002, 15: 567-585
- [57] 王辉, 周雄辉, 阮雪榆. 基于遗传算法的事例改写策略在注塑成本评估 CBR 系统中的应用研究[J]. 中国机械工程, 2002, 13(22): 1957-1960
- [58] 李明, Multi-Agent 的范例推理[J]. 重庆师范学院学报(自然科学版), 2001, 18(3): 57-59
- [59] 耿同焕, 肖明军, 邹翔等. 聚类算法在范例维护中的应用研究[J]. 计算机工程, 2005, 31(12): 166-168
-

-
-
- [60]Baumeister J, Atzmuller M, Puppe F. Inductive Learning for Case-based Diagnosis with Multiple Faults[C]. Advances in Case-Based Reasoning: 6th European Conference (ECCBR2002), Aberdeen, Scotland: Springer, 2002:28-43
- [61]Cardie C. Using Decision Trees to Improve Case-based Reasoning Learning [C].Proceedings of the Tenth International Conference on Machine Learning, 1993:5-16
- [62]Park Y J,Choi E,Park S H. Two-step Filtering DataMining Method Integrating Case-based Reasoning and Rule Induction[J].Expert Systems with Applications, 2009,36(1):861-871

附 录

表 25 历史案例分词情况汇总表

编号	案例标题	灾害类型	案例关键词	分词排名前五词汇	关键词 匹配度
1	湖南涟源 75 名学生疑似食物中毒事件	公共卫生	湖南、学生、中毒、营养餐、事件	营养餐、乳业、事件、学生、乙方	60.00%
2	湖南石门县新铺乡中小学生上呼吸道感染暴发疫情	公共卫生	湖南、中小学生、呼吸道感染、疫情	事件、病毒、中小学生、肠道、感染	40.00%
3	四川岳池顾县小学发生牛奶中毒事件	公共卫生	四川岳池、小学生、牛奶、中毒事件	学生、部门、事件、不良、小学	20.00%
4	辽宁现炭疽病疫情	公共卫生	辽宁、炭疽、疫情、传染病	炭疽、疫情、芽孢、感染、皮肤	40.00%
5	甘肃省白银市景泰县发生禽流感疫情	公共卫生	甘肃、禽流感、疫情、疫区防控	疫区、疫情、禽流感、防控、事件	80.00%
6	山东临沂大学学生中毒事件	公共卫生	临沂大学、学生中毒、腹泻	学生、大学、出现、腹泻、事件	60.00%
7	西藏山南地区加查县发生 O 型口蹄疫疫情	公共卫生	山南地区、O 型口蹄疫疫情、疫区防控	口蹄疫、疫情、事件、疫区、防控	80.00%
8	甘肃陇西学生群体性瘰病事件调查	公共卫生	甘肃陇西、学生、群体性、瘰病	症状、医院、群体性、瘰病、患者	60.00%
9	上海染色馒头事件调查	公共卫生	上海、染色馒头、调查组、食品安全	馒头、食品、超市、调查组、问题	60.00%
10	广州“瘦肉精”中毒事件调查	公共卫生	广州、瘦肉精、生猪、中毒	生猪、事件、猪肉、发病、业户	20.00%
11	青海全力救治“双黄连注射液”不良反应患者	公共卫生	青海、双黄连、注射液、救治	事件、双黄连、患者、注射液、救治	60.00%
12	张海超开胸验肺案	公共卫生	开胸验肺、职业病、尘肺病	职业病、尘肺、诊断、进行、粉尘	40.00%
13	北京发生冲撞天安门暴力恐怖袭击事件	社会安全	北京、天安门、暴力恐怖袭击事件	事件、伤员、救治、恐怖、袭击	40.00%

编号	案例标题	灾害类型	案例关键词	分词排名前五词汇	关键词匹配度
14	河北廊坊数千学生不满校园管理打砸食堂抗议	社会安全	河北廊坊、学生、打砸食堂	学生、事件、学校、食堂、大学城	40.00%
15	云南晋宁发生群体性事件	社会安全	云南、征地拆迁、群体性事件	征地、村民、事件、拆迁、原因	60.00%
16	河北承德技师学院群体斗殴事件	社会安全	河北承德、群体性斗殴事件	女生、某某、斗殴、发生、技师	20.00%
17	新疆鄯善发生暴力恐怖袭击事件	社会安全	新疆鄯善、暴力恐怖袭击事件	团伙、袭击、巡警、派出所、群众	20.00%
18	厦门快速公交高架路上着火	社会安全	厦门、公交车、高架路、起火	公交车、起火、客流量、事件、伤员	40.00%
19	新疆巴楚暴力恐怖事件	社会安全	新疆、暴力恐怖事件、团伙	暴徒、恐怖、暴力、团伙、工作人员	60.00%
20	广东汕头一内衣厂发生一起火灾	社会安全	汕头、内衣厂、火灾	火灾、事故、全力、查明、书记	20.00%
21	四川泸州群体事件	社会安全	四川、泸州、群体性事件	群体性、处置、警车、货车、司机	20.00%
22	辽宁盘锦警察击毙村民事件	社会安全	辽宁盘锦、警察、袭击村民	民警、开枪、事件、村民、没有	20.00%
23	钓鱼岛事件引发的群体性事件	社会安全	钓鱼岛、抗议游行、岛屿争端	海域、进行、渔船、抗议、岛屿	40.00%
24	江苏启东政府门前抗议活动	社会安全	江苏启东、抗议集会	市政府、抗议、民众、污水、市民	20.00%
25	贵州仁怀茅台园区发生群体性事件	社会安全	贵州怀仁、群体性事件	群众、安置、处置、社会、聚集	20.00%
26	陕西宝鸡车祸引群体事件	社会安全	陕西宝鸡、群体性事件、车祸	出租车、司机、罚金、殴打、民警	0.00%
27	四川什邡群体性事件	社会安全	四川什邡、灾后重建、群体性事件	项目、氧化酶、环保部、嘌呤、市民	0.00%
28	新疆和田劫机事件	社会安全	新疆和田、劫机事件、恐怖事件	和田、势力、事件、机场、机组	60.00%
29	广东中山沙溪镇群体性聚集事件	社会安全	广东中山、群体性事件	少年、接报、处置、社会、群众	0.00%
30	河南郑州强拆案	社会安全	河南郑州、征地拆迁、城管	火灾、拆迁、编组站、建筑、城管	40.00%

编号	案例标题	灾害类型	案例关键词	分词排名前五词汇	关键词 匹配度
31	济南狱警殴打老人事件	社会安全	济南、狱警、殴打老太太	狱警、丈夫、老太太、事件、随后	40.00%
32	江西宜春市奉新县精品店火灾	事故灾难	江西宜春、火灾、消防	事件、赶赴、事故、精品店、凌晨	0.00%
33	山东省乳山合和食品有限公司发生液氨泄漏	事故灾难	山东、液氨泄露、抢救	液氨、事故、事件、泄漏、抢救	40.00%
34	山西交口煤矿山体滑坡事故	事故灾难	山西交口、煤矿、山体滑坡	事件、煤业、山体、事故、受伤	40.00%
35	贵州省毕节市赫章公路塌方事故	事故灾难	贵州毕节、公路塌方、事故	事件、路段、路面、路基、省道	20.00%
36	广东省肇庆市广宁县发生拖拉机翻车事故	事故灾难	广东肇庆、拖拉机、事故	事故、事件、伤势、造成、伤员	0.00%
37	东北暴雪黑龙江牡丹江市厂房楼顶被压出“大洞”	事故灾难	牡丹江、暴雪、厂房、救援	降雪量、事件、暴雪、厂房、事故	40.00%
38	黑龙江黑河市发生非法采矿坍塌事件	事故灾难	黑龙江、非法、采矿、坍塌	事件、事故、嫌疑犯、刑拘、非法	40.00%
39	甘肃酒泉两客车侧翻相撞	事故灾难	甘肃酒泉、客车、事故	事故、人员、处置、分赴、企业	20.00%
40	哈尔滨至北京动车 D28 发生事故	事故灾难	动车、D28、事故	天窗、列车、铁路、线路、铁路局	0.00%
41	山东青岛黄岛输油管道泄露爆炸	事故灾难	青岛、中石油、泄漏、爆炸、管道	事故、管道、泄漏、暗渠、原油	40.00%
42	合六叶高速车祸	事故灾难	合六叶高速、车祸、泄漏	封堵、发生、车辆、泄漏、大雾	20.00%
43	湖北襄阳市南漳建设工地脚手架坍塌	事故灾难	襄阳、脚手架、坍塌、救助	事故、建设、脚手架、工地、事件	20.00%
44	新疆阿克苏地区新和县重大交通事故	事故灾难	新疆、阿克苏、面包车、交通事故	死亡、事件、伤者、面包车、车辆	20.00%
45	北京朝阳区汽配城大火	事故灾难	北京、朝阳、汽配城、火灾	火灾、仓库、人员、扑灭、汽配	40.00%
46	山西吕梁市石楼县交通事故	事故灾难	吕梁、三轮车、交通事故	三轮车、事故、事件、死亡、人员	40.00%
47	大连市甘井子区交通事故	事故灾难	大连、面包车、翻斗车、交通事故	翻斗车、事件、面包车、直行、路口	40.00%

编号	案例标题	灾害类型	案例关键词	分词排名前五词汇	关键词 匹配度
48	新疆八钢钢结构有限责任公司一车间发生液氨爆炸	事故灾难	新疆、液氨爆炸、钢结构	液氨、事故、伤者、钢结构、送往	40.00%
49	河南新乡县浴池坍塌致事故	事故灾难	河南、浴池、坍塌、救治	二层、事件、救治、搜救、浴池	40.00%
50	广东省广州白云区一住宅楼起火	事故灾难	广州、住宅楼、火灾、救治	化名、福建人、处置、某某、事件	0.00%
51	贵州毕节市金沙县黄水坝煤矿事故	事故灾难	毕节、黄水坝、煤矿事故	煤矿、事故、水坝、事件、事发	20.00%
52	广西岑溪炮竹厂爆炸事故	事故灾难	广西岑溪、炮竹厂、爆炸、事故	炮竹、事故、事件、受伤、造成	40.00%
53	四川广安一水泥制品厂花岗岩板材倒塌	事故灾难	四川广安、水泥厂、花岗岩、倒塌	事故、事件、花岗岩、领导、板材	20.00%
54	枝江市董市镇平湖村都云建材公司发生蒸汽釜爆炸	事故灾难	枝江市、蒸汽釜、尾矿、爆炸	事故、事件、主要、尾矿、市政府	20.00%
55	云南楚雄至大理高速公路一旅游车翻车事故	事故灾难	云南楚雄、旅游车、交通事故	驶往、边坡、驶出、客车、受伤	0.00%
56	陕西西安城西西晁村三层民房垮塌	事故灾难	陕西西安、民房垮塌、救治	搜救、人员、加盖、楼板、民房	20.00%
57	黑龙江鸡西市住宅楼爆炸	事故灾难	黑龙江鸡西市、住宅楼、煤气、爆炸	救治、伤员、事件、专家队、气罐	20.00%
58	贵州省惠水县发生煤矿透水事故	事故灾难	贵州、煤矿、透水、事故、抢险救援	事故、抢险救援、支队、煤矿、人员	40.00%
59	云南昆明市宜良县盘江打捞水葫芦船只翻沉事件	事故灾难	昆明、打捞、水葫芦、翻沉事故	水葫芦、搜救、河道、打捞、泄洪	40.00%
60	上海宝山区共和新路呼玛路口3人窨井内身亡	事故灾难	上海、宝山区、井盖、偷盗	电力、偷盗、事件、电缆、公司	20.00%
61	陕西澄城一硫磺矿发生燃烧事故	事故灾难	陕西、硫磺矿、燃烧事故、救治	硫磺、施救、事件、事故、发生	40.00%
62	北京市朝阳区大望桥路面塌陷案	事故灾难	北京市朝阳区、路面塌陷、抢险	塌陷、抢险、地面、管线、交通	60.00%
63	上海市“6.27 倒楼”案例	事故灾难	上海市、倒楼事件	危机、购房人、事件、合同、事故	0.00%
64	湖南省湘西自治州凤凰桥垮塌事故案	事故灾难	湖南湘西、凤凰桥、坍塌事故	事故、人员、伤员、施工方、桥墩	20.00%

编号	案例标题	灾害类型	案例关键词	分词排名前五词汇	关键词 匹配度
65	北京市西城区西工工程坍塌案	事故灾难	北京市西城区、工程坍塌事故	事故、公司、坍塌、处置、建筑	40.00%
66	北京市石景山区“12.25”供热事故案	事故灾难	石景山、供热事故、抢修处理	供热、事故、单位、抢修、处置	60.00%
67	河南杞县钴 60 案	事故灾难	河南杞县、核辐射危机	信息、政府、辐照、事件、危机	40.00%
68	重庆市綦江县虹桥特大垮塌事故	事故灾难	重庆綦江、虹桥垮塌事故	工程、事故、质量、起事、责任	20.00%
69	重庆市开县特大油气井喷事故	事故灾难	重庆开县、油气井喷事故	泥浆、群众、事故、井队、井喷	40.00%
70	北京市东三环京广桥路面坍塌案	事故灾难	北京市、路面坍塌事故	抢险、管线、事故、污水、恢复	20.00%
71	武汉市水果湖小学集体食物中毒事件	事故灾难	武汉市、水果湖、小学生、食物中毒	学生、事件、卫生厅、家长、发生	20.00%
72	临海桐岩岭隧道车辆火灾	事故灾难	宁波市、油罐车、火灾，隧道	隧道、油罐车、火灾、货车、堵截	60.00%
73	奥地利卡普伦滑雪胜地列车隧道火灾	事故灾难	奥地利、卡普轮、列车、隧道、火灾	列车、隧道、火灾、车厢、海拔	60.00%
74	新疆七五事件	事故灾难	新疆、恐怖暴力事件、针刺	事件、群众、暴徒、暴力、针刺	40.00%
75	内蒙古包头市应急处理盐酸泄漏事故	事故灾难	内蒙古包头、盐酸泄漏事故	盐酸、泄漏、事故、处置、货车	60.00%
76	中央电视台火灾事故报告	事故灾难	中央电视台、火灾事故、烟花爆竹	火灾、燃放、烟花、央视、事故	80.00%
77	余姚遭菲特重击水库难以承压泄洪	事故灾难	余姚、菲特、水库、泄洪	水库、停运、泄洪、水位、乡镇	40.00%
78	浙江宁波余姚堤防决口	事故灾难	余姚、提防、决口	堤防、险情、决口、抢险、排涝	40.00%
79	台风“菲特”登陆与“丹娜丝”形成双台风效应	事故灾难	菲特、丹娜丝、双台风	台风、紫菜、强台风、船只、船员	20.00%
80	宁波母亲河水位创新高 市民广场被淹不见底	事故灾难	宁波、余姚、姚江、水位	水位、抢险、大闸、水闸、江水	20.00%
81	浙江余姚市 7 成被淹 四个安置点电话均接不通	事故灾难	余姚、台风、内涝	积水、云系、内涝、台风、海葵	40.00%

编号	案例标题	灾害类型	案例关键词	分词排名前五词汇	关键词 匹配度
82	余姚 300 个安置点安顿民众 救援物资储备充足	事故灾难	余姚、民众、安置、救援	安置点、村民、矿泉水、物资、镇政府	20.00%
83	宁波电视台关于余姚卫星车被围堵事件的说明	事故灾难	宁波、记者、群众围堵	警察、记者、卫星、群众、特别报道	40.00%
84	山东威海海域沉没天津籍货船已致 7 人遇难	自然灾害	威海、货船、救助	救助、搜救、海上搜救、事件、航次	20.00%
85	台风“海燕”袭击广西、海南等地	自然灾害	台风、海燕	余间、倒塌、安置、房屋、台风	20.00%
86	海南省三亚市一广西籍货轮被台风吹入海	自然灾害	三亚、货轮、台风、救助	搜救、船员、遇难者、强台风、货船	60.00%
87	山东烟台渔船遇大风浪沉没	自然灾害	烟台、风浪、船舶、沉没	渔民、事件、遇险、渔船、船舶	20.00%
88	陕西省安康市旬阳县白柳镇佛洞村森林火灾	自然灾害	安康、森林火灾	火灾、事件、村民、森林、遇难	40.00%
89	强台风“菲特”在福建福鼎登陆	自然灾害	菲特、福建、台风	台风、截至、升格、警报、倒塌	20.00%
90	青海西宁市大通县城长宁镇新寨五村山体滑坡	自然灾害	西宁、山体滑坡	取土、滑塌、土体、事故、山体	40.00%
91	河北邯郸武安一山村房屋意外塌陷	自然灾害	武安、岩溶、塌陷	塌陷、岩溶、人员、顶板、发生	60.00%
92	“9.29”海南西沙沉船事故	自然灾害	西沙、渔船、沉船事故、搜救	搜救、渔民、海域、渔船、遗体	40.00%
93	陕西榆林绥德县连续两天发生坍塌事故	自然灾害	榆林、坍塌事故、救援	事故、抢救、坍塌、事件、组织	60.00%
94	湖南常德发生特大洪涝灾害	自然灾害	常德、暴雨、特大洪水、救援	洪水、暴雨、事件、围困、多名	60.00%
95	台风“天兔”致南方五省 34 人死亡 1 人失踪	自然灾害	台风、天兔、III级应急响应	台风、绝收、余间、农作物、响应	40.00%
96	云南省昭通市永善县发生山洪灾害	自然灾害	云南昭通、山洪、救援	细沙、灾害、山洪、事件、抢险	40.00%
97	宜宾暴雨引发地质灾害	自然灾害	宜宾、暴雨、地质灾害	暴雨、垮塌、灾情、出现、事件	20.00%
98	甘肃定西市岷县、漳县暴洪泥石流灾害	自然灾害	甘肃定西、泥石流、抢险	灾害、抢修、群众、乡镇、抢险	20.00%

编号	案例标题	灾害类型	案例关键词	分词排名前五词汇	关键词 匹配度
99	9·13 上海特大暴雨	自然灾害	上海、特大暴雨	暴雨、积水、雨量、降雨量、特大	40.00%
100	云南大理云龙县发生泥石流灾害	自然灾害	大理、泥石流	泥石流、灾害、碎屑、固体、破坏	20.00%
101	云南省迪庆州香格里拉 5.9 级地震	自然灾害	云南、5.9 级、地震、香格里拉	地震、灾区、5.9 级地震、发生、棉被	40.00%
102	青海省乌兰县突发山洪 21 人遇难	自然灾害	乌兰县、山洪、21 人遇难	事发、人员、搜救、山洪、暴雨	20.00%
103	台风尤特致广西受灾人数上升至 143 余万	自然灾害	广西、台风、尤特	农房、事件、云系、绝收、强台风	20.00%
104	东北地区发生严重的洪涝灾害	自然灾害	东北、洪涝灾害、救援	灾区、降水、高压、热带、暖湿气流	0.00%
105	西藏昌都地区左贡县、芒康县交界发生 6.1 级地震	自然灾害	西藏昌都、6.1 级地震、救援	地震、震级、构造、震源、边界	20.00%
106	山西省吕梁市石楼县山体滑坡事故	自然灾害	吕梁、山体滑坡、抢险	事件、山体、抢险救援、灾害、地质	40.00%
107	云南昭通市永善县发生山体滑坡事故	自然灾害	昭通、山体滑坡、抢险	细沙、灾害、山洪、事件、抢险	40.00%
108	甘肃天水暴洪灾害	自然灾害	甘肃天水、暴洪、救灾	安置、暴洪、群众、灾害、每户	20.00%
109	甘肃定西市岷县、漳县交界发生 6.6 级地震	自然灾害	甘肃定西、地震、抢险救灾	地震、6.6 级地震、灾区、发生、地震局	20.00%
110	辽宁朝阳三年来最严重伏旱	自然灾害	辽宁朝阳、三年、伏旱	万亩、面积、重旱、事件、受旱	0.00%
111	四川都江堰市发生特大型高位山体滑坡事故	自然灾害	都江堰、山体滑坡、救灾	灾害、滑坡、山体、高位、失踪	40.00%
112	四川西部特大暴雨灾害事件	自然灾害	四川西部、特大暴雨	暴雨、灾区、洪涝灾害、降雨量、棉被	20.00%
113	我国南方遭遇 2013 年以来最强高温	自然灾害	南方、热射病、高温	高温、日数、热射病、天气、气温	40.00%
114	黄山市发生百年不遇洪灾	自然灾害	黄山市、洪灾、暴雨、救灾	降雨量、水库、洪灾、暴雨、溢流	40.00%
115	我国南方发生大范围干旱灾害	自然灾害	南方、大范围、干旱、饮水困难	高温、资金、受旱、地区、饮水	20.00%

编号	案例标题	灾害类型	案例关键词	分词排名前五词汇	关键词匹配度
116	湖南省湘中强降雨	自然灾害	湖南中部、强降雨、暴雨	雨量、暴雨、小时、大暴雨、拦洪	20.00%
117	新疆维吾尔自治区阿克苏地区强降雨	自然灾害	新疆、阿克苏、强降雨、洪水	造成、死亡、洪水、牲畜、农作物	20.00%
118	福建广东两省遭遇洪涝灾害	自然灾害	福建、广东、洪涝灾害	暴雨、灾区、善款、降水、洪涝灾害	20.00%
119	四川宜宾遭受冰雹大风袭击	自然灾害	四川宜宾、冰雹、大风、农作物绝收	冰雹、灾害、大风、受损、绝收	60.00%
120	南方暴雨洪涝风雹灾害	自然灾害	南方、暴雨、洪涝、冰雹、灾害	风雹、灾害、洪涝、暴雨、冰雹	80.00%
121	广东发生特大暴雨灾害	自然灾害	广东、特大暴雨、救灾	暴雨、暖湿气流、截至、灾区、汛期	20.00%
122	四川理县发生雪崩	自然灾害	四川理县、雪崩、搜救	人员、搜救、雪崩、官兵、民兵	40.00%
123	贵州天柱县山体滑坡	自然灾害	贵州天柱山、山体滑坡	山体、板房、滑坡、工地、事故	40.00%