中图分类号：TP391　　论文编号：
学科分类号：520　　密　　级：公　开

安徽理工大学

硕 士 学 位 论 文

事件及其事件要素的抽取研究

作者姓名：轩小星
专业名称：计算机科学与技术
研究方向：人工智能、数据挖掘
导师姓名：廖 涛 副教授
导师单位：安徽理工大学
答辩委员会主席：任良勇

论文答辩日期：2015 年 5 月 30 日

安徽理工大学研究生处

2015 年 6 月 3 日

A Dissertation in Computer Science and Technology

Reaserch on events and event elements extraction

Candidate: Xuan Xiaoxing

Supervisor: Liao Tao

Computer Science and Engineering School

AnHui University of Science and Technology

No.168, Shungeng Road, Huainan ,232001,P.R.CHINA

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得安徽理工大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：郭小星 日期：2015年6月3日

学位论文版权使用授权书

本学位论文作者完全了解 安徽理工大学 有保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属于安徽理工大学。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 安徽理工大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。（保密的学位论文在解密后适用本授权书）

学位论文作者签名：郭小星 签字日期：2015年6月3日

导师签名：[illegible] 签字日期：2015年6月3日

摘　要

近年来，对于事件的研究一直以来深受学术界的高度重视，其中，从海量文本中进行事件抽取，从而获取人们感兴趣的信息和数据是大数据时代亟待解决的关键技术。事件抽取是信息抽取的重要组成部分，事件抽取就是从非结构化文本中抽取出用户感兴趣的事件，然后用结构化或半结构化的形式描述出来，供用户浏览、查询或者进一步分析利用。事件识别和事件要素的抽取是事件抽取的两大主要任务，本文主要针对这两个内容展开深入研究。

（1）事件识别其实是事件触发词的识别过程，针对当下语料库资源缺乏造成的数据稀疏问题，本文提出了基于扩展触发词表和多特征融合下的机器学习相结合的事件触发词识别方法。基于扩展触发词表的识别方法有较高的召回率但准确率却很低。基于机器学习的识别方法准确率有明显提高但召回率却低于前一种方法。鉴于此，本文把两种识别方法结合起来，根据计算得出的候选触发词的权重分布情况设定一个阈值，候选触发词的权重大于阈值时即认定为事件触发词，当小于阈值时，就用机器学习的方法进行识别判断。实验结果表明通过把两种方法进行结合使用，召回率和准确率得到兼顾，F值也比较理想。

（2）事件要素的抽取方面，基于监督学习的方法对语料库资源的依赖比较强，不少研究工作都受到了数据稀疏问题的困扰。本文提出了聚类（无监督）学习的事件要素抽取方法，该方法能有效的减少对语料库的依赖。聚类算法选用基于距离的典型的*k*-means算法，但是传统的*k*-means算法忽略了各个特征对聚类分析影响的差异。在实际应用中，各个特征对聚类的贡献是不等的，所以在聚类分析过程中，特征的权重必须考虑在内。利用特征选择算法即 ReliefF 算法对特征进行加权选择，然后对传统的*k*-means算法进行移植改进，使改进后的算法能够适用于事件要素的抽取工作。实验表明，改进后的聚类算法比传统算法的识别效果好。

图 6 表 10 参 54

关键词：事件；事件识别；事件触发词；事件要素抽取

分类号：（1--2）；

Abstract

In recent years, the research on event has deeply attached the academia' great importance. The event extraction from massive amounts of text, so as to get the information or data in which the people are interested, is the key technology to be solved of big data era. Event extraction is an important part of information extraction, the events which users are interested in are extracted from the unstructured text, and then described in the form of a structured or semi-structured, so that users can do browsing, querying or further analyzing conveniently. Events recognition and event elements extraction are the two major tasks of event extraction, this paper in-depth research on these two contents.

(1) Event recognition is the process of event denoter recognition, to the problems caused by the lack of data for the current sparse corpus resources, In this paper, we propose a event denoter recognition method which based on the combination of extended denoter vocabularies method and multi-feature fusion machine learning method. The recognition method based on extended denoter vocabularies has higher recall rate but low accuracy. However, when using the machine learning method alone, the precision rate is improved obviously but the recall rate is lower than the former method. In view of this, this paper combine the two kinds of identification methods, according to the calculated weight distribution of candidate denoter set a threshold, if the weight of the candidate denoter is greater than the threshold ,it can be determined as a event denoter, when less than the threshold, using machine learning method to recognize it. The experimental results show that we can obtain the rate of recall and precision balance and the F value is more ideal by combining the two recognition methods.

(2) In the aspect of event elements extraction, the method based on supervised learning stronger dependence on corpus resources, a lot of research works have been beset with sparse data problem. This paper proposes a clustering (unsupervised learning) event elements extraction method so that we can effectively reduce the dependence on corpus resource. we choose the typical clustering algorithm based on distance named k-means algorithm. However, traditional k-means algorithm ignores the different impact of individual characteristics to the cluster analysis. In practice, each feature's contribution to the clustering is unequal, so in the process of clustering analysis, feature weights must be taken into account. Feature selection algorithm named ReliefF algorithm is used to analyse the characteristics of

the weighted selection, then porting it into the traditional k-means algorithm so as to the improved algorithm can be applied to the work of elements extraction. The experimental results show that the improved clustering algorithm is better than traditional algorithm recognition results.

Figure 6 table 10 reference 54

Key Words: event, event recognition, event denoter, event elements extraction

Chinese books catalog:

目 录

Contents

插图或附表清单

1 绪 论

1.1 课题来源

本文来源于国家自然科学基金面上项目“事件本体形式化方法中的几个重要问题”（项目编号：61273328）。

1.2 研究背景和意义

近年来，“事件”的概念逐渐被各个知识处理领域所采用，包括计算机语言学、信息检索、人工智能、信息抽取、自动文摘以及自然语言处理领域等。人们寄希望于通过认识事件的相关信息来认识和了解整个世界。事件是包含了时间、地点、参与者等概念的语义单元，也是人类知识的基本单元。对于事件的研究一直以来深受学术界的高度重视，其中，从海量文本中进行事件抽取，从而获取人们感兴趣的信息和数据是大数据时代亟待解决的关键技术。然而，国内外对于该领域的研究仍然处于起步阶段。在信息爆炸式增长的当今时代，如何从海量数据文本中抽取出人们感兴趣的信息成为科研工作者需要克服的难题。正是在这种需求的驱动下，对于事件的研究在自然语言处理领域悄然兴起，而信息抽取技术成为该领域的一大研究热点，其中事件抽取是其三大主要任务之一。

由 DARPA（美国国防高级研究计划委员会）主办的 TDT[1]（话题识别与跟踪）评测会议的目的即是发展基于事件的信息组织技术。NIST（美国国家标准技术研究所）组织的 ACE[2]（自动内容抽取）评测会议的评测任务之一即是事件的识别与抽取，该任务的目的是利用自动抽取技术从非结构化的文本数据中抽取用户关注的事件信息，并转化为结构化形式呈现给用户，从而方便阅读或浏览。

事件抽取不仅给人们日常生活带来了便利，而且还是关系民生国计的大事。2009 年，国家重点基础研究发展计划的重要支持方向之一即是对突发性灾难事件的研究，希望通过钻研突发性灾难事件的相关技术，从而对该类事件的防御和应对能力均有所提高。例如对于交通事故，事故处理人员比较侧重于获知什么时间，哪个路段容易发生交通事故，通过事件抽取技术可以从众多事故类文本数据中识别和抽取出重要信息数据，为后期制定应对策略提供可靠的数据依据。可见，对事件抽取技术的研究不仅理论意义和实际意义重大，而且应用前景广泛。

1.3 国内外研究现状

国外，Ahn[3]于 2006 年提出了事件抽取的两个主要任务：事件触发词及其类别识别

和事件要素的识别。他整合了 Timbl 和 MegaM 这两种机器学习方法用于事件抽取系统中，该系统在 ACE2005 英文语料上进行了测试，识别结果的 F 值分别达到了 60.1%和 57.3%。H. L.Chieu 和 H. T. Ng[4] 将事件要素的识别看成是分类问题，把最大熵分类器引入到事件要素抽取的研究中。这套系统在 MUC 2002 评测的事件抽取任务中取得了不错的效果。很多研究都把事件抽取看成分类问题，而 Z. Chen[5]打破了这种思维模式，将事件及事件要素的识别看作序列标注问题。他选择中文独有的特征，采用最大熵隐马尔科夫模型，在 ACE 2005 中文语料上进行了测试，其 F 值达到了 70.3%。Ji 等[6]将范围扩大到一个话题集簇中的所有文档，而不局限于在单一文档中抽取事件；主要提出了一种基于规则的方法对触发词、事件参与者和角色进行判断。由于考虑了全局信息，即话题集簇中的所有相关文档，该方法取得了很好的效果。

国内，上海交通大学与德国研究机构联合开发了一种基于信息抽取的检索系统[7]，主要应用于投资和股票信息领域。清华大学周剑辉[8]通过机器学习方法建立抽取规则集，对金融领域事件抽取进行了深入的研究，实验结果的 F 值达到了 80%的理想状态。与之类似，近年来哈尔滨工业大学秦兵教授带领的团队[9]在中文事件抽取研究中也做出了很多成果。特别在音乐领域，他们开发了一套集成音乐事件抽取、音乐关系抽取等功能的信息抽取平台。清华大学吴平博等[10]对突发灾难性事件抽取方法做了专门研究。依据句法模板，他们构造出了抽取规则，最终建立了“突发灾难性事件抽取系统”。平均 F 值可以达到 70%左右。杨尔弘[11]提出通过语句聚类的方法获得事件的信息结构（事件模板），提出了以信息结构表达突发事件的有关信息、以特征项获取模式并进一步提取特定信息的思想。付剑锋[12]通过文本中句子的依存关系进行事件的抽取，用依存分析发掘触发词与其它词之间的句法关系，以此为特征在 SVM 分类器上对事件进行分类，最终实现事件识别。并且在文章[13]中，付剑锋又提出了利用特征加权的方法对事件发生的时间、地点、参与者等要素的识别。

可见，事件抽取的相关研究工作越来越受到国内外学者的关注。从以上研究来看，大体情况为：（1）目前比较成熟的抽取技术主要是针对英文文本中的事件实现的相关信息抽取，而针对中文事件的抽取工作才刚刚起步，目前的中文事件抽取系统还不够健壮，性能有待提高；（2）目前中文事件抽取的研究只面向特定的领域，例如：金融领域、音乐领域、灾难性事件等，抽取方法的可移植性较差；（3）基于机器学习是主流的事件抽取方法，该方法需要大规模的语料库作为训练集去训练分类器，但如今语料库资源的缺乏，使相关研究受到限制。

1.4 研究内容

本文主要面向中文事件抽取进行研究，鉴于机器学习方法存在的不足，事件的识别把扩展触发词表和机器学习两种方法结合起来，而事件要素的识别用无监督（聚类）的方法，从而减少了对语料库的依赖。具体研究内容如下：

（1）基于触发词扩展表和机器学习相结合的事件识别研究

本文选用对中文事件和事件要素标注较全面的 CEC 语料作为训练语料和测试语料，提出了两种事件识别的方法：触发词扩展表和多特征融合下的机器学习方法。基于触发词扩展表的事件触发词识别有较高的召回率但准确率却很低。而基于机器学习的事件触发词识别的准确率有明显提高但召回率却低于前一种方法。如此，本文把两种方法结合起来，扬长避短，可兼顾召回率和准确率。

（2） 基于改进型聚类算法的事件要素抽取研究

本文采用聚类（无监督）学习的方法对生语料中的事件要素实现直接抽取，减少了对语料库的依赖。利用特征选择算法即 ReliefF 算法对特征进行加权选择，从而对传统的 k-means 算法进行改进，使改进后的算法能够适用于事件要素的抽取。

两项研究内容相互关联。 首先，通过对文本事件的研究，给出事件及事件要素的形式化定义，利用触发词扩展表和多特征融合的机器学习相结合的方法识别出文本中事件及其事件类，划分出事件区域，每个区域内对应一件或一类事件的发生，进而采用基于改进型聚类算法的方法抽取出相应事件的事件要素。

1.5 论文结构

本文共分五章：

第一章是绪论部分，主要介绍了本文的课题来源、研究背景和意义以及国内外在事件抽取领域的研究现状，并论述了本文的主要研究内容和结构组织。

第二章主要介绍了事件抽取的相关知识，给出了事件及其事件要素的形式化定义，对比并评价了三大语料库资源，最后引出本文的研究路线和方法。

第三章研究了事件的识别和抽取，主要包括分类器模型和文本的表示模型的比较和选取，分别研究了基于扩展触发词表和基于多特征融合下的机器学习以及两种方法相结合的事件识别和抽取方法。

第四章研究了基于改进型聚类算法的事件要素的抽取方法。针对传统聚类算法存在的缺陷，用特征选择算法对其移植改进使其适用于本文的事件要素抽取工作。

第五章总结了本文的研究工作，并对今后要做的研究进行展望。

2 事件抽取相关知识概述

2.1 本章概述

本章对事件的相关知识及定义、语料库资源、事件及其事件要素的抽取方法进行概述。2.2 节综述了事件的相关知识，包括各领域对事件的不同定义，然后给出了事件及事件要素的形式化定义，以及事件触发词和事件类的相关定义。2.3 节对现有的三大语料库资源进行对比和评价，选定对事件和事件要素的标注较为全面的 CEC 语料库作为后续研究所用的训练语料和测试语料。2.4 节简要介绍了事件的抽取研究工作，引出了本文的研究路线和方法。

2.2 事件的研究及其定义

事件，反映着现实世界中的行为、运动及变化，是人类知识的基本单元。对于事件的相关研究，一直以来都受到学术界的特别重视。在哲学、认知科学、语言学、人工智能等领域的文献中，“事件”的有关概念随处可见。

哲学家认为，实体和事件共同构成了整个世界。实体通常具有具体的形态和结构，比如质量、体积等，能够被人们所感知和触摸。而事件通常指代一个行为或一系列状态的改变，是一个具体事实的体现且总随时间发生变化。早在公元前，古代著名的哲学家----亚里士多德，就已经给出了事件及事件之间的关系的有关论述。

对于认知科学家而言，他们对于事件的研究侧重于从两个方面去考虑：大脑的记忆原理和事件的结构。Zacks 和 Tversky 认为事件是一个具有起点和终点的时间片段[14]。Lindsay 和 Norman 等提出了一个由结点和连线构成的网络结构模型-----Elinor 模型[15]。其中结点代表概念、情景、事件等，连线表示二者在意义方面存在的联系。Nelson [16]认为事件是一个庞大的整体，且包括了对象和关系两个结构单元。

从语言学的角度考虑，语言学家给出了事件结构和事件的定义。Davidson 认为事件不但包含动词，还包括名词及一些修饰的成份。Chung[17]则认为事件是由三个部分组成的：谓词（句中的动词、形容词或部分指代词）、事件框架（谓词发生的时间段）、事件界（谓词发生的条件）。近年来，部分学术研究者将事件的结构与语言学中的 SVO（Subject-Verb-Object）结构相互应[18]：事件的参与者与语句中的主语互应，行为与语句中的谓语互应，事件的承受者与语句中的宾语互应。

在信息检索领域，把细化了的用于检索的主题称之为事件。由美国国防高级研究计划委员会主办的话题识别与跟踪（TDT）评测会议认为事件是小于话题的概念，一个话题由多个事件组成。“事件”被定义为“特定时间特定地点发生的事情”。会议指出话

题识别与跟踪主要有五项任务，而事件识别是其中一个十分关键的子任务。ACE 评测会议推动了事件抽取这一重大语言处理领域的向前发展，它认为事件通常是一种状态的改变，并将“事件”定义为包含参与者的特殊的事情。

在自动文摘领域，Filatova 等[19]通过调查问卷的方式得出事件单元等同于一个单独句子的结论，定义了“原子事件（Atomic Events）”的概念，将事件看作是<命名实体，事件项，命名实体>这样的一个三元组。周文等[20]将“事件”用多元组形式来表示，提出了基于事件和 FCA 的多文档自动文摘方法，在 DUC2005 的语料上进行的评测结果比较理想。

根据以上各领域对事件的相关研究可以看出，不同的应用领域对事件的定义各不相同。尽管如此,对事件的定义都离不开行为（一般由动词或动名词来描述）、事件的参与对象、事件发生的时间和地点等几个基本要素。综上所述，我们给出了事件及事件要素的形式化定义。

定义 2-1 （事件）：**事件（Event）**[21]指在某个特定的时间和环境下发生的、由若干角色参与、表现出若干动作特征的一件事情。形式上，事件可表示为e,定义为一个六元组:

$$e=(A,O,T,V,P,L)$$

其中，事件六元组中的元素称为事件要素，分别表示动作、对象、时间、环境、断言和语言表现。

A（动作）：事件的特征或其变化过程，是对方法、方式、程度、工具等的描绘，例如使用某物、赶赴某地等。

O（对象）：指事件的参与者。它是各对象的集合，参与事件的所有角色都是事件的对象要素。对象由主体和客体两部分组成，例如“消防员赶到现场”、“打败敌人”中的“消防员”是主体对象，而“敌人”是客体对象。

T（时间）：事件发生的时间段，从事件发生的起点到事件结束的终点，时间分为绝对时间段和相对时间段两类，例如“2008 年 5 月 12 日”是绝对时间，“30 分钟后”是相对时间。

V（环境）：事件发生的场所等。例如：在商场里购物。环境：商场。

P（断言）：断言是由三个条件构成，分别是：前置条件、后置条件和中间断言。触发事件发生的被称为前置条件；事件发生过程中所处的中间状态是中间断言；事件发生后各要素或状态的改变是事件的后置条件。

L（语言表现）：事件在语言方面的表现规律，包括中心词的集合、表现、搭配等。

下面举例说明各要素的语言表现形式：

例 2-1：昨日上午 9 时许，新 107 国道郑许路段，一辆大卡车与一辆小轿车发生追尾，小轿车司机李某当场死亡，其余 3 人不同程度受伤。

事件名称：追尾事故

动作：追尾，死亡，受伤

对象：李某，其余 3 人

时间：昨日上午 9 时许

环境：新 107 国道郑许路段

断言：

前置：人员正常，卡车完好，小轿车完好

后置：人员伤亡，卡车损坏，小轿车损毁

定义 2-2（事件触发词）：**事件触发词（Event Denoter）**是文本中可用来清晰的表示所发生事情的词语，即事件的动作要素。事件触发词一般是动词和名词。

例 2-2：2008 年 5 月 12 日，四川汶川县发生了 7.8 级地震。

例 2-3：9 日夜晚，郑（郑州）开（开封）高速公路上一辆大货车与一辆小轿车相撞，截至记者发稿时，这起事故已造成 3 人死亡，5 人受伤。

如下划线所示，例 2-2 中包含一个触发词即一个事件，例 2-3 中包含四个触发词即四个事件。

定义 2-3（事件类）：**事件类（Event Class）**指具有共同特征的事件的集合，用 EC 表示。

$$EC = (E, C_1, C_2, \dots, C_6)$$

其中，E 是事件的集合，称为事件类的外延；$C_i = \{c_{i1}, c_{i2}, \dots, c_{im}, \dots\}$称为事件类的内涵，是 E 中的每个事件在第 i 个要素上具有的共同特性的集合，c_{im} 是事件类中每个事件在第 i 个要素上具有的一个共同特性。例 2-4 为交通事故类事件的自然语言描述实例。

例 2-4：事件类：交通事故

动作：程度：严重、特大、重大

环境：路面

断言：前置：人和车辆正常行驶

后置：人伤亡、车辆受损

语言表现：

动作：触发词：碰撞、撞上、撞毁、追尾、车祸等

程度：严重、特大

环境：路面（马路、国道、高速路等）

断言：前置：无

后置：事故现场

整个事件类：在某时某地发生了交通事故；

2.3 基于事件的语料库

语料库（corpus）是存放语言材料的仓库，又叫语言数据库。现代语料库大致包括两种语料文本，一种是直接存放在计算机里的原始语料文本，另一种是经过加工后带有语言学信息标注的语料文本。其中，基于事件的标注语料库不仅可以用来辅助调查和统计，建立相应的统计模型，还可以对已有的基于事件的信息处理技术进行比较和评测。

目前已有的基于事件的语料库包括：自动内容抽取即ACE（Automatic Content Extraction）评测会议提供的ACE语料库[22,23]以及由美国高级研究发展学会（Advanced Research and Development Activity, ARDA）主办的问题回答系统中的时间和事件的识别即TERQAS（Time and Event Recognition for Question Answering Systems）会议的TimeBank语料库[24]。其中，ACE语料库提供了中英文二个版本，但它只对特定类型的事件（如：运输事件、生命事件以及交易事件等八大类事件）进行了标注，而忽略了同一文本中其他的事件，这种标注方法造成了语料库中事件的覆盖范围太小，无形中会约束后续的研究工作。而TimeBank 则采用了一种基于动词的标注模式，把句子中的动词都标注为事件触发词，这种标注模式把句子中表示状态的动词，比如“In Hong Kong , is always belongs to the seller's market.”中的 “is”也被标注为事件触发词，显然，这不符合人们对事件的定义。此外，TimeBank的标注重点是事件的时间要素，忽略了与事件相关的地点及参与者等关键信息的标注，而这些缺失的信息对于事件抽取方面的研究非常重要。再者TimeBank主要面向英文，由于中英文的语言表达方式存在巨大差异，所以它的标注方法对于中文事件是行不通的。

针对以上两种语料库存在的缺陷，上海大学语义智能实验室从互联网上搜集地震、交通事故、恐怖袭击、食物中毒以及火灾等五大主题类突发事件的新闻报道200篇作为生语料，开发了一个面向中文事件的标注工具----Event Annotator，使用XML语言对中文文本中的事件及其各事件要素进行标注，构建了面向中文事件的CEC（Chinese Emergency Corpus）[21]语料库，如表1所示。

表 1 CEC 语料库

Table1 CEC Corpus

事件类型	文本数	句子数	包含事件的句子数	事件	事件触发词	事件要素
交通事故	49	265	253	798	798	1201
地震	45	292	264	682	682	1026
火灾	31	260	224	496	496	783
食物中毒	45	288	260	701	701	1127
恐怖袭击	30	273	227	456	456	741
总数	200	1378	1228	3133	3133	4878

与ACE和TimeBank语料库相比，CEC语料库的规模虽小，但其对事件和事件要素的标注却是最全面的，因此本文在后续的研究中选用CEC语料库作为训练语料和测试语料。三者之间的对比如表2所示。

表 2 CEC 与 ACE 和 TimeBank 对比

Table2 Comparison of CEC，ACE and TimeBank

	CEC	ACE	TimeBank
支持语言	中文	中文/英文	英文
文本篇数	200	633/599	300
事件数	3133	2521/4090	7571
是否标注所有事件	是	否	是
是否标注事件要素	是	是	是
是否标注事件关系	是	否	是
标注模式	基于语义	基于特定事件	基于动词

2.4 事件的抽取研究介绍

事件的抽取主要包含两个步骤：一是对事件的识别，二是抽取出所识别事件的事件要素信息。

事件的识别是事件抽取的基础，而事件触发词是能清晰表示事件发生的词语，所以事件的识别即是能表征事件的触发词的识别过程。本文首先研究了基于扩展触发词表和基于多特征融合的机器学习两种触发词识别和抽取方法，经研究分析发现这两种方法各有利弊，为了各取其利避其弊从而达到更好的抽取效果，我们考虑将两种方法结合起来，先统计出训练语料中的事件触发词制成初始触发词表，然后根据扩展规则用《同义词林》对初始触发词表进行扩展得到触发词扩展表，进而利用触发词扩展表和机器学习相结合的方法对事件触发词进行自动识别抽取。实现触发词的识别后，划分出事件的范围，从而确定出文本中与事件触发词相关的事件要素，利用改进的聚类算法抽取出该事件的事件要素。

本文事件及其事件要素的抽取研究路线如图 1 所示。

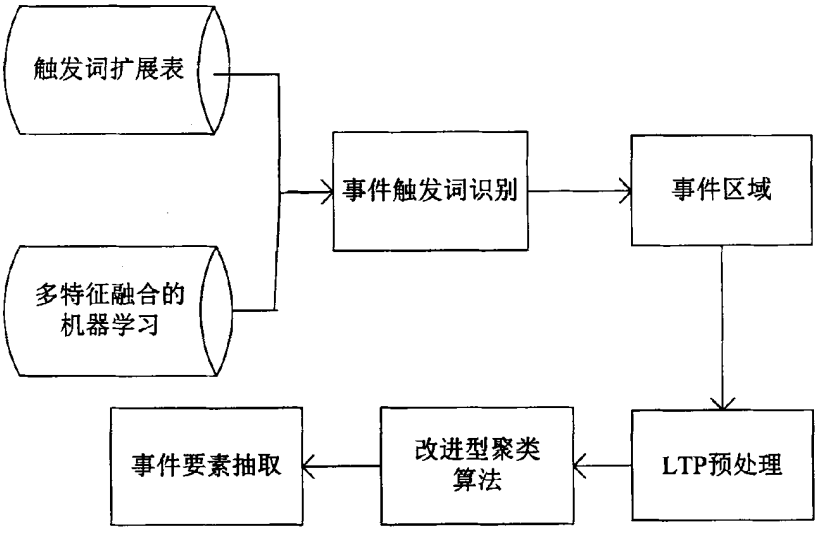

图 1 事件及其事件要素的抽取研究路线

Fig1 Event and event elements extraction research route

2.5 本章小结

本章主要介绍了事件的相关知识包括各领域对事件的定义，然后给出了事件、事件要素、事件触发词、事件类等的形式化定义。对现有的三种语料库资源进行了对比和评价，拟选取对中文事件信息标注较全面的 CEC 语料作为训练语料和测试语料。简要介

绍了事件抽取研究内容并给出了本文的研究路线和方法。本章内容是为后文事件的识别及事件要素的抽取工作提供理论基础。

3 事件的识别研究

3.1 本章概述

本章主要研究事件的识别即事件触发词的识别问题。3.2 节概述了事件识别的任务，明确识别抽取目标----事件触发词识别。3.3 节主要讨论文本表示模型和分类器模型的选取问题，经过比较，文本选用空间向量表示模型，分类器选用支持向量机（SVM）模型。3.4 节给出了事件触发词识别的研究现状，总结各研究方法的利弊，从而确定本文的研究方案----扩展触发词表和机器学习相结合的方法对事件触发词进行自动识别抽取。3.5 节介绍了基于扩展触发词表的事件抽取方法，而 3.6 节介绍了基于多特征融合的机器学习的事件抽取方法。3.7 节是把两种抽取方法相结合，保证了召回率和准确率。3.8 节给出了实验结果并对其讨论分析。3.9 节通过把本文研究方案与其他相关的研究工作进行讨论对比，从而总结出本文研究方法的利弊。

3.2 事件识别任务概述

事件的抽取任务主要分为两项：一是对事件的识别，二是抽取出已识别事件的事件要素。因此，事件识别是事件抽取的一个关键的子任务，是事件抽取的基础，它的效果的好坏直接影响到事件抽取的结果。事件触发词是能清晰表示事件发生的词语，所以事件的识别即是能表征事件的触发词的识别过程。事件、事件触发词及事件类的相关定义在本文第二章中已经给出。根据 ACE 评测任务，把事件识别作如下定义：

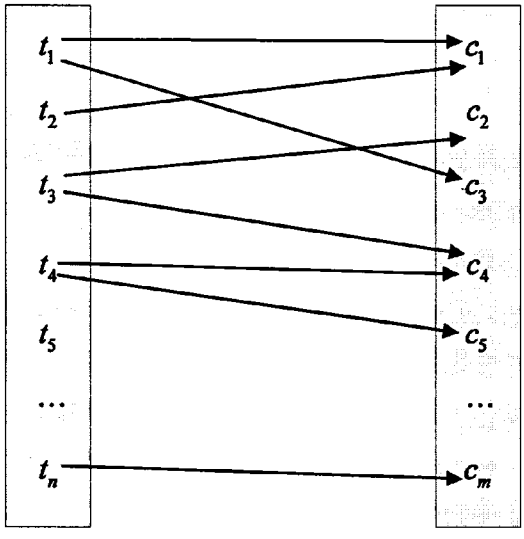

图 2 句子与事件类的映射关系

Fig 2 Mapping relationship between sentences and event classes

定义3-1（事件识别）[25]：事件识别（**Event Recognition**）是从包含事件触发词的句子或文本中发现现实世界中所发生的事件。

中文文本中的某个句子可能含有一个触发词（即一个事件）也可能含有多个触发词（即多个事件），对于事件识别任务，即是把这些触发词都识别出来，然后把相对应的句子划归到某个事件类中。设文本$T=\{t_1,t_2,...t_n\}$,其中t_i为文本T中的第i个句子，事件类$C=\{c_1,c_2,...c_m\}$，其中c_j为事件类C的第j个类别。事件识别实际上是要在t_i，i=1,2...n和c_j，j=1,2...m之间建立映射关系，如图2所示。

图 2 中，m是一个大于2 的数，它代表事件的类别数。所以事件识别问题可看成是一个多元分类问题。另外，句子与事件类之间是1对n的关系，$n \geqslant 0$。$n=0$表示一个句子不含有任何事件，n=1表示一个句子只含有一个事件，而n >1则表示一个句子中含有多个事件。例如：“昨日凌晨，新107国道郑许路段发生了一起交通事故，一辆大货车与一辆小轿车发生追尾，事故共造成2人死亡，3人受伤。”这一突发事故新闻报道中同时包含了“交通事故”、“追尾”、“死亡”、“受伤”四个事件，前两个事件属于交通事故类，后两个事件属于伤亡类，因此，一个含有事件的句子可能对应多个事件类。因为不存在一个事件既属于事件c_1又属于c_2的情况，也就是说事件类之间是相互独立的，所以可以把事件的识别看成多个二元分类问题，分类器可以选用支持向量机、最大熵或者k 最近邻等，本文选用支持向量机和最大熵这两个分类器。下文将对这两个分类器的模型及性能做详细介绍。

3.3 模型的选取

3.3.1 分类器模型

为了验证不同分类器对后文中各特征的区分能力，本文拟采用两种分类器模型：支持向量机（SVM）[26,27]模型和最大熵模型（ME）[28,29]。

（1）支持向量机（SVM）

支持向量机（SVM）是由Vapnik等于1995年提出的一种机器学习技术。SVM具备出色的学习性能，该技术也是当前机器学习方面的研究热点。近年来，SVM被广泛应用在文本分类技术中，并取得了重大的进展。SVM 的基本思想是以保证分类正确为前提最大化样本之间的间隔，从而最小化分类的实际风险和经验风险，它是从两类样本线性可分情况下的最优分类超平面发展起来的。

设线性可分的样本集$D=\{(\mathrm{x}_i,\mathrm{c}_i)\,|\,\mathrm{x}_i\in R^p,\mathrm{c}_i\in\{-1,1\}\}_{i=1}^n$，其中，$R^p$为样本空间，$x_i$

为R^p的向量，c_i为分类类别，n为样本数，则最优分类超平面可用以下方程表示：

$$w\cdot x+b=0 \tag{3.1}$$

且满足

$$c_i(\mathrm{w}\cdot \mathrm{x}+\mathrm{b})\geq 1,1\leq \mathrm{i}\leq \mathrm{n} \tag{3.2}$$

w为法向量，它的最优分类决策函数可表示为：

$$f(x)=\mathrm{sgn}(w\cdot x+b)=\mathrm{sgn}\left(\sum_{i=1}^{n}a_i^{*}c_i(\mathrm{x}_i\cdot \mathrm{x})+\mathrm{b}^{*}\right) \tag{3.3}$$

其中，sgn()为符号函数，a_i^{*}表示最优拉格朗日系数，b^{*}为最优阈值。针对线性不可分的情况，SVM引入了广义最优超平面，即在约束条件中允许训练样本增加松弛因子$\xi_i\geq 0$，约束条件放松为：

$$c_i(w\cdot x+b)+\xi_i\geq 1,1\leq i\leq n \tag{3.4}$$

对于非线性可分情况，SVM通过非线性特征的映射，将训练样本映射到高维特征空间中，并利用核函数$K(x_i,x_j)$计算高维向量的内积。常用的核函数包括：

1) 多项式核函数：$K(x_i,x_j)=(x_i\cdot x_j+1)^d$

2) 径向基核函数：$K(x_i,x_j)=\exp\left(-\gamma\left\|x_i-x_j\right\|^2\right)$

3) Sigmoid核函数：$K(x_i,x_j)=\tanh(\alpha x_i\cdot x_j+c)$

SVM 能有效处理非线性可分问题且具有很好的推广能力，算法最终将转化为二次型的寻优问题。SVM 最初是为二元分类问题设计的，针对该文事件识别任务，要求识别出多个事件类别，我们可以通过组合多个二元分类器来构造多元分类器，该方法能方便的处理一个样本属于多个类别的情况，对于本文中存在的一个事件句可能同时包含多个事件类的情况比较适用，也就是说比较适用于本文的研究。

（2）最大熵（ME）

最大熵模型的基本思想是排除未知因素为所有已知因素建立模型，即以不受任何未知因素的影响为前提找到一个满足全部已知事实的概率分布。该模型最显著的特点是它不要求特征必须具备独立性，所以我们可以随意添加对最后分类有用的特征，而不必考虑它们之间会产生相互影响。再者，ME 的训练效率相对较高。以上优点使其广泛应用

于句法分析与信息抽取等多个语言处理领域。

在讨论一个事件是否属于某个事件类的过程中会用到多种特征因素。假设 X 是这些因素构成的一个向量，变量 y 的值为 1（属于候选事件类）或者 0（不属于候选事件类）。$p(y|X)$指事件被识别为某事件类的概率，可以用上述思想来估计这个概率。最大熵模型要求$p(y|X)$在一定约束条件下，必须使得下面定义的熵取得最大值：

$$H(\mathrm{p})=\sum_{X,y}p(y|X)\log p(y|X) \tag{3.5}$$

这里的约束条件为最大熵模型的特征，用以下方式表述：

$$f_i(X,y)=\begin{cases}1, if(\mathrm{X,y})\text{ satisfies certain condition}\\ 0, else\end{cases},i=1,2,3,...,n \tag{3.6}$$

其中，$f_i(\mathrm{X,y})$为最大熵模型的特征，n 为所有特征的数目，可见这些特征描述了向量 X 和变量 y 之间的联系，输出的最终概率为：

$$p^*(y|X)=\frac{1}{Z(X)}\exp\left(\sum_i\lambda_i f_i(\mathrm{X,y})\right) \tag{3.7}$$

其中λ_i是每个向量的权重，且

$$Z(X)=\sum_y\exp\left(\sum_i\lambda_i f_i(X|y)\right) \tag{3.8}$$

3.3.2 文本表示模型

采用机器学习的方法识别事件，需要首先把文本转化为计算机可以读懂的格式，比如矩阵、向量等，然后句子中的事件才能用分类模型进行分类。常用的文本表示模型包括：概率模型（Probabilistic Model）、布尔模型（Boolean Model）、和向量空间模型（Vector Space Model，VSM）等[30]。

概率模型由 Spark Jones 和 Stephen Robertson 等人[31]提出，被广泛应用于信息检索领域。该模型基于概率排队理论，文本中关键词和文档之间的相互关系能够被准确的描述，但此模型过强依赖所处理的文本且处理方式过于简单，而且需要事先确定关键词和文档之间相关概率，从而限制了其广泛使用，但对检索系统的理论研究提供了依据。

布尔模型是基于经典集合论和布尔代数的[32] 首个被提出的文本表示模型。该模型在文本中构成特征项只有两种状态：每个特征项的权值为 1 或 0 相应的表示出现或不出现。布尔模型虽然表示方式简单、易于理解，但其表示能力非常刚性，特征项对文本语

义的重要程度如何得不到有效反映。

向量空间模型由 Salton 等首次提出，并于 1988 年又提出了有名的 TF-IDF 公式法[33]。该模型是目前文本表示的常用方法，是一种在自然语言处理领域被广泛应用的一种文本表示模型。该模型的优点是用向量空间中的向量运算去简化处理文本内容，在很大程度上降低了问题的复杂性，从而大大提高了文本处理速度。

通过对比以上三种常用的文本表示模型，发现向量空间模型在文本表示方面更具优势，因此本文选用向量空间模型来处理文本。

3.4 事件触发词识别的研究现状

近年来，在信息抽取领域，事件触发词的识别方法主要有三种：基于统计的方法[34,35]、基于规则的方法[36,37]和机器学习方法[38,39]。基于统计的方法是指人工统计出句子或文本中的所有触发词，建立一个较完整的触发词字典，通过此字典来判断其他词语是否为触发词。该方法简单易行，技术上要求不高，但它是一种典型的经验性方法，且要求训练语料规模足够大且足够经典，但事实上，由于非遍历性为首统计语料的限制，此方法并不能保证统计结果和测试结果的正确性，并且统计过程费时费力。基于规则的方法则是事先定义一些规则去寻找触发词，比如在文献[25]中，付剑锋经过研究验证得出了触发词一般是动词或名词的结论，就可以以此规则过滤掉其它词性的词语。在一定条件下该方法能有效的提高触发词的识别效率，减少工作量，但它是一个偏理论性的方法，只有在理想的情况下定义出涵盖所有语言特征的规则，才能保证该方法有效，而且规则的定义过程耗费大量的人力，如果规则定义的不够好，也可能过滤掉一些本身可以充当触发词的词，导致识别效果较低。中文语境和词性千变万化，但由于规则的有限性，这种理想化的任务几乎是不可能完成的。上述两种方法的性能在很大程度上依赖字典和规则的构建，进而依赖构建者的水平并且会耗费大量的人力和时间。伴随着机器学习的高速发展，基于机器学习的触发词识别能够基于训练集进行自动学习。机器学习方法引进了自动化模式，大大节省了人力物力的投入，但是，机器学习需要足够量的特征集训练分类器，即要求训练语料和测试语料必须满足一定的规模才能保证识别结果的精确率，机器学习也是一种统计学习方法，不可能照顾到每个实例。

由此可见，三种识别方法各有利弊，本文拟综合采用以上三种方法，各取其利。先基于统计的方法人工统计出训练语料中的事件触发词制成初始触发词表，然后根据触发词聚类规则用《同义词林》对初始触发词表进行扩展得到扩展的触发词表，进而利用触发词扩展表和机器学习相结合的方法对事件触发词进行自动识别抽取。整个过程融合了三种识别方法，能发挥各自优势提高识别效率且弥补了单一方法的局限性和不足，基于

统计方法的简单易行比较适用于事件触发词的初步统计从而建立初始事件触发词表，建立简单的聚类规则能实现触发词的有效扩展，扩展触发词表和机器学习的结合使用弥补了语料库资源缺乏造成单一使用机器学习带来的局限。

3.5 基于触发词扩展表的事件识别方法

3.5.1 事件触发词的聚类

（1）构建原始触发词表

利用对中文事件和事件要素标注较全面的 CEC 语料构建原始触发词表。在统计实验中，选取标注好的 CEC 语料作为训练语料，通过对 CEC 语料进行研究统计，进一步对五大主题类突发事件（地震、交通事故、恐怖袭击、食物中毒、火灾）进行细分，整理出了语料中出现频率较高且较重要的九类事件及各类事件的触发词，原始触发词表如表 3 所示。

表 3 CEC 中评测事件及其触发词统计表

Table 3 Evaluation of events and denoterstatistics in CEC

事件类型	数量	事件触发词	数量
地震	130	地震、震感、余震……	7
交通事故	142	车祸、追尾、撞车……	28
恐怖袭击	102	袭击、爆炸、劫持……	25
食物中毒	118	中毒、呕吐、恶心……	9
火灾	97	火灾、着火、燃烧……	18
伤亡	478	死亡、丧生、受伤……	33
损失	395	倒塌、损坏、救助……	98
救援	319	救治、施救、救助……	93
移动	287	赶到、赶赴、送往……	80
合计	2068		391

（2）扩展触发词表

由于受语料规模的限制，一些重要的触发词没有被统计到原始触发词表中，所以我们需要对触发词进行聚类，利用哈工大同义词林对原始触发词表进行扩展，本文使用的哈工大同义词林是《哈工大信息检索研究室同义词词林扩展版》。《同义词林扩展版》

具有五级结构即：大类、中类、小类、词群、原子词群。它提供五层编码，即大写英文字母表示大类，小写英文字母表示中类，两位十进制整数表示小类，第四、五级的编码与以上三级编码合并构成一个完整的编码，唯一的代表词典中出现的词语。具体的标记参见表 4。

表 4 词语编码表

Table 4 Word code table

编码位	1	2	3	4	5	6	7	8
符号举例	F	a	0	1	A	0	1	=\#\@
符号性质	大类	中类	小类		词群	原子词群		
级别	第1级	第2级	第3级		第4级	第5级		

在同义词林的基础上，结合人工检查的方法半自动的对触发词进行聚类从而对原始触发词表进行扩展。具体的触发词聚类规则如下：

1）从原始触发词表中找出各事件类型下的标志性触发词，并将其映射到同义词林中，得到其对应的词语编码。

2）对各触发词在同义词林中对应的词语编码进行统计，如果前四级与某事件类型下的标志性触发词的编码相同，则认为该触发词与标志性触发词具有相同或相近的意思，然后将其归并到该事件类下。

选择只比较词语编码的前四级是因为三级的范围过广，所包含词语的表达含义差别太大，与我们所定义的同类事件触发词的标准不符合，而五级分的过细，词语数量太少，所以四级是比较合适的。

3）最后，对于未归并入各事件类的事件触发词和在同义词词林中未查询到词语编码的事件触发词，我们进行人工聚类，得到扩展后的事件触发词表。

图 3 给出了根据聚类规则实现的事件触发词扩展策略流程图。

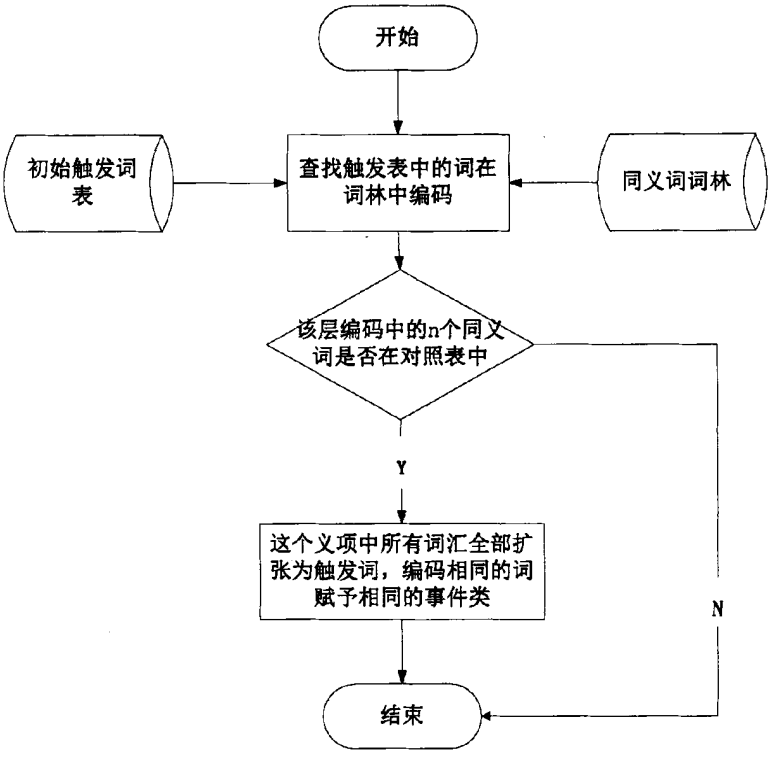

图 3 触发词扩展策略流程

Fig 3 Denoter expansion strategy process

按照以上聚类规则，得出了 CEC 语料库的事件触发词扩展表，部分结果如表 5 所示。

表 5 同类事件触发词扩展表（部分）

Table 5 Similar event denoter extension table(part)

事件类名称	事件触发词集合
相撞	撞 相撞 撞上 撞倒 碰撞 相碰 撞破 撞向 撞击 连撞 剐上 撞到 撞坏 撞出 撞穿 撞瘪
伤亡	死亡 伤亡 身亡 死伤 受伤 伤 负伤 受重伤 重伤 轻伤 轻微伤 丧生 危重 捅伤 遇难
救治	救治 医治 救护 就医 医治 急救 抢救 急诊 治疗 治病 诊治 诊疗 急救 注射 消毒 手术 输血

3.5.2 事件触发词的识别

首先，用分词工具对中文文本进行预处理，即分句、分词、词性标注，本文使用的

分词工具是哈工大语义实验室的 LTP 语言技术平台。然后，从预处理后的文本中筛选出名词、动词、动名词，因为由文献[25]的研究结果表明触发词的词性一般为名词、动词和动名词。通过词性筛选可以缩小候选触发词的范围。候选触发词可以用空间向量表示为：

$$W=\{(w_1,score_1),(w_2,score_2),...,(w_k,score_k)\} \tag{3.9}$$

式（3.9）中w表示候选触发词，score 代表触发词的权重。我们采用类似 TF*IDF 的方法来计算权重，计算公式如下：

$$score_i=TF(w_i)*IDF(w_i) \tag{3.10}$$

TF（Term Frequency）即词频，它反映触发词对整个事件的贡献程度，对于词w_i它的重要性表示为：

$$TF(w_i)=n_i/m_i \tag{3.11}$$

其中n_i为候选触发词w_i在所有训练语料中触发的某类事件的总个数，m_i为训练语料中该类事件的总个数。

IDF（Inverse Document Frequency）即逆向文件频率，反映词在训练语料中出现的频度，表示为：

$$IDF(w_i)=\log_2\left(\frac{N_i}{M_i}\right) \tag{3.12}$$

其中N_i为全部训练语料中句子总数，M_i为含有触发词w_i的句子总数。

我们可以根据权重 score 的取值情况设定一个阈值，把那些 score 小于阈值的触发词过滤掉。实验表明，该方法有很高的召回率，但由于一词多义等情况的存在造成准确率不高，为了提高识别的准确率，下面用基于机器学习的方法实现事件触发词的抽取。

3.6 基于多特征融合的机器学习的事件识别

基于机器学习的事件触发词的识别主要包括以下几个步骤：

（1）用分词工具对中文文本进行预处理，即分句、分词、词性标注，然后，从预处理后的文本中筛选出名词、动词、动名词；

（2）选取触发词的自身词性特征及词的上下文特征构造向量空间模型，从而创建训练集获得机器学习模型；

（3）用 SVM（支持向量机）和 ME（最大熵）机器学习方法对测试集进行分类。

本文中，根据触发词的共现特征，选取触发词和上下文的信息这两种语言特征建立

空间特征向量。所采用的特征有：词特征、词法特征、句法特征、语义特征及上下文特征，如表 6 所示。

表 6 触发词识别所采用的特征

Table 6 Characteristics of the denoterrecognition

特征名	特征描述
词特征	触发词本身
词法特征	触发词所对应的词性
句法特征	触发词所对应的依存关系
语义特征	触发词在字典中的释义
上下文特征	触发词左边 a 个词，右边 b 个词的词的特征、词法特征、句法特征及语义特征

由文献[8]研究表明核心词最近距离左边 3 个词语的位置和右边 2 个词语的位置，可以为核心词提供 90%以上的信息量，如果扩大词的范围，信息量不会显著增加反而会浪费更多不必要的计算。因此，我们以事件触发词为核心，取其左边 3 个词汇和右边 2 两个词汇（a=3，b=2）及它们的各个特征作为上下文特征来构造空间向量。以事件触发词为核心的特征向量可形式化表示为：

$$V=\begin{Bmatrix}\left(w_{i-3},f^1\left(w_{i-3}\right),\ldots f^k\left(w_{i-3}\right)\right),\ldots\ldots,\left(w_i,f^1\left(w_i\right),\ldots f^k\left(w_i\right)\right),\ldots\ldots\\ \left(w_{i+2},f^1\left(w_{i+2}\right),\ldots f^k\left(w_{i+2}\right)\right)\end{Bmatrix} \quad (3.13)$$

其中，w_i 表示事件触发词（词汇特征），$f^j\left(w_i\right)$ 表示词 w_i 的第 j 类特征（即词法、句法、语义等特征）。

例：消防员很快赶到 30 多公里外的灾区。句中事件触发词是“赶到”，所以该句子的特征向量表示为{ （‘NULL’，‘NULL’，‘NULL’，‘NULL’，‘NULL’，‘NULL’），（‘消防员’，‘n’,‘SBV’，‘1’，‘Ae01’），（‘很快’，‘d’，‘ADV’，‘1’，‘Eb23’），（‘赶到’，‘V’，‘HED’，‘Hf08’），（‘了’，‘u’，‘MT’，‘1’，‘Kd05’），（‘灾区’，‘n’，‘VOB’，‘1’，‘cb08’）}。因为在触发词左边只有两个依存词，所以左边第三个词是空的，用“NULL“来标记。

由于机器学习方法的引入使得触发词识别的准确率有了很大幅度的改善，但是，由于机器学习是一种偏统计学习的方法，不可能照顾到每一个实例，因此召回率有所下降，其实验结果见 3.8 节。为了兼顾准确率和召回率，我们考虑把触发词扩展表和多特征融

合下的机器学习两种抽取方法结合起来。

3.7 基于两种方法相结合的事件识别

基于扩展触发词表的事件触发词抽取是有较高的召回率但准确率却很低。而基于机器学习的事件触发词抽取的准确率有明显提高但召回率却低于前一种方法。如此，我们可以把两种方法结合起来，扬长避短。方法设计如下：

（1）先用基于扩展触发词表的方法构建候选触发词集，计算每个候选触发词的权重 score；

（2）根据 score 的取值情况设定一个阈值 threshold，阈值要设定的足够高，使得大于阈值的词一般都为单义词，从而保证抽取的准确率；

（3）如果候选触发词中存在 score 大于 threshold 的词，则把 score 最高的词确定为事件触发词；

（4）若不存在，使用 SVM 和 ME 机器学习的方法来抽取事件触发词。

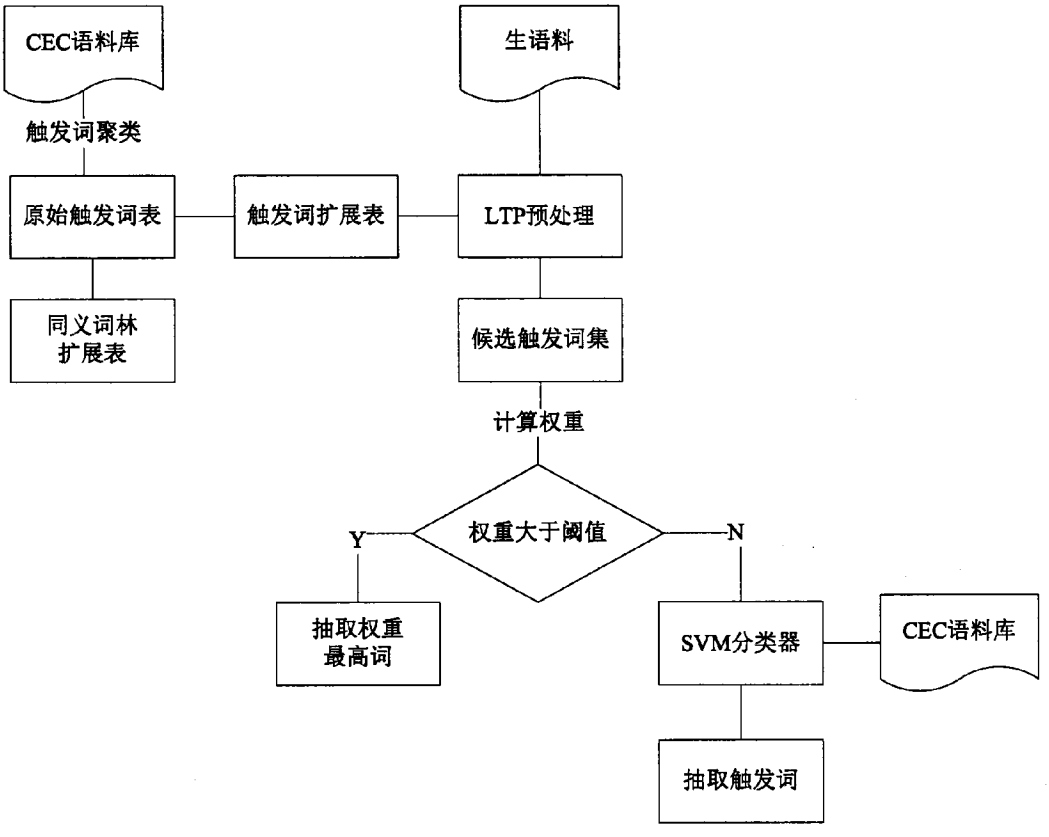

图 4 触发词识别过程

Fig 4 Denoter recognition process

如此，对于权重大于阈值的触发词一般为单义词，出现概率较低但对事件的贡献程度大，一旦出现一般就能表征某一类型事件的发生，如果这类词使用机器学习的方法来识别，由于实例比较缺乏，所以不具典型性，容易造成识别错误，导致召回率降低，因此这类词可直接进行查表识别。而对于小于阈值的那部分触发词，一般都有一词多义的情况，所以如果使用直接查表的方法极易造成事件的多标，导致精确率降低，因此使用机器学习的方法来解决。触发词识别过程如图 4 所示。

3.8 实验结果

实验采用三种评测方法来评价每个事件类的识别情况，分别为准确率 P、召回率 R 及 P 和 R 的加权几何平均值 F 指数。三种评测方法定义如下：

$$召回率(R)=\frac{正确的抽取结果}{所有可能的正确结果} \tag{3.14}$$

$$准确率(P)=\frac{正确的抽取结果}{所有的抽取结果} \tag{3.15}$$

为了综合评价系统的性能，通常还需计算召回率和准确率的加权几何平均值，即 F 指数，其计算公式如下：

$$F=\frac{(1.0+\beta^2)*P*R}{\beta^2*P*R} \tag{3.16}$$

其中，β是 R 和 P 的相对权重，$\beta=1$时，二者同样重要；$\beta>1$ 时，P 更重要些；$\beta<1$ 时，R 更重要些；β取值一般为 1、1/2、2。

（1）阈值的选择

阈值由候选触发词的权重 score 的取值情况来确定，下面举例说明阈值的选择问题。以“地震、余震、震感、震毁、损毁、受灾、震动、震撼”这 8 个可能触发地震类事件发生的候选触发词为例。根据 3.5.2 节已经介绍的类似 TF*IDF 的权重计算方法，可以算出以上 8 个词各自的权重，如表 7 所示。

表 7 地震类事件候选触发词的权重

Table 7 Comparison of experimental results of four extraction methods

候选触发词	地震	余震	震感	震毁	损毁	受灾	震动	震撼
权重 score	1.31	1.12	0.97	0.91	0.82	0.75	0.54	0.39

阈值大小选取的原则是：阈值要设定的足够高，使得大于阈值的词一般都为单义词。以上 8 个候选触发词中前 4 个词的权重相对较高且都为单义词，而后 4 个词中的“损毁”、

“受灾”可能触发地震类事件也可能触发其他类型的灾难性事件，比如火灾等，而“震动”、“震撼”不是单义词。以震动为例，“这个消息对大家的震动很大”、“大地再次震动起来”，这两句中的“震动”语义不同。所以，后 4 个词需要用机器学习的方法进行识别确定是否为地震类事件的触发词。综上，我们可以选择阈值为 0.9。权重大于阈值 0.9 的前 4 个词可以直接查取触发词扩展表确定为事件触发词和其对应的事件类，而权重小于阈值的后 4 个词需要用机器学习的方法进行识别。当阈值选择过小时，会造成事件的多标，从而影响识别的准确率，而当阈值选择过大时，易造成识别错误导致召回率降低，因此，阈值选择要适中。

表 8 五种抽取方法实验结果对比

Table 8 Comparison of experimental results of four extraction methods

试验方法	P	R	F
扩展触发词表	0.324	0.805	0.572
SVM	0.793	0.599	0.637
ME	0.785	0.562	0.625
扩展触发词表+ SVM	0.732	0.674	0.703
扩展触发词表+ME	0.703	0.682	0.691

（2）五种识别方法实验结果对比

为了验证五种识别方法在突发事件领域的抽取效果，我们采用 CEC 语料为训练语料和测试语料，整理出了 CEC 语料中出现频率较高且较重要的 9 类事件进行评测，采用语料的 3/4 作训练集，1/4 作为测试集。使用 SVM 和 ME 机器学习方法对测试集进行分类。实验结果如表 8 所示。

从以上实验结果可以看出，当词的权重较大即候选触发词为单义词的时候，使用基于扩展触发词表的方法能取得很高的召回率，但是准确率却很低。但当候选触发词的权重较小即存在一词多义的现象时，就不能单单依靠查表的方法来识别，而是要融合词本身和其他的多种特征进行学习来判定，即使用机器学习的方法。将两种抽取方法相结合的方法既保证了准确率又兼顾了召回率，达到了较为理想的效果。

（3）各特征的对比实验

机器学习采用了多特征融合的方法，这些特征包括：词自身特征、语义特征、词法

特征、句法特征及上下文特征。为了验证各特征对事件识别的贡献，分别进行了以下六组实验：

1) 单纯以词（Word）为特征进行事件识别；

2) 以词和语义（Word+Semantic）作为事件识别的特征；

3) 以词和词法（Word+POS）作为事件识别的特征；

4) 以词和句法（Word+DR）作为事件识别的特征；

5) 以词和上下文（Word+ Contextual）作为事件识别的特征；

6) 以上全部五种特征（All Features）作为事件识别的特征。

表 9 不同特征组合下的事件识别结果

Table 9 Event recognition results of different feature combinations

Features	扩展触发词表+SVM			扩展触发词表+ME		
	P	R	F	P	R	F
Word	0.662	0.601	0.633	0.667	0.603	0.636
Word+Semantic	0.688	0.623	0.651	0.676	0.628	0.647
Word+POS	0.701	0.658	0.675	0.697	0.647	0.663
Word+DR	0.721	0.671	0.694	0.706	0.665	0.681
Word+ Contextual	0.733	0.714	0.720	0.727	0.710	0.718
All Features	0.785	0.769	0.775	0.781	0.757	0.761

从表 9 可以看出，仅以词为特征，采用 SVM 和 ME 分别得到 0.631 和 0.623 的 F 值，加入语义特征后，F 值分别提高到了 0.652 和 0.647，以词和词法为特征可以得到 0.675 和 0.663 的 F 值，以词和句法为特征将 F 值提高到 0.694 和 0.681，而以词和词的上下文为特征时，进一步将 F 值提高到了 0.720 和 0.718，综合所有特征得到了 0.775 和 0.761 的 F 值。根据表 9 的实验结果，我们可得出结论如下：

1) 分类模型不同，得到的事件识别结果也有些许差异，从本文来看，SVM 的效果略好于 ME，但无论哪种分类模型下，在加入了单个类型特征（语义、词法、句法、上下文）后，分类效果均有所提高，多特征融合后取得的识别效果最好；

2) 选择词和句法特征以及词和上下文特征，分类效果好于词和语义及词和语法作为特征，也就是说，句法特征和上下文特征更能体现词汇所在句子的语法结构，二者有更好的区分能力；

3) 以单独采用词自身为特征的分类结果为基准，在事件识别时选择的有效特征越

多，分类的性能越好，事件识别的效果也越好。由于事件识别本身的特殊性，一个句子中包含的词汇的信息量非常有限，因此我们需要在有限的词汇中发掘尽量多的有效特征来实现事件的区分，尤其是发掘词汇的语法、句法以及上下文特征。

3.9 相关工作对比分析

本文的研究工作受到文献[40]的启发，但在语料库选取和机器学习方面做了改进，语料库选取对中文事件标注较全面的 CEC 语料，机器学习方面不止关注触发词本身特征而是融合其他多种特征进行学习并训练分类器，所以取得了较之理想的 F 值。文献[41] [42]等分别只针对生物和音乐领域事件进行抽取，抽取方法适用领域有限，而本文抽取方法可跨领域使用，更具普遍性。Grishman[43,44]、Hardy[45]、于江德[46]及赵妍妍[40]等也提出了基于触发词驱动的识别模型，该机器学习模型将每个词作为实例来训练并判断其是否为事件触发词，这样引入大量的反例从而导致正比例严重不平衡；并且，由于受语料规模的限制，事件类别的多元分类存在一定的数据稀疏问题。本文通过对触发词进行扩展然后再结合多特征融合的机器学习进行训练及判断，能有效解决这一问题，取得较理想的实验结果。

3.10 本章小结

本章提出了基于扩展触发词表和多特征融合的机器学习相结合的事件抽取方法，该方法先从训练语料中统计出事件触发词制成初始触发词表，然后用《同义词林》对原始触发词表进行扩展得到触发词扩展表，利用基于扩展触发词表的事件触发词识别方法能得到较高的召回率但准确率却不高。机器学习中融合词特征、词法特征、句法特征、语义特征及上下文特征等多种特征训练分类器，用机器学习的方法进行事件触发词抽取准确率较高但召回率受到影响，本文将以上两种抽取方法相结合使用，实验表明，该方法兼顾了事件抽取的召回率和准确率，取得了较理想的 F 值。

4 基于改进型聚类算法的事件要素抽取研究

4.1 本章概述

事件要素抽取是事件抽取的子任务之一。针对传统聚类算法的不足，本章提出了一种适用于事件要素抽取工作的改进型聚类算法。4.2 节对事件要素的抽取任务进行了简要概述。4.3 节先分别介绍了特征选择算法（ReliefF 算法）和传统的聚类算法（*k*-means 算法），然后对两种算法进行了移植改进得到了改进型的聚类算法。4.4 节介绍了改进型聚类算法在事件要素抽取研究中的应用。4.5 节给出了实验结果并与传统的聚类算法结果进行对比分析

4.2 事件要素抽取任务概述

上文中完成了对事件触发词的识别，我们把含有触发词的单句或分句划分为一个事件区域，事件要素的抽取都是在事件区域中进行的。通过上文事件的识别任务，人们可以清晰的获知发生了什么事情，若想进一步知道的与事件密切相关的其它信息，则需要对其事件要素进行识别抽取。例如对于发生的一起交通事故，大家侧重希望知道它发生的时间、地点、涉及的肇事者和受害者等详细情况。利用这些事故信息，可以进而分析出交通事故较易在哪个时间段以及哪个地段发生，从而相关部门可以加以有效的防范。

事件要素是与事件相关的实体以及实体的属性[47]，通常包括事件发生的时间、地点、参与者等，比如：在交通事故类事件中，其事件要素包括事故发生的时间、地点、肇事者以及受害者等；而对于一个死亡事件，它的要素则包括死亡的时间、地点、死亡对象等。事件要素抽取的任务即是从事件区域中识别出与事件相关的实体以及实体的属性并抽取。

目前，事件要素的抽取工作几乎都是选用基于机器学习的方法，而机器学习方法主要采用了监督学习的方法。为了获取事件要素信息，这种学习方法需要大规模人工标注的熟语料作为训练集，然而现如今语料库资源的缺乏导致最终的识别效果并不理想。文献[40,48-50]在事件抽取的研究工作中都受到了语料稀疏问题的困扰。因此，我们应该探索出一种对语料库依赖尽量小的新的事件要素识别方法。本文采用聚类（无监督）学习的方法从生语料中直接抽取事件要素，减少了对语料库的依赖。

k-means 算法[51]是一种基于距离的典型的聚类算法，采用距离作为相似性的指标，即两个对象相似度的大小取决于它们距离的远近。而本章所讨论的事件要素的抽取需要把同一事件里的不同要素进行聚类，不同要素所归属哪个类簇需要通过计算它们的相似度来决定，因此*k*-means 算法模型比较适用于事件要素的抽取研究 。但是，传统的

k-means 算法隐含的前提是：假定样本矢量的各维特征对聚类分析的贡献效果是均等的，即忽略各个特征对聚类分析影响的差异，然而在实际应用中，各个特征在聚类过程中的作用差别较大，即对聚类的贡献是不等的。所以在聚类分析过程中，特征的权重必须考虑在内。为了解决这一问题，我们提出了一种改进型的 *k*-means 聚类算法，利用特征选择算法即 ReliefF 算法对特征进行加权选择，然后将两个算法进行移植改进，使改进后的算法能够适用于事件要素的抽取中。

基于改进型聚类算法的事件要素抽取的研究方法和路线如图 5 所示。

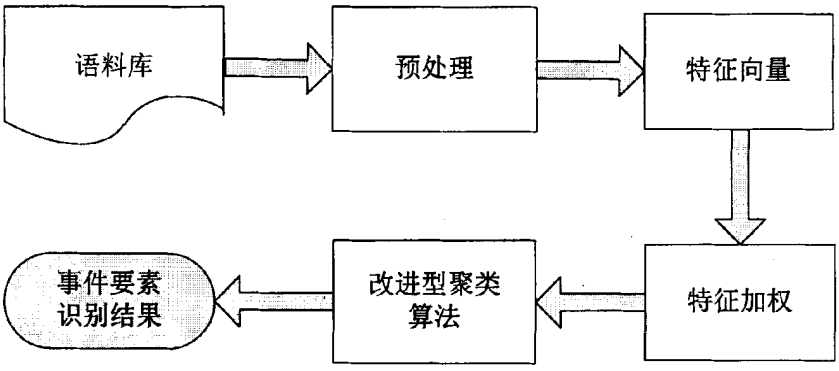

图 5 事件要素抽取研究路线图

Fig 5 Event elements extraction research roadmap

4.3 基于特征加权的改进型聚类算法

作为数据挖掘和市场分析领域的重要方法，聚类分析的目的就是依据特定的准则将未作任何标记的样本划分为若干类簇，并且要求相似度大的尽可能的在同一类别中，而相似度小的应在不同的类别中。*k*-means 算法的基本思路是：首先假设目标函数可微，将数据进行初始化处理划分为数个类簇并计算各个类簇的中心；随后，在迭代过程中调整各样本所属的类簇，直到其所属类簇不再发生变化，即其收敛。此时，目标函数达到最优，得到最终聚类结果。

4.3.1 ReliefF 算法介绍

特征选择是市场分析及数据挖掘等领域的关键技术之一，其根本任务是从待分析数据的众多特征信息中选择出最有效的特征。科研领域较为常见的特征选择方法为ReliefF 算法，该算法于 1992 年被提出，其作用是依据某个特征集合中的每一个特征在聚类分析时的重要程度赋予其相应的权重。

待进行聚类分析的n个对象为矢量记为$W=\{W_1,W_2\cdots W_n\}$，矢量中$W_i=\left[W_{i,1},W_{i,2},\cdots,W_{i,m}\right]^T$表示第$i$个对象的$m$个特征值。各特征在整体中所占权重的权值用$\alpha$表示，$\alpha$为$m\times1$的矩阵。

针对样本$W_i(i=1,2\cdots m)$做如下处理：

（1）在W_i的同类簇中寻找出数量为R的与W_i相邻最近的样本，记为$x_j(j=1,2,\cdots R)$;

（2）在与W_i的不同类簇的每一个子集中寻找出数量为R的与W_i相邻最近的样本$y_{lj}(j=1,2,\cdots R)$，其中$l\neq class(W_i)$。

记$diff_hit$表示W_i与x_j之间特征上的差异，且其为$m\times1$的矩阵，表达式为：

$$diff\_hit=\sum_{j=1}^{R}\frac{\left|W_i-x_j\right|}{\max(W)-\min(W)} \tag{4.1}$$

其中$\max(W)$和 $\min(W)$分别表示矢量W中的最大值和最小值。

记$diff_miss$表示W_i与y_{lj}之间特征上的差异，且其为$m\times1$的矩阵，表达式为:

$$diff\_miss=\sum_{l\neq class(W_i)}\frac{P(l)}{1-P(class(W_i))}\sum_{j=1}^{R}\frac{\left|W_i-y_j\right|}{\max(W)-\min(W)} \tag{4.2}$$

其中，$P(l)$表示第l类簇出现的概率，是该类簇中样本数量与全部样本数量的比值。

上文中提到的矩阵α为各位特征的权值，在 ReliefF 算法中，α的值可以通过式(4.3)不断更新

$$\alpha=\alpha-diff\_hit/R+diff\_miss/R \tag{4.3}$$

矩阵α的值不再发生变化时，则为特征集中的各个特征所对应的权重。

4.3.2 基于特征加权的改进型 k-means 算法

在传统k-means算法中隐含了一个特定的前提：样本矢量的各维特征对聚类分析的贡献效果是均等的，即忽略各个特征对聚类分析影响的差异。然而在实际应用中，各个特征在聚类过程中的作用差别较大：某些特征处于支配地位，对类簇的形成起决定性的作用，同时另外一些特征则处于从属地位，对类簇形成无关紧要，一些时候甚至会引入干扰而对类簇的形成带来干扰，破坏类簇的形成。因此，在聚类分析中，特征的权重必须考虑在内。

为了克服这种缺陷，使聚类效果更好，本文提出了一种基于特征加权的改进型k-means算法。改进后的聚类算法能充分考虑各维特征在聚类时的贡献，并通过 ReliefF

算法计算出各维的权重。为与传统k-means算法进行比较，接下来首先说明传统方法，并在其基础之上详细描述改进型方法。

矢量$W=\{W_1,W_2\cdots W_n\}$为将进行聚类的n个对象，$W_i=\left[W_i^r,W_i^c\right]^{\mathrm{T}}$ 为矢量W中的第i个样本的特征值，W_i^r和W_i^c分别为数值特征与类属特征，具体表达式：

$W_i^r=\left[W_{i,1}^r,W_{i,2}^r\cdots W_{i,t}^r\right]$，$W_i^c=\left[W_{i,t+1}^c\cdots W_{i,m}^c\right]$。

将W按照一定标准初始化为数目为k的聚类，并记为$Q=\{Q_1,Q_2\cdots Q_k\}$，为k个聚类原型，且$Q_i=\left[Q_{i,1}^r,\ldots,Q_{i,t}^r,Q_{i,t+1}^c,\cdots,Q_{i,m}^c\right]$。在$k$-mean算法中，目标函数为：

$$f(Q)=\sum_{i=1}^{k}\left(\sum_{j=1}^{n}\sum_{l=1}^{t}\left|W_{j,l}^r-Q_{i,l}^r\right|^2+\alpha\sum_{j=1}^{n}\sum_{l=t+1}^{m}\delta\left(W_{j,l}^c,Q_{i,l}^c\right)\right) \tag{4.4}$$

其中，加号左边为Euclid距离的平方，右边为类属特征的相异匹配度，表达式为：

$$\delta(u,v)=\begin{cases}0,u=v\\1,u\neq v\end{cases} \tag{4.5}$$

虽然在目标函数中，可用α调节上述两项在$f(Q)$中所占比重，但两项内部的各维对聚类的影响仍然是均等的，这也是传统k-means算法的缺陷所在。

为解决此问题，需要对传统k-means算法进行改进。设$\alpha^r=\left[\alpha_1^r,\cdots,\alpha_t^r\right]^{\mathrm{T}}$为各维数值特征所对应的权值，$\alpha^c=\left[\alpha_{t+1}^c,\cdots,\alpha_m^c\right]^{\mathrm{T}}$为为各维类属特征所对应的权值，在此基础之上，对原目标函数可以作如下调整：对各维特征均乘上其对应的权值。调整之后的目标函数为：

$$f(Q)=\sum_{i=1}^{k}\left(\sum_{j=1}^{n}\sum_{l=1}^{t}\alpha_l^r\left|W_{j,l}^r-Q_{i,l}^r\right|^2+\sum_{j=1}^{n}\sum_{l=t+1}^{m}\alpha_l^c\delta\left(W_{j,l}^c,Q_{i,l}^c\right)\right) \tag{4.6}$$

与传统算法一样，当改进后的目标函数$f(Q)$取得最小值时聚类结果就是最优的。由式（4.6）可知改进型算法的关键问题在于各特征权值的求取，而上节中已经讨论ReliefF算法可以自动更新权值直到类簇收敛。因此，可以引入该算法对改进后的目标函数中的权值进行求取。

改进后的特征权值分为两类：数字的权值，类属的权值。在使用ReliefF算法时应分别处理，区别对待。针对α^r，其求法与ReliefF算法一致：

$$\alpha^r=\alpha^r-diff\_hit^r/R+diff\_miss^r/R \tag{4.7}$$

由式（4.1），（4.2）可知，$diff_hit^r$及$diff_miss^r$均为$t\times1$矩阵，其意义上节已述。而$diff_hit^c$及$diff_miss^c$为$(m-t+1)\times1$矩阵：

$$diff\_hit^c = \sum_{j=1}^{R} \delta\left(h_j^c, W_i^c\right) \tag{4.8}$$

$$diff\_miss^c = \sum_{l \neq class(W_i)} \frac{P(l)}{1 - P\left(class(W_i)\right)} \sum_{j=1}^{R} \delta\left(m_{l,j}^c, W_i^c\right) \tag{4.9}$$

在重新定义了 $diff\_hit^c$ 及 $diff\_miss^c$ 后，类属特征各维权值计算方法如下：

$$\alpha^c = \alpha^c - diff\_hit^c / R + diff\_miss^c / R \tag{4.10}$$

此时，通过 ReliefF 算法不断更新改进型目标函数中各特征中的各维权值，直到各特征不再变化时得到收敛效果即最优的聚类结果。

4.4 基于改进型聚类算法的事件要素识别

事件要素的识别在事件抽取中至关重要，接下来本文将上述改进 aa 型聚类算法应用于事件要素的识别中，提出一种基于改进型 *k*-means 算法的事件要素识别法。其基本思想方法是：

（1）在某些特定类型的事件中，其语境类似，譬如空难事故的报道中都会出现“机场”、“航班”、“机组人员”、“黑匣子”、“航空公司”等词汇。可以使用这种特征把该类事件包含的全部要素聚类，且将不同要素聚类在相异的类簇之中；

（2）在聚类的过程当中，不同特征对聚类结果的贡献各有不同。通过 ReliefF 算法对传统 *k*-means 算法的目标函数进行改进，对各个特征进行加权处理，以达到良好的聚类结果。

为区别事件中各要素，借鉴事件触发词识别方法，在描述词语特征时本文选取了其词性、句法、语义以及位置等特征，文本表示模型仍采用向量空间模型，。针对上下文特征的方面，文献[14]研究得出以中心词汇的前后各一个词作为上下文可有效避免要素间的干扰，本文亦采取同样方法处理上下文特征。综上所述，用于事件要素抽取的特征为：

1）中心词 w_i，w_i 词性，w_i 自身的语义，w_i 所在语句中的句法；

2）w_i 左边词汇 w_{i-1}，w_{i-1} 词性，w_{i-1} 自身的语义，w_{i-1} 在所在语句中的句法；

3）w_i 右边词汇 w_{i+1}，w_{i+1} 词性，w_{i+1} 自身的语义，w_{i+1} 在所在语句中的句法；

4）事件触发词 d，d 所属事件类，d 词性，d 自身的语义，d 在所在语句中的句法；

5）w_i 和 d 二者位置特征。

在确定了以上特征之后即可进行聚类分析。显而易见，在聚类分析时核心词 w_i 对应的特征对聚类效果的影响要大于 w_{i-1}，w_{i+1} 等词对应特征的影响，因此需要区别对待这些特征。本文采用 4.3.1 节中所述的 ReliefF 算法对各个特征进行加权处理，经过若

干次迭代后，得到特征集合中任一特征在聚类分析中所对应的权重。权重矢量为α：

$$\alpha = \alpha - diff\_hit/R + diff\_miss/R \tag{4.11}$$

随后，将式子（4.11）中的各特征权重融入传统k-means算法，将其进行改进，形成基于特征加权的改进型k-means算法，如4.3.1 节所述。利用该算法，对本节确定的事件要素抽取的特征进行聚类，可得到事件要素识别的结果。

4.5 实验结果

首先，采用ReliefF 算法对各个特征进行加权处理，经过25 次迭代后，得到特征集合中各个特征在聚类分析中所对应的权重，如图6 所示。

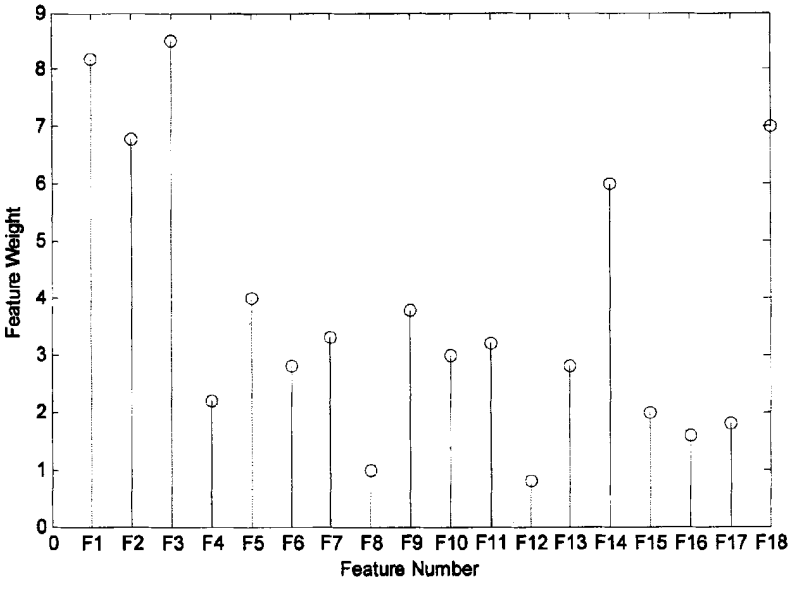

图6 各特征对应的权重

Fig 6 Weight of each feature

其中，F1—F4 分别表示中心词w_i以及它的词性、句法、语义特征；F5—F8 分别表示w_i左边词汇w_{i-1}以及它的词性、句法、语义特征；F9—F12 分别表示w_i右边词汇w_{i+1}以及它的词性、句法、语义特征；F13—F17 分别表示事件触发词d，d所属事件类，d词性、句法及语义特征；F18 表示中心词w_i和d二者位置特征。从图中可以很清晰的看出，各个特征对聚类的贡献是不一样的，其中，中心词w_i本身所有特征、事件触发词d所属事件类、w_i和d二者位置关系对聚类贡献较大；w_i左边词汇w_{i-1}和右边词汇w_{i+1}的贡献基本相当。

为了验证特征加权的有效性，我们比较进行了两组实验：用传统的k-means聚

类算法和基于特征加权的改进型 *k*-means 聚类算法分别对事件要素进行识别抽取，仍采用准确率 P、召回率 R 及 P 和 R 的加权几何平均值 F 指数对抽取结果进行评测，实验结果如表 10 所示。

表 10 聚类算法改进前后实验结果对比

Table 10 Comparison of experimental results of before and after the improvement of clustering algorithm

传统 *k*-means			改进型 *k*-means		
P	R	F	P	R	F
0.663	0.625	0.642	0.706	0.657	0.681

从以上结果可以看出，基于特征加权的改进型聚类算法可以有效的提高事件要素识别的聚类效果，F 值优于用传统的聚类算法的识别结果。

4.6 相关工作对比分析

文献[40,48-50]采用机器学习的方法，由于该方法需要大规模标注好的熟语料作为训练集，在事件抽取的研究工作中都受到了语料稀疏问题的困扰。文献[52-54]采用基于规则的方法抽取灾难性事件或突发事件的事件要素信息，由于手工构造规则费时费力，他们在规则构造过程中都加入了机器学习的方法。但是基于规则的方法在特定数据集或特定领域的抽取效果比较好，如果数据集发生变化，性能差异就较大，也就是说基于规则的方法移植性比较差。另外，规则较多时不同规则之间可能会发生冲突，会影响匹配效果和最终的识别结果。本采用无监督聚类的事件要素识别方法，可以有效摆脱对语料库资源的依赖，移植性也较好。

4.7 本章小结

在事件识别的研究工作中，比较主流的方法是基于机器学习的方法，该方法鲁棒性好、可移植性好，但是需要大规模的标注语料训练分类器，而现如今语料库资源匮乏，该方法的使用也受到了一定的限制。本文采用聚类（无监督）学习的方法从生语料中直接抽取事件要素，减少了对语料库的依赖。利用特征选择算法即 ReliefF 算法对特征进行加权选择，对传统的 *k*-means 算法进行改进，使改进后的算法能够适用于事件要素的抽取中。实验表明，该方法有较好的识别效果，可直接用于事件要素的抽取。

5 总结与展望

5.1 总结

人类是以“事件”为单位了解和认识现实世界的，对事件的研究一直都深受学术界的重视。事件的识别与抽取目的是利用自动抽取技术从非结构化的文本数据中抽取用户关注的事件信息，并转化为结构化形式呈现给用户，从而方便阅读或浏览。

本文对事件及其事件要素的抽取工作进行了研究，主要有以下两方面研究成果：

（1）提出了基于扩展触发词表和机器学习相结合的事件触发词抽取研究

本文提出了一种事件识别的方法：基于扩展触发词表和多特征融合下的机器学习相结合的方法。实验结果表明，单独基于扩展触发词表的事件触发词识别有较高的召回率但准确率却很低。机器学习方法融合多种特征训练分类器完成事件触发词的识别和抽取，构造的特征向量中加入了词特征、词法特征、句法特征、语义特征及上下文特征这五类特征，实验表明融入的特征越多识别的效果越好，将这多种特征融合一起时识别的结果最好。单独使用该识别方法时准确率有明显提高但召回率却低于前一种方法。如此，本文把两种识别方法结合起来，根据计算得出的权重分布情况设定一个阈值，候选触发词的权重大于阈值时即认定为事件触发词，当小于阈值时，就用机器学习的方法进行识别判断。实验结果表明通过把两种方法进行结合使用，召回率和准确率得到兼顾，F 值也比较理想。

（2）提出了基于改进型聚类算法的事件要素抽取研究

本文采用聚类（无监督）学习的方法从生语料中直接抽取事件要素，可以减少对语料库的依赖。传统的 *k*-means 算法忽略了各个特征对聚类分析影响的差异，在实际应用中，各个特征对聚类的贡献是不等的，所以在聚类分析过程中，特征的权重必须考虑在内。利用特征选择算法即 ReliefF 算法对特征进行加权选择，然后对传统的 *k*-means 算法进行改进，使改进后的算法能够适用于事件要素的抽取工作中。实验表明，改进后的算法比传统算法的识别效果好。

5.2 展望

本文主要研究了事件抽取技术的几个相关问题，主要包括：基于扩展触发词表的事件触发词抽取识别、基于多特征融合的机器学习的事件触发词识别、以及以上两种方法结合下的事件触发词识别、基于改进型聚类算法的事件要素抽取等几方面的内容。尽管已经取得了一些阶段性的成果，但仍存在一些不足，有些内容需要进一步的研究和改进：

（1）语料库资源有待进一步的开发。现如今信息抽取的主流技术仍是基于机器学习的方法，但是语料库资源的缺乏限制了该方法广泛有效的使用，构建大规模的语料库有助于机器学习方法在信息抽取方向的发展；

（2）在事件和事件要素识别方面，忽略了指代消解的问题。比如：“他们”（指代事件参与者）、“当时”（时间指代）、“当场”（地点指代）等等，这些指代信息对理解文本理解和事件信息很重要，对该类信息的识别和抽取需要进行深入的研究；

（3）本文对事件要素的抽取工作都是在事件区域内进行的，可是事件要素也很能隐含在事件区域外，如何根据上下文关系对这部分要素进行识别是一项挑战性的工作。

参考文献

[1] Allan J, Carbonell J G, Doddington G, et al. Topic detection and tracking pilot study: Final report. Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[2] ACE.The ACE 2005 (ACE05) evaluation plan. In http: //www.nist.gov/speech/tests/ace /ace05/doc/ace05-evalplan.v3.pdf. 2005.

[3] David Ahn. The stages of event extraction[C]. In Proceedings of the Workshop on Annotations and Reasoning about Time and Events, Sydney, 2006:1-8.

[4] Chieu H L, Ng H T. A maximum entropy approach to information extraction from semi-structured and free text[J]. AAAI/IAAI, 2002: 786-791.

[5]Chen Z, Ji H. Language specific issue and feature exploration in Chinese event extraction[C].Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Association for Computational Linguistics, 2009: 209-212.

[6] Heng Ji , Ralph Grishman. Refining Event Extraction through Cross-Document Inference[C]. In Proc. ACL-08: HLT, 2008: 254–262.

[7] 李芳,盛焕烨等. 多语种投资信息抽取系统地实现[J]. 上海交通大学学报，2004(1): 22-25.

[8] 周剑辉. 基于规则自动获取的金融事件抽取研究[M]. 清华大学硕士论文,2003.

[9] 宋凡. 音乐领域典型事件抽取技术的研究[M]. 哈尔滨工业大学硕士论文, 2009.

[10] 吴平博,陈群秀,马亮. 基于事件框架的事件相关文档的智能检索研究[J].中文信息学报，2003,17(6):25-30.

[11] 杨尔弘. 突发事件信息提取研究[D]. 北京:北京语言大学, 2005.

[12] 付剑锋, 刘宗田, 付雪峰, 周文, 仲兆满. 基于依存分析的事件识别[J]. 计算机科学， 2009, 36(11):217-219.

[13] 付剑锋, 刘宗田, 刘炜, 单建芳. 基于特征加权的事件要素识别[J]. 计算机科学， 2010, 37(3):239-241.

[14] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing and Management, 2003, 24(5):513-523.

[15] 潘云鹤,耿卫东.面向智能计算的记忆结构理论综述[J].计算机研究与发展.2000,31(12):37-42.

[16] Nelson K, Gruendel J. Event knowledge: structure and function in development[M]. Erlbaum, Hillsdale, NJ ,2001.

[17] Chung S, Timberlake A. Tense, aspect, and mood[J]. Language Typology and Syntactic

Description, 1999(3):202–258.

[18] Chang J. Event Structure and Argument Linking in Chinese[J]. Language and Linguistics, 2003,4(2):317-351.

[19] Hatzivassiloglou V, Filatova E. Domain-independent detection, extraction, and labeling of atomic events[M]. 2003.

[20] 周文.若干知识表示模型及其相关方法研究[D]. 上海：上海大学，2007.

[21] 刘宗田,黄美丽,周文,仲兆满,付剑锋,单建芳,智慧来.面向事件的本体研究[J]. 计算机科学, 2009, 36(11):189-192.

[22] Consortium L.D. ACE (Automatic Content Extraction) English Annotation Guidelines for Events, 2005. http://www.ldc.upenn.edu/Projects/ACE/.

[23] Consortium L.D. ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Entities, 2005. http://projects.ldc.upenn.edu/ace/docs/Chinese- Entities-Guidelines_v5.5.pdf.

[24] Pustejovsky J, Hanks P, Sauri R, et al. The timebank corpus[C].Corpus linguistics. 2003: 40.

[25] 付剑锋.面向事件的知识处理研究[D]. 上海：上海大学，2010.

[26] Chen X. Why did John Herschel fail to understand polarization? The differences between object and event concepts[J]. Studies in History and Philosophy of Science Part A, 2003, 34(3): 491-513.

[27] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval[M]. New York: ACM press, 1999.

[28] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing and Management, 2004, 24(5):513-523.

[29] Cristianini N, Shawe-Taylor J. An introduction to support vector machines[M]. 2000.

[30] 廖涛. 面向事件的文本表示及其应用研究[D]. 上海：上海大学，2014.

[31] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3): 293-300.

[32] 李素建, 刘群, 张志勇. 语言信息处理技术中的最大熵模型方法[J]. 计算机科学. 2002, 29(7):108-111.

[33] Berger AL, Della Pietra SA, Della Pietra VJ. 1996. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 2004,22(1):39-71.

[34] Buyko E, Faessler E, Wermter J, et al. Event extraction from trimmed dependency graphs[C]. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, 2009:19-27.

[35] Vlachos A, Buttery P, Séaghdha D O, et al. Biomedical event extraction without training data[C].Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing:

Shared Task. Association for Computational Linguistics, 2009:37-40.

[36] Cohen K B, Verspoor K, Johnson H L, et al. High-precision biological event extraction with a concept recognizer[C]. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, 2009:50-58.

[37] Le Minh Q, Truong S N, Bao Q H. A pattern approach for biomedical event annotation[C]. Proceedings of the BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, 2011: 149-150.

[38] Björne, Heimonen J, Ginter F, et al. Extracting complex biological events with rich graph-based feature sets[C]. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, 2009:10-18.

[39] Vlachos A, Craven M. Biomedical event extraction from abstracts and full papers using searchbased structured prediction[J]. BMC bioinformatics, 2012,13(11):S5.

[40] 赵妍妍,王啸吟,秦兵, et al. 中文事件抽取中事件类别的自动识别[C]. 中国辽宁沈阳:第三届学生计算语言学研讨会,2006:240-245.

[41] 钟大悠,周德宇. 基于多类型特征的生物事件触发词识别[EB/OL].北京：中国科技论文在线. http://www.paper.edu.cn/html/releasepaper/2013/12/739/, 2013-12-24.

[42] 丁效,宋凡,秦兵. 音乐领域典型事件抽取方法研究[J].中文信息学报,2011,(02):15-20.

[43] Ji H, Grishman R. Refining event extraction through unsupervised Cross-document inference[C]. Proceedings of ACL-08. HLT Columbus, USA: HLT, 2008:54-262.

[44] Grishman R, Westbrook D, Meyers A. NYU's English ACE 2005 system description[C]. Proceedings of ACE 2005 Evaluation Workshop. Washington, US, 2005:05-19.

[45] Hardy H, Kanchakouskaya V, Strzalkowski T. Automatic event classification using Surface text features[C]. Proceedings of AAAI Workshop on Event Extraction and Synthesis. Boston, USA: American Association for Artificial Intelligence, 2006:31-27.

[46] 于江德, 肖新峰, 樊孝忠. 基于隐马尔可夫模型的中文文本事件信息抽取[J].微电子学与计算机,2007, 24(10):92-94.

[47] Doddington G R, Mitchell A, Przybocki M A, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation[C].LREC. 2004:837-840.

[48] 赵妍妍, 秦兵, 车万翔,等.中文事件抽取技术研究[J].中文信息学报,2008,22(1): 3-8 .

[49] Chen Z, Ji H. Language specific issue and feature exploration in Chinese event extraction[C].Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers.

Association for Computational Linguistics, 2009: 209-212.

[50] Chen Z, Ji H. Can one language bootstrap the other: a case study on event extraction[C].Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. Association for Computational Linguistics, 2009: 66-74.

[51] 李洁,高新波,焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报,2006,01:89-92.

[52] 孙荣,周文,刘宗田. 用规则抽取句子中事件信息[J].小型微型计算机系统, 2011, 11:2309-2314.

[53] 霍娜,吕国英. 基于规则匹配的灾难性追踪事件信息抽取的研究[J]. 电脑开发与应用, 2012, 06:7-13.

[54] 蒋德良. 基于规则匹配的突发事件结果信息抽取研究[J]. 计算机工程与设计, 2010, 14:3294-3297.

致谢

首先，我要感谢我的导师廖涛副教授。这篇硕士论文是在廖老师的悉心指导下完成的。在整个论文的选题、内容规划、研究路线上廖老师都给了细心的指导。在学习上，廖老师要求特别严格，大到论文的选题，小到文章中的每一句话和每一个标点符号，都以规范、严谨、简练、精确的科研作风和求真务实的科研精神对我严格要求，在生活上，廖老师宽厚待人。在此，谨向我的导师廖老师致以最衷心的感谢。

其次，我要感谢我亲爱的室友杜源和好友刘佳、石玲玲、赵洪宋等，感谢她们研究生三年来的陪伴，无论学习上还是生活上我们都互帮互助、共同进步。她们活泼开朗的性格时刻感染着我，给我带来了很多欢乐。感谢王凯、岳雨俭等实验室同学在学习和生活上给我的帮助和关心，自由融洽的实验室氛围也是我学习的极大动力。

最后，我要感谢我的父母和亲人，无论我做什么样的决定，你们的爱和鼓舞给了我坚持下去的勇气和信心。此时想起了吴汝伦的《百字铭》：且挨过三冬四夏，暂受些此痛苦，雪尽后再看梅花。再多的言语都无以为谢，谨以此文献给你们，祝你们平安顺遂，健康长寿！

作者简介及读研期间主要科研成果

1. 作者简介

轩小星，女，河南省漯河市，1987 年 10 月生。2012 年 7 月毕业于南阳师范学院通信工程专业；2012 年 9 月考入安徽理工大学计算机科学与工程学院攻读硕士学位。

主要研究方向：人工智能、数据挖掘

2. 读研期间发表的论文

[1] LIAO T, LIU Z, XUAN X. Research on Event Co-occurrence Network Structure Based Method for Chinese Text Representation[J]. Journal of Computational Information Systems, 2013, 9(14): 5535-5542.

[2] Liao T, Xuan X, Liu Z, et al. Important events extraction based on event co-occurrence network text representation method[C]. Progress in Informatics and Computing (PIC), 2014 International Conference on. IEEE, 2014: 37-41.

[3] 轩小星,廖涛. 中文事件触发词的自动抽取研究[J]. 计算机与数字工程,2015,03:457-461.