

北京机械工业学院

硕士学位论文

公安领域案件信息抽取系统设计与实现

姓名：乔春庚

申请学位级别：硕士

专业：计算机应用

指导教师：肖诗斌

20070201

## 摘要

近年来，随着公安刑侦专业信息及案件新闻报道等信息的急剧增长，以及犯罪分子作案手段的多样性和隐蔽性的增加，在如此多的资源信息中如何获得破案线索信息，如何提高破案效率，如何快速有效的找到相关案件信息，已经成为刑侦工作迫切解决的问题。

本文在对公安领域案件相关文本的特点进行分析的基础上，进行公安领域案件信息抽取的研究。主要包括：基于未标注语料的领域词汇自动抽取，从语料中找到和案件相关的词汇，为下面的研究提供了基础；规则与统计相结合的命名实体的识别与研究；模式自动获取研究；案件信息抽取系统的模型研究，在此基础上获得案件信息描述。

我们的研究是面向实际应用的，是在对文本进行分析和词性识别基础上的研究。由于我们掌握的资源有限，因此，在研究过程中，对每一项研究内容，我们都分析了现有资源，采取了一种最简单、最有效的方法。

本文对信息抽取进行了初步研究，实验结果还比较粗糙。在实际的应用中，领域词汇抽取的准确度、精确的命名实体识别、信息结构的表达方法等还都有待进一步的研究。我们所积累的一些经验和资源，也可以作为进一步研究的基础。

**关键词：**公安领域；信息抽取；命名实体识别；模式获取

## **ABSTRACT**

Lately, corresponding with the rapid growth of information of Police Criminal Investigations and news reports, with the increase of criminals involved's diversity and concealment. In so many informations, how to obtain information resources? How to improve the detectin efficiency? How quickly and effectively to find relevant information on cases? The criminal investigation has become a pressing issue.

This thesis focuses on extracting information regarding the cases-information Extraction, based on analyses of various tests in the field of public security-related cases. The study consists of the following tasks: terms automatic extraction based on no-tagging corpus, we find term from corpus in the field of public, the results can be used in the further; the recognition of named entities on statistics and rules; examining the means of automatic pattern acquisition for information; and the research of models of case-information system, so we can obtain the description of cases' event.

This research is customized to practical uses, it is primarily constructed on the basis of Part-Of-Speech (POS) tagging. Due to the limited resources at our disposal, therefore, in the course of the study, each a study, we analyzed the existing resources, taken one of the most simple, most effective method.

This thesis provides the foundation research for information extraction. The empirical results are rather rough at this point. Accuracy of the terms in specific field, precise application-oriented named entities recognition, and the expression method of information structurization etc., requires further more detailed research. What we acquired via the experiences and resources involved with this work can provide the fundamental basics for other related research projects.

**Key Words:** Police Field; Information Extraction; Named Entities Recognition;  
Pattern acquisition

## 学位论文版权使用授权书

本人完全了解北京机械工业学院关于收集、保存、使用学位论文的规定，同意如下各项内容：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以赢利为目的的前提下，学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：乔春庚

2007年3月20日

-----  
(注：非保密论文无需签字)

经指导教师同意，本学位论文属于保密，在        年解密后适用本授权书。

指导教师签名：

学位论文作者签名：

年    月    日

年    月    日

## 硕士学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

签名：乔春庚

2007年3月20日

## 第1章 绪论

### 1.1 选题背景

在信息技术高速发展的时代中,信息的获取、处理和应用已经成为了经济、军事、科学、文化等各个领域发展的关键。其中,信息的获取是这三个步骤的开端,在信息技术领域中具有尤其重要的地位。

随着计算机和互联网技术的迅猛发展,各领域的信息量呈指数增长。如何高效地获取有用信息,成为有效利用信息的关键。信息抽取(Information Extraction, 简称IE)技术<sup>[1][2]</sup>,是自然语言处理领域中一种新兴的技术。该技术通过抽取、过滤无关信息,使文本信息以用户关心的形式进行再组织,实现高效重组。将结构松散的自然语言信息,通过抽取转为结构严谨、语义明确的表现形式,利用计算机进行高效存储并加以利用。

近年来,随着公安刑侦专业信息及相关信息等资源信息的急剧增长,如果让侦查人员费时费力去查卷宗,对于一般的案件就勉为其难,还会增加破案成本。在如此多的资源信息中如何获得破案线索信息,如何提高破案效率,如何快速有效的找到相关案件信息,已经成为刑侦工作迫切解决的问题。

在案件侦破过程中,刑侦破案的重要手段就是把调查到的各种线索联系起来,查出犯罪分子作案规律,将几个相对独立的案件放到一起进行调查,就是串并案。这样,只要破一案,就能带一批案件,大大提高破案效率。为了给侦查员减轻很多负担,降低了破案成本,我们应用信息抽取技术,通过计算机自动对案件涉及到的人、事件、地点、时间等信息抽取出来,找出这些实体间的关系,并通过可视化的方式展现,以供查询和修改,从而能大大提高侦查员的办案效率。

### 1.2 研究背景

#### 1.2.1 信息抽取概述

信息抽取 (Information Extraction) 是指从一段文本中抽取指定的一类信息 (例如事件、事实)、并将其 (形成结构化的数据) 填入一个数据库中供用户查询使用的过程。

信息抽取系统的主要功能是从文本中抽取出特定的事实信息。比如, 从新闻报道中抽取出恐怖事件的详细情况: 时间、地点、作案者、受害者、袭击目标、使用的武器等; 从经济新闻中抽取出公司发布新产品的情况: 公司名、产品名、发布时间、产品性能等; 从病人的医疗记录中抽取出症状、诊断记录、检验结果、处方等等。通常, 被抽取出来的信息以结构化的形式描述, 可以直接存入数据库中, 供用户查询以及进一步分析利用。

### 1.2.2 信息抽取与信息检索

与信息抽取密切相关的一项研究是信息检索, 但信息抽取与信息检索存在差异, 主要表现在三个方面:

- 功能不同: 信息检索系统主要是从大量的文档集合中找到与用户需求相关的文档列表; 而信息抽取系统则旨在从文本中直接获得用户感兴趣的事实信息。
- 处理技术不同: 信息检索系统通常利用统计及关键词匹配等技术, 把文本看成词的集合, 不需要对文本进行深入分析理解; 而信息抽取往往要借助自然语言处理技术, 通过对文本中的句子以及篇章进行分析处理后才能完成。
- 适用领域不同: 由于采用的技术不同, 信息检索系统通常是领域无关的, 而信息抽取系统则是领域相关的, 只能抽取系统预先设定好的有限种类的事实信息。

另一方面, 信息检索与信息抽取又是互补的。为了处理海量文本, 信息抽取系统通常以信息检索系统 (如文本过滤) 的输出作为输入; 而信息抽取技术又可以用来提高信息检索系统的性能。二者的结合能够更好地服务于用户的信息处理需求。

一般来说, 信息抽取系统的处理对象是自然语言文本尤其是非结构化文本。但广义上讲, 除了文本以外, 信息抽取系统的处理对象还可以是语音、图像、视频等其他媒体类型的数据。在这里, 我们只讨论狭义上的信息抽取研究, 即

针对自然语言文本的信息抽取。

### 1.2.3 信息抽取研究现状

美国纽约大学开展的 Linguistic String 项目开始于 60 年代中期并一直延续到 80 年代。该项目的主要研究内容是建立一个大规模的英语计算语法，与之相关的应用是从医疗领域的 X 光报告和医院出院记录中抽取信息格式，这种信息格式实际上就是现在我们所说的模板 (Templates)。另一个相关的长期项目是由耶鲁大学 Roger Schank 及其同事在 20 世纪 70 年代开展的有关故事理解的研究。由他的学生 Gerald De Jong 设计实现的 FRUMP 系统是根据故事脚本理论建立的一个信息抽取系统。该系统从新闻报道中抽取信息，内容涉及地震、工人罢工等很多领域或场景。该系统采用了期望驱动 (top-down, 脚本) 与数据驱动 (bottom-up, 输入文本) 相结合的处理方法。这种方法被后来许多信息抽取系统采用。

从 20 世纪 80 年代末开始，信息抽取研究蓬勃开展起来，这主要得益于消息理解系列会议 (MUC, Message Understanding Conference)<sup>[6][6][7]</sup> 的召开。

从 1987 年开始到 1998 年，MUC 会议共举行了七届，它由美国国防高级研究计划委员会 (DARPA, the Defense Advanced Research Projects Agency) 资助。MUC 的显著特点并不是会议本身，而在于对信息抽取系统的评测。只有参加信息抽取系统评测的单位才被允许参加 MUC 会议。

MUC 系列会议对信息抽取这一研究方向的确立和发展起到了巨大的推动作用。MUC 定义的信息抽取任务的各种规范以及确立的评价体系已经成为信息抽取研究事实上的标准。

目前，正在推动信息抽取研究进一步发展的动力主要来自美国国家标准技术研究所 (NIST) 组织的自动内容抽取 (ACE, Automatic Content Extraction) 评测会议。

与 MUC 相比，目前的 ACE 评测不针对某个具体的领域或场景，采用基于漏报 (标准答案中有而系统输出中没有) 和误报 (标准答案中没有而系统输出中有) 为基础的一套评价体系，还对系统跨文档处理 (Cross-document processing) 能力进行评测。这一新的评测会议将把信息抽取技术研究引向新的高度。

### 1.2.4 信息抽取的一般过程

按照 MUC 的任务规定<sup>[4]</sup>，一个完整的信息提取过程包括如下 5 个阶段：

1. 命名实体 NE (Named Entities)：提取文本中相关的命名实体，包括人名、机构/公司名称的识别。

如：国家财政部/Org 部长 项怀诚/Person。

2. 实体关系 ER (Entity Relations)：提取命名实体之间的各种关系（事实）等，例如 Location\_of, Employee\_of, Product\_of 等关系。

如：Post\_of (部长, 项怀诚), employee\_of (国家财政部, 项怀诚)。

3. 脚本模板 ST (Scenario Template)：提取指定的事件，包括参与这些事件中的各个实体、属性或关系。

如：召开会议 (Time<...>, Spot<...>, Convener<...>, Topic<...>)

航天器发射事件（其涉及的运载工具、负载、时间和场地）。

4. 共指 Coreference (Identity descriptions)：代词、名词共指分析。

5. 模板合并 Template Merger：把相同的事件合并成为一个。

### 1.2.5 信息抽取的关键技术

信息抽取技术包括各种对象（实体、实体间的关系、事件）的识别、把识别出来的对象有机关联起来两个方面。这两个方面包含了信息的表达、语言的分析（词法、句法、语义）、知识的获取与推理等技术。显然，不同的应用对信息抽取的要求不同，不同的事件主题会有不同的关注点，不同的信息描述侧面会有不同的信息模式。在信息模式中，融合了实体及实体间关系等信息，信息抽取技术的研究集中在识别对象以及对象间的关联模式的获取上。所定义的识别对象不同，识别的难易程度也不同。下面介绍几种关键技术：

#### 1. 句法分析技术

通过句法分析得到输入的某种结构表示，如完整的分析树或分析树片段集合，是计算机理解自然语言的基础。在信息抽取领域一个比较明显的趋势是越来越多的系统采用部分分析技术，这主要是由于以下原因造成的。

首先是信息抽取任务自身的特殊性，即需要抽取的信息通常只是某一领域中数量有限的事件或关系。这样，文本中可能只有小部分与抽取任务有关。并且，对每一个句子，并不需要得到它的完整的结构表示，只要识别出部分片段间的某些特定关系就行了，得到的只是完整分析树的部分子图。

其次是部分分析技术在 MUC 系列评测中的成功。

SRI 公司在参加 MUC-4 评测的 FASTUS 系统中开始采用层级的有限状态自动机 (Cascaded Finite-State Automata) 分析方法。该方法使 FASTUS 系统具有概念简单、运行速度快、开发周期短等优点, 在多次 MUC 评测中都居于领先地位。

### 2. 篇章分析与推理

一般说来, 用户关心的事件和关系往往散布于文本的不同位置, 其中涉及到的实体通常可以有多种不同的表达方式, 并且还有许多事实信息隐含于文本之中。为了准确而没有遗漏地从文本中抽取相关信息, 信息抽取系统必须能够识别文本中的共指现象, 进行必要的推理, 以合并描述同一事件或实体的信息片段。因此, 篇章分析、推理能力对信息抽取系统来说是必不可少的。

初看起来, 信息抽取中的篇章分析比故事理解中的篇章分析要简单得多。因为在信息抽取中只需要记录某些类型的实体和事件。但是, 大多数信息抽取系统只识别和保存与需求相关的文本片段, 从中抽取零碎的信息。在这个过程中很可能把用以区分不同事件、不同实体的关键信息给遗漏了。在这种情况下要完成篇章分析是相当困难的。

除此之外, 目前尚缺乏有效的篇章分析理论和方法可以借鉴。现有篇章分析理论大多是面向人、面向口语的, 需要借助大量的常识, 它们设想的目标文本也比真实文本要规范, 并且理论本身也没有在大规模语料上进行过测试。

### 3. 知识获取技术

不同的领域、不同的主题对信息提取的内容是不一样的, 支持提取的知识也有差别。每一个具体的信息提取任务都期望有相关领域的知识资源, 包括词典、模式集合及相应的匹配规则。于是一个 IE 系统初始建立、向其他领域的可移植性成为 IE 中的一个热点研究问题。从根本上讲是 IE 系统如何获取知识的问题。在不同的信息抽取系统中知识库的结构和内容是不同的, 一般都由通用知识和领域知识两部分组成。通常有词典或概念常识库: 存放通用词汇以及领域词汇的静态属性信息, 并表达概念间的关系; 抽取模式库, 每个模式可以有附加的(语义)操作; 有关篇章分析和推理的规则库、模板填充规则库等。

知识库的建立有三种方式: 手工编制、半自动方式、自动获得。

手工编制相对简单一些, 人工工作仍然是主体, 只是为移植者提供了一些图形化的辅助工具, 以方便和加快领域知识获取过程。后两种采用有指导的、

无指导的或间接指导的机器学习技术从文本语料中自动或半自动获取领域知识，人工干预程度较低。

### 4. 命名实体的识别

命名实体 (Named Entity) 是文本中重要的信息元素。狭义地讲，命名实体是指现实世界中的具体的或抽象的实体，如人、组织、公司、地点等，通常用唯一的标志符 (专有名称) 表示，如人名、组织名、地名等。广义地讲，命名实体还可以包含时间、数字表达式等 (本文中用“实体名”或者“实体词”来代表狭义上的命名实体，包括人名、地名、组织名；而用“命名实体”来代表广义上的命名实体)。实际研究中，命名实体的确切含义，需要根据具体应用来确定，比如，在具体应用中，可能需要把住址、网址、电子信箱地址、电话号码、舰船编号、会议名称等作为命名实体。有些词属于专门领域中的实体名，例如药名、医学条件、轮船名字、以及参考目录等，也应该把其归入考虑范围内。

命名实体的识别按照方法的不同，大体可以分为三类：基于规则的方法；基于统计的方法；统计与规则相结合的方法。其中后两种方法目前占主导地位。

#### 1. 基于规则的方法

基于人工组织规则的方法的典型代表是纽约大学 (New York University) 的 Proteus 系统、IsoQuest Inc. 的 NetOwl 系统、曼彻斯特科技大学 (University of Manchester Institute of Science and Technology) 的 FACILE 系统。这三个系统都参加了 MUC-7 的测试，并且都采用了模式匹配的方法，不同之处是这些系统采用的模式 (或规则) 外在表达方式不同。

我们可以发现，对于这类采用人工组织规则的系统，主要存在以下缺点：

- 人工组织规则的代价非常昂贵，并主要依赖于有经验的计算语言学家。
- 当把此系统移植到不同领域时，需要大量的人工修改工作。
- 当把此系统移植到新的语种时，这些规则需要重新书写和组织。
- 语言学家书写规则的经验 and 所花费人力劳动的大小对性能的影响很大。

#### 2. 基于统计的方法

统计方法主要是针对命名实体语料库来训练某个字作为命名实体组成部分的概率值，并用它们来计算某个候选字段作为命名实体的概率，其中概率值大于一定阈值的字段为识别出的命名实体。

基于统计的方法有助于克服基于规则的方法中的知识获取瓶颈，又因为使

用工具的自动化程度较基于规则的方法要高,使这种方法越来越受到人们的关注。用于命名实体识别的统计方法主要有隐马尔可夫模型(HMM)<sup>[8]</sup>、最大熵模型(ME)<sup>[9]</sup>、决策树(decision tree)<sup>[10]</sup>、基于转换的学习方法(Transform-based Learning)<sup>[11]</sup>、推进方法(Boosting)<sup>[12]</sup>、自展的方法(Bootstrapping)<sup>[13]</sup>、基于分类的语言模型(Class-based Language Model)<sup>[14]</sup>、支持向量机器的方法(Support Vector Machine)<sup>[15]</sup>。

现在的系统使用单纯统计方法的很少,或多或少都应用了一些规则。

### 3. 统计与规则相结合的方法

就命名实体识别方法而言,基于规则的方法主观性强,可移植性不好,而且歧义是语言的一个固有特点,是基于规则的方法必须面对的问题,此外规则很难覆盖所有的语言现象,因此在做语言处理时希望机器具有学习能力。另一方面,人类语言的运用并不纯粹是一个随机过程,单单使用基于统计的方法将使状态搜索空间非常庞大,借助规则知识及早剪枝是一个比较有效的方法。所以目前几乎没有单纯使用统计模型而不使用规则的命名实体识别系统,在很多情况下是使用混合方法,即统计模型+规则进行识别,这是一个不可避免的趋势。

规则与统计相结合的办法,可以通过概率计算减少规则方法的复杂性与盲目性,而且可以降低统计方法对语料库规模的要求。目前的研究基本上都是采取规则与统计相结合的方法,不同之处仅仅在于规则与统计的侧重不同而已。

### 5. 实体间关系的识别

实体间关系的识别是在MUC-7上提出的。不同的主题表现出来的实体间的关系是不相同的。比如公司与其负责人关系、地理位置关系、雇佣关系等。进行信息提取时,通常是事先指定要抽取的关系(ACE规定了7种实体间的关系)。

目前实体间关系研究范围只局限于二元关系,识别技术属于模式获取范畴。具有代表性的方法包括:

1. 基于种子实例的自适应的(bootstrapping)获取方法<sup>[21][26]</sup>
2. 基于分类的获取方法<sup>[22]</sup>

将关系的抽取转化为一个分类问题,对句子中所含的实体对,使用一个分类器决定哪些是要提取的关系。分类器的选择可以有多种方式,如SVM、Winnow、核方法。所有这些方法都是有指导的学习方法,需要有大规模的标注了实体间关系语料支持。

3. 自动获取关系模式方法<sup>[23][24]</sup>

与原有方法最大的区别是关系的确定不再基于已有的模板，而是自动的从大规模语料库中自动获取关系和关系的类型，其核心技术是对实体对进行聚类。基本过程如下：

1. 标注所有的命名实体
2. 获取任意两个命名实体同现的上下文
3. 计算这些上下文的相似性
4. 对命名实体对进行聚类
5. 对聚类结果中的每一类选择关系名，形成关系

经过在ACE数据集上的测试，在PER-GEP和COM-COM的关系识别上，该方法获得了较好的效果（F值分别达到了80%，75%）。主要问题是低频的实体关系不容易发现。

#### 6. 指代的消解

指代的消解是信息提取中较为困难的一个任务，由于受到资源的限制，现在的指代的消解多集中在人称代词的消解上。用于解决指代的方法包括：

1. 规则法：人工总结出指代消解规则，按照这些规则来实现指代消解。
2. 简单的同现方法：通过角色同现消解英语中的It代词。
3. 统计模型<sup>[19]</sup>：利用统计特征，包括距离、是否人称代词、字符串匹配否、是否有限名词、是否指示代词、单复数是否对应、语义类是否一致、是否别名、是否同位语等，构造统计模型。
4. 聚类法：选定一些对实现指代消解能产生作用的特征，对一些对象进行聚类，聚为一类的对象就认为是同指对象。
5. 分类法：选定一些对实现指代消解能产生作用的特征，选定一些训练语料，采用某种机器学习方法训练出一个指代消解分类器，并用该分类器完成新来文本中的指代消解。
6. 决策树的指代消解法<sup>[20]</sup>：对训练集采用CS算法学习出决策树，对新的指代进行判断。

这些方法对指代处理的效果差别不大，识别正确率在50%-70%的范围之内。

#### 7. 事件识别

事件的识别技术是IE中的核心技术。其研究包括：

- 表达事件的模式获取<sup>[26]</sup>
- 模板中关于槽值填充的规则获取<sup>[27]</sup>

## ■ 事件信息的表示方法研究。

从信息提取的流程可知,信息的表达形式通常是事先规定好的模式。模式的形式可以是:词袋(bag-of-word)、词项(word item)、词汇、短语或更复杂的结构的正则表达形式、谓词论元结构(predicate-argument structure)、依存链(dependency chain)形式、子树(Subtree)模型。

目前,大多数IE系统的模板形式都采用谓词论元结构或依存链的形式表达事件及其关系。这两种表达是建立在句法分析的基础上的。谓词论元结构表达的是一种句法关系,比如主语-动词(subject-verb)、宾语-动词(object-verb)形式。这种表达对于信息提取来说,存在两个问题:一是覆盖性问题(仅是分句范围内),另一是嵌套的实体只能作为一个整体的论元表达;依存链的表达形式是使用依存树中始于谓词节点的路径表达,这样可以打破分句的界限和嵌套实体的问题。在模式的自动获取中召回率比谓词论元结构的提高了5个百分点。但这种方法对于上下文环境的依赖性很大,精确率很难提高。在ACE的任务描述中,对事件的抽取内容可以认为是事件的论元结构。

### 8. 领域词汇的抽取

一篇文章的关键词,是指能够反映文章主要内容以及文章所涉及的领域的词语或短语。一个领域的关键词,指该领域的术语,专业名词,放宽一点,也可以是经常应用于该领域的词语或短语。比如一提起“借贷”,就让人们想起金融领域。领域关键词提取,指使用计算机在领域语料库中自动提取关键词,必要时可以经过人工筛选。

最早的关于术语抽取的研究是 H.P.Luhn 所作的工作,到目前为止,已经有很多学者参与术语抽取工作的研究,并且取得了一定的成果。Salton, Yang & Yu 通过简单的加权两个相邻字的方法来抽取术语;Damerau 使用互信息来测量两个词之间的联合强度,取得了一些效果;Dunning 和 Cohen 则创建和使用了 log-likelihood 参数,避免了一些低频词的遗漏,从而较有效的弥补了互信息的不足;Patrick & Dekang 将互信息和 log-likelihood 两个参数相结合进行术语抽取,取得了一定的成功。在中文领域词汇抽取方面,东北大学的陈文亮等人提出了基于 Bootstrapping 的领域词汇自动抽取方法<sup>[28]</sup>;北京大学的惠志方也提出了信息科学技术领域术语自动识别策略<sup>[30]</sup>。

## 1.2.6 信息抽取系统的通用体系结构

Hobbs 曾提出一个信息抽取系统的通用体系结构<sup>[3]</sup>，他将信息抽取系统抽象为“级联的转换器或模块集合，利用手工编制或自动获得的规则在每一步过滤掉不相关的信息，增加新的结构信息”。

Hobbs 认为典型的信息抽取系统应当由依次相连的十个模块组成：

1. 文本分块：将输入文本分割为不同的部分-文本块。
2. 预处理：将得到的文本块转换为句子序列，每个句子由词汇项（词或特定类型短语）及相关的属性（如词类）组成。
3. 过滤：过滤掉不相关的句子。
4. 预分析：在词汇项（Lexical Items）序列中识别确定的结构，如名词短语、动词短语、并列结构等。
5. 分析：通过分析小型结构和词汇项的序列建立描述句子结构的完整分析树或分析树片段集合。
6. 片段组合：如果上一步没有得到完整的分析树，则需要将分析树片段集合或逻辑形式片段组合成整句的一棵分析树或其他逻辑表示形式。
7. 语义解释：从分析树或分析树片段集合生成语义结构、意义表示或其他逻辑形式。
8. 词汇消歧：消解上一模块中存在的歧义得到唯一的语义结构表示。
9. 指代消解或篇章处理：通过确定同一实体在文本不同部分中的不同描述将当前句的语义结构表示合并到先前的处理结果中。
10. 模板生成：由文本的语义结构表示生成最终的模板。

当然，并不是所有的信息抽取系统都明确包含所有这些模块，并且也未必完全遵循以上的处理顺序，比如 6、7 两个模块执行顺序可能就相反。但一个信息抽取系统应当包含以上模块中描述的功能。

### 1.2.7 信息抽取的发展趋势

信息抽取经过二十多年尤其是最近十多年的发展，已经成为自然语言处理领域一个重要的分支，其独特的发展轨迹——通过系统化、大规模地定量评测推动研究向前发展，以及某些成功启示，如部分分析技术的有效性、快速 NLP（Natural Language Processing）系统开发的必要性、知识工程研究以及软件工程技术的重要性等等，都极大地推动了自然语言处理研究的发展，迫使 NLP 研究

人员面向实际的应用重新考虑他们的研究重点，开始重视解决以前曾被忽视的一些深层问题，如语义特征标注、共指消解、篇章分析等等。

目前，影响信息抽取技术广泛应用的两个最主要的因素是：系统性能和系统可移植能力。因此，今后信息抽取研究将紧紧围绕如何克服和解决这两个问题展开，重点解决知识获取、篇章分析、高效句法分析等问题，不断提高信息抽取系统的性能、增强其可移植能力。

未来的信息抽取系统将是动态（Dynamic）的、开放域（Open Domain）的，前景光明。

### 1.3 论文结构

本文共分六章：

第一章绪论，介绍信息抽取的基本概念、研究内容、发展现状以及相关技术等，为我们下面的研究奠定理论基础。

第二章基于未标注语料的领域词汇自动抽取，介绍了领域词汇抽取系统的设计与实现，并以公安领域文本为语料进行了实验。

第三章命名实体识别，介绍公安领域案件信息相关实体的抽取方法。

第四章公安领域案件信息的模式获取，主要介绍公安领域案件信息二元实体关系抽取的方法，并以实例说明其实际应用价值。

第五章公安领域案件信息抽取系统，介绍本研究所开发的实验系统，包括系统结构及工作流程。

第六章全文总结，概括本文的观点和结论，介绍了本文的研究背景、研究内容、研究方法及本文特点。最后提出了今后工作的展望。

## 第2章 基于未标注语料的领域词汇自动抽取

### 2.1 引言

领域词汇集中体现和承载了一个学科领域的核心知识，领域词汇的抽取是很多自然语言处理应用的一个起始点，对于信息检索、多语检索、文档分类、词典编辑以及双语对齐等语言信息处理研究具有重要的理论和现实意义。所以，在本论文中，我们把其作为信息抽取系统的一部分重点研究。

在论文中，我们提出了一种基于未标注语料的领域词汇自动抽取方法。该方法主要分三步进行：首先，在分词的基础上，计算语料库中词语之间结合的“紧密程度”，根据阈值过滤，得到多词词串，加入分词词典；然后，利用新的分词词典对语料进行分词分句，统计词语的特征；最后，我们利用 SVM 机器学习方法，根据词语的特征对词语进行分类得到领域词汇。

### 2.2 基本概念

1. 领域词汇  $t$ ：一个学科领域中使用、表示该学科领域内概念、特征或关系的词语。领域词汇可以是词，也可以是短语，可以只在一个学科领域中存在，也可并存于多个学科领域中。如：犯罪嫌疑人、法庭、作案工具、案发时间等词就是公安领域词。领域词集  $T$  是领域词的集合，即  $t \in T$ 。
2. 一般词语  $c$ ：一个学科领域中除了领域词汇之外的词语都叫做一般词语。所有学科领域中一般词语的并集构成了一般词语的全集  $C$ 。一般词语的全集加上所有学科领域中的领域词汇构成语言交际的词语的全集  $A$ ，即  $A = T \cup C$ 。
3. 领域种子词  $d$ ：在自动抽取之前，我们需要先指定一些领域词，我们称之为领域种子词。领域种子词集  $D$  是领域种子词的集合，即  $d \in D$ ，从上面的定义我们可以看出  $d \in T$ ， $D \subset T$ 。

## 2.3 领域词汇抽取模型

基于未标注语料的领域词汇自动抽取主要包括三个阶段(见图 2.1): 预处理阶段、特征统计阶段和 SVM 分类阶段。本模型首先对分词词典进行了改进, 计算词语间的结合紧密度, 获得结合紧密的词串加入分词词典; 然后从领域种子词集出发, 从未标注的语料中统计词语的特征, 得到具有特征的候选词集; 最后对候选词集进行 SVM 分类, 得到领域词集。整个过程是一个循环的过程, 得到的领域词汇重新加入种子词集, 进行特征统计, 再次进行分类, 直到不出现新的领域词汇为止。

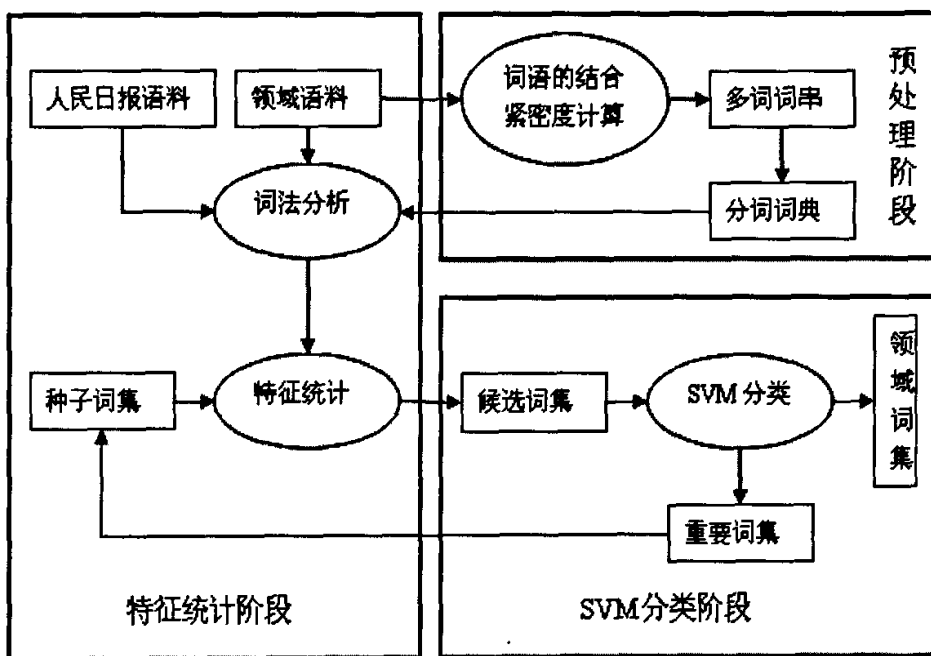


图 2.1 领域词汇抽取框图

### 2.3.1 预处理阶段

预处理阶段主要是在分词的基础上, 计算两个相邻词语的结合紧密程度, 得到两词词串, 加入分词词典。这是一个循环的过程, 利用新的分词词典重新对语料进行分词, 计算结合紧密度, 直到不再有词语结合为止。最后, 我们得到的分词词典包含领域语料中经常出现的词串。

在我们的算法中,我们使用了两个参数来衡量两个相邻词语之间的结合紧密程度,这两个参数是互信息(mi)和 log-likelihood(log L),他们的定义如下。

$$mi(x, y) = \log\left(\frac{p(x, y)}{p(x) \cdot p(y)}\right)$$

$$(p(x, y) = \frac{C(x, y)}{C(*)}, p(x(y)) = \frac{C(x(y))}{C(*)}) \quad (2.1)$$

在公式 2.1 中,  $x$ 、 $y$  为词语项,  $C(x)$  ( $C(x, y)$ ) 为  $x$  ( $(x, y)$ ) 在该语料中出现的频数。 $mi(x, y)$  在  $x$ 、 $y$  共现次数较多时, 值就比较大; 反之, 在低频时就会恶化。为解决此问题, 我们使用第二种方法 log-likelihood, 对于低频情况, 这种方法更健壮(公式 2.2)。

$$\log L(x, y) = \log L\left(\frac{k1}{n1}, k1, n1\right) + \log L\left(\frac{k2}{n2}, k2, n2\right)$$

$$- \log L\left(\frac{k1+k2}{n1+n2}, k1, n1\right) - \log L\left(\frac{k1+k2}{n1+n2}, k2, n2\right) \quad (2.2)$$

这里,  $\log L(p, k, n) = k \log p + (n - k) \log(1 - p)$ ,  $k1 = C(x, y)$

$n1 = C(x, *) \approx C(x)$ ,  $n2 = C(-x, *) = C(*, *) - C(x, *) \approx C(*) - C(x)$

$k2 = C(-x, y) = C(*, y) - C(x, y) \approx C(y) - C(x, y)$

这种方法和互信息方法一样, 在  $x$ 、 $y$  共现次数较多时, 值就比较大; 对两个出现频率都较大的词时, 但共现次数较少时, log-likelihood 值也比较大。为解决两种方法的弊端, Patrick & Dekang<sup>[31]</sup>将两种方法结合用于英文术语抽取取得了不错的结果。我们将其用于中文相邻两词结合紧密度计算, 实验表明, 该方法取得了不错的效果。计算结合紧密程度见公式 2.3, 其中 *MinMutInfo* 为设定的互信息最小阈值。

$$S(x, y) = \begin{cases} \log L(x, y), mi(x, y) \geq MinMutInfo \\ 0 \end{cases} \quad (2.3)$$

具体计算结合紧密度算法如下:

1. 统计领域语料库, 就是要在分词的基础上统计语料中单词词频  $C(x)$  和相邻两词的共现频率  $C(x, y)$ 。

2. 然后计算相应的统计数据  $mi(x, y)$ 、 $\log L(x, y)$ 、 $S(x, y)$ ，然后选取符合条件的双词项  $(x, y)$ ，作为候选词项。
3. 将候选词项加入分词词典，对语料重新分词，转向第1步继续执行，直到不再出现新的双词结合为止。

这样，语料中的结合紧密度大的相邻词语就一直结合，直到没有新的词语出现为止（见下图 2.2）。

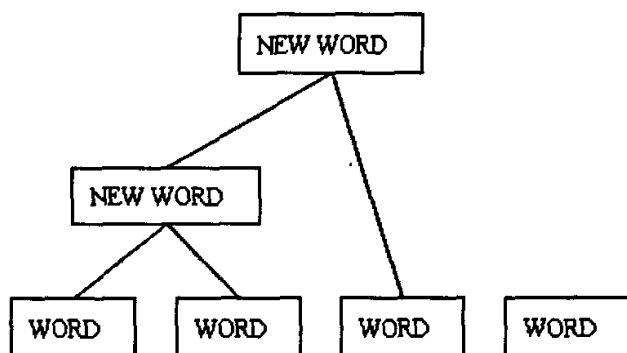


图 2.2 相邻词语结合紧密度

我们将所有符合条件的新词加入分词词典，为下一步特征统计阶段做准备。

### 2.3.2 特征统计阶段

我们总结了领域词汇的六个特征，对语料经过分词词典进行简单的词法分析，统计每个词汇的这六个特征，将结果存入候选词集。在本文中，统计单元是一个句子。首先，我们先定义几个参数：

频数  $C_w$ ：表示在语料中包含词  $w$  的句子数。注意： $w$  无论在句子中出现多少次，都只记为一次。

频数  $C_X$ ：表示在整个语料中包含词集  $X$  中任意元素  $x$  的句子数。注意： $X$  中元素  $x$  无论在句子中出现多少次，都只记为一次。

共现频数  $C_{w,X}$ ：表示  $w$  与  $X$  中任意元素  $x$  在同一句子中共现的句子数。注意：在同一句子中，无论同时包含多少个  $w$  和  $x$ ，都只记为一次。

下面我们就具体阐述这六个特征：

#### 1. IDF 值

我们对人民日报语料进行统计，因为人民日报语料中词汇的出现次数在一

定程度上反映了该词汇出现的普遍程度,所以我们计算每一个词汇的IDF值,以此来表征该词汇的领域区分程度。IDF值称为反文档频率,一个词在越多的文档中出现它的IDF就越小,反之就越大,公式为:  $\log(N(f_i)/N)$ 。其中,  $N(f_i)$  是词语  $f_i$  的训练文本数,  $N$  是总训练文本数。IDF值越大,说明包含该词的文本数越多,所以该词的领域区分能力越低,该词越普通;反之,该词是领域词汇的可能性越大。

## 2. 频数 $C_x$

我们对领域语料进行统计,计算每一个词的频数。频数越大,包含该词的句子数越多,该词是领域词汇或普通词汇的可能性都有,但是该词 IDF 值越小的话,说明该词是领域词汇的可能性越大。所以我们把词频同样作为一个特征。

## 3. 与种子词的共现频数 $C_{w,X}$

我们对领域语料进行统计,计算每一个词  $w$  和种子词集  $X$  中任意元素  $x$  共现的句子数。共现频数越大,该词是领域词汇的可能性就越大,但是也不排除该词是普通词汇的可能。

## 4. 支持度 $R$

支持度公式为:  $R = C_{w,X} / C_w$ 。支持度越大说明词  $w$  和种子词集  $X$  中的元素基本上都在一个句子中出现,单独出现次数较少。

## 5. M 评价<sup>[32]</sup>

评价公式如下:  $M_w = \log_2 C_{w,X} \times (C_{w,X} / C_w)$ 。其中,  $w$  表示词,  $M_w$  值越大,该词是领域词汇的可能性越大。

## 6. T 评价<sup>[33]</sup>

评价公式如下:  $T_w = \frac{P(w, X) - P(w)P(X)}{\sqrt{\frac{P(w, X)}{N}}}$ 。其中,  $w$  表示词,  $X$

表示种子词集,  $P(w, X)$  表示同现概率,  $P(w, X) = C_{w,X} / N$  (使用极大似然估计来计算),  $P(w)$  表示  $w$  出现的概率,  $P(X)$  表示  $X$  出现的概率,  $N$  表示句子总数。  $T_w$  值越大,该词是领域词汇的可能性越大。

我们统计了词汇的以上六个特征,将结果存入候选词汇,为下一步 SVM 分类阶段做准备。

### 2.3.3 SVM 分类阶段

一个词汇，或者是某一领域的词汇，或者是通用词汇。因此，我们把领域词汇的抽取过程看成一个二元分类问题。如图 2.1 所示，我们使用 SVM 机器学习方法对候选词集中的词汇进行分类，将分类结果中的领域词汇加入种子词集，然后返回第二阶段继续执行，直到不出现新的领域词汇为止。

#### 1. SVM 基本原理

支持向量机 (SVM) <sup>[94]</sup> 从本质上讲是一种前向神经网络，根据结构风险最小化准则，在使训练样本分类误差极小化的前提下，尽量提高分类器的泛化推广能力。从实施的角度，训练支持向量机的核心思想等价于求解一个线性约束的二次规划问题，从而构造一个超平面作为决策平面，使得特征空间中两类模式之间的距离最大，而且它能保证得到的解为全局最优解。

##### ■ 线性可分情况

设  $x_1, x_2, \dots, x_l$ ，其中  $x_i \in R^d, i=1, \dots, l$ ，是  $l$  个  $d$  维训练样本。每个样本对应的标记为  $y_1, y_2, \dots, y_l$ ，其中  $y_i \in [1, -1], i=1, \dots, l$ ，表明该向量属于两类中的哪一类。

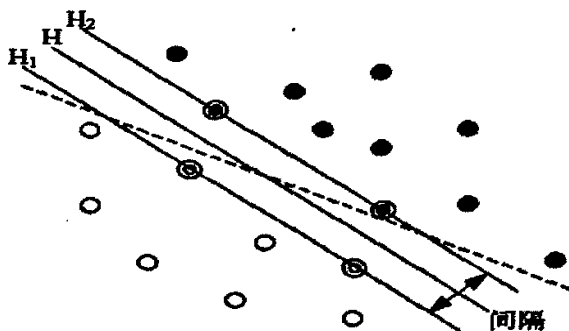


图 2.3 线性可分情况下的最优分界面

若超平面能将训练样本分开，则：

$$\left. \begin{aligned} w \cdot x_i + b &> 0, \text{ 若 } y_i = 1 \\ w \cdot x_i + b &< 0, \text{ 若 } y_i = -1 \end{aligned} \right\} \quad (2.4)$$

适当调整  $w$  和  $b$ ，可以将公式 2.4 改写为

$$\left. \begin{aligned} w \cdot x_i + b &\geq 1, \text{若 } y_i = 1 \\ w \cdot x_i + b &\leq -1, \text{若 } y_i = -1 \end{aligned} \right\} \quad (2.5)$$

或者

$$y_i(w \cdot x_i + b) \geq 1, \forall i \in \{1, \dots, l\} \quad (2.6)$$

根据统计学理论, 最优分界面不仅能将两类样本正确分开, 而且使得分类间隔最大, 如图 2.3 所示, 实心点和空心点代表两类样本,  $H$  为分界面,  $H1$ 、 $H2$  分别为过各类中离分界面最近的样本且平行于分界面的面, 他们之间的距离叫分类间隔, 虽然途中虚线也能将样本分开, 但是他的分类间隔比  $H$  小。

分界面  $w \cdot x + b$  的分类间隔为:

$$d(w, b) = \min_{\{x_i | y_i = 1\}} \frac{w \cdot x_i + b}{|w|} - \max_{\{x_i | y_i = -1\}} \frac{w \cdot x_i + b}{|w|} \quad (2.7)$$

由公式 2.5 可得:

$$d(w, b) = \frac{1}{|w|} - \frac{-1}{|w|} = \frac{2}{|w|} \quad (2.8)$$

所以最大化分类间隔  $d(w, b)$  的问题, 就转化为在约束条件 (2.6) 下最小化  $\frac{|w|^2}{2}$ , 由拉格朗日乘法, 问题等价于在约束条件

$$\left. \begin{aligned} \alpha_i &\geq 0 \\ \sum_i \alpha_i y_i &= 0 \end{aligned} \right\} \quad (2.9)$$

下最小化

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i, j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \quad (2.10)$$

每个拉格朗日乘数  $\alpha_i$  对应一个训练样本  $x_i$ , 对应的  $\alpha_i > 0$  的样本就被称为“支持向量”。最后得到的分类函数为:

$$f(x) = \text{sgn}(w \cdot x + b) = \text{sgn}\left(\sum_{i=1}^{Ns} \alpha_i y_i x_i \cdot x + b\right) \quad (2.11)$$

其中  $Ns$  是支持向量的个数。

### ■ 线性不可分情况

如果训练样本线性不可分,那么上一节的优化问题将变得无解。为此,可以放宽条件 2.5,引入  $\xi_i \geq 0, i = 1, \dots, l$  得

$$\left. \begin{aligned} w \cdot x_i + b &\geq 1 - \xi_i, \text{若 } y_i = 1 \\ w \cdot x_i + b &\leq \xi_i - 1, \text{若 } y_i = -1 \end{aligned} \right\} \quad (2.12)$$

如果  $x_i$  被错分,则  $\xi_i > 1$ , 因此总的错分类数小于  $\sum_i \xi_i$ 。在目标函数中加入一项对错分类进行惩罚,折中考虑最大分类间隔和最少错分样本,即改求  $\frac{|w|^2}{2} + C(\sum_i \xi_i)$  最小,就得到了线性不可分情况下的支持向量机。其中  $C > 0$  是控制惩罚程度的常数。由拉格朗日乘法,问题等价于在约束条件

$$\left. \begin{aligned} 0 &\leq \alpha_i \leq C \\ \sum_i \alpha_i y_i &= 0 \end{aligned} \right\} \quad (2.13)$$

下最小化

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \quad (2.14)$$

### ■ 非线性支持向量机

至此,对支持向量机的讨论都仅限于线性分界面的情况。对于非线性划分问题,可以通过一个非线性变换  $\Phi: R^n \mapsto H$  将它转化为某个高维空间  $H$  中的线性划分问题。一般来说,这种非线性变换的形式可能非常复杂,难于实现。但是注意到在上面的问题中,不论是优化的目标函数(2.10)还是分类函数(2.11)都只涉及到向量的点积运算,即  $\Phi(x_i) \cdot \Phi(x_j)$  的形式。如果存在一个“核函数”  $K$ , 满足

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \quad (2.15)$$

那么就能用原空间中的函数来实现变换空间中的点积,从而绕开映射  $\Phi$  的具体形式。

根据泛函分析中的有关理论,只要核函数  $K(x, y)$  满足 Mercer 条件,它就对应于某一变换空间中的点积,也就是说,存在映射  $\Phi(x)$ , 使得 (2.15) 成立。常见的满足 Mercer 条件的核函数有多项式核函数  $K(x, y) = (x \cdot y + 1)^p$ , 高斯径向基函数  $K(x, y) = e^{-|x - y|^2 / 2\sigma^2}$ , 以及 sigmoid 函数

$K(x, y) = \tanh(kx \cdot y - \sigma)$ 。此时的分类函数为：

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N_s} \alpha_i y_i \Phi(x_i) \cdot \Phi(x_j) + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{N_s} \alpha_i y_i K(x_i, x_j) + b\right) \quad (2.16)$$

## 2. 领域词汇 SVM 分类

我们使用 SVM light 开源工具对语料进行训练和分类。我们选取种子词集中的部分词汇作为训练的正例，选取通用词表中的部分词汇作为训练的反例。然后，我们对候选词集进行分类。这是一个循环的过程，我们将分类结果中的领域词汇加入种子词集，然后返回第二阶段继续向下执行，直到不出现新的领域词汇为止。

## 2.4 实验结果与分析

本实验对犯罪案件文本进行领域词汇抽取，领域语料来自互联网上关于犯罪案件的新闻报道，共计 33,796 篇。同时，在进行特征统计时，我们使用了人民日报语料，其由 31,174 篇文本构成。

我们使用上面的方法，对数据进行测试，我们共得到犯罪案件中的领域词汇 8,750 个（部分领域词汇实例见附录 B），全面准确率达到 66.21%，由于领域词汇的定义和范围不好确定，所以召回率还没有办法计算。

在我们的下面的研究中，我们基本用到了下面三个评测公式：

### 1. 准确率

系统准确率 = 系统识别出的正确的个数 / 系统识别出的总个数

### 2. 召回率

系统召回率 = 系统识别出的正确的个数 / 本来应该有的个数

### 3. F 值（其中，R 是召回率，P 是准确率）

$$F = \frac{(\beta^2 + 1)RP}{\beta^2 R + P}$$

在所抽取的 8,150 个词语中，我们做了如下的统计：

### 1. 对 SVM 分类后词语做了前 K 个术语的准确率分析

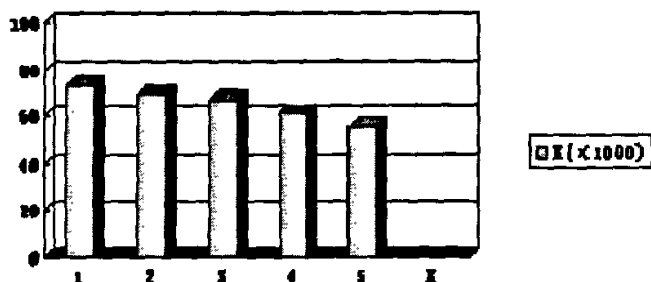


图 2.1 犯罪案件中的领域词汇前 K 个的准确率分析

从图 2.1 中可以看出,在抽取前 3000 个词时,我们的准确率达到 70%,而前 1000 个词的准确率更是达到了 73.3%,考虑到我们的数据是在开放性测试中得到的,这就证明了我们的方法是比较成功的。

## 2. 对 SVM 分类后词语做词串长度的准确率分析

表 2.1 犯罪案件中的领域词汇词串长度分析

词串长度 (字)	2	3	4	5	6	Average
准确率	71.42	50.18	70.11	55.75	59.37	66.21
(%)	(2,763)	(1,279)	(3,426)	(342)	(320)	

从表 2.1 所做的统计,我们可以看出,在犯罪案件的领域词汇中,二字词和四字词出现的频率是最高的,其识别的准确率也比较高,说明我们的算法是很有效的。其中,三字词出现的频率也比较高,其准确率相对较低,对系统的准确率影响较大,分析其原因在,在计算词语结合紧密度的时候,和参数选择有很大关系,把许多不是词语的相连两个词结合了(如“我们说、该案件、使用了”等词);另外一方面,在 SVM 分类时,关于这些词串的反例样本不够,使得分类时不能很好的区分。在以后的研究中,我们将对这两方面做比较深入的分析。

## 2.5 本章小结

在我们的领域词汇自动抽取方法中,我们首先对语料进行了统计分析,计算了词语间的结合紧密度,把结合紧密度较高的词串加入分词词典作为一个词处理,这样就减轻了领域词汇抽取中系统的复杂度;然后,我们统计了词串

的特征,利用 SVM 分类的方法对词串进行分类,从而把领域词汇与普通词汇分开。实验结果表明,我们的方法是比较成功的,我们的准确率达到了 66.21%,前 1,000 个词语的准确率达到了 73.3%。

在实验中,我们只对犯罪案件领域词汇进行了分析。我们的研究中,把领域词汇抽取当作一个二元分类问题,训练的语料是某一领域的语料,还是针对某一个领域进行的处理。在以后的研究中,我们的重点是,把领域词汇抽取当作一个多元分类问题,训练语料也不再是某一领域的语料,而输出却能将各个领域的领域词汇抽取出来。这需要很好的统计各个领域的领域词汇的特征,并且还需要 SVM 分类中的大量正例、反例信息。总之,我们在这里提出了一种全新的思路,利用分类的思想进行领域词汇的抽取,在今后的研究中,还有很长的路要走。

## 第3章 命名实体识别

### 3.1 引言

命名实体 (Named Entity) 是文本中重要的信息元素。在第一章中, 我们已经介绍了其基本概念和识别方法, 这里不再作详细说明。我们对公安领域的文本进行了总结与分析, 在我们的系统中, 主要识别以下实体: Email、Tel、身份证、车牌号码、银行帐号、时间信息、人名、地名、机构名、案件名称。这些实体是公安领域中的比较重要的实体, 也是破案所必须的实体。

在对 Email、Tel、身份证、车牌号码、银行帐号、时间信息这几个实体的识别上, 因为它们都包含了数字字母信息, 规则也比较明确, 所以我们采用了规则的方法, 利用正规表达式制定规则进行识别。其中, 时间信息相对比较复杂, 这里我们重点介绍时间信息的识别。

人名、地名、机构名的实体识别是研究的比较多, 也比较深入的。我们使用的 TRS 分词工具中包含了对人名、地名、机构名的识别, 它是基于 HMM 模型的方法, 在这里我们不做深入的分析。但是, 在案件中, 案件的地点信息和地名还是有一定区别的, 这里我们重点讨论地点信息的识别。

在案件名称的识别上, 我们采用了基于规则与统计相结合的方法, 该方法用于识别案件名称, 我们对案件文本进行了总结与分析, 建立了相应的知识库, 在规则的基础上识别出候选案件名, 再利用统计的方法进行过滤。

{

### 3.2 时间信息识别

在信息抽取中, 时间信息始终是与事件息息相关的内容。时间表达式的识别与发现, 已经成为信息抽取评测的一项内容。在案件信息表达中, 时间信息是非常重要的一个参数, 时间信息大致可以分为以下两类:

1. 案件事件发生的时间
2. 案件报道的发出时间

这些时间信息在信息抽取中起着重要的作用, 因此准确的信息抽取必须按照明确的时间进行。

### 3.2.1 时间表达式的类型

通过对所处理数据中的时间表达形式的分析,我们得到了如下几类时间表达式:

#### 一、 精确的时间表达式

##### 1. 精确的点时间表达式

- a) 含有“年”、“月”、“日”、“时”、“分”、“秒”的表达式。比如:“2006-12-11”、“1998年”、“4月2日”、“去年7月”、“6:30:29”、“8点35分”、“10点30”等。
- b) 直接用时间词表达的表达式。比如:“昨天”、“今年”、“本月”等。
- c) 表示星期的表达式。比如:“周一”、“上周二”、“星期日”、等。

##### 2. 精确的段时间

- a) 表示时间经历的长度的时间表达式。比如:“三分钟”、“8小时”、“两天”、“一周”、“五个月”、“二十年”等。
- b) 表示一个范围的段时间表达式。比如:“15到20天”、“15分钟到20分钟”等。
- c) 表示年代、世纪、千年和公元前的段时间表达式。比如:“九十年代”、“下个世纪”等等
- d) 由两个点时间表达的段时间表达式。比如:“四月十八日至十九日”、“星期一到星期二”、“九点三十分至九点四十分”等

在点时间的表达中,从相对性角度来看,有绝对的时间表达和相对的时间表达,如“2004-03-11”即为绝对时间,而“上周二”为相对的时间。

#### 二、 模糊的时间表达式

1. 表示“过去”、“现在”、“将来”。比如:“迄今”、“目前”、“以往”、“此前”、“届时”、“后来”、“过去”、“现在”、“将来”等。
2. 表示季节。比如:“春,春季,夏,夏季,秋,秋季,冬,冬季”等。
3. 季度和半年、周末、早晨、下午和晚上。比如:“第一季度,第二季度,第三季度,第四季度,上半年,下半年”、“周末”、“早上,上午,早晨,中午,下午,傍晚,晚间,白天,日间”等。
4. 未指明具体时间长度的段时间。比如:“90年代晚期”、“长期”、“几年”、“数小时”、“多年”等。

#### 三、 指代的时间表达

比如：“当年”、“当月”、“当日”等。

在上述的时间表达中，我们对相关文本集中时间词语进行了统计，在案件中使用频率最高的时间表达形式有两种：

1. 精确的点时间表达式
2. 指代的时间表达

在我们的研究中，只对这两种情况进行讨论，其他情况暂时不做考虑。

### 3.2.2 时间信息的识别策略

在中文信息处理中，对汉语时间短语的研究较少，对时间信息的处理仍然处于起步阶段。不同的应用对表达时间信息颗粒度的要求不同，因而对时间表达式的界定也不尽相同。根据应用，确定时间表达式的形式、范围，在此基础上进行识别，这是面向应用的研究，也是需要进一步深入的研究。在我们的研究中，我们只考虑上节中提到的在案件中使用频率最高的两种时间表达式的识别。

为了得到时间表达式，这里，我们采取了以下两种策略：

#### 1. 词典方法

对于那些用法固定的表达式，特别是那些描述相对日期的词语（如“今年”、“昨天”等），我们把它们收录到词典中，通过对词表的匹配进行识别。

#### 2. 有限自动机方法

对于绝对时间（如“1999年8月”、“十一月五日”等），我们制定一定的规则，建立一个正规式。再把一个正规式编译为一个NFA（不确定的有限自动机）进而转换为相应的DFA（确定的有限自动机），这个NFA或DFA就是识别正规式所表示语言的句子的识别器。

在这里，我们使用了词法分析器的生成系统LEX。LEX源程序的核心是识别规则，它由正规式和动作组成。它的工作原理是将正规式转换成有限自动机。关于LEX的介绍，大家可参阅编译原理<sup>[36]</sup>，在这里，我们简单介绍一下时间的正规式表示方法：

D1 ( 一 | 1 )

D2 ( 二 | 2 )

.....

D0 ( 零 | 0 )

$$DE \quad (\{D1\} | \{D2\} | \{D3\} | \{D4\} | \{D5\} | \{D6\} | \{D7\} | \{D8\} | \{D9\} | \{D0\})$$

$$Dy \quad (\{D1\}\{D9\} | \{D2\}\{D0\})\{DE\}\{DE\}(\text{年})$$

上面给出的是识别时间信息的部分正规式（详细的正规式集合见附录 C）。其中，D0.....D9 表示数字，DE 表示数字的集合，Dy 表示完整的“年”时间表达式（这里只考虑了 1900—2099 之间的年份）。

### 3.2.3 时间信息的规范化

文本中的时间表达式形式很多，为了能够找到文本中时间变化的信息，按照时间线索进行案件推理，一个基础性的工作就是对时间表达式进行规范化处理。

所谓规范化处理，就是要将所有的时间表达式表示成为统一的、显示的格式。规范化处理涉及：

#### 1. 时间规范形式的表达

我们利用上节的方法，得到时间表达式，为了方便处理，我们把时间表达式规范为“\*\*\*\*年\*\*月\*\*日\*\*时\*\*分\*\*秒”的格式。

#### 2. 基准时间的确定，以便规范化相对时间信息

对于案件报道来说，得到基准时间不难做到。初始基准时间可以从每篇报道的日期戳记直接得到，但是一般来说，报道的日期比案发时间要晚，所以这里，我们做一个简单的处理，把案件报道正文中第一个出现的时间作为基准时间。

#### 3. 时间指代词的消解，以便找到对应的精确显式的时间表达

时间指代词的识别是利用词典的方法进行识别的。我们在建立词典的时候，同时表达出其语义含义，在进行规范化时就变得容易了。例如，“当天”，我们可表示为“0D”；“明年”我们表示为“+1Y”；“上个月”我们表示为“-1M”。前面的数字表示偏移量，后面的单位表示时间进制单位“年、月、日”。

### 3.2.4 实验结果与分析

在本节中，我们讨论了时间表达式的识别和规范化处理。实验使用的语料由 1,000 篇犯罪信息文本构成，通过和人工标注的结果的对比，准确率和召回率分别达到了 93.25% 和 51.48%。实验结果表明，该方法的准确率还是比较高的，但是召回率比较低。下面我们对原因进行分析：

1. 系统准确率比较高,因为我们是利用词典和正规式生成规则的方法进行的识别,凡是符合规则的都能识别出来。但是其中也有一些错误,主要集中在“日”的识别上,比如在“这次审判历时8日”,这句话中,“8日”就识别为时间,但在我们的要求上,这是不需要的。
2. 系统召回率比较低,主要原因是,我们只定义了我们需要的时间表达式的规则,对于其他的时间表达式我们没有考虑,所以召回率比较低。

时间短语是命名实体的一种类型,由于我们对时间短语的识别和规范化处理研究还处于起步阶段,主要考察了时间短语的形式和特点。在此,我们提出了两种策略进行时间短语的识别,并进行了小规模实验,为进一步深入的研究奠定基础。

### 3.3 地点信息识别

近几年,国内有关中文地名识别的研究比较多,既有基于规则的方法,也有基于统计的方法,达到了很好的识别效果。我们使用的 TRS 分词系统中,包含了地名的识别,其使用了 HMM 模型的方法。但是,分词识别出来的地名并不是我们真正需要的作案地点信息,在犯罪案件中,地点信息和地名还是有一定区别的,我们在地名识别的基础上,制定了一些规则,来获得地点信息,效果还是很理想的。

#### 3.3.1 地点信息与地名的区别

##### 3.3.1.1 关于地名的定义

- 地名相关的指代词标记为地名的一部分,例如:

<LOC>北京市</LOC>

<LOC>朝鲜半岛</LOC>

<LOC>长江流域</LOC>

- 两个或多个地名连续出现时要分开标记,例如:

<LOC>中国</LOC><LOC>广东</LOC>

<LOC>科</LOC><LOC>伊</LOC>边境

- 其它例子

<LOC>四川</LOC>话

<LOC>美</LOC>籍华侨

有关地名定义的详细介绍见附录 A。

### 3.3.1.2 案件信息中地点和地名的区别

在公安领域案件信息中，地点和地名有以下几点区别：

1. 地点是连续出现的地名的集合，例如：“吉林省四平市梨树县梨树镇霍家店村”，这句话中，地名识别后为：“<LOC> 吉林省</LOC><LOC> 四平市</LOC><LOC> 梨树县</LOC><LOC> 梨树镇</LOC><LOC> 霍家店村</LOC>”，而地点识别后应该是“<STA> 吉林省四平市梨树县梨树镇霍家店村</STA>”。
2. 公安领域案件信息中，地名后缀为“街，路，道，巷，里，町，庄，村，弄，堡”等字时组成的地点是我们真正需要的，而单单出现“省、市”对我们来说没什么意义。换句话说，我们需要的地点是精确的地名的集合。
3. 由于案件信息中的地点信息包含比较精确的地名，而地名识别并不能把这些地名识别出来，例如：“白草洼”、“西大望”、“亮马厂”等地名并不能被识别出来，这样我们就要做进一步处理。

### 3.3.2 地点信息的识别策略

本系统的地点信息识别是在分词和地名识别的基础上，通过地名库和简单规则相结合的方法实现的。在我们的地名库中收集了一些比较精确的地名。另外我们还收集了地名前导词 600 个，地名后导词 502 个，非地名后导词 111 个，常用地名后缀 120 个。

地名前导词：在、从、向、来到、到达、离开、进入、抵达、前往、赴、来自、经过等；

地名后导词：使馆、首府、历任、援助、郊区、建交、失守、地处、申办、各级、选手等；

非地名后导词：话、籍、语、文、侨等；

地名后缀：庄、洋、区、堡、州、城、县、村、宫、山、乡、路、街、王国、港、河、沟等。

本系统中用所用的地名识别规则有以下六条：

1. 地名前导词+长度小于等于 3 的字符串+地名后缀，且字符串中的词的

- 词性不能为虚词和动词，则长度小于等于3的字符串+地名后缀组成新的地名。例如：来到(LOCQDC)+鹤雀+楼(LOCHZ)→鹤雀楼(LOC)；
2. 地名前导词+长度小于等于3的字符串+地名后导词组，且长度小于等于3的字符串中的词的词性不能为虚词和动词，则长度小于等于3的字符串为地名。例如：从(LOCQDC)+遵化+出发(LOCHDC)→遵化(LOC)；
  3. 地名+长度小于等于3的字符串+地名后导词组，且长度小于等于3的字符串中的词的词性不能为虚词和动词，则长度小于等于3的字符串为地名。例如：北京市(LOC)+白草+洼(LOCHDC)→北京市(LOC)+白草洼(LOC)；
  4. 地名+非地名后导词，该地名不是地点，否则识别为地点。例如：河北(LOC)+人→(河北不是STA)；
  5. 地名+地名组成地点，例如：北京市(LOC)+海淀区(LOC)→北京市海淀区(STA)；
  6. 地点+地名组成新地点，例如：北京市海淀区(STA)+中关村(LOC)→北京市海淀区中关村(STA)；

### 3.3.3 实验结果与分析

实验使用的语料由 100 篇犯罪信息文本构成，我们分别对地点和地名进行了识别，实验结果如下：

表 3.1 地点、地名实验结果

	测试 文本	总个数	识别出的 个数	正确	错误	准确率 %	召回率 %	F <sub>1</sub>
地点识别	100	523	484	468	16	96.50	89.64	93.07
地名识别	100	1525	1419	1307	112	92.07	85.74	88.91

从实验结果中，我们可以发现，地点识别的准确率和召回率都有一定程度的提高，分析原因在于：

1. 我们的规则还是很有效的，地点识别的规则中，结合了识别出的地名，同时过滤了我们不需要的地名，并且还能识别出一些新的地名。
2. 地点识别中的错误识别主要是因为最初地名识别的错误，从而导致了地点的识别错误。
3. 地点的召回率比较低，是因为地点识别中过滤掉了部分不需要的地点，

还有个原因是规则定的不恰当，还有修改的必要。

### 3.4 案件名称识别

在公安领域信息中，案件名称有着举足轻重的作用。因此，如何准确的识别出文本中的案件名称是一个非常重要的研究课题。在对公安领域文本进行了深入地分析和研究的基础上，总结出了案件名称的结构特征及其上下文信息，建立了用于识别案件名称的知识库。在知识库的基础上，首先对案件名称进行模板识别，然后进行结构分析和上下文分析，并利用禁用词库对案件名称进行排歧，从而识别出候选案件名称。我们再使用统计方法对识别出的候选案件名称计算权值，过滤权值比较低的实体，这样能大大提高系统的准确率。初步实验结果表明，在封闭测试中犯罪案件名称抽取的精确率可以达到 95.26%，召回率可达 89.14%；在开放测试中精确率可以达到 84.47%，召回率可达 75.56%。

#### 3.4.1 案件名称的特征

犯罪案件名称众多，规律各异，长短不一，所以案件名称的识别有一定的困难。

1. 没有明确规范的案件名称定义，随着时代的发展，会有新的案件名称不断出现。
2. 案件名称的用词比较自有、分散。案件名称中既可以有名词，也可以有动词。例如：“贩卖|毒品|案”。
3. 案件名称的长短不一，短的如“杀人案”，长的如“中国长城资产管理公司广州办事处诉中山大松通用机电有限公司及火炬集团(担保方)借款合同纠纷一案”。
4. 案件名称有时同一些介词、动词、方位词之类的指示词出现，但有些指示词可以作为案件名称的组成。例如“制造‘3.25’杀人案”，其中“制造”为指示词；“非法制造贩卖毒品案”，其中“制造”为案件名称的一部分。
5. 有的案件名称中包含标点符号，识别上产生困难。例如“杀人、抢劫案”，“平安证券有限责任公司（简称：平安证券）偷税案”。
6. 案件名称结尾一般有案件名称特征词出现，例如“案”就是一个明显的特征词。

其中，1、2、3、4、5增加了案件名称识别的难度，并可能产生歧异；4、6有助于案件名称的识别。

3.4.2 规则与统计相结合的案件名称识别

规则与统计相结合的案件名称识别主要包括三个过程（见图 3.1）：文本预处理；基于规则的识别；基于统计的识别。文本预处理阶段就是对原始文本经过简单的分词、分句之后，得到初加工文本。基于规则的识别就是利用建立的知识库对案件名称进行识别。对识别出的案件名称，通过统计方法计算权值，过滤权值较低的案件名，最后得到识别结果。

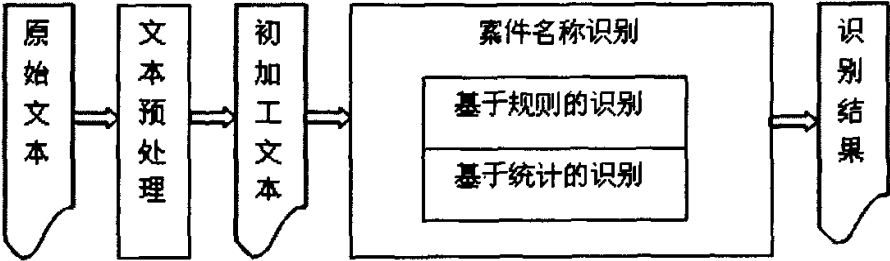


图 3.1 案件名称识别策略

3.4.2.1 文本预处理

原始文本首先进入分词系统，该分词系统进行了改造，添加了部分公安领域词词典。部分公安领域词词典中收集了公安领域中出现频率比较高的词汇。经过分词，然后进行分句后，我们得到初加工文本。

3.4.2.2 基于规则的案件名称识别

1. 知识库的建立

本文使用的语料来自互联网上，共 10,026 篇真实文本。其中案件名称 21,500 个。在这些资源的基础上，我们利用机器统计和人工筛选相结合，建立了如下知识库，用来指导案件名称的识别。

1. 关键词库

案件名称结尾有案件名称特征词出现。所以，在我们的系统中，对于案件名称的识别是从确定案件名称的右边界开始。以“案”字结尾的这些词可以提供准确的案件名称右边界信息，将这些信息收集整理，建立案件名称关键词库，作为案件名称识别的触发条件。

## 2. 前缀库

案件名称有时同一些介词、动词、方位词之类的指示词出现,例如“审理”、“判决”、“抓获”,“涉嫌”等。为了确定案件名的左边界,我们建立了案件名称前缀库。但是前面提到的特征 4 已经表明,有些词既可以是案件名称的前缀,也可以是案件名称的一部分。所以在确定案件名称前缀库时,一定要确保该词不可以作为案件名称的一部分,例如“制造”、“提供”、“参与”、“组织”等词都不能作为案件名称前缀。

## 3. 关键词前词禁用词库

我们经过实验发现,在案件名称中,有一些词不能出现在案件名称关键词前面,我们收集了这些词,建立了案件名称关键词前词禁用词库。如果关键词前词在禁用词库中,表明这不是案件名称,不用继续识别,从而可以大大减轻系统的工作量,也可以提高系统准确率。

## 4. 禁用词库

有一些词不能作为案件名称的组成部分,主要是一些介词,例如“随着”、“因而”、“此外”、“其余”等等,我们建立案件名称禁用词库。在识别过程中,碰到库中的词则中断案件名称的识别,认为当前识别对象不是案件名称成分。

## 2. 基于知识库的识别

在识别系统的核心部分“案件名称识别”中,我们首先借助于知识库的进行规则识别(如图 3.2 所示)。

在识别过程,有些词既可以作为案件名称的左边界,又可以作为案件名称的一部分,但是当这些词与某些词搭配出现时,那它只能作为案件名称的左边界,例如“与[案件名称]有关”,此时“与”就能确定为案件名称的左边界。所以,我们首先对案件名称进行常用模板分析,凡是符合模板的,就识别为案件名称。我们利用机器学习和人工选择相结合的方法,定义了 35 套模板,模板形式为:[案件名前词][案件名称][与前词搭配的尾词]。例如:“和[案件名称]有牵连”,这样“[和]……[有牵连]”,就定义为一套模板。

对于那些左边界比较清晰的案件名称,我们经过案件名称构成分析以及上下文分析,同时利用禁用词库对识别出的案件名称进行排歧,将最后结果保存到识别结果中。识别方法如下:首先,我们根据关键词库确定案件名称的右边界;然后从案件名称右边界开始向左搜索,判断关键词的前词是否在前词禁用

词库中，如果在的话，就停止向左扫描，这不是案件名称；继续向左扫描，直到该词在前缀词库中或者到一句话的开头，此时识别出候选的案件名称；再从左向右判断案件名称的组成词是否在禁用词库中，如果在的话，就删去该词，重复执行，如果不在的话，就停止。这样我们就利用知识库识别出了候选案件名称。

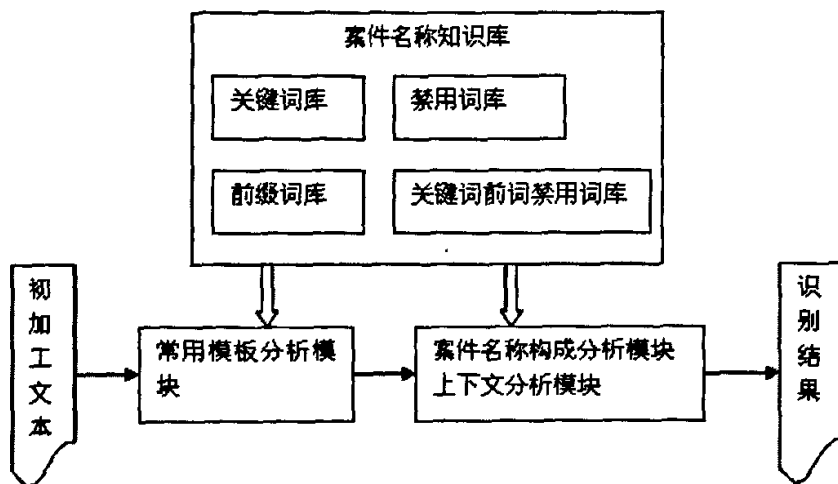


图 3.2 基于知识库的案件名称识别

### 3.4.2.3 基于统计的案件名称识别

在利用规则对案件名称进行识别后，我们发现一些噪音，组成这些噪音的词大都是些比较常见的词，我们利用统计方法计算识别出的候选案件名称的权重，从而过滤掉那些权值比较低的案件名称。我们的训练语料来自人民日报，共 26,987 篇文本。

在信息检索中最常用的确定一个词在文档中重要性的方法是 TF\*IDF 的方法。TF 即该词在一篇文档中出现的频率，IDF 称为反文档频率，一个词在越多的文档中出现它的 IDF 就越小，反之就越大。公式如下：

$$W(f_i, d) = TF(f_i, d) * IDF(f_i) = N(f_d) * \log(N(f_i) / N) \quad (3.1)$$

其中， $W(f_i, d)$  是特征  $f_i$  在文本  $d$  中的权重， $N(f_i)$  是出现  $f_i$  的训练文本数， $N$  是总训练文本数， $N(f_d)$  是文本  $d$  中出现  $f_i$  的次数。

TF\*IDF 方法识别原理：(1) 一方面，文档集中包含某一词条的文档越多，说明它区分文档类别属性的能力越低，其权值越小；反之，其权值越大。在犯罪

案件名中出现的词，一般都是犯罪信息领域中的词，而我们的训练语料来自人民日报，所以包含词条的文档较少，其IDF值较大。(2)另一方面，某一文档中某一词条出现的频率越高，说明它区分文档内容属性的能力越强，其权值越大。在犯罪信息文本中，犯罪案件名称中出现的词条一般都是文本中比较重要的词，其出现的次数也比较多，所以其TF值较大。根据上面两个原因，组成案件名称的词汇的TF\*IDF值比普通词汇的大很多。我们利用该方法计算案件名称的权重，然后过滤权重比较低的噪音，实验表明，该方法能大大提高识别的准确率。

**TF\*IDF方法识别：**对于根据规则识别出的案件名称，我们计算每一个词的TF\*IDF值，然后取平均值，再和设定的阈值进行比较，如果大于阈值，那就是案件名称，如果小于阈值，就过滤掉。根据统计，我们取阈值value=3.1。

### 3.4.3 实验结果和分析

本文使用的语料库由10,026篇犯罪信息文本构成，我们从中随机选出1000篇犯罪信息文本进行封闭测试，实验结果如下表3.1：

表 3.1 封闭测试的实验结果

	测试 文本	案件名称 个数	识别出的 个数	正确	错误	准确率 %	召回率 %	F <sub>1</sub>
进行统计 识别前	1000	2118	2105	1897	208	90.11	89.57	89.83
进行统计 识别后	1000	2118	1982	1888	94	95.26	89.14	92.10

同时，我们还对300篇文本进行了开放性测试，实验结果如下表3.2：

表 3.2 开放测试的实验结果

	测试 文本	案件名称 个数	识别出的 个数	正确	错误	准确率 %	召回率 %	F <sub>1</sub>
进行统计 识别前	300	540	531	415	106	78.15	76.85	77.49
进行统计 识别后	300	540	483	408	75	84.47	75.56	79.76

我们对识别结果中的错误进行了整理分析，发现错误主要有以下几种类型：

#### 1. 标点符号导致的识别错误

在识别的开始，我们对文本进行了分句处理，所以当案件名称中包含标志一句话结束的标点符号时，就导致了识别错误。例如：“由广州市人民检察院提起公诉的郑洪钧等36名被告人走私‘红油’（香港地区专用的添加红色染色

剂的免税柴油，俗称红油）案”，在这段话中，案件名称识别为“俗称红油）案”，这就是由于逗号导致的错误。

#### 2. 分词错误导致的识别错误

在文本预处理阶段，对文本进行分词，由于分词错误，也可能导致识别错误。例如：“南充市中级人民法院近日分别对方吉和罗小林两起涉黑犯罪大案做出一审判处”，在这句话中分词为“分别|对方|吉|和|罗|小|林|两起|涉|黑|犯罪|大|案”，由于分成了“对方”，所以导致了识别错误。

#### 3. 有些案件名称没有明显左边界，导致不能正确识别

如“我国加入世贸后知识产权案大幅增加”、“公交分局立扒窃案 160 起”、“使这个银行连续 12 年无经济罪案发生”，这些句话中，没有明显的左边界导致识别错误。

在公安领域的信息抽取问题中，如何正确识别文本中出现的犯罪案件名称是一个非常重要的问题。本文在对公安领域中犯罪案件名称的结构及其在文本中的出现的上下文进行了深入研究的基础上，建立了四个识别用的知识库，并且使用TF\*IDF方法对识别出的案件名称进行过滤。经过初步实验，结果表明我们的识别策略是很有效的。

## 3.5 本章小结

在本章中，主要论述了命名实体识别的方法，主要介绍了犯罪案件中的时间信息、地点信息和案件名称的识别。我们认为，对于命名实体的识别现在已经具有了相当成熟的技术，然而从评测的结果来看，汉语命名实体的识别还远不能满足应用的需求，因为这里存在着技术、资源、应用需求之间的有机结合问题。

命名实体识别是集中体现自然语言处理技术的一个研究点，基于理性主义的规则方法和基于经验主义的统计方法在这一任务中都有所体现。每一种方法都有其优势与局限。因此，在命名实体的识别研究中，我们根据具体的应用需求，对每一种实体采用不同的方法进行识别。特别是在案件名称的识别上，我们没有可以用的资源，因此，案件名称的识别过程同样是资源建设的过程，可以为以后的研究奠定基础。

## 第4章 公安领域案件信息的模式获取

### 4.1 引言

模式（Pattern）是待抽取信息的一种抽象表达方式，它体现了特定信息的组成元素，这些元素也是人们对信息的关注焦点。利用模式从文本中获得感兴趣的信息，是信息抽取系统普遍采用的方法。在信息抽取系统中，事件、关系等信息通常是通过模式来表达的，因而信息模式的获取是 IE 的关键技术之一，也是该领域的一个热点研究内容。

模式是面向领域的，不同的领域、不同的特定信息，其模式是不同的。比如从药品说明书中抽取药品的属性，与从新闻报道中抽取恐怖活动的有关信息，所用的模式是不同的。

模式的获取方式有三种：手工的、半自动的和自动的。由于手工的和半自动的获取方法费时、费力，很难适应 IE 系统向不同领域的移植，因此自动的模式获取方式已经成为构造信息抽取系统、提高系统可移植性的主要手段，越来越多的研究集中在无指导学习的自动获取方式上。

根据所处理对象加工深度的不同，组成模式的成分可以有不同的形式：

1. 在进行了词法分析、命名实体识别时，通常由词语、实体等项构成
2. 在进行了浅层句法分析后，由浅层分析获得的语块等成分构成
3. 在进行了完全句法分析后，由表达句法结构的关系构成

模式的表达方法主要包括：根据模板描述的词项的固定搭配、谓词论元结构、依存树等。英语中的模式获取工作大都在句法分析的基础上进行，以句法分析结果为模式的组成单位，可以从结构上保证模式的完整性和易识别性。对于汉语，句法分析还远不能满足应用的要求，因此本研究是在进行了文本分词、词性标注、命名实体识别等处理的基础上进行的。

### 4.2 模式的获取

我们的模式获取研究有所不同，我们没有以句法分析的结果作为模式的表达形式。在英语中，描述内容与句法结构间存在着较强的对应关系，而在汉语

中,关于事件的描述通常由若干小句组成,事件的参与成分散布在各个小句中,目前我们也还没有一个可以使用的分析器对构成事件的句子进行句法分析,因此我们采用词项作为模式的表达形式。

#### 4.2.1 事件信息表达的特点

经过对语料的分析我们发现,汉语对于事件信息的表达与英语有着较大的差别。在英语中,语法在语言表达中起着严格的约束作用,若一句话是对一个事件进行描述的话,则其相应的论元被约束于一个句子结构中,这样使用句法分析的结果,可以比较方便的找到关于事件的参与成分,因此,英语中的模式通常使用谓词—论元结构、依存树结构等表达。对于汉语,事件表达的特点是:

##### 1. 句法结构的灵活性

汉语句子的词序及语法规则都有较大的灵活性,造成汉语句法分析很困难。

##### 2. 事件参与成分的跨句性

事件的参与成分分布在若干个句子之内。例如,对于犯罪案件来说,其参与成分在一个句群中而不是在一个句子内部出现。关于犯罪案件的时间、地点、作案人情况等,在几个小句中分述。

##### 3. 事件描述用语的稳定性

对于犯罪案件的报道,尽管不同的来源、不同的撰稿人,但其表达的方式有着良好的一致性。例如:“[机构名]审理了[案件名]”、“[机构名]抓获了[案件名]的犯罪嫌疑人[人名]”等结构在案件的报道中经常出现。

#### 4.2.2 模式及其抽取过程

对于模式来说,不同的表达方法包括了不同的上下文信息。根据我们对公安领域案件信息描述的特点分析,事件的参与成分集中表现在命名实体上。从目前我们所拥有的处理资源状况来看,我们还没有办法将模式建立在部分句法分析的结果上,也不能与利用句法结构获取的模式相比较,因此,仍然遵循我们进行该研究的基本思想,即在现有的资源上定义我们的研究任务,为此我们将模式的表达定义在词语与实体类信息的组合上。

**基本模式:**包含关键词语在内的项串,其中项可以是实词、短语、词语语义类或命名实体名。

**组合模式:**包含基本模式在内的几个模式的有序排列。

由于事件组成成分具有动态结合性的特点，组合模式的提出便是对此特点的。利用组合模式可以将若干个分句中的有关信息关联起来，从而找到与特定信息搭配的相关信息，从而对案件信息进行描述。

### 4.2.3 模式的分类

在公安领域案件信息中，从形式上，我们可以将模式分为两类：

1. 关系类模式
2. 事件类模式

关系类模式，只是针对一句话中的两个实体之间关系而言。所以，关系模式也就是实体间的二元关系模式。例如“[机构名]nt 审理了 [案件名]cn”，这句话中[机构名]是[案件]的审理机构；而“[案件名]cn 的被告是 [机构名]nt”，这句话中[机构名]是[案件]的被告。

事件类模式，可以看成是多个关系类模式的组合模式，其中的各个命名实体是该事件的参与成分。我们可以对得到的二元关系模式进行合并，从而得到事件类模式。

下面我们将主要对实体关系的抽取做详细的介绍。

## 4.3 实体关系抽取

随着互联网上的文本数量的增加，从中抽取出各种关系并将之结构化存储是进一步的数据查询和数据挖掘的前提。最初构造模式库的方法，是组织经验丰富的专家，手工写出模板。这种方法虽然取得了比较好的效果，但是人力的耗费非常巨大。而要从一个大的文档集合中抽取出所有的二元关系，通过手工方法费时费力，对于不同类别的二元关系，如果都要通过手工方法为之获取相应的二元关系模式，那是不可能的。因而，如果能够实现关系模式获取的自动化，对用户的要求很低，就很有意义。

对于本系统，我们只考虑一个句子中的两个实体之间的关系，而不考虑跨越句子的实体之间的关系。也就是说实体关系抽取问题的输入文本是一个句子以及句子中被正确识别的实体，所谓正确识别，不但要识别出实体的类别，还要正确确定实体的边界，这些我们在上一章中已经介绍过了。

本系统的实现方法主要借鉴了Snowball模型<sup>[36]</sup>。

### 4.3.1 基本定义

为了研究和叙述的方便，在此，我们给出如下定义：

**二元关系类别：**一个文本片段中，包含某两个命名实体，并且两个实体出现的顺序相同，那么我们说这是一种类别的二元关系对。

**二元关系：**两个实体之间内部的语义关系。

**模式：**上面我们已经说明，我们采用词项作为模式的表达形式，我们采用的模式表示方式是一个五元组 $\langle left, tag1, middle, tag2, right \rangle$ 。其中  $tag1$  和  $tag2$  是命名实体类别标签， $left$  表示  $tag1$  左边的上下文， $middle$  表示  $tag1$  和  $tag2$  之间的上下文， $right$  表示  $tag2$  右边的上下文。

**实例模式：**由一个同时含有某个二元关系对的两个对象的文本片段生成的，实际上是一个二元关系对的一次文本出现的形式化表示。实例模式只能概括生成它的文本片段，不具备更宽的概括作用，因而不能用来指导抽取其它的二元关系。

**泛化模式：**泛化模式是由一个或多个实例模式生成的，具有一定的概括作用，可以用来指导抽取其它的二元关系。

这里，为了提高模式在表达上的灵活性，我们把实体的上下文都用一个词向量表示，向量中的每个词都带有该词相对于它所在上下文的权重，我们只考虑其中的领域词汇或者是表征者两个实体间关系的名词或谓语动词，我们忽略了没有实际意义的虚词、助词等词汇。该权重是通过统计该词在它的上下文中出现的频率，并通过归一化处理后得到的。

泛化模式和实例模式都采用上面相同的模式表示方式。例如，文本片段“在奉节县永安镇发生了一起入室抢劫杀人案”中含有一个地点类命名实体和一个案件名称类命名实体。其中，词“在”（左边的上下文）位于地点类命名实体“奉节县永安镇”之前，该命名实体后紧跟的是词“发生”和“一起”（中间的上下文，其中“了”忽略）和一个案件名称类命名实体“入室抢劫杀人案”。从该文本片段中可以生成一个实例模式为： $\langle \langle \text{在}, 1 \rangle, \text{地点}, \langle \text{发生}, 1 \rangle \langle \text{一起}, 1 \rangle, \text{案件名称}, \{\} \rangle$ 。多个由同类文本片段（这里，说两个文本片段是同类文本片段，是指每个文本片段中含有一个地点类命名实体和一个案件名称类命名实体，而且地点类命名实体在案件名称类命名实体之前）生成的实例模式，可以生成一个泛化模式。如 $\langle \langle \text{在}, 0.2 \rangle, \text{地点}, \langle \text{发生}, 0.5 \rangle \langle \text{一起},$

0.5>}, 案件名称, {}> 就是由该种实例模式生成的泛化模式。

#### 4.3.2 实体关系抽取模型

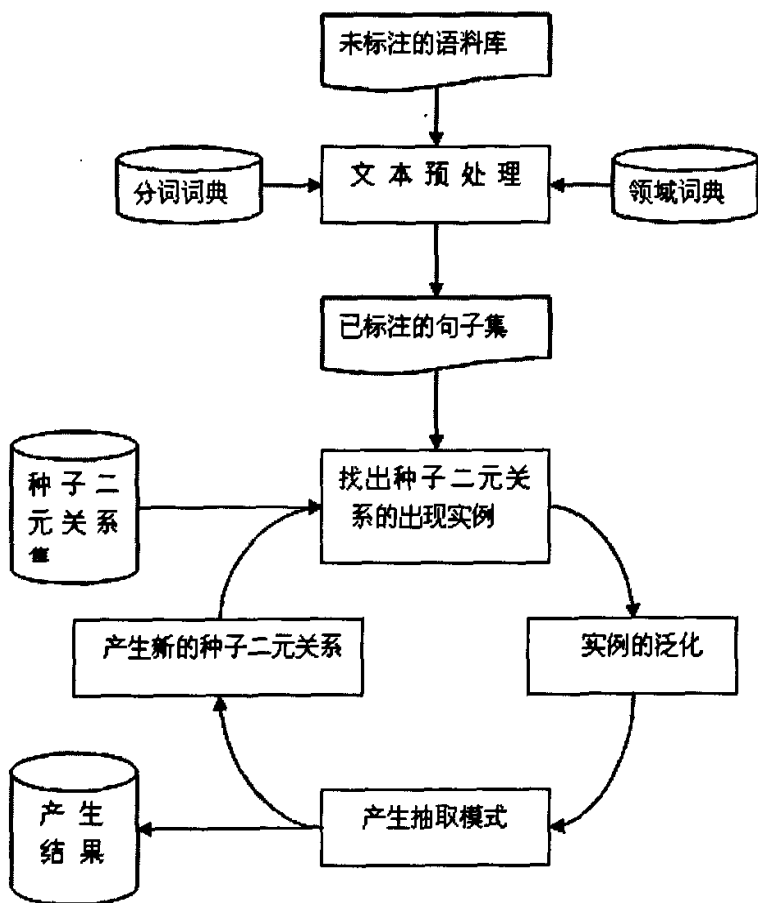


图 4.1 实体关系抽取模型

如上图 4.1 所示，整个过程大概分为四个部分：一、语料的预处理，通过预处理我们得到标注好的句子集  $D$ ；二、通过人工给出的种子二元关系  $R$ ，找到出现种子二元关系的出现文本，并用实例模式进行表示，形成实例模式集  $O$ ；三、对实例模式集合进行处理，通过聚类的方法生成泛化模式，并对生成的模式进行选择评价，从而得到二元关系的抽取模式  $P$ ；四、从标注的句子集中，找出更多的与抽取模式相匹配的二元关系，计算二元关系的可信度，将可信度高的二

元关系加入到种子二元关系集中。我们利用产生的新的种子二元关系重新循环，直到一定的结束条件为止，结束条件可以是到达循环的次数或是不再出现新的二元关系。

#### 4.3.2.1 文本预处理

在系统中，首先对无标注的训练语料进行分句、分词、词性标注和实体（如：人名、地名、机构、案件名称等）识别，然后无用的实体或不包含特征词的句子被过滤掉，得到标注好的句子集。

实验中使用的分词是 TRS 公司提供的，关于实体的识别我们前面已经详细论述，这里不再重复。

#### 4.3.2.2 实例模式的获取

首先，到已标注的句子集合  $D$  中找出同时含有  $R$  中任意一个二元关系对的句子，形成一个句子集合。实验中，一个句子生成一个实例模式。

由句子生成一个实例模式的方式：设实例模式  $T_p = \langle l_p, t1, m_p, t2, r_p \rangle$  由句子  $(wl1, wl2, \dots)$  [地点实体]  $(wm1, wm2, \dots)$  [案件名称实体]  $(wr1, wr2, \dots)$  生成，则  $l_s = \{(wl1, 1), (wl2, 1), \dots\}$ ， $m_s = \{(wm1, 1), (wm2, 1), \dots\}$ ， $r_s = \{(wr1, 1), (wr2, 1), \dots\}$ ， $t1 = \text{“地点实体”}$ ， $t2 = \text{“案件名称实体”}$ 。

#### 4.3.2.3 实例模式的泛化

我们得到了实例模式集合后，为了生成一个泛化模式，我们分两步进行：首先将实例模式集合  $O$  中的实例模式聚类，然后由聚类形成的每个实例模式子类生成一个泛化模式。

##### 1. 实例模式的聚类

这里，我们采用一种简单的 K-means 聚类方法。计算两个模式之间相似度的公式为：

$$\text{sim}(T_p, T_s) = \begin{cases} l_p \times l_s + m_p \times m_s + r_p \times r_s, & \text{if}(t1 = t1' \& \& t2 = t2') \\ 0, & \text{otherwise} \end{cases} \quad (4.1)$$

其中， $T_p = \langle l_p, t1, m_p, t2, r_p \rangle$ ， $T_s = \langle l_s, t1', m_s, t2', r_s \rangle$ ，两个泛化模式之

间或两个实例模式或泛化模式和实例模式之间的相似度都可采用上面的公式来计算。其中  $l_p \times l_r$  为两个词语之间的相似度，关于相似度的计算，我们利用了哈工大信息检索实验室的同义词词林和领域词汇，在此不做详细说明。

## 2. 泛化模式的生成

设实例模式集合  $O$  通过聚类后分成了  $m$  个子类  $O_1, O_2, \dots, O_m$ ，每个子类实例模式中的两个命名实体标签分别相同，这些实例模式之间的相似度超过预先规定的阈值。对其任意一个  $O_i = \{ \langle l_1, t_1, m_1, t_2, r_1 \rangle, \dots, \langle l_n, t_1, m_n, t_2, r_n \rangle \}$  ( $i = 1, \dots, m$ )，从中生成一个泛化模式为  $p = \langle l, t_1, m, t_2, r \rangle$ ，则  $l$  中词  $w$  的权重为  $l_1, l_2, \dots, l_n$  中词  $w$  的权重的和的平均值， $m$  和  $r$  中的词的权重同理计算。若某个  $O_i = \{ \langle l_1, t_1, m_1, t_2, r_1 \rangle, \dots, \langle l_n, t_1, m_n, t_2, r_n \rangle \}$  中实例模式的数目小于某个预先设定的一个泛化模式应该覆盖的最小实例模式数目阈值  $N_{\min}$ ，则舍弃该泛化模式，因为这种情况说明生成的泛化模式不具备好的覆盖度。

### 4.3.2.4 二元关系的生成

到已标注的句子集合  $D$  中找到同时含有  $tag1$  类和  $tag2$  类命名实体的句子，生成一个句子集合。为其中每个句子生成一个与之对应的实例模式，生成一个实例模式集合。对该集合中的每个实例模式，求其与每个泛化模式之间的相似度，若有一个或多个泛化模式与该特例模式之间的相似度大于事先设定的阈值，则从与该实例模式对应的句子中识别出的两个分别属于  $tag1$  类和  $tag2$  类的命名实体就组成了一个二元关系。在生成该二元关系的过程中记下能够生成它的那些模式及这些模式各自的可信度，并据此算出该二元关系的可信度。最后，将那些可信度在事先设定的阈值之上的二元关系放到最终的二元关系集合中。

我们定义模式  $P$  的可信度的计算公式为：

$$Conf(P) = \frac{\text{在模式P指导下，生成的正确二元关系数目}}{\text{在模式P指导下，生成的总的二元关系数目}} \quad (4.2)$$

设  $P = \{P_i\}$  是能够指导获得二元关系  $r$  的模式的集合，假设模式  $P_i$  能够指导获得二元关系  $r$  的概率为  $Prob(P_i)$ ，则  $P = \{P_i\}$  能够指导获得二元关系  $r$  的概率

为  $Prob(r) = 1 - \prod_{i=0}^{|P|} (1 - Prob(P_i))$ 。而  $Prob(r)$  可以看作是对  $Conf(r)$  的粗略估计，

$Conf(P_i) \times Sim(SP_i, P_i)$  可以看作是对  $Prob(P_i)$  的粗略估计。故而我们定义一个二元关系  $r$  的可信度计算公式为：

$$Conf(r) = 1 - \prod_{i=0}^{|P|} (1 - Conf(P_i) \times Sim(SP_i, P_i)) \quad (4.3)$$

其中,  $SP_i$  是在模式  $P_i$  的指导下所发现的二元关系  $r$  所在的句子所对应的实例模式,  $Sim(SP_i, P_i)$  是  $SP_i$  和  $P_i$  之间的模式相似度。

### 4.3.3 实验结果与分析

#### 1. 实验设置

我们给出几个<案件名称, 人名>的二元关系类别的二元关系种子。

测试最终得到的训练语料库包括从 2000 个文本得到的 62540 条句子, 经过领域过滤之后得到 10723 条领域相关语句, 组成了系统的训练语料库。

#### 2. 评价内容和评价方法

理想的评价方法是在标准的数据集上, 按照统一的评价准则进行评价。我们不具备这样的条件。在模式获取评价中常用的两种方法, 一是通过人工检查, 给出模式获取的准确性; 另一种方法是利用间接的方法对获取模式的性能进行评价。这是一种自动评价的方法。根据我们处理问题的特点, 进行自动评价比较困难。所获得的某类二元关系的全面性和准确性程度借用信息检索中的两个指标准确率和召回率来评测。其中, 关于识别的出的正确数目和错误数目, 我们通过人工检查。

#### 3. 实验结果

我们利用给出的种子模式和相关参数阈值的设定, 从测试集合中抽取出一个二元关系集合。我们从二元关系集合中, 随机抽取 50 个二元关系, 利用人工的方法进行检查, 得到的准确率为 80%。

关于全面性的评测, 我们从二元关系集合中, 找出 50 个正确的二元关系, 然后让系统从测试集合中运行, 我们人工数出其中有 32 个二元关系属于 50 个中, 所以其全面性为 64%。

#### 4. 结论

本文在 Snowball 模型的基础上, 利用给出的种子关系进行扩展, 得到了更多的二元关系。我们在模板相似度匹配时, 使用了同义词词林来计算词语间的相似性, 使得性能有一定程度的提高。同时, 我们还加入了模式的可信度判断, 避免了坏模式对系统的影响。

#### 4.4 案件事件信息描述

在本系统中，事件信息单一，主要是抽取和案件相关的信息，我们利用抽取的二元关系进行组合，得到案件事件的信息描述，从而达到刻画案件信息的目的。

在这里我们不进行详细的描述，我们只给出部分的案件信息描述的实例（间下图 4.2）。

案件信息描述	
案件名称:	马加爵杀人案
案发时间:	2月23日
抓获时间:	3月15日晚7时30分
嫌疑人:	马加爵
被害人:	邵瑞杰、龚博、杨开红、唐学李
原告:	李文杨、唐先和、邵渭清、黄莹梅、杨绍权、马存英
审理机构:	云南省高级法院
.....	

图 4.2 部分公安领域案件信息描述

#### 4.5 应用实例

关于案件信息二元关系抽取，对刑侦工作人员了解案件线索、获得破案突破点、总结犯罪规律都有着重要的作用。这里，我们介绍一个实际工作中的应用实例。利用抽取到的二元关系，进行数据挖掘，从而建立智能化的搜索引擎。我们以一个模拟例子来说明。

假定现在数据库中存在有 1000 份数据，有四份相关数据如下：

数据 1：1998 年，甲单位在某地抓获毒犯张三，缴获海洛因共 100 克，张三供认，接头人为李四（在逃），供货商外号刀疤。

数据 2：2000 年，乙单位抓获岩帅贩卖鸦片 2000 克。供认王五供货，李四接头。

数据 3: 2002 年, 丙单位查获冰毒 100 克, 获的情报供货商可能是王五, 接头人可能是李四。

数据 4: 2003 年, 调研处情报, 某毒枭刀疤可能姓王。

当我们输入: “刀疤”, 进行智能化检索时, 我们利用抽取到的二元关系进行数据挖掘, 可以得到如下信息:

■ 情报可信度分析:

1. 刀疤可能是王五 (可信度 80%)
2. 王五、李四可能是一个供应链 (可信度 85%)
3. 刀疤可能在边境一线贩毒, 海洛因、冰毒和鸦片都有可能 (可信度 65%)

■ 情报关联分析:

1. 关联人物: 张三 (关联度 90%)
2. 关联人物: 岩帅 (关联度 80%)

可以看出, 只有智能化搜索引擎才能进行关联性的检索, 也才不会遗漏有用的情报, 同时进行自动关联, 这正是由于我们进行了关系抽取的结果。

■ 情报统计分析:

根据用户的需求, 通过统计数据, 对情报进行初步的分析, 例如: 输入人名“张三”或绰号“刀疤”, 系统在所有的情报资料中进行统计, 并进行关系抽取, 可以分析出一些嫌疑人、嫌疑地点等等。

## 4.6 本章小结

本章利用所处理问题的文本特点, 提出了一种自动获取二元关系及其模式的方法。该方法以 Snowball 模型为基础, 我们对其中的一些技术进行了改进, 利用同义词词林进行模式相似度计算, 同时引入了关系和模式可信度的计算, 过滤掉了坏模式。实验表明, 系统性能达到了比较好的效果, 但是实际应用中还需进一步提高。我们在二元关系模式基础上获得组合模式, 从而用来描述案件信息。与同类研究相比, 我们利用相对简单的技术, 在浅层分析和较少知识资源支持的条件下, 取得了一定的效果, 同时我们也认识到该方法的局限性所在, 在今后的研究中将加入句法分析, 使系统性能进一步提高。

## 第5章 公安领域案件信息抽取系统

在本章中，主要介绍公安领域案件信息抽取系统的整体框架。本系统采取了分层模式，能够方便、高效的输出各个层次的中间结果，以供不同的应用。

### 5.1 系统结构

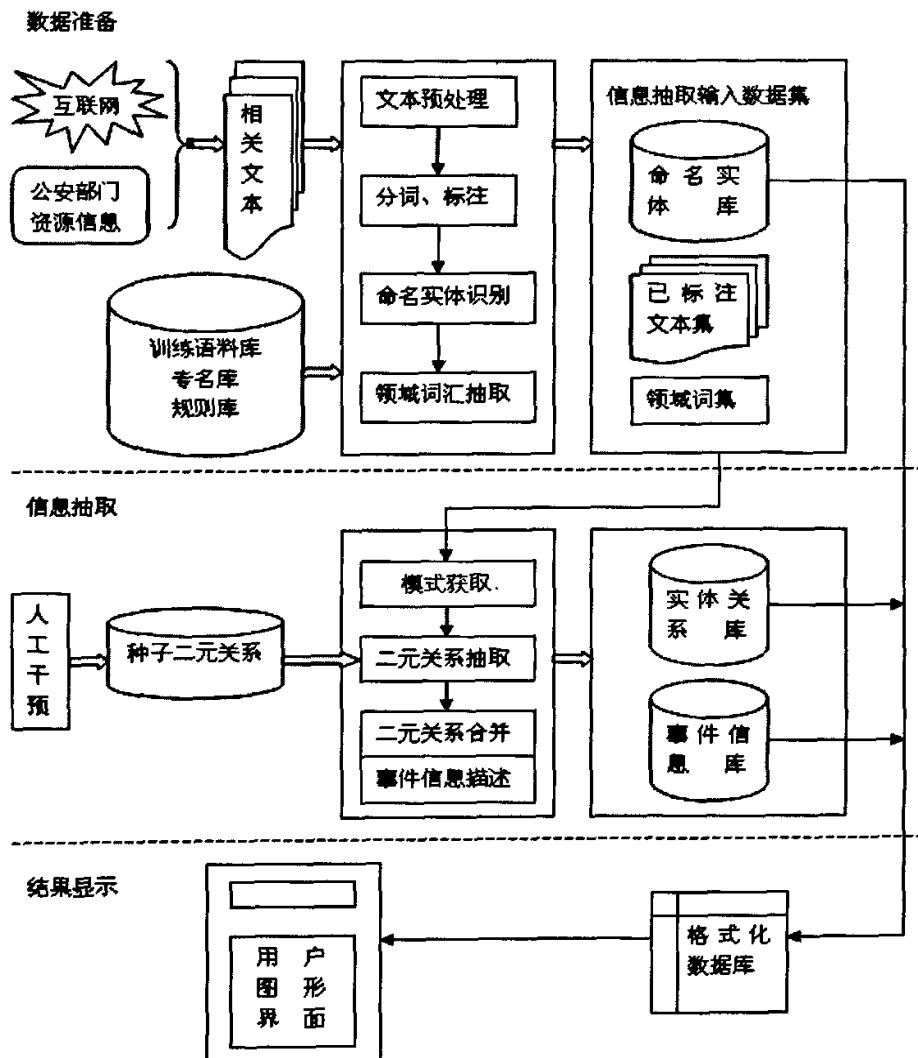


图 5.1 公安领域案件信息抽取系统结构图

公安领域案件信息抽取系统的系统结构如图 5.1 所示。本系统由三个部分组成，数据准备、信息抽取、结果显示。

数据准备部分是整个系统的基础。在此部分，将对待处理文本进行基础加工，包括文本分词、标注、命名实体识别、领域词汇抽取等。加工结果将作为信息抽取部分的输入。

信息提取部分是系统的核心部分，进行二元关系的抽取，并对二元关系进行合并，从而获得事件信息描述。

结果显示部分将信息抽取的结果存入数据库中，并以可视化的方式显示。

### 5.2 工作流程

本系统的工作流程如下：

1. 通过互联网和公安部门信息资源获得相关文本；
2. 文本预处理，包括：去除和修改乱码，整理句子中的间断，去除文档中的多空格和多回车换行相连的结构，统一段落格式；
3. 利用 TRS 分词系统，对文本进行分词和词性标注；
4. 进行命名实体识别，产生标注文本库和命名实体库；
5. 按标点符号分割文本为实验句子，形成实验用句子库；
6. 利用 SVM 方法进行领域词汇抽取，得到领域词典；
7. 根据定义的模式的表现形式，分别形成句子的向量表示，根据二元关系的抽取方法进行二元关系抽取，获得二元关系库；
8. 对二元关系进行合并，得到事件信息描述，存储到事件信息库；
9. 把结构化的数据以可视化的方式显示出来。

### 5.3 本章小结

本章对在研究中开发的公安领域案件信息提取系统的结构和工作流程作了介绍。信息抽取系统的信息结构是与文本集合相关的。当文本集合不断地动态变化时，可以反复进行上述的工作流程。信息抽取的过程，也是一个资源积累的过程，但这些资源还显得比较粗糙，许多还需要人工干预。随着资源的积累，可以进行资源的分析，从中提取普遍模式，形成信息抽取的基础性资源，并进一步发掘资源的自动建设方法，找到资源建设和应用系统开发间良好的结合点。

## 第6章 结论与展望

### 6.1 全文总结

信息抽取是文本信息处理的重要任务之一，是自然语言处理的研究热点。自1987年MUC评测会议以来，在评测的驱动下，信息提取技术和资源建设取得了长足进展。汉语的信息抽取由于受到基础加工、资源建设的限制，还处在起步阶段。

本文以公安领域案件信息为处理对象，在分析案件信息文本特征的基础上，进行案件信息的抽取。该研究是在加工精度不高、知识资源不够丰富的基础上进行的，考察我们现有的技术对信息抽取能够支持的程度，为信息提取的资源建设和深入研究奠定基础。

#### 6.1.1 本论文的研究内容

本课题通过对信息抽取的基本概念、信息抽取系统的体系结构、信息抽取的各个过程以及涉及到的各种关键技术的研究，并且结合公安部具体要求，具体研究内容如下：

1. 了解信息抽取的基础理论和应用领域，比较国内外研究现状；了解信息抽取过程涉及的各个阶段，以及相关的技术。
2. 深入研究命名实体识别，比较其各种方法的优劣，找出识别每种实体的简单有效的方法。
3. 深入研究领域词汇抽取及其各种实现方法，深入研究领域词汇的自动获取以及各种机器学习算法，并且比较各种机器学习算法，把领域词汇的抽取和信息抽取系统紧密结合起来。
4. 尝试性研究实体关系识别和模式的自动获取。

#### 6.1.2 本论文的工作基础

本研究涉及以下数据和资源：

1. 已加工的数据
  - 人民日报标注语料(1998年1月份)：该语料进行了分词、词性标注、命名实体识别

- 同义词词林（哈工大信息检索实验室）
- 2002 年人民日报语料：31174 篇文本
- 2. 实验数据  
实验数据来源是互联网上有关公安领域的新闻报道，共 34700 篇文本。
- 3. 软件资源
  - TRS 分词、标注系统
  - SVM\_Light 工具

### 6.1.3 本论文的研究特色

本研究具有以下特色：

1. 搭建了分层的公安领域案件信息抽取系统，能够输出各层次的中间结果。
2. 在命名实体识别上，对不同的实体采用不同的方法，在案件名称识别上，采用规则与统计相结合的方法，利用  $TF*IDF$  方法，统计每一个词出现的频率，从而对计算的权值进行过滤，识别案件名称。
3. 领域词汇的抽取，采用 SVM 自动分类的方法。并且把领域词汇的抽取应用到信息抽取系统中。
4. 二元实体关系抽取，我们在 Snowball 模型的基础上进行了改进，加入了领域词汇和同义词词林进行模式相似度计算，提高系统性能。
5. 在信息抽取系统中，大部分使用了机器学习和统计的方法，大大提高了系统的可移植性；同时比较各种算法的复杂度，使用比较简单有效的方法，提高系统的性能。

## 6.2 进一步工作的方向

1. 命名实体识别作为信息抽取的基本任务，目前已经具有成熟的技术和良好的资源支持。命名实体识别系统如何适应应用要求，是系统的适应性问题，也是命名实体识别研究必须考虑的一个问题。在我们的系统中，案件名称识别并没有在信息抽取的评测任务之内，但在我们的系统中却是最重要的实体。任何一个领域都有其自己的独特的实体，如何识别这些实体是一个长期的、复杂的研究课题。
2. 如何合理、有效地使用资源一直是我们的一个追求目标。在资源的约

束条件下，我们还不能运用一些现有的技术和方法。因此资源和语料库的建设工作也是必不可少的，特别是特定领域的。

3. 模式是实现信息抽取的主要手段。因此，探索不同信息对模式表达的要求，使得模式的表达和获取、信息的抽取之间能够更有效、更合适地结合起来，为特定信息提取服务，是需要进一步深入研究的。

## 致谢

值此论文完成之际，我要向所有帮助、支持我的人表示感谢！

首先，我衷心地感谢我的导师肖诗斌高级工程师。将近三年的研究生学习生活，肖老师一直给予我悉心地指导和无私地关怀。他渊博的知识给我打开了一个新的研究视野；他积极、执着、不畏艰难的事业追求，豁达、平和的人生态度将使我终身受益。

在这里，我还要感谢孙丽华老师，无论在课题研究还是实际工作项目中，孙老师总是毫无保留的帮助和激励着我。从她身上，我不仅学到了许多知识，而且她对研究和工作的态度、她对别人的真诚与热情将永远感染着我。

感谢北京信息科技大学中文信息研究中心的吕学强博士，吕博士在我的论文发表和课题研究中都提出了许多宝贵的建议。

感谢和我一起并肩学习的师兄、师弟、师妹们：陈言敏、王振华、陈志玮、余小军等等。在生活和学习中，他们经常帮助我，他们用其良好的品质感染着我。能够与他们一起度过这段美好的时光，是我一生最大的财富。

最后，我要感谢我的父母、我的弟弟，是他们无私的关爱、默默的支持，才有了现在的我。

感谢所有帮助和支持我的人们，感谢将近三年的研究生生活。我将以饱满的工作热情、乐观的人生态度和严谨勤奋的求学精神去面对新的工作、新的挑战！

2007 年 1 月

## 参考文献

- [1] Zhang Y M, Zhou J F, A Trainable Method for Extracting Chinese Entity Names and Their Relations. In Proceedings of the Second Chinese Language Processing Workshop, Hong Kong, 2000
- [2] Walter Daelemans, Jakub Zavrel, Kovander Sloot, and Antalvanden Bosch, TiMBL: Tilburg Memory Based Learner, version 3.0, Reference Guide Reference: ILK Technical Report. <http://ilk.kub.nl/ilkpapers/ilk0001.ps.gz>, 2000
- [3] Hobbs J, The Generic Information Extraction System. In Proceedings of the Fifth Message Understanding Conference (MUC-5), 1993
- [4] Nancy Chinchor, MUC-7 Information Extraction Task Definition. [http://www.itl.nist.gov/iad/894.02/related\\_projects/muc/proceedings/ie\\_task.htm](http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ie_task.htm), 1998
- [5] A. Borthwick, J. Sterling, E. Agichtein, R. Grishman, NYU: Description of the MENE Named Entity System as Used in MUC-7. In Proceedings of the Fifth Message Understanding Conference (MUC-7), 1998
- [6] G. R. Krupka, K. Hausman, Description of the NetOwl Extractor System as Used for MUC-7. In Proceedings of the Fifth Message Understanding Conference (MUC-7), 1998
- [7] W.J. Black, F. Rinaldi, D. Mowatt, FACILE: Description of the NE System Used for MUC-7. In Proceedings of the Fifth Message Understanding Conference (MUC-7), 1998
- [8] 郑家恒, 张辉, 基于HMM的中国组织机构名自动识别. 计算机应用, 2002
- [9] 陈宁昱, 周雅倩, 黄萱菁, 吴立德, 利用未标注语料改进实体名识别性能. 中文信息学报, 2004
- [10] 秦文, 苑春法, 基于决策树的汉语未登录词识别. 中文信息学报, 2003
- [11] 郑家恒, 谭红叶, 基于变换的中文姓名识别技术探讨. 中文信息处理国际会议论文集, 1998
- [12] 刘椿年, 宋霞, 基于boosting的半结构化信息抽取. 北京工业大学学报, 2005
- [13] Cheng Niu, Wei Li, Jihong Ding, Rohini K. Srihari, A Bootstrapping Approach to Named Entity Classification Using Successive Learners. ACL, 2003
- [14] 张华平, 刘群, 基于角色标注的中国人名自动识别研究. 中文信息学报, 2002
- [15] Hongqiao Li, Chang-Ning Huang, Jianfeng Gao, Xiaozhong Fan, The Use of SVM for Chinese New Word Identification. In IJCNLP-04, 2004
- [16] Mitkov R, Anaphora resolution: the state of the art. University of Wolverhampton, 1999
- [17] Niyu Ge, John Hale, Eugene Charniak, A Statistical Approach to Anaphora

- Resolution. In Proceedings of the Sixth Workshop on Very Large Corpora, 1998
- [18] Joseph F. McCarthy, Wendy Lehnert, Using Decision Trees for Coreference Resolution. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995
- [19] 王厚峰, 何婷婷, 汉语中人称代词的消解研究. 计算机学报, 2001
- [20] 许敏, 王能忠, 马彦华, 汉语中指代问题的研究及讨论. 西南师范大学学报(自然科学版), 1999
- [21] C. Aone, M. Ramos-Santacruz, Rees: A large-scale relation and event extraction system. In Proceedings of the 6th Applied Natural Language Processing Conference, 2000
- [22] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel, Algorithms that learn to extract information-BBN: Description of the SIFT system as used for MUC. In Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998
- [23] D. Zelenko, C. Aone, A. Richardella, Kernel methods for relation extraction. J. Mach. Learn. Res, 2003
- [24] A. Culotta, J. Sorensen, Dependency tree kernels for relation extraction. In Proceedings of ACL, 2004
- [25] 车万翔, 刘挺, 李生, 实体关系自动抽取. 第一届全国内容安全与信息检索学术会议, 2004
- [26] 袁毓林, 用动词的论元结构跟事件模板相匹配——一种由动词驱动的信息抽取方法. 中文信息学报, 2005
- [27] 郑家恒, 王兴义, 李飞, 信息抽取模式自动生成方法的研究. 中文信息学报, 2004
- [28] 陈文亮, 朱靖波, 姚天顺, 张宇新, 基于Bootstrapping 的领域词汇自动获取. Proc. of JSCL, 2003
- [29] 郑家恒, 杜永萍, 刘昌钰, 基于语料的动态获取专业词汇方法初探. 计算机工程, 2002
- [30] SUI Zhifang, CHEN Yirong, HU Junfeng, WU Yunfang, YU Shiwen, The Research on the Automatic Term Extraction in the Domain of Information Science and Technology. In Proceedings of ACL, 2002
- [31] Patrick Pantel & Dekang Lin, Astatistical Corpus-Based Term Extractor. Cannadian Conference on AI 2001, 2001
- [32] Ellen Riloff, Rosie Jones, Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), 1999
- [33] Christopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing (141-177). MIT Press, 1999
- [34] J.C. Burges, a Tutorial on Support Vector Machines for Pattern Recognition. Bell Laboratories, Lucent Technologies, 1997

- [35] 吕映芝, 张素琴, 蒋维杜, 编译原理. 清华大学出版社, 2003
- [36] E Agchtein, Gravano, Snowball: Extracting relations from Large Plaintext Collections[C]. In: Proceedings of the 5th ACM International Conference on Digital Libraries, 2000

## 附录 A 地名实体词语识别标准

地名: ns

安徽/ns, 深圳/ns, 杭州/ns, 拉萨/ns, 哈尔滨/ns, 呼和浩特/ns,  
乌鲁木齐/ns, 长江/ns, 黄海/ns, 太平洋/ns, 泰山/ns, 华山/ns,

① 国名不论长短, 作为一个切分单位。

中国/ns, 中华人民共和国/ns, 日本国/ns, 美利坚合众国/ns, 美国/ns

② 地名后有“省”、“市”、“县”、“区”、“乡”、“镇”、“村”、“旗”、“州”、“都”、“府”、“道”等单字的行政区划名称时, 不切分开, 作为一个切分单位。

四川省/ns, 天津市/ns, 正定县/ns, 海淀区/ns, 东升乡/ns, 双桥镇/ns,  
大阪府/ns, 北海道/ns, 长野县/ns, 开封府/ns, 宣城县/ns

③ 地名后的行政区划有两个以上的汉字, 则将地名同行政区划名称切开, 不过要将地名同行政区划名称用方括号括起来, 并标以 ns。

[芜湖/ns 专区/n]ns, [宣城/ns 地区/n]ns, [内蒙古/ns 自治区/n]ns,  
[香港/ns 特区/n]ns, [广西/ns 环江/ns 毛南族/nz 自治县/n]ns,

④ 地名后有表示地形地貌的一个字的普通名词, 如“江、河、山、洋、海、岛、峰、湖”等, 不予切分。

鸭绿江/ns, 亚马逊河/ns, 喜马拉雅山/ns, 珠穆朗玛峰/ns, 地中海/ns

⑤ 地名后接的表示地形地貌的普通名词若有两个以上汉字, 则应切开。也要将地名同该普通名词用方括号括起来, 并标以 ns。

[台湾/ns 海峡/n]ns, [华北/ns 平原/n]ns, [帕米尔/ns 高原/n]ns

⑥ 地名后有表示自然区划的一个字的普通名词, 如“街、路、道、巷、里、町、庄、村、弄、堡”等, 不予切分。

中关村/ns, 长安街/ns, 学院路/ns, 景德镇/ns, 吴家堡/ns

⑦ 地名后接的表示自然区划的普通名词若有两个以上汉字, 则应切开。也要将地名同自然区划名词用方括号括起来, 并标以 ns。

[米市/ns 大街/n]ns, [蒋家/nz 胡同/n]ns, [陶然亭/ns 公园/n]ns

⑧ 大小地名相连时的标注方式为:

北京市/ns 海淀区/ns [南/f 大街/n]ns [蒋家/nz 胡同/n]ns 24/m 号/q

## 附录 B 部分公安领域词汇

Domain_Word	Domain_Word	Domain_Word
案犯	案件	案卷
案例	案情	案由
案子	暗杀	扒窃
罢免	败类	败诉
班房	颁布	办案
办理	帮凶	绑匪
绑架	包庇	包藏
保释	保释金	保证人
报案	报复	报警
报警器	报请	报送
暴力	暴徒	暴行
爆炸	爆炸性	备案
被捕	被告	被告人
被害	被害人	被害者
逼供	匕首	笔录
庇护	边防部队	边境地区
辩护	辩护人	表决权
剥夺	剥夺政治权利	驳回
搏斗	捕获	捕杀
不法之徒	不服	步枪
部长	部队	财产
财产权	财物	裁定
裁定书	裁决	裁判
裁判权	裁判员	残害
惨案	惨剧	惨遭
藏匿	藏身	操纵
草案	策划	查办
查出	查处	查封

## 附录 C 用于识别时间表达式的部分规则集

### DATE\_ITEM:

**Dy** ({D1} {D9}| {D2} {D0})? ({D1}| {D2}| {D3}| {D4}| {D5}| {D6}| {D7}| {D8}| {D9}| {D0})(年)

**Dm** ({D1}| {D2}| {D3}| {D4}| {D5}| {D6}| {D7}| {D8}| {D9}| {Da})| ({Da} ({D1}| {D2}))| ({D1} ({D0}| {D1}| {D2}))(月)

**Dd** ({D1}| {D2}| {D3}| {D4}| {D5}| {D6}| {D7}| {D8}| {D9}| {Da})| ({D1}? {Da}| {D1})({D1}| {D2}| {D3}| {D4}| {D5}| {D6}| {D7}| {D8}| {D9})| ({D1}| {D2}| {D3})({Da}| {D0})| ({D2}{Da}? ({D1}| {D2}| {D3}| {D4}| {D5}| {D6}| {D7}| {D8}| {D9}))| ({D3}{Da}? {D1})(日)

### NO\_DATE:

**nodate** (每| 近)({Dy}| {Dm}| {Dd})| ((每| (个)? (周| 星期) ({D1}| {D2}| {D3}| {D4}| {D5}| {D6}))

### REL\_DATE:

**Date** ({Dy} {Dm} {Dd})| ({Dy} {Dm})| ({Dm} {Dd})| {Dy}| {Dm}| {Dd}| ((周| 星期) ({D1}| {D2}| {D3}| {D4}| {D5}| {D6}))

## 个人简历 在读期间发表的学术论文与研究成果

### 个人简历:

乔春庚, 男, 1981 年 7 月生。

2000 年 7 月毕业于河北工业大学 计算机科学与技术专业 获学士学位。

### 已发表论文:

- [1] 乔春庚, 肖诗斌, 孙丽华, 施水才. 规则与统计相结合的案件名称识别. 第三届学生计算语言学研讨会, 2006