

浙江大学计算机学院

硕士学位论文

公安犯罪案件文本挖掘关键技术研究

姓名：程春惠

申请学位级别：硕士

专业：计算机应用技术

指导教师：何钦铭

20100101

摘要

由于信息技术的快速发展,公安信息系统中积累了海量的业务信息。面对着日益庞大的公安信息量,迫切需要应用人工智能的相关技术,对数据进行深层次的分析并研究各类信息的规律和关系,以更好地打击犯罪、防控犯罪。因此,将数据挖掘技术有效地应用于犯罪分析是目前公安工作的迫切需要。文本挖掘技术是近几年来数据挖掘领域的一个新兴的分支。而在海量案件信息中,除了规范化程度很强的数据库数据外,还有大量的案件叙述性文本描述。对这些大量的案件文本进行相关文本挖掘技术研究和应用是非常有意义的。

本文主要针对公安领域中大量的犯罪案件文本信息,对其中的案情文本挖掘相关技术进行研究和应用。本文论文的工作包括以下几点:

(1) 在文本预处理方面。结合实际应用需要,对公安业务中的一些术语建立专业词库;同时针对案件文本的特征,提出了具有针对性的特殊预处理方法。

(2) 在案件特征选择方面。根据实际应用的需求,研究了六种特征选择算法,并通过比较六种特征选择算法,确定了对案情文文本挖掘有利的特征选择算法。

(3) 在案件分类挖掘方面。提出了案件属性信息抽取方法和同义词语义分析方法,并在此基础上提出了改进的案件相似度计算方法;根据犯罪案件文本类别不均衡的特征,改进了朴素贝叶斯中的多变量贝努里模型,提出了面向不均衡类别的改进朴素贝叶斯案件文本分类方法。

(4) 在应用系统设计方面。设计和实现了一个典型三层 C/S 结构的犯罪案件文本挖掘系统,实现了相似犯罪案件文本检索模块和犯罪案件文本分类模块。

关键词: 文本挖掘, 文本分类, 文本相似度计算, 数据挖掘, 犯罪挖掘, 中文分词, 特征选择

Abstract

Due to the rapid development of information technology, public security information system has accumulated vast amounts of business information.

In the face of increasingly large amount of police security information, we urgently need AI related technologies which analysis of the data in-depth, research the laws of various kinds of information and relationships in order to better combat crime, crime prevention and control. Therefore, data mining technology is effectively applied to crime analysis is the urgent need for public security work.

Text mining technology is a emerging branch of data mining for the past few years. In the massive case information, in addition to a strong degree of standardization of the database data, there are a large number of cases of narrative text descriptions. Text mining technology research and application on the massive text-case information is very meaningful.

In this paper, we do some research and application of text mining technology on the massive text-cases. This paper's work includes the following:

(1) In the text pre-processing aspects. Combination of practical application, this paper establishes professional police terminology thesaurus and explores the special text preprocessing method according to the feature of case text.

(2) In the case feature selection aspects. According to the needs of practical applications, this paper researches the six kinds of feature selection algorithm. And by comparing the six kinds of feature selection algorithm, this paper determines the most useful feature selection algorithm to criminal text mining.

(3) In the criminal-case text mining aspects, this paper proposes an improved case-texts similarity calculation method based on the cases-attribute information extracted, combined with the synonyms semantic analysis method; This paper also proposes a improve criminal text classification Of unbalanced classes method based on Naive Bayes. An improved model based on multi-variate Bernoulli model of Naive Bayes is proposed due to the unbalanced distribution of criminal case categories.

(4) Design and implementation of the criminal case text mining system. This paper constructs the criminal case text mining system base on a typical C/S structure. The system implements the similar criminal case-texts retrieval module and text classification model.

Key words: Text Mining, Text Categorization, Text Similarity Computing, Data Mining, Crime Data Mining, Chinese Word Segmentation, Feature Selection

图目录

图 2.1 文本挖掘的一般流程	7
图 2.2 文本相似度计算的流程图	11
图 2.3 文本分类一般流程图	14
图 3.1 犯罪案件文本挖掘的一般流程	18
图 4.1 犯罪特征数据挖掘系统框架图	30
图 4.2 犯罪案件文本挖掘子系统体系结构	32
图 4.3 蜘蛛爬虫类图	34
图 4.4 案情网页 XML 格式的主要框架示例	35
图 4.5 标题, 时间, 内容正则表达式	35
图 4.6 案情通告内容抽取类图	36
图 4.7 相似案件文本检索模块结构设计图	37
图 4.8 相似案件文本检索模块组件图	38
图 4.9 案件属性信息抽取系统实现	40
图 4.10 数据源选择为“犯罪数据库”的相似犯罪案件文本检索结果	41
图 4.11 数据源选择为“网站文本数据库”的相似犯罪案件文本检索结果	42
图 4.12 案件文本分类模块结构设计图	43
图 4.13 案件文本分类模块组件图	44
图 4.14 数据源为“犯罪数据库”的犯罪案件文本自动分类结果	48
图 4.15 数据源为“犯罪数据库”的测试案件文本内容	49
图 4.16 数据源为“网站文本数据库”犯罪案件文本自动分类	50
图 4.17 数据源为“网站文本数据库”的测试案件文本内容	50

表目录

表 4.1 样本集一采用多项式模型，多变量贝努里模型和改进的多努里模型的
分类结果 46

表 4.2 五个样本集采用多项式模型，多变量贝努里模型和 46
改进的多变量贝努里模型的分类结果 46

表 4.3 改进多变量贝努里模型的比较试验结果 47

致谢

时光飞逝，日月如梭，转眼间两年半研究生学习生涯过去了。在此论文完成之际，我首先要衷心感谢我的导师何钦铭教授。何老师严谨的治学态度、渊博的知识理论以及诲人不倦的育人精神让我铭记。从何老师身上学到不仅仅是如何勤勤恳恳做学问，更重要的是学会如何踏踏实实做人，这些都将使我一生受益无穷。我还要感谢老师为我们精心营造了一个开放自由的研究环境，在这里，大家相互交流、相互帮助、相互鼓励，得到共同进步的机会。

最后，感谢评阅、评议论文和答辩委员会的各位专家学者在百忙的工作中能给予指导。

程春惠

2010年1月于浙大

第1章 绪论

1.1 背景和研究意义

随着经济的发展和信息技术的深入应用,公安信息系统中积累了海量业务信息。案件信息达数百万条,且每年以100至120万条速度递增。目前公安部门面临的一个主要问题就是如何对日益增长的包含涉案人员,涉案物品,户籍,简要案情文本等信息数据的大量案件进行准确和有效的分析。虽然,目前随着公安信息化的发展,信息系统得到了推广和应用,在打击违法犯罪,维护社会稳定方面发挥了一定作用。但是对信息的处理还基本上停留在查询,统计等传统方法上。面对大量的数据,依靠现有的技术和系统,很难发现其中的隐藏的联系并找出对破案有价值的线索。

因此,面对复杂的犯罪形势,面对日益庞大的公安信息量,迫切需要应用人工智能相关技术,对数据进行深层次的分析、研究各类信息的规律和关系、进一步挖掘各类信息的作用,以更好地打击犯罪、防控犯罪。因此,将数据挖掘技术有效地应用于犯罪分析是目前公安工作的迫切需要。

数据挖掘技术是从海量的数据中抽取或挖掘隐含的、事先未知、潜在有用的信息和知识的重要方法和途径^[1]。数据挖掘已经在很多领域中得到了成功的应用,例如在金融业、零售业、医疗、电信、航空等领域都已经得到了广泛的应用^[2]。同样,传统的数据挖掘技术如关联分析、分类、预测、聚类分析都很好地应用于公安犯罪信息领域。利用数据挖掘技术可以从大量犯罪记录中有效和快速地发现犯罪趋势、破案线索等,从而能为公安部门提供有效的决策支持。

在海量案件信息中,除了规范化程度很强的数据库数据外,还有大量的案件叙述性文本描述,包括犯罪数据库中的自由文本案情描述和公安内部网络上的案情公告。对这些大量的案件叙述性文本进行相关数据挖掘技术研究和应用是非常有意义的。

文本挖掘作为近几年来数据挖掘领域的一个新兴的分支,涵盖多种技术,包

括信息抽取、信息检索、自然语言处理和数据挖掘技术^[3]。主要着力于帮助用户从来源于web或者数据库中的大量的非结构化或者半结构化的数字化文本文档中获得用户感兴趣或者有用的模式^[4]。

目前,文本挖掘在多个领域中得到了应用,包括:在信息检索领域中的应用^[5]、在科技情报中的应用^[6]、在互联网信息统计中的应用^[7]、在专利文献信息中的应用^[8]、在医学领域中的应用^[9]等。

面对海量的案情文本信息,文本挖掘技术是非常有用的技术,它能从这些海量的案情文本信息中挖掘中隐藏的、对公安业务人员有用的信息。例如:通过文本聚类技术能够挖掘出相似案件从而有利于破案;通过信息抽取技术能够从文本中自动抽取出人名、地名、作案工具、作案物品等信息;通过文本分类技术与公安内部网络信息检索技术相结合,有利于公安业务人员快速定位有用的案情。因此,文本挖掘能够为公安业务提供有效的决策支持,不仅能提高犯罪信息分析的质量和效率,还能有效支持公安系统更好地打击犯罪、防控犯罪、提高公安快速响应能力与作战能力。

1.2 公安犯罪文本挖掘的发展和现状

目前,国内外都在进行深入地研究和探讨公安信息领域中的数据挖掘理论方法和技术研究。传统的数据挖掘技术如关联分析、分类、预测、聚类分析都很好地应用于公安犯罪信息领域。具体的应用如:

- 聚类分析,可用来分析识别具有相似犯罪行为的犯罪嫌疑人^[10];
- 独立点分析,分析数据中出现的一些反常或不满足规则的特例,通常用于网络入侵检测等犯罪分析^[10];
- 关联规则,发现数据库中的频繁项集并挖掘出隐藏在数据库中的关联规则,通常用于网络入侵检测。从入侵者的交互历史中获取关联规则,从而预测未来可能的网络攻击^[11]。
- 社会关系网络分析,通过构建由犯罪嫌疑人之间的角色和关系组成的社会关系网络,分析该网络可挖掘出关键人物以及犯罪团伙等^[12]。

- 趋势预测，通常是通过建立连续值函数的模型来预测数据趋势，预测各类案件的发生趋势，从而用来辅助决策并提供实时的预警^[11]。

而针对大量非结构化或半结构化的案件叙述性文本描述的文本挖掘技术也得到了广泛的研究和应用。以下是国内外犯罪文本挖掘发展和现状的情况。

1.2.1 国外犯罪文本挖掘的发展和现状

目前，在国外，文本挖掘技术也很好地应用于公安犯罪信息领域。以下是具体的文本挖掘技术在犯罪领域中的应用：

(1) 信息抽取(Entity extraction)

2002 年，Michael Chau, Jennifer J.Xu 等人将信息抽取技术用于从案件叙述性文本中自动识别出人名、地名、作案手段、作案工具等^[13]。信息抽取一般作为犯罪数据挖掘的基础，信息抽取后可使用其他数据挖掘方法进行犯罪分析^[11]。信息抽取技术也在 2003 年，Hsinchun Chen, Wingyan Chung, 等人在 COPLINK 项目^[14]中得到了应用。

(2) 文本比较(record comparison algorithm)

2006 年，Wang, G., Chen, H. 等人利用字符串比较方法检测以往数据库案件文本记录中相同的诈骗信息。从而实现同一犯罪诈骗的识别^[15]。

(3) 文本分类

文本分类算法是给定类别体系的前提下，根据文本的内容自动判别文本的类别。2007 年，S. Appavu alias Balamurugan, Ramasamy Rajaram 将基于决策树的文本分类方法应用于 e-mail 分类系统中，在截获的电子邮件中通过文本自动分类发现含有犯罪行为的邮件，从而挖掘出犯罪嫌疑人或者犯罪组织结构^[16]。

此外，文本挖掘还有文本自动摘要，模式识别^[17]，文本聚类等技术在犯罪领域中的研究和应用。

1.2.2 国内犯罪文本挖掘的发展和现状

目前，国内研究学者也对犯罪文本挖掘进行了研究。具体的犯罪文本挖掘相

关研究有:

- 邮件的自动分类。通过对可疑人员的电子邮件进行监控,对截获的电子邮件数据进行处理,实现对犯罪组织的结构挖掘^[18]。
- 文本分类和聚类在出入境管理部门的应用。根据入境人员的犯罪记录将入境人员分为高度危险分子、普通危险分子和一般人员等,从而有利于公安部门决定重点审查对象^[19]。

此外还有基于社会网络的犯罪组织关系挖掘^[20]、文档自动摘要^[21]等相关犯罪文本挖掘技术的研究和应用。

1.3 本文研究内容和贡献

本文的研究目的是:研究犯罪案件文本挖掘相关技术,并将文本挖掘相关技术应用于公安信息领域,从而为串并案业务人员提供帮助、为公安业务提供有效的决策支持,以提高公安快速响应能力与作战能力。

本文的研究对象是案件文本。案件文本数据源主要来自于两部分:一是来自现有犯罪数据库中的自由文本案情描述;二是来自公安内部网络上的案情公告。

本文研究的主要关键技术有:中文分词技术,特征表示,特征选择,信息抽取,文本相似度计算,文本分类。具体研究内容、贡献以及创新如下:

(1) 中文分词和预处理

结合实际应用需要,加入公安领域词汇,并对中科院 ICTCLAS 分词组件的结果进行修正;针对案件文本特征,提出了具有针对性的特殊预处理方法。

(2) 特征选择

通过比较六种特征选择算法,选择并确定了对案情文文本挖掘有利的特征选择算法。

(3) 信息抽取

通过分析犯罪案件应用的实际需要,将案件文本进行属性信息抽取,抽取如下属性信息:作案时间、作案地点、涉案人员、作案手段、作案工具、损失物品、损失金额。

(4) 文本相似度计算

研究犯罪案件文本相似度计算,提出了案件文本之间相似度的计算方法,以作为相似案件文本检索的基础;本文还针对犯罪案件文本的特征,提出了案件属性信息抽取方法和同义词语义分析方法,并在此基础上提出改进的案件文本相似度计算方法。

(5) 文本分类

研究文本分类的相关算法,设计针对犯罪案件文本特征的文本分类算法。案件文本分类可以方便用户快速定位到相关的案件类别信息,并有利于相关案件的串并。本文针对案件文本以及文本类别的特征,提出面向不均衡案件类别的改进的朴素贝叶斯文本分类算法。

(6) 设计和实现文本挖掘系统

实现基于案件属性信息抽取的相似犯罪案件文本检索组件以及犯罪案件文本自动分类组件。

1.4 论文组织形式

第一章:绪论。介绍公安信息领域中数据挖掘和文本挖掘的背景和相关研究,探讨了公安犯罪数据挖掘相关技术的发展和现状。

第二章:文本挖掘技术概述。介绍了文本挖掘的一般流程,中文分词技术,特征选择,文本相似度计算,文本分类相关技术。

第三章:介绍了犯罪案件文本挖掘关键技术。中文分词和预处理:结合实际应用需要,对中科院 ICTCLAS 分词结果进行改进。针对案件文本特征,提出了具有针对性的特殊预处理方法;提出了案件属性信息抽取方法和同义词语义分析方法,并在此基础上提出改进的案件文本相似度计算方法;提出面向不均衡案件类别的改进的朴素贝叶斯文本分类算法。

第四章:犯罪案件文本分类系统的设计与实现。基于第三章介绍的犯罪案件文本挖掘关键技术的方法,设计并开发了犯罪案件文本挖掘系统,主要包含相似犯罪案件文本检索组件和犯罪案件文本分类组件,并给出了实验分析结果以及系

统运行的示例。

第五章：结论和展望。

第2章 文本挖掘相关技术概述

本章将对本文研究中涉及到的技术进行概述，包括：文本挖掘的一般流程、中文分词技术的常用方法、特征表示方法、特征选择的常用算法、文本相似度计算的一般流程和常用算法、文本分类的一般流程和常用算法等。

2.1 文本挖掘的一般流程

文本挖掘一般包含中文分词与处理，特征选择，文本挖掘相关技术，结果评价等几个过程^[22]，具体的一般流程如下图 2.1:

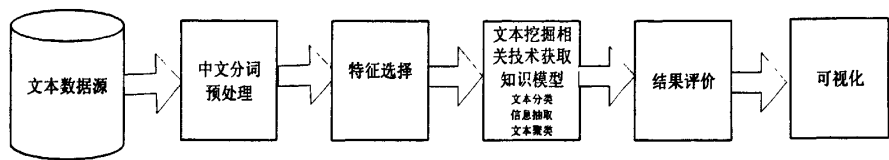


图 2.1 文本挖掘的一般流程

首先对文本数据库的文本进行中文分词和预处理。中文分词和预处理后，将文本表示成计算机能够理解的数字形式，最常用的是向量空间模型方法。根据处理速度和精度的要求，对文本中的特征进行特征选择。然后采用文本挖掘方法获取隐藏的知识模型，文本挖掘方法有文本分类，信息抽取，文本聚类等方法。接着是对文本挖掘方法的评价，最后将文本挖掘获取的知识模型以可视化的形式输出，从而实现指导人们日常实践和工作。

2.2 中文分词

中文分词技术是能将中文本中的词语正确切分开的一种技术。分词技术是计算机处理中文文本的第一步，因此文本挖掘的基础。目前的中文分词算法主要分为三大类：基于词典的方法，基于统计的方法和基于规则的方法^[23]。

1. 基于词典的分词算法

这种方法的主要思想是：按照一定的策略从文本中取词条，并将待分析的词条与词典中的词进行匹配。如果在词典中找到该词，则匹配成功，否则匹配不成

功。按照扫描方向的不同,该分词方法可以分为正向匹配和逆向匹配;按照长度的不同,可以分为最大匹配和最小匹配^[24]。

其中正向最大匹配算法思想^[24]是:每次按照从左到右的顺序从文本中取长度为词典中最大词长的子串,与词典中的词进行匹配,如果成功,则该子串为词,算法接着匹配余下的文本。如果不成功,则子串长度逐次减一进行匹配。逆向最大匹配算法的基本原理与正向最大匹配算法类似,只是分词的扫描方向不同,该算法是从右向左取子串。

基于词典的分词算法优点是简单易实现,缺点是,精确度不高,词典构造困难。

2. 基于统计的分词算法

目前基于统计的分词算法的主要思想是,首先切分出词表匹配的所有可能的词,通过运用统计语言模型和决策算法决定最有的切分效果。较为常见的算法是,基于互信息的概率统计算法, N-Gram 算法,基于组合度的汉语分词决策算法等等^[23]。

它的优点在于能够发现可以发现所有的切分歧义,缺点在于需要大量的标注预料,分词速度较慢^[23]。

3. 基于规则的分词算法

基于规则的分词方法的基本思想是在分词的同时进行句法、语法分析,利用上下文内容所提供的句法信息和语义信息来对文本进行分词^[23]。这种分词方法优点在于它可以在实例中进行自动推理和证明,可以实现歧义处理和自动补充未登录词,缺点在于需要大量的语言知识,并且汉语语言知识非常复杂,很难讲各种语言信息组成计算机可以理解的形式^[23]。

通常情况下,对于一个成熟的分词系统,是采用混合型的分词算法,即采用几种分词算法相结合的方法。

本文采用中科院分词系统 ICTCLAS^[25]的 .net 版本 NICTCLAS 对文本进行分词,该系统的功能有:中文分词、词性标注、未登录词识别。该系统是基于层叠隐马模型的结合词典方法和统计方法的汉语词法分析系统^[25]。中科院分词正确率

可高达 97.58%，基于角色标注的未登录词识别能取得高于 90% 召回率，其中中国人名词的识别召回率接近 98%，分词和词性标注处理速度为 543.5KB/s^[26]。ICTCLAS1.0 版本开源，并已经广为流传，一些商业项目也使用它来分词。

2.3 特征表示

要进行文本挖掘，首先要将文本表示成计算机能够理解和处理的数字形式，才能进行分析和处理。因此，文本表示是文本挖掘的重要前提。常用的文本表示有向量空间模型，概率模型和语言模型^[27]。

常用的文本表示模型是向量空间模型 (VSM)。空间向量模型采用了独立性假设，将文本看成是相互独立的词条组($T_1, T_2, T_3, \dots, T_n$)构成，而($W_1, W_2, W_3, \dots, W_n$)为对应每个词条的权值。一个文本 d 表示成的向量模型如下：

$$V(d) = (t_1, w_1(d); \dots; t_i, w_i(d); \dots; t_n, w_n(d)) \dots \dots \dots \text{公式(2.1)}$$

其中 t_i 为词条项， $w_i(d)$ 为 t_i 在 d 中的权值， n 是特征项的维数。这样每个文本就被映射到多维空间中的一个点。权值一般采用布尔型或者词频型。布尔型考察特征词是否在文本中出现，如果出现则权值为 1，反之为 0。词频型考察特征词在文本中出现的次数，权值则为特征词出现的次数。

2.4 特征选择

在进行文本挖掘的时候，特征向量维度通常都非常大，这样往往会降低文本挖掘的效率和质量，因此需要进行特征选择从而降低向量维度。特征选择是从原始特征向量空间中选择部分最能反映模式类别统计特征的相关特征。用于特征选择的方法主要有：文档频率方法 (DF)、互信息(MI)、信息增益(IG)、X2 统计量 (CHI)、期望交叉熵(ECE)、文本证据权(WET)等^[28]。这些方法的基本思想都是通过设定一个阈值，然后对每一个特征词计算其统计度量值，最后取度量值于大于阈值的那些特征词作为有效的特征词。

下面公式中： t 代表特征项， C_j 代表第 j 个类别， m 为类别数， P 代表概率， $P(t, C)$ 为包含词 t 且属于类别 C 的文档在文档集合中出现的概率， $P(t)$ 表示词 t 出

现的概率, $P(C_j)$ 表示类别 C_j 在文档集合中出现的概率。 $P(C_j | t)$ 表示在出现词 t 的情况下, 文档属于第 j 类的概率。 $P(t | C_j)$ 表示在词 t 在类别 C_j 出现的概率。

1. 文档频率 (DF) 的计算公式:

$$DF(t) = \frac{\text{出现特征词 } t \text{ 的文档数}}{\text{训练集中的文档数}} \dots\dots\dots \text{公式(2.2)}$$

通过设定阈值, 选取 DF 较大的且大于阈值的特征词构成文本的向量。

2. 互信息量 (MI) 的计算公式:

$$MI(t) = \sum_j^m p(C_j) \log \frac{p(t | C_j)}{p(t)} \dots\dots\dots \text{公式(2.3)}$$

MI 互信息量度量了类和特征词之间的关联信息。

3. 信息增益 (IG) 的计算公式:

$$IG(t) = p(t) \sum_{j=1}^m p(C_j | t) \log \frac{p(C_j | t)}{p(C_j)} \dots\dots\dots \text{公式(2.4)}$$

公式(2.4), 不考虑特征词不出现的情况。特征词的信息增益值越大, 对分类越重要^[29]。通过设定阈值, 选取信息增益较大的且大于阈值的特征词构成文本的向量。

4. X^2 统计量 (CHI) 的计算公式:

$$T^2(G, t) = \frac{(AD - CB)^2 \times N}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \dots\dots\dots \text{公式(2.5)}$$

其中, A 表示属于类别 C_j 且包含特征词 t 的文档频率, B 表示不属于类别 C_j 但包含词条 t 的文档频率, C 表示属于类别 C_j 但不包含词条 t 的文档频率, D 表示既不属于类别 C_j 也不包含 t 的文档频率, N 表示训练集的文本数。 x^2 统计值越大的特征项与类别之间的独立性就越小, 对分类的贡献就越大^[29]。通过设定阈值, 选取 x^2 统计值越大的词作为特征词。

5. 期望交叉熵(CE) 的计算公式:

$$CE(t) = p(t) \sum_{j=1}^m p(C_j | t) \log \frac{p(C_j | t)}{p(C_j)} \dots\dots\dots \text{公式(2.6)}$$

交叉熵反映了文本类别的概率分布和出现了某个特征词的条件下文类别的概率分布之间的距离。词 t 的交叉熵越大, 对文本分类分布的影响也越大。

6. 文本证据权(WET)的计算公式:

$$WET(t) = p(t) \sum_{j=1}^m p(C_j) \left| \log \frac{p(C_j|t)(1-p(C_j))}{p(C_j)(1-p(C_j|t))} \right| \dots\dots\dots \text{公式(2.7)}$$

2.5 文本相似度计算

文本相似度计算，是将文本转化成计算机所能处理的数据形式，并用相似度算法计算两个文本的相似程度。文本相似度计算在信息检索，文本聚类，文本分类中都起着重要的作用。

2.5.1 文本相似度计算的一般流程

文本相似度计算一般包含中文分词与处理，特征选择，将文本表示成向量空间模型，结合 IF-IDF 计算向量的权重，再利用相似度计算算法计算文本的相似度，最后得到结果等几个过程，具体的一般流程如下图 2.2:

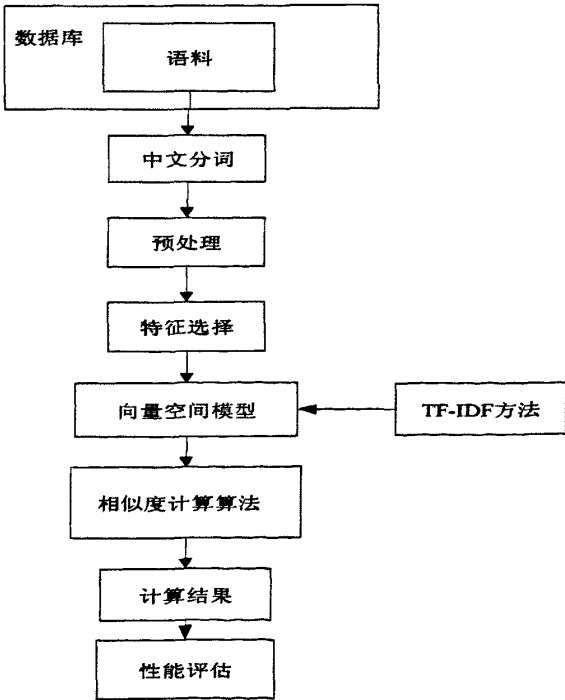


图 2.2 文本相似度计算的流程图

由上图 2.2 所示, 要对文本进行相似度等分析, 首先需要将文本转化为计算机所能处理的数据形式, 通常采用 2.3 节特征表示中的向量空间模型。经过中文分词, 预处理, 以及特征选择后, 将文本表示成向量空间模型。这样, 每个文本就被映射都为空间中的一个点。计算两个文本之间的相似度就可以转化为计算两个点之间距离。

在向量空间模型的权值计算方法中, 比较常用的是 TF-IDF 方法。该方法综合考虑了词频 TF(Term Frequency)和逆文档频率 IDF(Inverse Document Frequency)。其计算公式:

$$w_{ij} = TF_{ij} \times IDF_j \quad \dots\dots\dots \text{公式(2.8)}$$

其中, w_{ij} 文档 D_i 中的第 j 个特征值权重。 TF_{ij} 表示单词 Term T_i 在文档 D_j 中的出现的次数。IDF 是逆文档频率。计算公式如下:

$$IDF_j = \log \frac{N}{DF_j} \quad \dots\dots\dots \text{公式(2.9)}$$

其中, DF_j 表示单词 T_j 的文档频率, 也就是单词 T_j 出现的文档的数目。 N 表示文章的总数。

利用 TFIDF 方法计算每个文档的特征项向量, 这样便可以计算文档的相似度。

2.5.2 文本相似度计算常用算法

当文档 D_1 和文档 D_2 分别对应特征项向量 V_1 和 V_2 后, 这两个文档的相似度就转换为两个向量之间相似度。向量 V_1 和 V_2 的相似度计算主要方法如下:

1. 余弦计算法:

$$\text{sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1| |v_2|} \quad \dots\dots\dots \text{公式(2.10)}$$

2. Jaccard 系数

$$\text{sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1|^2 + |v_2|^2 - v_1 \cdot v_2} \quad \dots\dots\dots \text{公式(2.11)}$$

3. 内积

$$sim(v_1, v_2) = v_1 \cdot v_2 \dots\dots\dots \text{公式(2.12)}$$

4. Dice 系数

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1|^2 + |v_2|^2} \dots\dots\dots \text{公式(2.13)}$$

其中内积 $v_1 \cdot v_2$ 为标准向量点积，定义为 $\sum_{i=1}^n v_{1i} v_{2i}$ ，分母中的范数 $|v_1|$ 定义为 $|v_1| = \sqrt{v_1 \cdot v_1}$ 。

余弦算法是最具有代表性的文本相似度算法，本文使用的是余弦算法。

2.6 文本分类

文本分类是对于待分类文本根据其内容，由计算机根据某种自动分类算法，把文本判定为预先定义好的类别。自动文本分类已被广泛地应用于邮件过滤、新闻过滤、用户偏好预测、文档组织等多个领域^[30]。

2.6.1 文本分类的一般流程

文本分类过程包含：语料库构建、中文分词与预处理、特征选择、文本特征向量表示、文本分类和分类性能评估阶段。文本分类的一般流程图如下图 2.3:

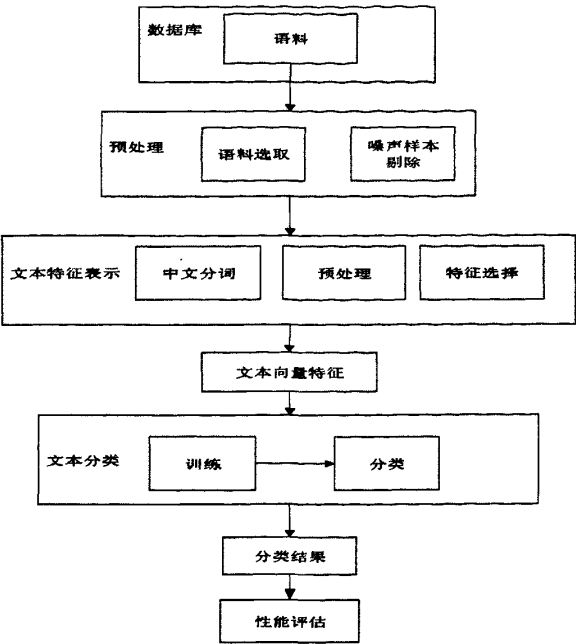


图 2.3 文本分类一般流程图

由上图 2.3 所示，文本分类一般包含训练和分类两个过程。得到对训练集进行训练从而获得文本分类所需的知识，再利用得到的知识对未知的文本进行分类。

2.6.2 文本分类常用算法

常用的文本分类算法有朴素贝叶斯方法、k-近邻方法、支持向量机方法、决策树、神经网络方法等^[31]。以下将重点讲述前三种算法：

1. K 近邻算法(K-Nearest Neighbor, KNN)

K 近邻算法考察和待分类文本最相似的 K 个训练样本点，根据这 K 个训练样本点的类别来判断待分类文本的类别值，即取未知样本 x 的 K 个近邻，分析比较这 K 个近邻多数属于哪一类，就把 x 归为哪一类^[32]。算法过程如下：

- 1) 将各个训练经过文本分词，预处理，特征词提取后，表示成特征向量；
- 2) 同样将待新文本表示成特征向量；
- 3) 选出训练文本集中与新文本最相似的 K 个训练文本。相似算法采用两向

量的余弦算法，余弦算法的公式见公式(2.10)。

4) 将文本分到 K 个训练样本点出现频率最大的类别；

2. 贝叶斯算法(Naive Bayes, NB)

贝叶斯算法的主要思想是在给定待分类文本的条件下，计算其属于各个类别的条件概率，然后选择其中条件概率最高的类别作为该文本所属的类别^[30]。算法过程如下：

1) 对于某个测试文本 d ，计算该文本属于类别 C_j 的概率：

$$p(C_j|d) = \frac{p(C_j)p(d|C_j)}{p(d)} \dots\dots\dots \text{公式(2.14)}$$

$$p(d) = \sum_{j=1}^{|C|} p(C_j)p(d|C_j) \dots\dots\dots \text{公式(2.15)}$$

其中 $p(C_j)$ 表示类别 C_j 的概率。 $p(d|C_j)$ 表示文档 d 属于类别 C_j 的概率。

2) 最后，将文本分到概率 $p(C_j|d)$ 最大的那个类别中。

$$V_{\max} = \arg \max_{q \in c} p(C_j|d) \dots\dots\dots \text{公式(2.16)}$$

朴素贝叶斯分类方法是目前公认的一种简单有效的分类方法，并且它在文本分类领域表现出令人满意的性能^[31]。本文使用的文本分类算法是在朴素贝叶斯方法上进行改进的。

3. 支持向量机(Support Vector Machine, SVM)

算法支持向量机(SVM)是建立在统计学习理论上发展而来的一种机器学习方法，它基于小样本学习，结构风险最小化等统计学习原理，将原始数据集压缩到支持向量集合，学习得到分类决策函数^[33]。其基本思想是构造一个超平面作为决策平面，使正负模式之前的间距最大。

2.7 本章小结

本章内容首先介绍了文本挖掘的一般流程，然后对项目所涉及到的主要技术进行概述。主要介绍了以下几种技术：

(1)中文分词，目前的中文分词算法主要分为三大类：基于词典的方法，基于统计的方法和基于规则的方法。

(2) 特征表示，主要介绍了常用的向量空间模型。

(3) 特征选择，主要介绍了文档频率方法 (DF)、互信息(MI)、信息增益(IG)等六种特征选择方法。这些方法的基本思想都是通过设定一个阈值，然后取度量值于大于阈值的那些特征词作为有效的特征词。

(4) 文本相似度计算，主要文本相似度计算的的一般流程，介绍了向量空间模型和 TF-IDF 权重计算方法，以及包含余弦算法在内的文本相似度计算常用的算法。

(5) 文本分类 介绍了文本分类的一般流程，以及常用的分类算法，包括朴素贝叶斯方法、k-近邻方法，支持向量机方法等。

第3章 犯罪案件文本挖掘关键技术

本章在介绍犯罪案件文本数据源及犯罪案件文本挖掘的一般流程的基础上分析犯罪案件文本挖掘的关键技术。主要包含适合犯罪案件文本的中文分词方法、犯罪案件文本的特殊预处理方法、在传统文本相似度计算方法基础上的改进犯罪案件文本相似性比较方法、在传统文本分类方法基础上的改进犯罪案件文本分类方法。

3.1 犯罪案件文本数据源

1、犯罪案件文本数据来源

本文的案件文本主要来自于两部分：一是来自现有犯罪数据库中的自由文本案情描述。二是来自公安内部网络上的案情公告。

犯罪数据库的自由文本案情描述，主要来自案件基本情况主表。该表是由业务人员在日常工作中记录下来的案件信息，其中有“报警内容或简要案情”的自由文本案情描述。例如：“2006年1月23日，报案人张三年来分局报案称：在本区××镇××村自己家中，22日晚停的一辆轻便二轮摩托车被人偷走，价值人民币1000多元。”

公安内部网络上的案情公告是由各地公安部门发布在网上的、即时的案件信息，为各地公安情报分析人员进行案件串并提供线索。例如：“2009年1月1日上午，张三报案称：1月1日晚8点至2日上午8点，城关镇某大酒店二楼餐厅收银台门被撬，抽屉内财物被盗。被盗现金22779元，还有香烟，总计价值3万多元。”

2、犯罪案件文本的特征

不管是犯罪数据库的自由文本案情描述还是公安内部网络的案情公告，这些案件文本都具有以下的特征：

(1) 文本篇幅短小，属于短文本类型。现有案件文本长度主要在50~200字之间，属于短文本类型。由于文本短小，单词出现频率低。

(2) 包含大量案件属性信息。一个案件文本主要包含以下属性信息：作案时间，作案地点，涉案人员，作案手段，作案工具，损失物品，损失金额等。

3.2 犯罪案件文本挖掘的一般流程

由于犯罪数据库的自由文本案情描述及公安内部网络的案情公告，具有以上共同的特征，因此两种数据源在文本挖掘过程中具有共性。

根据公安部门案件串并分析的需要，本文重点研究为串并案业务人员提供文本挖掘的相关功能。犯罪案件文本挖掘的一般流程如下图 3.1:

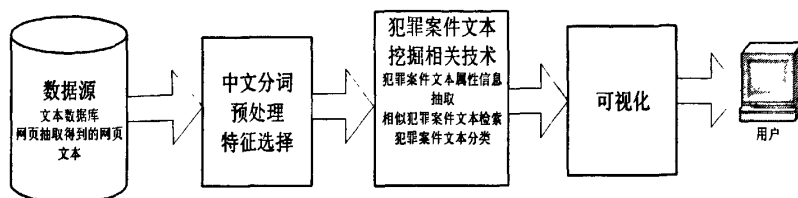


图 3.1 犯罪案件文本挖掘的一般流程

犯罪案件挖掘的一般流程与传统的文本挖掘一般流程相似。在犯罪案件文本挖掘相关技术中，主要包含三种技术：犯罪案件文本属性信息抽取，相似犯罪案件文本检索，犯罪案件文本分类。具体描述如下：

(1) 犯罪案件文本属性信息抽取是从案件文本中自动识别出人名，地名，作案手段，作案工具等属性，抽取的结果主要用于相似犯罪案件文本检索。

(2) 相似犯罪案件文本检索实现了检索出与给定案件文本相似的案件文本集的方法，并对这些案件文本集按相似度大小排序。

(3) 犯罪案件文本分类在给定犯罪案件类别分类体系的前提下，实现根据案件文本的内容自动判别文本所属的案件类别的方法。

3.3 中文分词及预处理

3.3.1 中文分词及改进

本文采用中科院分词系统 ICTCLAS^[25]的.net 版本 NICTCLAS 对文本进行分词，NICTCLAS 分词系统具有词性标注功能。本项目还根据公安领域特征对

ICTCLAS 分词结果进行了改进。

加入公安专业词汇。由于公安领域的很多词汇如“故意伤害”、“使用假证”、“非法持有假币”等词在案件文本中出现频繁，具有语义特征，但是 NICTCLAS 分词组件却无法精确切分出这些词。因此，本项目对该分词进行了改进，建立了针对公安领域的专业词汇的词库。将自定义词库的词加载到分词组件中去，有效地改进分词的效果。

3.3.2 预处理

1. 去停用词

停用词是在文本集合中出现频率很高，但是无用的词，如“的”、“了”、“还是”、“或者”等。去除这些词可以降低特征词的维度，同时可以提高挖掘效果。

此外，对于案件文本分类，案件文本常常包含出现频率非常高但对文本分类无关的词，例如“犯罪嫌疑人”、“受害人”、“价值”、“报案”等，即公安领域的专用停用词。因此，本文增加了针对公安案件领域专用停用词的处理功能，建立了公安案件领域专门停用词表，即公安领域的专用停用词。

2. 去除单字

一般来说，单字对于案件相似度比较的作用不大。需要去除单字。

3. 根据词性剔除对案件中无用的词。

一个案件文本主要包含以下信息：作案时间、作案地点、涉案人、作案手段、作案工具、损失物品、损失金额等。根据 ICTCLAS 分词的词性标注信息，剔除与案件属性无关的词性，如拟声词、副词、介词、连词等。

而对于文本分类，还需要剔除案件文本属性中对案件文本分类无用的属性，具体的是：作案时间属性、涉案人员属性，损失金额属性，这些属性对于一个案件属于哪一个案件类是没有作用的。这些属性所对应的词性分别为时间、人名和数词等。

3.4 相似犯罪案件文本检索

相似犯罪案件信息检索功能是：给定案件文本，计算该案件与数据库中每个案件文本之间的相似度，返回与给定案件文本相似的案件文本集合，为案情的串并提供支持。因此，相似犯罪案件文本检索就是在案件文本集合中寻找与待检索案件文本最相近的案件文本，因此其流程与传统的文本相似度计算的一般流程相似。

在本文中，案件文本特征表示采用 2.3 节特征表示中的向量空间模型（VSM）方法，一个特征向量为：

$$V(d) = (t_1, w_1(d); \dots; t_i, w_i(d); \dots; t_n, w_n(d)) \dots\dots\dots \text{公式(3.1)}$$

其中 t_i 为词条项， $w_i(d)$ 为 t_i 在 d 中的权值， n 是特征项的维数。

由于案件文本具有文本短的特点，词在一个文本中出现的频率低。所以本文在量化方法上，对文档中各个词组的权值计算，采用的是布尔模型方法。在该方法中文档被表示成特征词是否在文档中出现的二进制向量。如果特征词在文档中出现，则向量值为 1，反之为 0。

3.4.1 基于同义词的语义分析

语义分析是自然语言处理领域的一个概念，语义分析主要是对单词、词组、句子、句群所包含的意义和在语言使用过程中所产生的意义进行分析，它包含了词与词之间的同义和蕴含关系。

同义词的使用避免了重复用词，丰富了语言的表达，但也分散了同一概念的频率。例如“偷窃”的同义词有“扒窃”、“窃取”等，当交替使用“偷窃”的同义词来表达这一概念时，在案件文本相似度比较的时候可能导致一些相似案件文本被不正确地过滤。因此，将表达同一概念的多个同义词转化成表达这个概念的代表词，例如，在案件文本相似度比较时，将“扒窃”和“窃取”都转化成“偷窃”，就可以将原有特征提取从词的层面上上升到了主题概念的层面。

目前，国内大部分对同义词的应用大多都是基于哈尔滨工业大学信息检索研

究室的《同义词词林》扩展版。然而由于该同义词林比较庞大、分散,不适合针对专门的公安领域的研究。

为此,我们建立了专门的公安领域的同义词词典,该同义词词典主要是通过参考大量的分词结果,人工建立的词典。

3.4.2 案件文本属性信息抽取

由于案件文本与一般的文本不同,包含了大量的案件细节信息。案件文本属性信息抽取的功能为:给定一个案件文本,抽取出案件的如下属性信息:作案时间、作案地点、涉案人员、作案手段、作案工具、损失物品、损失金额。

属性信息抽取算法以及实现如下:

(一)属性抽取算法。

其主要思想是:对文本进行中文分词,预处理后,根据分词结果得到的词性标注信息,并结合专门的关键词库,抽取出上述案件属性信息。属性抽取的目的是得到结构化的文本信息。

(二)属性信息抽取的流程。

属性信息抽取的主要流程是:首先,对案件文本进行中文分词,预处理后得到特征词,然后信息抽取算法根据特征词的词性标注信息,并结合各个属性关键词库抽取出各个属性的信息,最后将这些结构化的属性信息存放在数据库中^[34]。信息抽取具体算法见(三)属性信息抽取实现。

(三)属性信息抽取实现

在抽取的过程中,将属性的抽取分成两种,一种是需要结合属性关键字词库抽取的属性,包含手段、物品、工具;另一种不需要结合关键词库抽取的属性,包含:时间、人、地点、金额。

1) 第一种情况的抽取方法:首先判断特征词是否为相应的词性(手段为动词,物品和工具为名词,其次判断该词是否在相应的词典里(属性关键词词库包含三个词典:手段词典、物品词典、工具词典)。该抽取方法难点在于属性关键字词库的建立。属性关键词库的建立,是通过大量的分词训练,人工提取属性关键词建立

词典。

2) 第二种情况的抽取方法:

首先判断特征词是否为相应的词性。如果是则做以下处理:

1、时间抽取: 在实验的过程中发现, 一个完整的时间, ICTCLAS分词组件倾向于进行细的切分。如: “2009年1月1日”, 被切分成“2009年/t 1月/t 1日/t” 。需要将连续的词性为时间的词进行合并。

2、人名抽取: 将 ICTCLAS 分词的结果中词性为人名的词抽取为人名。

3、金额抽取: 选取词性为数词的词, 并且该词的下一个词是“元”, “余元”, “万元”。最后将两个词数词和量词进行合并操作成为一个新词作为金额。

3.4.3 基于信息抽取和同义词分析的改进案件文本相似度计算

计算文档相似度的方法很多, 具有代表性的是余弦算法, 其定义如下。假设文档 D_1 和文档 D_2 分别对应了特征项向量 v_1 和 v_2 , 则这两个文档的相似度:

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|} \dots\dots\dots \text{公式(3.2)}$$

其中内积 $v_1 \cdot v_2$ 为标准向量点积, 定义为 $\sum_{i=1}^n v_{1i} v_{2i}$, 分母中的范数 $|v_1|$ 定义为 $|v_1| = \sqrt{v_1 \cdot v_1}$ 。

由于在本文中, 一个特征向量的权重计算采用布尔方法。因此公式 (3-2) 等效于:

$$sim(v_1, v_2) = \frac{C}{C_1 C_2} \dots\dots\dots \text{公式(3.3)}$$

其中, C 表示文本 D_1 和文本 D_2 相同的特征词总数的个数, C_1 表示文本 D_1 特征词总数的个数, C_2 表示文本 D_2 特征词总数的个数。

由于案件文本与一般的文本不同, 案件文本有作案时间、地点、涉案人、作案手段、作案工具、损失物品、涉案金额, 其它等属性, 各个属性的重要程度不一样。我们分别赋予这些属性一个权值, 表示其重要性程度, 这些值是介于 0 和

1 之间的一个小数，且所有特征项的系数值总和为 1。这些属性的权值的确定主要是根据业务人员根据经验判定各个属性的重要性而确定的。由于各属性重要程度不同，因此需要将案件文本进行属性信息抽取。案件文本经过属性信息抽取后，将特征词归入到相应的属性中。

在计算每一个属性的相似度公式的时候，引入同义词分析方法，考虑一个词是否出现在同义词词典中，如果出现则将该词转化成表示同一个概念的多个同义词的代表词。

基于公式(3-3)，计算两个案件文本 D_1 和文本 D_2 各属性的相似度 $Sim_i(1 \leq i \leq 7)$ ，见公式：

$$sim_i(v_1, v_2) = \frac{C_i}{C_{1i}C_{2i}} \dots\dots\dots \text{公式(3.4) 其中}$$

C_i 是两个案件属性 i 中具有相同特征词的总数的个数。 C_{1i} 是案件文本 D_1 属于属性 i 的特征词总数。 C_{2i} 是案件文本 D_2 属于属性 i 的特征词总数。

上式是计算两个案件文本中每一个属性的相似度的公式。得到了各属性的相似度以后，分别乘以各自的权值系数 k_i ，最终就得到了整条案件文本之间的相似度^[34]，见公式（5-5）：

$$sim(v_1, v_2) = \sum_{i=1}^8 (sim_i * k_i) \dots\dots\dots \text{公式(3.5)}$$

$$\text{其中, } \sum_{i=1}^8 k_i = 1, 1 \leq i \leq 8, 0 \leq Sim_i \leq 1.$$

输入一个案件，对其他案件进行相似性的计算，得到案件之间的相似度大小，进行排序，最后返回和与该案件相似性度从高到低的高于阈值的相似案件。

为了调高相似度计算的速度，系统进行了系统属性抽取的预处理，这样大大地提高了改检索的速度。

3.5 犯罪案件文本分类

犯罪案件文本分类模块是在给定犯罪案件类别分类体系的前提下，实现了根据案件文本的内容自动判别文本的类别的功能。犯罪案件文本分类与一般文本分

类过程相似,基本包括了:中文分词与预处理、特征选择、基于特征的文本分类等几个主要过程。中文分词和预处理方法见3.3节中文分词及预处理中具体描述。

3.5.1 犯罪案件文本类别的特征

犯罪案件文本具有3.2节所描述的特征:文本短小,包含大量案件细节信息。除此之外,犯罪案件文本类别具有在一定区域内的不同时期,某一案件类别所包含的文本数占该时期总文本数的比例基本接近、各类别文本数目分布比例不均衡等特点。

3.5.2 犯罪案件文本分类特征选择

文本分类的一个特点是特征词向量的维度很大。为了降低向量维度,需要对文本进行特征选择,同时不损伤分类准确率。目前文本分类中,用于特征选择的方法主要有:文档频率方法(DF)、互信息(MI)、信息增益(IG)、 X^2 统计量(CHI)、期望交叉熵、文本证据权等^[28]。本文作者通过实验分析比较了上述6种特征选择方法的效果,发现IG和CHI具有相对较好的特征选择效果。本文实验部分:4.5节案件文本分类模块设计和实章现中的4.5.3节实验和结果分析中具体展开比较了IG和CHI两种特征选择算法,最后得到CHI特征选择在维度的降低和分类准确率两个方面都表现出比IG略高的性能的结论。因此,本文最后选取 X^2 统计量(CHI)特征选择算法进行特征选择。

3.5.3 朴素贝叶斯文本分类

朴素贝叶斯算法的主要思想是:在给定待分类文本的条件下,首先根据训练集计算各个特征词属于每个类别的先验概率以及类别的先验概率。当新文本到来时,用特征词集合表示文本,然后根据特征词的先验概率计算文本属于各个类别的后验概率,然后选择其中后验概率最高的类别作为该文本所属的类别^[11]。

(一)训练阶段

1. 准备训练集

训练集是由已经入库的人工已经标注好类别的案件文本组成。

2. 分词及预处理

使用 3.3 节中文分词与预处理方法对训练集所有案件文本进行分词并预处理。

3. 进行训练

1) 首先, 计算每个类别的类先验概率 $p(C_j)$:

$$p(C_j) = \frac{N_{C_j}}{N} \dots\dots\dots \text{公式(3.6)}$$

其中 N_{C_j} 是 C_j 类的训练文本数, N 是总的训练集数。

2) 其次, 计算特征词 w_i 属于类别 C_j 的概率 $p(w_i|C_j)$ 。

$p(w_i|C_j)$ 根据不同的模型有不同的计算方法, 其计算方法将在第 3.5.4 节中详细描述。

(二)分类过程

1. 准备测试集

训练集是由已经入库的人工已经标注好类别的案件文本组成。

2. 预处理

使用 3.3 节中文分词与预处理方法对测试集所有案件文本进行分词并预处理。

3. 进行分类

朴素贝叶斯算法有一个很强的假设, 即文本中的特征词之间是相互独立的^[1]。

对于某个测试文本 d , 计算该文本属于类别 C_j 的概率:

$$p(C_j|d) = \frac{p(C_j) \prod_{i=1}^n p(w_i|C_j)}{\sum_{k=1}^{|C|} p(C_k) \prod_{i=1}^n p(w_i|C_k)} \dots\dots\dots \text{公式(3.7)}$$

其中, $p(C_j) = \frac{N_{C_j}}{N}$, N_{C_j} 是 C_j 类的训练文本数, N 是总的训练集数。 $p(w_i|C_j)$

表示特征词 w_i 属于类别 C_j 的概率, 根据不同的模型有不同的计算方法, 其计算方法将在第 3.5.4 节中详细描述。由于对于不同的类别, 后验概率 $p(C_j|d)$ 的分母

$\sum_{r=1}^{|C|} p(C_r) \prod_{i=1}^n p(w_i|C_r)$ 值都相同。所以求最大后验概率值等效于求以下 V_{\max} :

$$V_{\max} = \arg \max_{c_j \in c} p(C_j) \prod_{i=1}^{i=n} p(w_i | C_j) \dots\dots\dots \text{公式(3.8)}$$

最后将文本分到概率 $p(C_j | d)$ 最大的那个类别中。

3.5.4 面向不均衡类别的改进朴素贝叶斯案件文本分类

贝叶斯算法有多变量贝努里模型和多项式模型两种模型，两种模型的区别在于文本文档是如何表示的^[35]。本文针对犯罪案件具有类别分布不均衡的特点，即训练集各个类别所包含的文本数目差异较大的特点，提出了改进的多变量贝努里模型。

(一)多项式模型

在该模型中，文本被表示为特征词词频的向量模型。对一个给定类别 C_j ，对每个特征词 w_i ，计算特征词 w_i 属于类别 C_j 的概率，公式如下：

$$p(w_i | C_j) = \frac{1 + \sum_{k=1}^{|D|} N(w_i, d_k)}{|V| + \sum_{s=1}^{|V|} \sum_{k=1}^{|D|} N(w_s, d_k)} \dots\dots\dots \text{公式(3.9)}$$

其中， $|D|$ 表示类别 C_j 中的文本总数， $|V|$ 为特征词的总数， $N(w_i, d_k)$ 表示特征词 w_i 在文档 d_k 的词频。

(二)多变量贝努里模型

在该模型中，文本被表示为特征词在该文本中是否出现的二进制向量模型。对一个给定类别 C_j ，对每个特征词 w_i ，计算特征词 w_i 属于类别 C_j 的概率，公式如下：

$$p(w_i | C_j) = \frac{1 + \sum_{k=1}^{k=|D|} B(w_i, d_k)}{|C| + |D|} \dots\dots\dots \text{公式(3.10)}$$

其中， $|C|$ 表示中类别种数； $|D|$ 表示在类 C_j 的总的文本数； $B(w_i, d_k) \in \{0,1\}$ ，当特征词 w_i 在文本 d_k 中出现则为 1，否则为 0。为避免 $P(w_i | C_j)$ 等于 0，分子进行了加权。

(三) 基于类别分布不均衡的改进多变量贝努里模型

在案件文本分类中, 由于案件文本具有文本短小、出现单词频率低等特点, 因此多变量贝努里模型适用于案件文本的表示。然而, 由于犯罪案件具有类别分布不均衡的特点, 多变量贝努里模型在类别分布不均衡的时候, 即当训练集各个类别所包含的文本数目差异较大时, 将会影响分类的准确率。该影响的产生主要源自于两种情况:

第一种情况: 从公式(3.10)中可以看到, 假设考虑训练集中两个所含文本数目差异较大的类别, 并且二者的 $\sum_{k=1}^{k=|D|} B(w_i, dk)$ 均等于 0 (即特征 w_i 在两个类别中均未出现) 时, 特征词 w_i 的 $p(w_i | C_j)$ 会因为分母中所含文本数目小的类别的 $|D|$ 过小, 而使得所含文本数目小的类别的 $p(w_i | C_j)$ 值偏大;

第二种情况: 当 w_i 在两个类别中均出现, 且特征词 w_i 在某一类别出现的文档数除以该类别的总文档数的值相同时, 也可能出现所含文本数目小的类别的 $p(w_i | C_j)$ 偏大的情况。

上述两种情况均会降低多变量贝努里模型的文本分类的准确率。实验结果也表明, 多变量贝努里模型的文本分类偏向训练集文本数少的类别, 使得所含文本数目小的类别查全率高但查准率低, 从而降低了分类的准确率。

为此, 提出了以下的改进方法:

$$p(w_i | C_j) = \frac{\frac{|D|}{|D_{max}|} + \sum_{k=1}^{k=|D|} B(w_i, dk)}{|C| + |D|} \quad \dots\dots\dots \text{公式(3.11)}$$

这里, $|D_{max}|$ 表示最大的类别文档总数, 其余变量表示和公式(3.10)相同。 $|D_{max}|$ 计算公式为: $|D_{max}| = \arg \max_{q \in c} |D_{C_j}|$ 。其中, $|D_{C_j}|$ 表示类别 C_j 的文档总数。由于, $0 < \frac{|D|}{|D_{max}|} \leq 1$ 不仅能够满足避免 $p(w_i | C_j)$ 为 0 的加权要求, 而且当类别 C_j 所含文本数目较小时, $\frac{|D|}{|D_{max}|}$ 也较小, 从而改善了上述第一种情况, 即改善了当两个所含文本数目差异大的类别的 $\sum_{k=1}^{k=|D|} B(w_i, dk)$ 均等于 0 的时候, 包含文本数目小的类别的 $p(w_i | C_j)$

偏大而导致分类准确率的降低的情况;同时公式(3.11)比公式(3.10)更逼近 $p(w_i|C_j)$ 的真实值,这样较好地改善第二种情况。实验结果也表明,相比多变量贝努里模型,改进的多变量贝努里模型大大提高了分类准确率。

历史的实验结果表明通常多项式模型相比多变量贝努里模型在文本分类中表现出更高的准确率^[35]。

然而本文犯罪案件文本分类模块设计与实现的实验部分 4.5.3 节实验和结果分析中通过对三种模型分别进行实验,实验结果表明,在案件文本分类中,改进的多变量贝努里模型不仅大大提高了多变量贝努里模型的分类准确率,同时也表现出比多项式模型相当甚至略高的分类准确率。第四章所描述的系统采用了改进的多变量贝努里模型对案件文本进行分类。

3.6 本章小结

本章首先介绍了犯罪案件文本的数据源,犯罪案件文本挖掘的一般流程,接着主要研究了犯罪案件文本挖掘的关键技术,提出了以下主要技术:

(1) 结合实际应用需要,对中科院 ICTCLAS 分词组件的结果进行修正和整理。针对案件文本特征,提出了加入公安领域词汇的改进的中文分词和具有针对性的特殊预处理方法。

(2) 建立了公安领域专用的同义词词典。提出了属性信息抽取方法和同义词语义分析方法,并在此基础上提出了改进的案件相似度计算方法。

(3) 根据犯罪案件文本类别不均衡的特征,改进了朴素贝叶斯中的多变量贝努里模型,提出了面向不均衡类别的改进朴素贝叶斯案件文本分类方法。

第4章 犯罪案件文本挖掘系统的设计和实现

本章在介绍浙江省公安厅犯罪特征数据挖掘系统的基础上,重点分析其中的犯罪案件文本挖掘子系统的结构和功能,以及相应功能的实现技术,包括:公安内部网页案情通告抽取的设计与实现,相似犯罪案件文本检索模块的设计与实现,犯罪案件文本分类的设计与实现。

4.1 犯罪特征数据挖掘系统概述

犯罪特征数据挖掘系统以犯罪数据库和公安内部网的海量数据为基础,以挖掘出有效模型、为公安系统的业务决策提供有效的支持为目标,在整理分析现有犯罪信息分析方法的基础上,设计相关的数据挖掘工具,提供面向犯罪信息研判的犯罪数据挖掘工具,包括分类、预测、关联等工具,涉及挖掘任务包括:案件特征、发案趋势、案件串并等。

系统框架结构:系统的框架结构如图 4.1 所示。本文将在介绍浙江省公安厅犯罪特征数据挖掘系统的框架结构后,重点分析下图中的用红色加粗显示的以及犯罪案件文本挖掘子系统的结构,功能以及设计和实现。

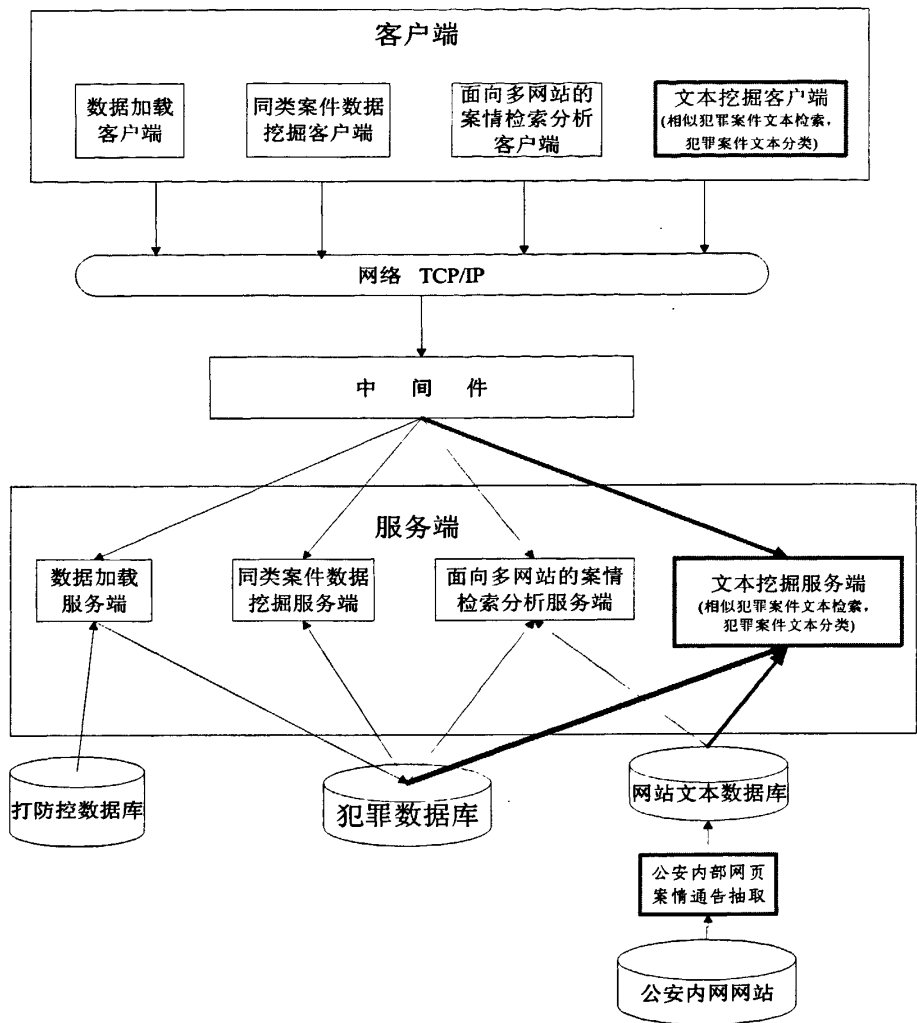


图 4.1 犯罪特征数据挖掘系统框架图

在该系统框架中，主要包含四大模块：数据加载模块，同类案件数据挖掘模块，文本挖掘模块，面向多网站的案情检索分析模块。其中：

1. 数据加载模块

由于项目相关数据可能会涉及到分散的数据库，且数据量极为庞大，因此需要立足业务需求和技术需要，在打防控主干系统应用库基础上，整合清洗现有数据，制定出若干规则，提取出有代表性的数据放入数据仓库后进行挖掘，既避免了集成超大容量数据库的难度，又节省了挖掘时间，同时也不会影响到业务数据

库的使用。项目使用 SQL-Server2005 作为数据仓库服务器。

2. 同类案件数据挖掘模块。

包含 3 个子模块, 即属性分析子模块、同类案件特征变化趋势分析子模块和同类案件分析子模块。

(1) 属性分析子模块

该模块针对特定案件, 从原有的属性集中自动筛选出与案件分类相关度高的属性名称以及对应属性的权重, 帮助业务人员快速确定与案件分类关联度高的若干个属性以及这些属性的重要程度。

(2) 同类案件特征变化趋势分析子模块

对于一定时间范围和时间粒度的特定类型的案件进行分析, 发现其中某些属性取值的变化规律, 及时发现案件的变化规律。

(3) 同类案件分析子模块

用 k-modes 聚类算法, 对于特定时间和区域范围内的特定类型案件, 系统自动发掘出其中相似的案件子集, 供业务人员进行分析和比较。

3. 文本挖掘模块

文本挖掘模块是本文主要研究的内容。主要包含两大模块: 相似犯罪案件文本检索子模块和犯罪案件文本分类子模块。

(1) 相似犯罪案件文本检索子模块

相似犯罪案件文本检索实现自动发掘与给定案件文本相似的案件文本集的方法。为案件串并提供有效帮助, 为公安业务人员提供有效的决策支持。

(2) 犯罪案件文本分类子模块

通过文本分类自动地将案件快速地分类, 这样可以方便用户快速定位到相关的案件类别信息, 并有利于相关案件的串并。

4. 面向多网站的案情检索分析模块

公安内部专用网络的网站中存在大量的案件信息(大部分为案件新闻), 信息研判人员经常要浏览这些网站, 并分类记录其中潜在的有用信息, 以便随时提取分析。目前, 基本采用人工方法对这些信息进行浏览、保存和管理, 效率不高。为

了给业务人员提供方便的浏览、检索以及分析的手段，本模块对信息抽取、检索和文本聚类分析等方面做了相应的研究，实现了跨网站自动收集所需要的网页的功能，再从这些网页中通过一定策略抽取结构化的新闻信息；为用户提供快速检索和浏览的方式；通过聚类分析实现潜在信息的发现。

4.2 犯罪案件文本挖掘子系统体系结构

犯罪案件文本挖掘是浙江省公安厅刑事犯罪特征数据挖掘系统的重要组成部分，主要为串并案业务人员提供文本挖掘相关功能。该系统的文本挖掘过程如下图：

系统采用三层 C/S 结构，业务层中主要实现了案件文本分类模块，相似案件文本检索模块。数据层中主要实现公安内部网页抽取模块的功能。系统体系结构如下图 4.2 所示。

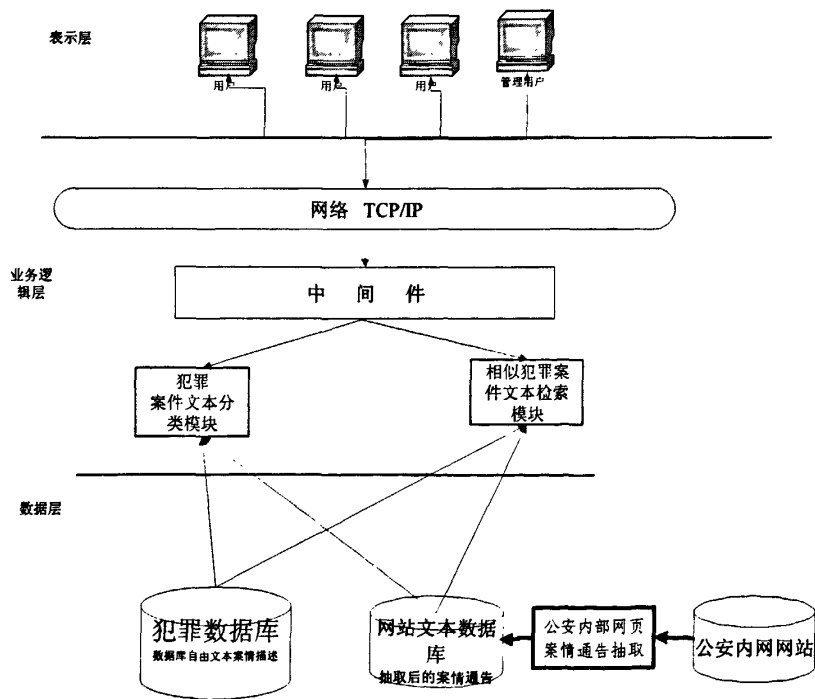


图 4.2 犯罪案件文本挖掘子系统体系结构

系统体系结构采用典型的三层 C/S 模型，分为数据层、业务逻辑层和表示层。

在数据层中是犯罪案件文本挖掘的数据源, 含有现有犯罪数据库中的自由文本案情描述, 以及公安内部网站的案情通告。从图中可以看到数据层中实现了公安内部网页案情通告抽取模块。如何抽取公安内部网站上的案情通告, 将在 4.3 节公安内部网页案情通告抽取的设计和实现中详细展开。

业务逻辑层中包括两大模块, 包括犯罪案件文本分类模块, 相似犯罪案件文本检索模块。

相似案件检索模块实现检索出与给定案件相似的案件文本集的功能, 并按相似度大小排序。主要为串并案业务人员提供相似案件检索功能, 为案件串并提供有效帮助。

犯罪案件文本分类模块实现在给定犯罪案件类别分类体系的前提下, 根据案件文本的内容自动判别文本的类别的功能。通过文本分类自动地将案件快速地分类, 这样可以方便用户快速定位到相关的案件类别信息, 并有利于相关案件的串并。

用户层中为不同用户提供了两类客户端程序: 普通客户端和管理客户端。普通用户主要对文本挖掘结果进行检索和浏览。管理员通过管理客户端管理数据库和数据加载, 属性配置等操作。

4.3 公安内部网页案情通告抽取的设计和实现

公安内部网页案情通告抽取模块定时按需从公安内部网络中抓取网页, 抽取网页中的案情通告文本, 并把结果存储在数据库中。主要用到了网络蜘蛛和网页抽取的技术。

4.3.1 案情网页抓取

在本项目的应用场景中, 需要下载的主要是公安内部网络中的案情网页。如何抓取这些网页, 有不同的抓取方法。最常用的是“蜘蛛爬虫技术”。

将 web 的网页看做是一个有向图, 从给定起始网页集合开始, 找到网页中的链接, 按照深度优先或者广度优先的策略进行遍历, 下载相应的网页, 由此循环

下去，直到符合某些抓取终止条件。这个过程就如同成蜘蛛(spider)在蜘蛛网(web)上爬行(crawl)。一般情况下按照广度优先搜索得到的网页集合要比深度优先搜索得到的集合重要。

一般情况下，任何搜索引擎都不可能将 web 上的网页搜索完全，通常是在其他条件下的限制下决定搜索过程的结束，例如磁盘满，或搜集时间过长^[36]。

本文实现了基于广度优先策略的蜘蛛爬虫组件，spider 从某个公安内部网站的首页开始爬行抓取案情网页。实现了案情网页的抓取，并将下载网页到指定路径。下图 4.3 是蜘蛛爬虫组件类图。

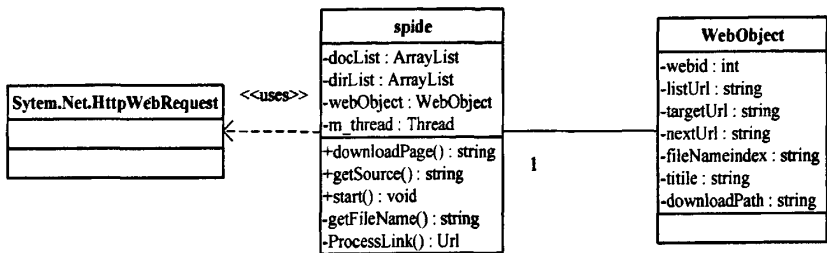


图 4.3 蜘蛛爬虫类图

由上图 4.3，spider 是蜘蛛爬虫的关键类，实现了对网页的抓取，并将网页下载到指定的目录。downloadPage()函数功能是下载网页到指定的路径，getSource()函数功能是取得网页源代码，以便分析网页中的 url。Start()函数是蜘蛛开始爬行，从开始网页 url 开始爬行，最后实现网页的抓取和下载。

4.3.2 案情通告内容抽取

为了抽取出案情网页中的案情通告，需要使用网页信息抽取技术。网页信息技术发展至今出现了各种抽取策略。由于同一公安网站中的案情网页的生成往往采用统一的模版，因此，本文仅选取了对网站模版稳定性要求高的正则表达式网页信息抽取方法。正则表达式的策略是：将网页看做字符串，利用编写好好正则表达式从字符串中抽取特定模式的内容。

通过观察分析，案情网页的内容具有统一的结构，图 4.4 是案情网页的 xml 格式的主要框架示例：

```
<body >
  <!--主体部分开始--> <!--信息统计条-->
  <table>
    <tbody>
      <tr>
        <td> <!--文档标题-->
        <div>抓获冒充电信工作人员招摇撞骗违法嫌疑人，请串并! </div>
        </td>
      </tr>
      <tr>
        <td>
          <div><!--发布时间-->
            2009-07-27 17:08
          </div>
        </td>
      </tr>
      <tr>
        <td><!--显示正文信息-->
        <p>2009年1月1日，嫌疑人张三事先电话联系受害人李四，
          然后冒充电信工作人员上门服务，以推销电信卡</p>
        </td>
      </tr>
    </tbody>
  </table>
  <!--主体部分结束-->
</body>
```

图 4.4 案情网页 XML 格式的主要框架示例

针对该类案情网页的统一模版，编写抽取标题，时间和内容的正则表达式如下图 4.5:

标题正则式	<code>[?<=<!--文档标题-->[\s]*<div[^>]*>)[\s\S]*?(</div>)</code>
时间正则式	<code>[?<=<!--发布时间-->[\s]*)\d\d\d\d\d\d\d\d\d\d</code>
内容正则式	<code>[?<=<!--显示正文信息-->)[\s\S]*<!--主体部分结束--></code>

图 4.5 标题，时间，内容正则表达式

其中，在标题正则表达式中： `(?<=<!--文档标题-->[\s]*<div[^>]*>)`和`</div>`是模式的前缀和后缀。其中`(?<=exp)`表示匹配 `exp` 后面的后置。`\s` 匹配任意空白字符，`\S` 匹配任意非空白字符，所以`[\s\S]`匹配任意字符。`[\s\S]*`表示`[\s\S]`匹配零次或者多次。`[\s\S]*?`的`?`表示对`[\s\S]*`采用非贪婪模式匹配。

正则表达式默认采用的贪婪模式，即尽量匹配多的字符。使用了非贪婪模式的时候，对于可匹配可不匹配的表达式，尽可能不匹配。如串

<div>context1</div><div> context2</div>，贪婪模式的正则表达式“<div>[\\s\\S]*</div>”匹配结果只有 1 个，为“<div> context1</div><div>context2</div>”。而以上非贪婪模式匹配结果 2 个，分别是“<div> context2</div>”和“<div> context2</div>”。

时间正则式中，/d 匹配一个数字。\\d\\d\\d\\d\\d\\d\\d\\d 表示匹配例如 2006-01-01 的数字。可抽取出时间年-月-日。

案情通告内容抽取类设计。类图如下图 4.6:

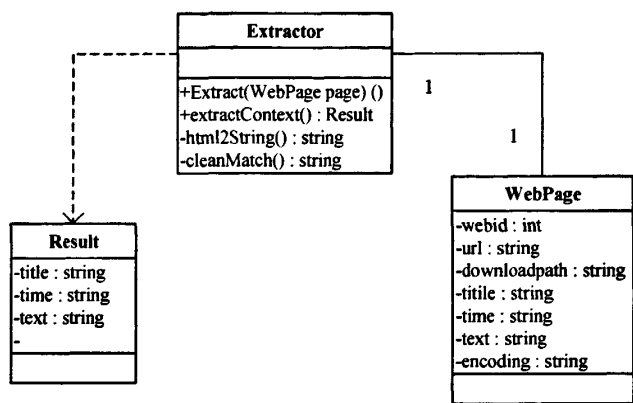


图 4.6 案情通告内容抽取类图

由上图 4.6，Extractor 是实现内容抽取的类，构造函数传入待抓取的网页。Html2String 函数是将 html 文件内容以 string 形式读取出来。extractContext 函数，是抽取关键函数，传入目标标题，时间，内容的正则表达式，返回抽取标题，时间和内容结果。cleanMarch 函数是对抽取后的结果进行清洗。

4.4 相似犯罪案件文本检索模块设计与实现

相似案件检索模块实现检索出与给定案件相似的案件文本集，并将相似案件文本集按相似度大小排序的功能。相似犯罪案件文本检索模块的具体算法见 3.4 节。接下来将详细介绍相似案件文本检索模块的结构设计，类设计以及最后的实验和结果分析。

4.4.1 相似案件文本检索模块结构设计

该模块实现了 3.4 节相似犯罪案件文本检索的算法。具体实现了基于案件属性信息抽取并结合同义词分析的相似案件文本检索的功能，以及为了加快检索速度的属性信息预处理功能。该模块主要包含两大块：属性抽取预处理模块和检索模块。属性抽取预处理是为了加快检索的速度，事先将数据库中的所有案件进行属性抽取，并将抽取结果存放在数据库中。图 4.7 是相似案件文本检索模块结构设计图：

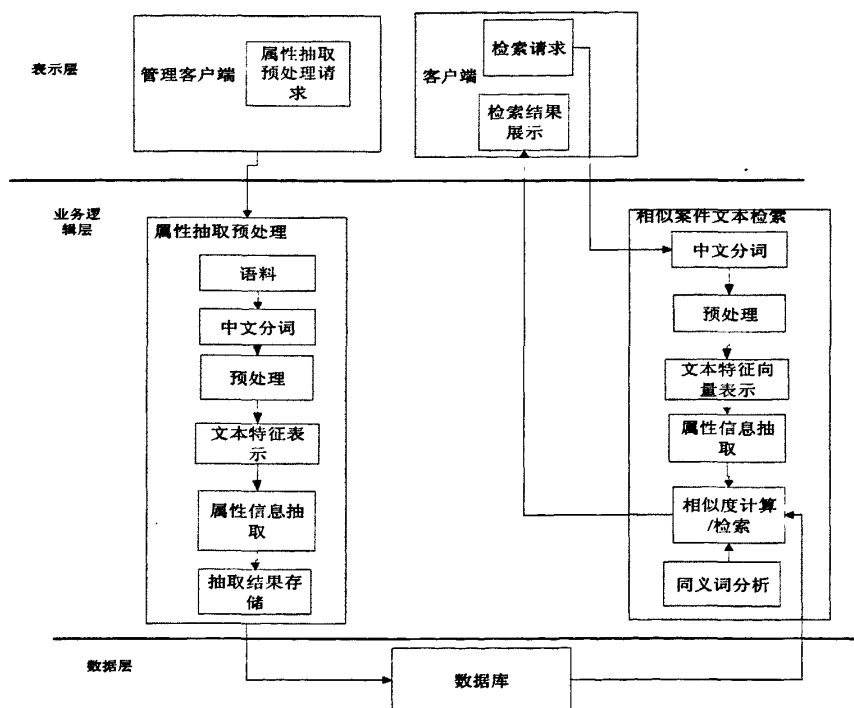


图 4.7 相似案件文本检索模块结构设计图

由图 4.7 所示，左边属性抽取预处理模块，首先将语料中的每个语料文本均做如下处理：中文分词，预处理后得到文本特征表示，再根据 3.4.2 节中案件文本属性信息抽取的算法实现属性信息抽取，最后将抽取后的机构化的属性信息保存在数据库中。

图 4.7 右边的相似案件文本检索模块，将给定的待检索的案件文本进行类似处理，中文分词，预处理，属性信息抽取后，利用 3.4.3 节基于信息抽取的改进

案件文本相似度计算方法计算该文本与数据库中的每个文本的相似度。最后将相似度高于一定阈值的相似案件文本集合以相似度大小从大到小的顺利排列的结果展示给客户端。

4.4.2 类设计

下图 4.8 描述了相似案件文本检索模块的主要组件组成，并描述了组件之间的依赖关系。

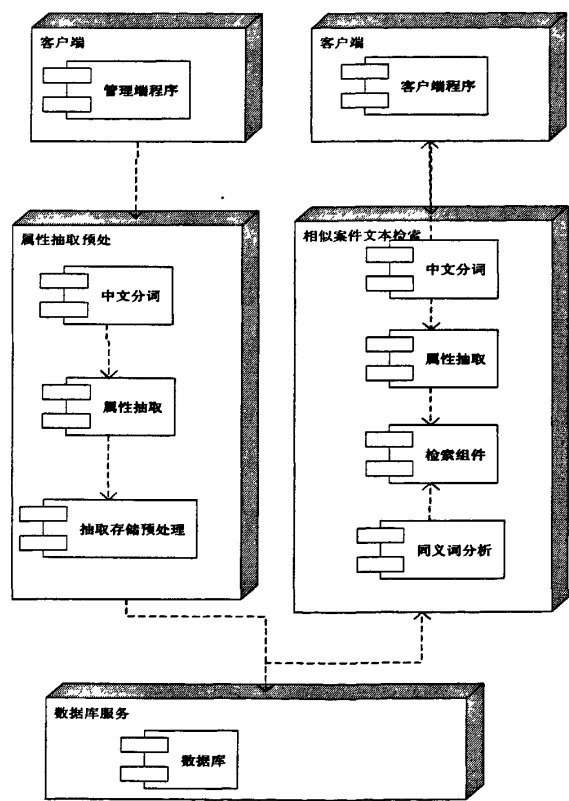


图 4.8 相似案件文本检索模块组件图

(一)中文分词组件

接口：ITokeniser，提供分词函数接口。对输入的文本分词，返回分词结果列表。

类：Tokeniser，实现接口 ITokeniser。封装了分词器 ICTCLAS，加入了公安

领域的词汇, 并进行去除停用词, 去除单字, 根据词性去除对案件属性无用的词操作。

辅助类: **StopWordHandler**。从停用词表文件中读取停用词, 用于对停用词的判断。

(二)属性信息抽取组件

类: **CaseAttributeExtract**。制定抽取属性信息规则, 实现抽取各个属性的算法。

实现将给定案件文本抽取出相应的作案时间, 作案人, 作案地点, 金额, 手段, 物品, 工具。

(三)抽取预处理组件

类: **CaseAttributeExtractPreSave**。事先对数据库中的案件文本进行属性信息抽取预处理, 并将抽取结果存放在数据库中。

(四)同义词分析组件

类: **SynonymWordHandler**。读取同义词表, 判断词是否有同义词, 如果存在同义词, 则得到改组同义词的代表词。

(五)检索组件

类: **SimilarCaseSearch**。实现给定案件文本, 计算该案件与数据库中每个案件文本之间的相似度, 返回与给定案件文本相似的案件文本集合, 并按相似度大小从大到小顺序排序。

4.4.3 实验和结果分析

本文使用 C#. net 编程环境在 Windows xp 操作系统上实现了属性信息抽取方法和同义词的语义分析方法, 并在此基础上的改进的相似犯罪案件文本检索功能。

案件属性信息抽取: 输入案件编号 (或文本), 将某案件的文本信息抽取出案件的如下属性信息: 作案时间、涉案人员、作案地点、作案手段、作案物品、作案工具、损失金额。

由于公安保密性, 案件属性信息抽取实验使用的案件文本是在真实案件的文

本基础上,修改了时间、人名、地点得到的案件文本:“2006年1月1日清晨1时许10分许,犯罪嫌疑人张三在义桥村1幢1号101室插片开门进入,放在卧室内的一条裤子及包内的现金失窃,总价值约2000元”。案件属性信息抽取的应用如下图4.9:

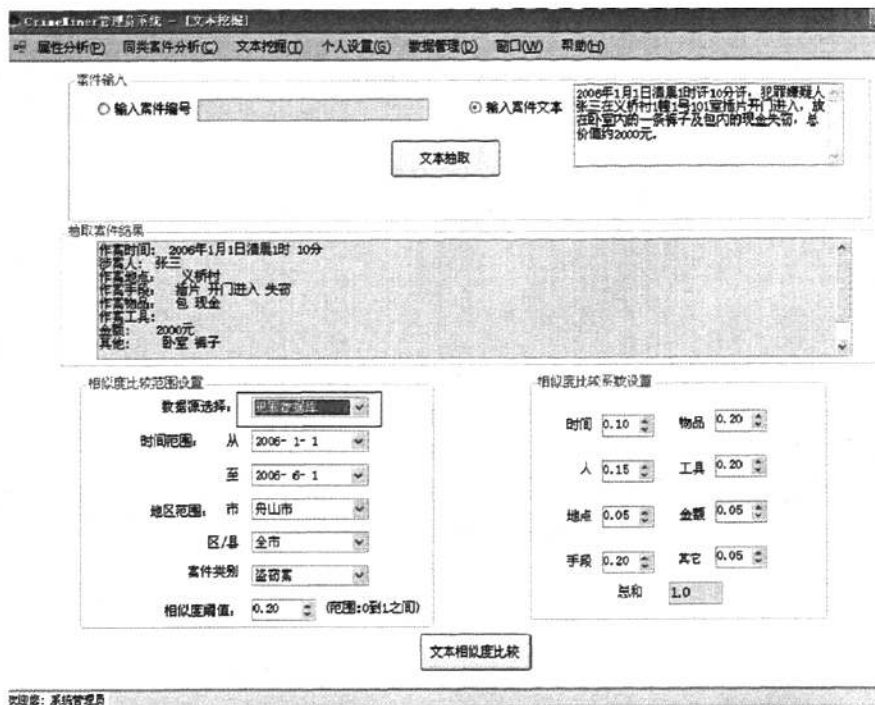


图 4.9 案件属性信息抽取系统实现

由上图 4.9 可以看到,该案件文本,经过案件属性抽取后,得到各属性如下:

“作案时间: 2006 年 1 月 1 日清晨 1 时 10 分

涉案人: 张三

作案地点: 义桥村

作案手段: 插片 开门进入 失窃

作案物品: 包 现金

作案工具:

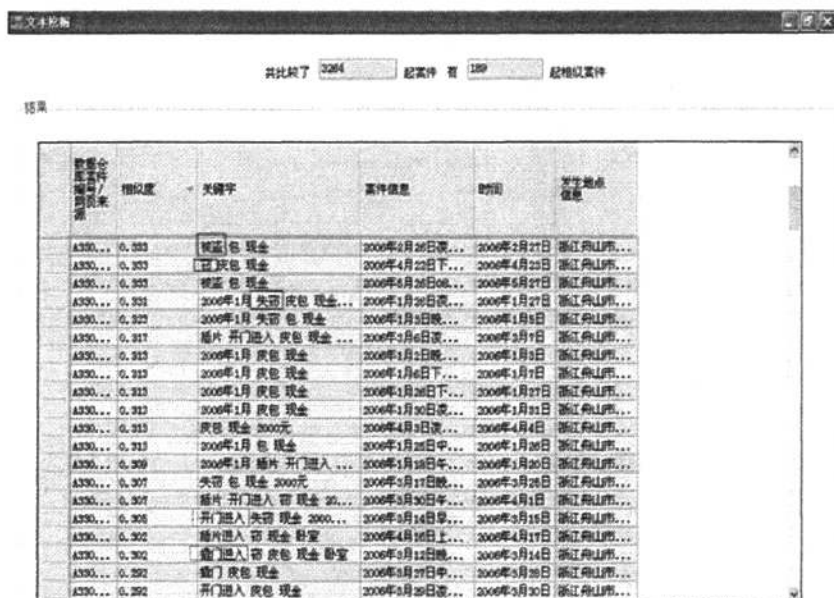
金额: 2000 元

其他: 卧室 裤子”

案件属性信息抽取后, 设置相似度比较范围和相似度比较系数, 即可进行相似度比较。

相似犯罪案件文本检索: 首先设置相似度比较范围, 其中数据源选择有两种: “犯罪数据库” 和 “网站文本数据库”。由图 4.10 中可以看到, 数据源选择 “犯罪数据库”, 时间选择 “2006-1 到 2006-6-1”, 地点选择 “舟山市”, 案件类别选择 “盗窃案”, 相似度阈值选择 “0.2”, 其次设置相似度设置各个属性的权重。给定案件文本将与犯罪数据库中的符合时间范围, 地区范围, 案件类别的案件进行比较, 最后将检索出相似度值大于相似度阈值的相似案件文本集, 并按相似度值大小从大到小的顺利排列。

- 数据源选择为 “犯罪数据库” 的相似案件检索结果如下图 4.10:



数据源/案件编号	相似度	关键词	案件信息	时间	发生地点
A330... 0.353	0.353	被盗 包 现金	2006年4月28日晚...	2006年4月28日	浙江舟山市...
A330... 0.353	0.353	被盗 包 现金	2006年4月22日下午...	2006年4月22日	浙江舟山市...
A330... 0.353	0.353	被盗 包 现金	2006年4月28日晚...	2006年4月28日	浙江舟山市...
A330... 0.321	0.321	2006年1月 失窃 皮包 现金...	2006年1月20日晚...	2006年1月20日	浙江舟山市...
A330... 0.323	0.323	2006年1月 失窃 包 现金	2006年1月5日晚...	2006年1月5日	浙江舟山市...
A330... 0.317	0.317	插片 开门进入 皮包 现金...	2006年1月6日晚...	2006年1月6日	浙江舟山市...
A330... 0.313	0.313	2006年1月 皮包 现金	2006年1月2日晚...	2006年1月2日	浙江舟山市...
A330... 0.313	0.313	2006年1月 皮包 现金	2006年1月6日下午...	2006年1月6日	浙江舟山市...
A330... 0.313	0.313	2006年1月 皮包 现金	2006年1月26日下午...	2006年1月26日	浙江舟山市...
A330... 0.313	0.313	2006年1月 皮包 现金	2006年1月30日晚...	2006年1月30日	浙江舟山市...
A330... 0.313	0.313	皮包 现金 2000元	2006年4月3日晚...	2006年4月4日	浙江舟山市...
A330... 0.313	0.313	2006年1月 包 现金	2006年1月25日中午...	2006年1月25日	浙江舟山市...
A330... 0.309	0.309	2006年1月 插片 开门进入...	2006年1月18日中午...	2006年1月18日	浙江舟山市...
A330... 0.307	0.307	失窃 包 现金 2000元	2006年3月17日晚...	2006年3月17日	浙江舟山市...
A330... 0.307	0.307	插片 开门进入 窃 现金 20...	2006年3月30日中午...	2006年4月1日	浙江舟山市...
A330... 0.306	0.306	开门进入 失窃 现金 2000...	2006年3月14日晚...	2006年3月15日	浙江舟山市...
A330... 0.302	0.302	插片进入 窃 现金 卧室	2006年4月16日上午...	2006年4月17日	浙江舟山市...
A330... 0.302	0.302	偷门进入 窃 皮包 现金 卧室	2006年3月12日晚...	2006年3月14日	浙江舟山市...
A330... 0.292	0.292	偷门 皮包 现金	2006年3月27日中午...	2006年3月28日	浙江舟山市...
A330... 0.292	0.292	开门进入 皮包 现金	2006年3月28日晚...	2006年3月30日	浙江舟山市...

图 4.10 数据源选择为 “犯罪数据库” 的相似犯罪案件文本检索结果

- 数据源选择为 “网站文本数据库” 的相似案件检索结果。数据源选择 “网站文本数据库” 时间选择 “2009-1 到 2009-6-1”, 地点选择 “舟山市”, 案件类别选择 “盗窃案”, 相似度阈值选择 “0.2”。下图 4.11 是相似犯罪

案件文本检索结果:

文本挖掘

共比较了 240 起案件 页 4 起相似案件

结果

数据源/网页来源	相似度	关键词	案件信息	时间	发生地点
浙江省研判...	0.260	1月 窃 包 现金	2009年3月9号19时...	2009年3...	浙江省舟山市
浙江省研判...	0.251	植片 偷门 被追 夜包 现金...	2009年2月份以来...	2009年2...	浙江省舟山市
浙江省研判...	0.250	包 现金	一、被害人: 3009...	2009年2...	浙江省舟山市
浙江省研判...	0.250	包 现金	2009年02月10日12时...	2009年2...	浙江省舟山市

图 4.11 数据源选择为“网站文本数据库”的相似犯罪案件文本检索结果

基于同义词的语义分析：从之前图 4.10 中的结果集中可以看到结果集中包含有“失窃”的同义词“窃”、“被盗”的案件文本，以及“开门进入”的同义词“撬门进去”的案件文本。

4.5 犯罪案件文本分类模块设计与实现

犯罪案件文本分类模块实现在给定犯罪案件类别分类体系的前提下，根据案件文本的内容自动判别文本的类别。犯罪案件文本分类的具体算法见 3.5 节。接下来将详细介绍案件文本分类模块结构设计，类设计以及最后的实验和结果分析。

4.5.1 案件文本分类模块结构设计

案件文本分类模块的主要过程是：首先对语料进行选择，然后对语料文本进

行中文分词，预处理和特征选择后，将文本表示成文本向量特征。本文的文本分类算法选取的是朴素贝叶斯文本分类算法，分类过程分成训练和分类两个过程。

案件文本分类模块结构如下图 4.12 所示。

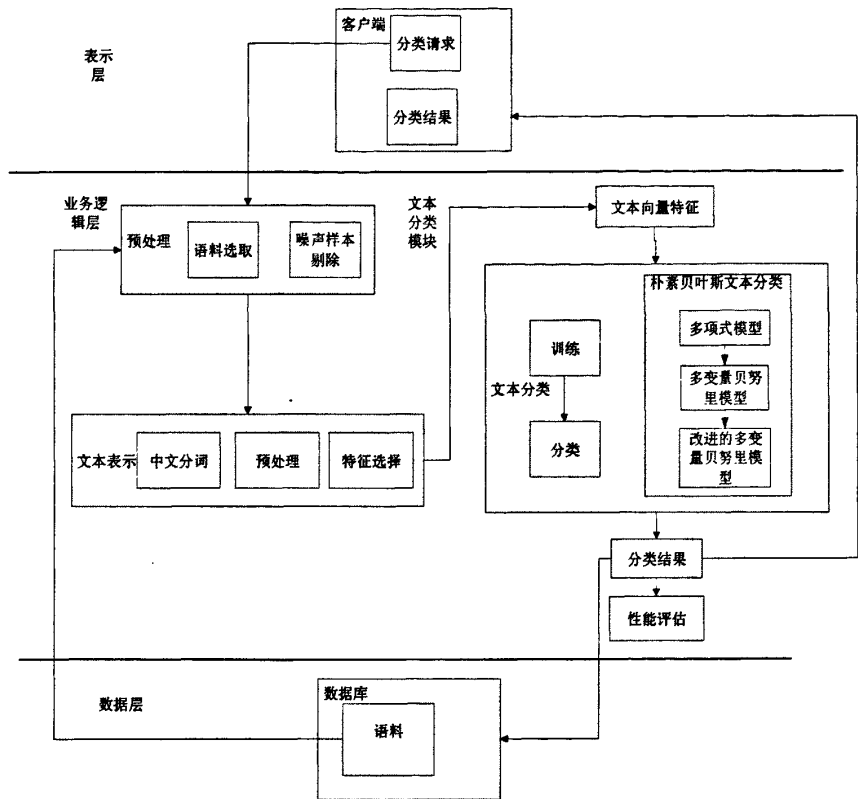


图 4.12 案件文本分类模块结构设计图

由上图 4.12 所示，实验部分对多项式、多变量贝努里模型、改进的多变量贝努里模型三种模型分别进行了实验。实验结果证明，改进的多变量贝努里模型不仅大大提高了多变量贝努里模型的分类准确率，同时也表现出比多项式模型相当甚至略高的分类准确率。最后系统采用了改进的多变量被努里模型对案件文本进行分类。

4.5.2 类设计

下图 4.13 描述了相似案件文本检索模块的主要组件组成，并描述了组件之间

的依赖关系。

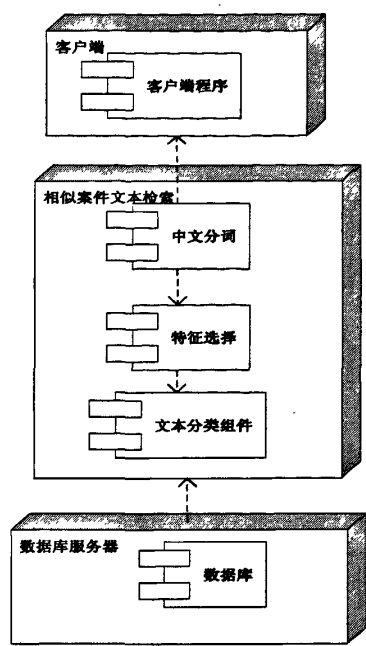


图 4.13 案件文本分类模块组件图

(一)中文分词组件

该组件与 4.4.2 节中介绍的中文分词组件同

(二)特征选择组件

积累：WordSelect

类：类 DFSelect，类 MISelect，类 IGSelect，类 CHISelect，类 ECESelect，类 WETSelect。 分别实现了文档频率方法（DF），互信息(MI)，信息增益(IG)，X2 统计量（CHI），期望交叉熵(ECE)，文本证据权(WET)方法

辅助类：TrainingDataManager

(三)文本分类组件

接口：IClassifier

类：BayeClassifier.

辅助类：TrainingDataManager.

4.5.3 实验和结果分析

本文使用 C#. net 编程环境在 Windows xp 操作系统上实现了文本分词、预处理、特征选择、朴素贝叶斯犯罪案件文本分类，并对比了改进的多变量贝努里模型和两种常用模型的案件文本分类结果。

(一)实验环境和实验数据

实验将某省打防控信息数据库中的案件文本信息作为数据源。使用了五种分别来自该省五个不同的城市的案件文本信息数据集，共 23875 条训练集文本信息和 16180 条测试集文本信息。每个数据集的训练集来自于该城市在一定时间段里抽取的人工已经标注好类别的案件，测试集来自于该城市与训练集不同时间段里抽取的人工已经标注好类别的案件。由于在犯罪案件中，通常盗窃案占有所有案件的比例通常达到 60% 以上，所以重点研究盗窃案，选取盗窃案的五个子类别作为分类类别体系。

(二)评价方法

本文采用查准率、查全率和综合考虑了查准率和查全率的 F1 测试值这三个指标对实验结果进行度量评价。

查全率： $Re = \frac{\text{分类的正确文本数}}{\text{应用的所有文本数}}$ 公式(4.1)

查准率： $Pr = \frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}$ 公式(4.2)

F1 测试值： $F1 = \frac{Pr \times Re \times 2}{Pr + Re}$ 公式(4.3)

为了全面衡量分类系统性能，我们还使用了宏观 F1 值作为评价指标，计算如下

$Macro_F1 = \sum_{i=1}^n \frac{Ni}{N} \times F1_i$ 公式(4.4)

(三)实验结果

对案件文本进行分词及预处理后，对样本集一，分别计算多项式模型，多变量贝努里模型和改进的多变量贝努里模型的分类结果，使用查准率，查全率，F1 值和宏观 F1 值作为实验的评测标准。结果见表 4.1。

表 4.1 样本集一采用多项式模型，多变量贝努里模型和改进的多努里模型的分类结果

案件类别	文本数		多项式模型			多变量贝努里模型			改进的多变量贝努里模型		
	训练集	测试集	查准率	查全率	F1	查准率	查全率	F1	查准率	查全率	F1
盗窃摩托车案	975	965	96.57%	96.47%	96.52%	97.52%	93.88%	95.67%	96.98%	96.68%	96.83%
盗窃自行车案	468	755	92.80%	92.31%	92.56%	90.70%	91.78%	91.24%	91.78%	93.24%	92.50%
扒窃案	364	290	83.90%	84.48%	84.19%	75.61%	84.48%	79.80%	80.62%	88.96%	84.59%
盗窃路财案	80	81	85%	20.98%	35.55%	8.35%	80.24%	15.13%	76.59%	44.44%	56.25%
入室盗窃案	1847	2213	93.32%	96.02%	96.65%	98.47%	67.14	79.84%	95.10%	94.89%	95.00%
宏观 F1	3554	4304	92.84%			84.17%			93.54%		

从表 4.1 中可以看到，用传统的多变量贝努里模型贝叶斯分类方法对案件文本进行分类的宏观 F1 值为 84.17%，通过改进的多变量贝努里模型的宏观 F1 值为 93.54%。改进的多变量贝努里模型的宏观 F1 值比传统的多变量贝努里模型的宏观 F1 值提高了 9.37%。同时，改进的多变量贝努里模型的宏观 F1 值也比多项式模型的宏观 F1 值提高了 0.7%。在表 4.1 中还可以看出，对于训练集文本较少的盗窃路财案，改进的多变量贝努里模型的 F1 值比传统的多变量贝努里模型提高了 41.12%，比多项式模型提高了 20.7%。改进的多变量贝努里模型表现出了对于训练文本数少的类别具有相对较高的 F1 值，平衡了传统多变量贝努里模型由于偏向训练文本数少的类别而使得训练文本数少的类别查全率高但查准率低，和多项式模型对于训练文本数少的类别查准率高但查全率低的矛盾。表 4.2 为对五个样本集分别采用三个模型进行实验，使用宏观 F1 值作为实验的评测标准。

表 4.2 五个样本集采用多项式模型，多变量贝努里模型和改进的多变量贝努里模型的分类结果

样本集	训练文本数	测试文本数	多项式模型	多变量贝努里模型	改进多变量贝努里
样本集一	3554	4304	92.85%	84.17%	93.54%
样本集二	6614	3044	95.15%	90.61%	95.44%
样本集三	3732	1964	90.66%	76.59%	93.11%
样本集四	2966	1977	93.97%	76.21%	94.51%
样本集五	7009	4891	96.70%	85.71%	96.56%

从表 4.2 中可以得到，五个样本集改进的多变量贝努里模型宏观 F1 平均值为

94.63%，比多变量贝努里模型的宏观 F1 平均值提高 11.98%，比多项式模型的宏观 F1 平均值提高了 0.76%。实验结果证明了，改进的多变量贝努里模型大大改善了多变量贝努里模型的分类准确率，同时也表现出比多项式模型相当甚至略好的分类准确率。

最后，对改进的多变量贝努里模型，分别采用特殊的预处理和两种特征选择进行试验。使用宏观 F1 值作为实验的评测标准。得到的分类结果见表 4.3，其中 NP 表示没有采用预处理；P 表示采用了预处理；P-IG 表示采用了预处理，并用 IG 进行特征选择；P-CHI 表示采用了预处理，并用 CHI 进行特征选择。

表 4.3 改进多变量贝努里模型的预处理和特征选择比较试验结果

城市	NP	P	P-IG	P-CHI
样本集一	93.63%	93.54%	93.54%	93.47%
样本集二	95.14%	95.44%	95.48%	95.88%
样本集三	91.15%	93.11%	93.04%	93.92%
样本集四	93.34%	94.51%	94.36%	94.73%
样本集五	96.61%	96.56%	96.59%	96.78%
平均值	93.97%	94.63%	94.60%	94.95%

采用了特殊的预处理后，向量的维度平均降低了 68.36%。从表 4.3 的数据可以看出采用预处理相比没有采用预处理，五个样本集宏观 F1 值的平均值提高了 0.66%，实验证明采用预处理后不仅降低了维度还提高了分类准确率。在预处理的基础上，采用 IG 和 CHI 两种特征选择算法。实验结果证明，特征选择不仅降低了向量维度，而且不损失分类准确率。P-IG 相比 P，五个样本集的向量维度平均降低了 48%；P-CHI 相比 P，五个样本集的向量维度平均降低了 62%。从表 4.3 的数据看到，P-IG 相比 P，宏观 F1 值的平均值几乎相当。P-CHI 相比 P，宏观 F1 值的平均值提高了 0.32%，说明采用了 CHI 特征选择方法提高了分类的分类准确率。P-CHI 相比 P-IG，宏观 F1 值的平均值提高 0.35%。实验结果表明，在犯罪文本分类中，CHI 特征选择在维度的降低和分类准确率两个方面都表现出比 IG 略高的性能。

(四)客户端结果

- 数据源为“犯罪数据库”的文本分类应用：

以下是文本分类在数据源为“犯罪数据库”的应用。在提供关键字检索的基础上,对检索出的相应的案件文本结果集进行自动文本分类的功能。

为了提高文本分类的速度,应首先并对犯罪数据库中的案件进行文本自动分类预处理,将分类后得到的类别存放在表中的“案件类别”字段中。

考虑到公安案件文本的保密性,不能显示真实的案件文本。本文在犯罪数据库的某个真实案件文本的基础上,修改时间、人名、地点,得到测试案件文本:

“2006年5月24日17时30分至20时之间,李四(女,38岁)停放在海山路×号院子内的一辆车牌为浙L14456黑色骑式新世纪牌摩托车被案犯开锁盗走,现价值5000余元”。由于不能显示真实的案件文本,为保护真实数据,我们将搜索关键字限定为“李四”。使得结果只出现这个测试案件文本。

该应用实现和结果如下图4.14所示。

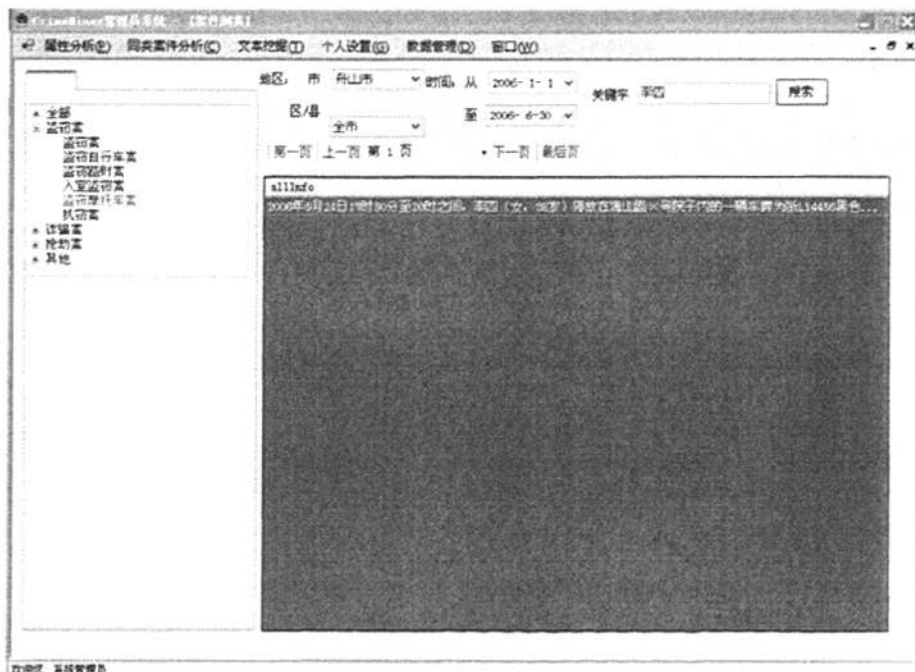


图 4.14 数据源为“犯罪数据库”的犯罪案件文本自动分类结果

点击内容,打开相应的案情公告如图4.15所示:

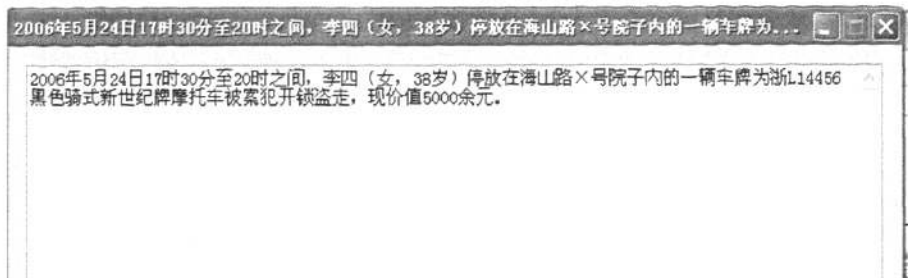


图 4.15 数据源为“犯罪数据库”的测试案件文本内容

由上图 4.15 可以看到, 该案件文本被自动分类到“盗窃摩托车案”类别中。

● 数据源为“网页文本数据库”的文本分类应用:

以下是文本分类在数据源为“网页文本数据库”的应用。本文将文本分类的功能与面向多网站的案情检索分析模块相结合。在多网站案情检索的基础上, 对检索的案情公告文本结果集进行自动文本分类。

为了提高检索和文本分类的速度, 首先对从公安内部网站中下载下来并抽取出案情公告信息存放在网站文本数据库中, 然后对所有的案情公告进行文本自动分类预处理, 将分类后得到的类别存放在表中的“案件类别”字段中。

考虑到公安案件文本的保密性, 不能显示真实的案件文本。本文在《浙江省研判平台—串并信息》网站上下载的某个真实案件文本的基础上, 修改时间, 人名, 和地点, 得到测试案件文本: “2009 年 1 月 1 日上午, 张三报案称: 1 月 1 日晚 8 点至 2 日上午 8 点, 城关镇某大酒店二楼餐厅收银台门被撬, 抽屉内财物被盗。被盗现金 22779 元, 还有香烟, 总计价值 3 万多元”。标题为: “某大酒店收银台门被撬, 请串并”。

由于不能显示真实的案件文本, 为保护真实数据, 我们将搜索关键字限定为“张三”。检索范围选择“题目和内容”。使得结果只出现这个测试案件文本。

该应用实现和结果如下图 4.16 所示。

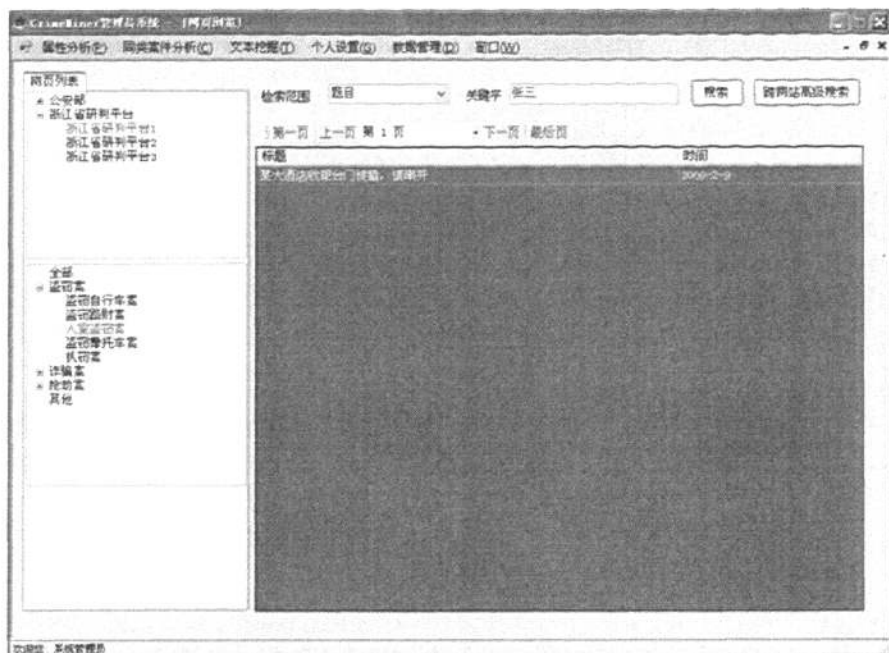


图 4.16 数据源为“网站文本数据库”犯罪案件文本自动分类
点击标题，打开相应的案情公告如下图 4.16 所示：

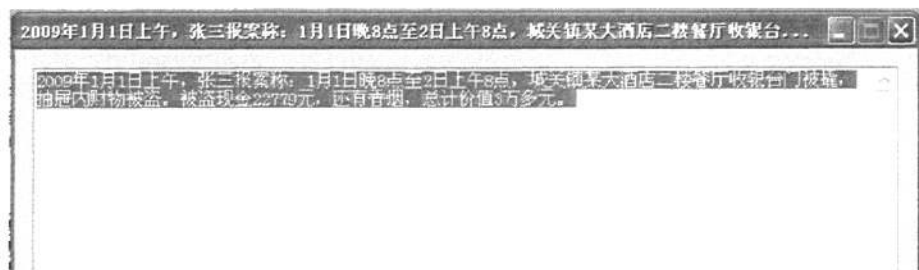


图 4.17 数据源为“网站文本数据库”的测试案件文本内容
由上图 4.16 可以看到，该案件文本被自动分类到“入室盗窃案类别”中。

4.6 本章小结

本章首先介绍了浙江省公安厅犯罪特征数据挖掘系统，其次介绍了文本挖掘子系统的系统结构，文本挖掘子系统采用三层 C/S 结构，业务层中主要实现了案件文本分类模块，相似案件文本检索模块。数据层中主要实现了公安内部网页抽取模块。

在设计和实现方面，首先介绍了公安内部网页抽取模块的设计和实现。

其次重点介绍了相似犯罪案件文本检索模块和犯罪案件文本分类的设计与实现。包括两个模块的结构设计，类设计，以及实验和结果分析。

在相似犯罪案件检索模块：实验实现了属性信息抽取，同义词语义分析，相似案件检索的功能。通过实验实现了相似犯罪案件检索在犯罪数据库以及公安内部网站案情通告两种数据源中的应用。

在犯罪案件文本分类模块：通过实验结果证明了，改进的多变量贝努里模型大大提高了多变量贝努里模型的分类准确率，同时也表现出比多项式模型相当甚至略好的分类准确率。并且在犯罪文本分类中，CHI 特征选择在维度的降低和分类准确率两个方面都表现出比 IG 略高的性能。实验还把文本分类应用于犯罪数据库以及公安内部网站案件通告两种数据源中。

第5章 总结和展望

本文对公安内部数据库和网络中大量的非结构化或半结构化的案件文本,进行了相关文本挖掘技术的研究和应用,应用和实现了诸如中文分词,特征选择,信息抽取,文本相似度计算,文本分类的技术,并对其中的一些技术结合实际应用进行了修正改进。主要贡献有:

(1) 在文本预处理方面。结合实际应用需要,对公安业务中的一些术语建立专业词库,并对中科院 ICTCLAS 分词组件的结果进行修正;同时针对案件文本的特征,提出了具有针对性的特殊预处理。

(2) 在案件特征选择方面。根据实际应用的需求,研究确定了案件文本信息抽取的 7 种特征算法;并通过比较六种特征选择算法,确定了对案情文文本挖掘有利的特征选择算法。通过实验结果表明在犯罪文本分类中,CHI 特征选择在维度的降低和分类准确率两个方面都表现出比 IG 略高的性能。

(3) 在案件分类挖掘方面。建立了公安领域专用的同义词词典。提出了案件属性信息抽取方法和同义词语义分析方法,并在此基础上提出了改进案件相似度计算方法;根据犯罪案件文本类别不均衡的特征,改进了朴素贝叶斯中的多变量贝努里模型,提出了面向不均衡类别的改进朴素贝叶斯案件文本分类方法。通过实验结果证明了,改进的多变量贝努里模型大大提高了多变量贝努里模型的分类准确率,同时也表现出比多项式模型相当甚至略好的分类准确率。

(4) 在应用系统设计方面。设计和实现了一个典型三层 C/S 结构的犯罪案件文本挖掘系统,实现了相似犯罪案件文本检索模块和犯罪案件文本分类模块。

当然由于时间有限能力不足,在一些方面也留下了缺憾,以下是总结工作中遗留的问题以及对未来的展望:

(1) 在文本预处理方面。由于公安领域术语庞大,很难建立完善的专业词典。实验中的跟公安专业词典是通过人工进行的,由于人手不足,时间有限等限制,无法建足够庞大的公安专业词典。公安专业词典有待继续添加和改进。

(2) 在案件特征选择方面。案件属性信息抽取算法大部分是根据词典匹配的方法进行抽取的,这对各属性的词典依赖程度很大。该方法存在着词典构造困难,无法识别未登录词等问题。希望在以后的工作中能使用基于统计和规则等其他算法来改进案件属性抽取算法。同义词语义分析中,,本文采用建立公安领域专用的同义词词典,同样存在词典不够完善的问题,以往在以后的工作中能添加和改进。

(3) 案件文本分类方面。本文的案件文本分类算法在测试集中大部分案件文本有明确类别的情况下准确度高。然而在实际中,有很多案件文本是很难划分到某一个具体的类别中,业务人员一般将这些案件文本放在“其他类别”中。在这种情况下,案件文本分类的准确度得到了一定程度的较低和损坏。因此在未来的工作中,将进行改进案件文本分类算法,更好地提高实际应用中的准确率。

参考文献

- [1] 周雪忠, 吴朝晖. 文本知识发现: 基于信息抽取的文本挖掘[J]. 计算机科学, 2003, 30(1): 63-66
- [2] 梁循. 数据挖掘: 建模, 算法, 应用和系统[J]. 计算机技术与发展, 2006, 16(001): 1-4.
- [3] 谌志群, 张国焯. 文本挖掘与中文文本挖掘模型研究[J]. 情报科学, 2007, 25(007): 1046-1051.
- [4] Sheng T, Wu Y F, Li Y X. Deploying approaches for pattern refinement in text mining[C]. Hong Kong, China: Proceedings of the Sixth International Conference on Data Mining, 2006: 1157-1161.
- [5] 鹿小明. 文本挖掘及其在信息检索中的而应用[J]. 情报资料工作, 2004(4): 26-28.
- [6] 王卫平, 郭长旺. 文本挖掘在科技情报中的应用[J]. 中国科技产业, 2004(012): 35-37.
- [7] 李颖, 阎保. 平文本挖掘在互联网信息统计中的研究与设计[J]. 微电子学与计算机, 2005, 22(001): 62-65.
- [8] 赵蕴华, 桂婕等. 基于深度标引的专利文本挖掘框架研究[J]. 数字图书馆论坛, 2008(11): 1-7.
- [9] 杨志豪. 面向生物医学领域的文本挖掘技术研究[D]. 大连理工大学博士学位论文, 2008.
- [10] Hsincbun C, Wingyan C, Jennifer J X. Crime data mining: a general framework and some examples[J]. IEEE Computer, 2004, 37(4): 50-60.
- [11] Lee W, Stolfo S, Mok K. A data mining framework for building intrusion detection models[C]. Oakland, CA : Proceedings of IEEE Symposium on Security and Privacy, 1999: 120-132.

- [12] Xu J, Chen H. Criminal network analysis and visualization[J]. Communications of the ACM, 2005, 48(6):107.
- [13] Chau M, Xu J, Chen H. Extracting meaningful entities from police narrative reports[C]. Los Angeles, California, USA: In Proceedings of the National Conference for Digital Government Research, 2002: 271-275.
- [14] Chen H, Chung W. Crime data mining: an overview and case studies [C]. Boston, MA :Proceedings of the 2003 annual national conference on Digital government research, 2003: 1-5
- [15] Wang G, Chen H, Atabakhsh H. Automatically detecting criminal identities deception: A Adaptive Detection Algorithm[C]. IEEE Transactions on Systems Man and Cybernetics Part a Systems and Humans, 2006, 36(5): 988-999.
- [16] Appavu A B, Rajaram R. Suspicious E-mail detection via decision tree: A data mining approach[J]. CIT. Journal of computing and information technology. 2007,15(2):161-166.
- [17] Wu T, Pottenger W M. A semi-supervised algorithm for pattern discovery in information extraction from textual data[J]. Lecture notes in computer science, 2003: 117-123.
- [18] 徐冰, 郭绍忠, 黄永忠. 基于朴素贝叶斯分类算法的活跃网络结构挖掘[J]. 计算机应用, 2007,27(006):1548-1550.
- [19] 夏咏梅. 基于文本挖掘的分类与聚类技术[J]. 情报探索, 2005, 3(3).
- [20] 杨莉莉, 杨永川. 基于社会网络的犯罪组织关系挖掘[J]. 计算机工程, 2009, 35(015): 91-93.
- [21] 刘莉. 数据挖掘技术在《公安科技信息》数据库中的运用[J]. 中国人民公安大学学报: 自然科学版, 2006, 12(001) 59-62.
- [22] 袁军鹏, 朱东华, 李毅. 文本挖掘技术研究进展[J]. 计算机研究应用, 2006, 23(002): 1-4.
- [23] 孙铁利, 刘延吉. 中文分词技术的研究现状与困难[J]. 信息技术, 2009(007): 187-189.
- [24] 余战秋. 中文分词技术及其应用初探[J]. 电脑知识与技术, 2004: 81-83.

- [25] 刘群, 张华平, 俞鸿魁等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004, 41(8): 1421-1429.
- [26] 中科院. 中文自然语言处理开放平台[EB/OL].
http://www.nlp.org.cn/project/project.php?proj_id=6.
- [27] 许洪波, 程学旗, 王斌等. 文本挖掘与机器学习[J]. 信息技术快报, 2005, 3(2): 1-3.
- [28] 周茜, 赵明生, 赵明生. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3): 17-23.
- [29] 王维娜, 康耀红, 伍小芹. 文本分类中特征选择方法研究[J]. 信息技术, 2008, 32(012): 29-31.
- [30] Karl M S. Techniques for Improving the Performance of Naive Bayes for Text Classification [C]. Mexico: In Proceedings of CICLing, 2005: 682-693.
- [31] Yuan P, Chen Y, Jin H. MSVM-kNN: Combining SVM and k-NN for Multi-Class Text Classification[C]. Huangshan: IEEE International Workshop on Semantic Computing and Systems, 2008: 133-140.
- [32] 吕琳, 刘玉树. 文本自动分类技术和算法研究综述[J]. 计算机科学, 2004 B09(31):24-26.
- [33] 徐学可. 网页文本分类及其在搜索引擎中的应用[D]. 北京工业大学硕士学位论文, 2008.
- [34] 徐亚娟. 基于公安业务信息的文本挖掘技术研究与实现[D]. 浙江大学硕士学位论文, 2008.
- [35] McCallum A, Nigam K. A comparison of event model for Naive Bayes text classification[C]. Pittsburgh, USA: In AAAI-98 Workshop on Learning for Text Categorization, 1998(752): 41-48.
- [36] 李晓明, 闫宏飞. 搜索引擎原理, 技术与系统[M]. 北京: 科学出版社, 2005: 1-248.

个人简历

个人简历:

程春惠，女，1984 年 12 月生。

2007 年 7 月毕业于福建省福州大学计算机科学与技术系。

2007 年 9 月入浙江大学就读计算机应用硕士研究生。

2007.9~2009.4 参加浙江省重大科技攻关项目《长三角公安信息资源共享与开发利用中若干关键技术与典型应用研究》项目中第二子课题《基于公安业务信息的刑事犯罪特征数据挖掘技术的研究与应用》

2009.5.27 论文《面向不均衡类别的朴素贝叶斯犯罪案件文本分类》被计算机工程与应用

作者：[程春惠](#)
学位授予单位：[浙江大学计算机学院](#)

本文读者也读过(6条)

1. [赵俊杰](#) [基于文本挖掘技术的论文抄袭判定研究](#)[学位论文]2009
2. [刘华](#) [数据挖掘技术在公安犯罪行为分析中的应用研究](#)[学位论文]2008
3. [许军](#) [基于公安信息的数据挖掘应用研究](#)[学位论文]2006
4. [赵小玲](#) [基于粗糙集的web文本挖掘研究](#)[学位论文]2009
5. [李明](#) [数据清洗技术在文本挖掘中的应用](#)[学位论文]2008
6. [徐亚娟](#) [基于公安业务信息的文本挖掘技术研究与实现](#)[学位论文]2008

引用本文格式：[程春惠](#) [公安犯罪案件文本挖掘关键技术研究](#)[学位论文]硕士 2010