分 类 号 _____ 学号 M201372456

学校代码 10487 密级 _____

华中科技大学

硕士学位论文

面向刑事案件的精细分类与串并案分析技术研究

学位申请人： 夏 明

学 科 专 业： 计算机科学与技术

指 导 教 师： 周 可 教授

答 辩 日 期： 2016.5.26

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering

Research on Fine-grained Classification and Cluster Analysis for Criminal Cases

Candidate : Xia Ming

Major : Computer Science and Technology

Supervisor : Prof. Zhou Ke

Huazhong University of Science & Technology

Wuhan, Hubei 430074, P.R.China

May, 2016

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于　保密□，　在_____年解密后适用本授权书。

　　不保密□。

（请在以上方框内打“√”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘 要

随着信息技术的高速发展，公安领域的情报信息系统也面临着海量数据，主要是文本数据带来的巨大挑战，传统的手工处理方式已经难以满足业务上的需求，必须采用更加自动化、智能化的文本挖掘技术来提高办案效率。

面向刑事案件文本，重点研究案件精细分类和串并案分析这两个刑侦人员普遍关注的问题。

提出了基于朴素贝叶斯和关键词共现图谱的两级分类方法 TLC-NBK，该方法根据案件文本长度短、词频低、类别分布具有层次性和不均衡性的特点，首先在文档频率 DF 方法的基础上引入了词性特征，提出双因子评估算法进行特征选择，然后利用面向不均衡类别的多变量贝努利模型进行朴素贝叶斯分类，实现了一级案件类别的快速、准确划分；在第一级分类器的基础上，针对其所属的二级案件类别分别构建以文档集为基本单位的关键词共现向量，以关键词间的共现关系代替词频计算权重，并提出了逆类别频率因子对共现权重进行修正，最后采用简单向量距离算法实现二级案件类别的精细分类。此外，还利用同义词网技术消除了领域同义词对分类结果的干扰。

提出了基于案件特征的密度聚类方法，实现了系列案件的串并分析。该方法首先结合规则和字典从非结构化的案情描述信息中抽取出结构化的案件特征；接着定义了案件文本间的特征相似度计算公式，综合考虑了精细案件类别、案发时间和案发地点对案件特征相似度的影响，并采用层次分析法决策各维度的权重值；最后，借鉴经典密度聚类算法 OPTICS 的思想，提出了特征密度聚类算法 OPTICS-FD，能够有效的分析出系列案件的密集簇，辅助刑侦人员破案。

最后，通过实验对双因子评估算法、两级分类器、案件特征抽取和串并案聚类进行了测试。结果表明，在刑事案件文本挖掘领域，相比于传统方法，TLC-NBK 方法的准确率和召回率分别提升了 7.53%和 12.99%；OPTICS-FD 算法的缩减率与召回率分别达到了 66.52%和 91.25%，更好的支持了刑侦人员进行决策。

关键词：精细分类，串并案分析，两级分类器，双因子评估，特征相似度

Abstract

With the rapid development of information technology, the system of public security intelligence is facing great challenges brought by the huge amount of data, mainly the text data. The traditional manual approach has been difficult to meet the needs of the business, so it is very urgent to use more automated and intelligent text mining technology to improve work efficiency.

A new system has been achieved, which focuses on refined case classification and mining implicit cases that police concerns a lot.

A two-level classifier based on Naïve Bayes and keywords co-occurrence graph is proposed. The method takes the features of case text which includes short length, low frequency and hierarchical but imbalance distribution of categories. The method has two steps. At first, it brings the characteristics of word nature into the widely used formula DF, and proposed double-factor evaluation method to select features. And then uses Multivariate Benoit Model for uneven categories to achieve the first-level classification fast and accurately. Secondly, based on the result of first-level classification, we build keywords co-occurrence vectors for every document set of two-level categories, the weight of the keyword in vectors is calculated by words' co-occurrence that has been corrected by inverse class frequency factor. At last, we use simple vector distance algorithm to achieve the second-level classification. In addition, we use the synonyms network technology to eliminate the interference of the field synonyms.

A method based on feature density, which is used to mine implicit cases, is proposed. This method includes three steps. First of all, we extracted the structured features of case text from unstructured case description information. Secondly, we defined a formula to calculate feature similarity between case texts, which takes time of the crime, scene of the crime and two-level classification into consideration and used Analytic Hierarchy Process to decide the weight of these three features. Thirdly, we proposed a feature density clustering algorithm based on OPTICS, named OPTICS-FD. And this algorithm could assist investigators solve the case by finding the clusters of implicit cases efficiently.

Finally, experiments on double-factor evaluation, two-level classifier, case feature extraction and implicit cases clustering were tested. The results show that, compared to the

traditional ways, our methods have enhanced three indicators include accuracy, recall and F-measure.

Keywords: Fine-grained classification, Cluster analysis for series of cases, Two-level classifier, Double-factor evaluation, Feature similarity

目 录

1 绪论

1.1 课题研究背景

随着信息技术的高速发展，公安领域的情报信息系统也面临着海量数据带来的巨大挑战。由于刑事案件对于社会稳定和百姓生活具有深刻的影响，是公安机关工作的重中之重。面对日益增长的大量案件数据（主要是文本数据），传统的人工处理方式已经难以满足业务上的需求。如何快速、有效的对案情进行分析、挖掘和解读是目前公安领域面临的主要问题。

经过在某地区公安局长达一年的深入工作与研究发现，案件类别作为案件文本最重要的一个特征，其分类的准确率和精细度是后续分析工作的关键。例如串并案分析和人案关联等工作都是围绕着案件类别为核心开展的。此外，刑侦人员对刑事案件类别的划分提出了严格的要求：在确保一级案件类别（如盗窃、诈骗等）分类准确率的前提下，进一步细化二级案件类别（如撬门入室盗窃案、摩托车盗窃案、丢包诈骗、信用卡诈骗等）。刑事案件文本除了严格的精细化分类要求外，本身也具备了一些不同于一般文本的特征：文本篇幅短小，词频较低；案件类别的分布具有层次性和不均衡性。刑事案件的文本长度一般都在 30 到 150 个字之间，属于短文本类别，并且，即使是关键性的案情描述信息通常也只出现了一次；现实生活中发生的刑事案件主要以“两抢一盗”和诈骗案为主，其他类别的案件相对较少，类别间的分布具有不均衡性；此外，案件类别本身还具有层次性特征，每一个一级案件类别都包含了若干个二级案件类别，例如抢劫案可细化为持刀抢劫、持枪抢劫、迷药抢劫和捆绑抢劫等等。

除了案件类别这一最重要的特征外，刑事案件文本还包含了案发时间、案发地点、作案手法和工具等信息，如何从非结构化的案情描述中抽取这些特征，也是刑侦部门自动化办案的重点和难点。

案件特征抽取成功后，可以将其应用在串并案分析、人案关联、公安领域知识库构建等方面，更有效的支持办案人员进行决策。其中，串并案分析作为打击系列犯罪案件的重要方法，可以挖掘案件之间的内在联系，减轻分析人员的工作量，提高破案效率。特别是将精细分类后的案件类别作为一个关键维度，再结合案件发生的时空维

度，将其应用在串并案分析后，可以有效的分析和预测犯罪主体的行为模式和活动特征，辅助刑侦人员快速破案。

因此，本文在深入理解公安刑侦工作业务需求的基础上，以案件类别的精细化分类问题为核心，结合自动化案情特征抽取技术对系列案件进行串并分析。在此过程中，对刑事案件的文本挖掘技术进行了深入的研究和探讨，提出了若干切实可行的算法与方案，并在实际的刑侦工作中取得了良好的效果。首先根据刑事案件文本的类别特征，提出了基于朴素贝叶斯和关键词共现图谱的两级分类方法 TLC-NBK（Two-Level Classifier based on Navie Bayes and KeyGraph），相比于传统分类算法，该方法实现了一、二级案件类别的准确划分，满足了刑侦人员的精细化分类要求；然后，采用基于规则和字典相结合的自动化信息抽取技术，对案件文本中的其他案情特征，如案发时间和案发地点等进行快速、准确的抽取；最后，结合精细化分类后的案件类别和自动化抽取得到的案件时空特征进行串并案分析，在经典密度聚类算法 OPTICS 的基础上，本文提出了案件特征相似度计算方法，改进后的算法称之为 OPTICS-FD （OPTICS Based on Feature Density），实现了系列案件的有效挖掘与发现，推进了公安刑侦工作的自动化进程，有效的支持了刑侦人员打击犯罪和防控犯罪的能力。

1.2 国内外研究现状分析

随着科技的迅猛发展，刑事犯罪的方法和手段也变得更加多样化，这就要求刑侦人员不断加强打击力度，通过更加智能化、信息化的工作方式提高办案效率。目前，国内外的学者越来越关注公安领域的数据挖掘技术研究，主要包括案件分类[1-3]与聚类分析[4-6]、关联分析[7,8]、信息抽取[9-11]、异常检测[12]、群体（团伙）发现[13,14]和人脸检测与识别[15,16]等。数据挖掘的对象也逐渐从文本扩展到图片、视频等多媒体数据，但主要还是以文本为主。

2010 年程春惠[17]提出了一种基于朴素贝叶斯的案件分类方法，根据案件分布的不均衡特征，对多变量贝努利模型进行了改进。并且将案件的属性信息与同义词的语义分析方法相结合，改进了案件相似度的计算方法，有效的提高了案件分类的准确率。

2012 年 X Shang 和 Y Yuan[18]将社交网络分析（SNA）技术应用于犯罪团体的挖掘与发现。他们根据犯罪网络中的隐性知识来挖掘嫌疑人之间的关系和特征，并构建了

一个综合指标模型。

2014 年 J Hosseinkhani[19]将文本挖掘技术应用在网页犯罪信息的研究中，首先从邮件、博客和网页中抽取案件特征，然后分析其中的犯罪热点并预测犯罪趋势。

2014 年 KR Rahem 和 N Omar[20]采用基于规则的信息抽取技术从新闻中提取出与毒品相关的信息，包括贩毒者藏匿毒品的方法，毒贩的国籍，毒品的种类及其在当地的数量和价格等信息。

2014 年杨军[21]结合 PCA 技术对隐藏朴素贝叶斯分类器进行了改进，对案件性质、发生时间与地点、犯罪历史等信息进行分析，从而发现犯罪规律和特征。

2015 年韩彦斌[22]开发出一种针对警务应用的移动终端人像采集系统，能够对非约束条件下的人像进行快速、灵活、准确的采集。

2015 年吴文浩[23]提出了一种多时间尺度密度的聚类算法，并将其应用于犯罪案件的时空特征分析中。

2015 年 R Niu[24]利用自然语言处理技术分析了微博中的犯罪信息，首先抽取犯罪文本的词性和语义特征，然后采用支持向量机方法来训练分类模型，最后结合 SMART 方法以提高分类的准确率。

根据上述的研究可以发现，针对公安领域的数据挖掘工作在不断的深化，不过虽然目前取得了不错的进展，但是仍具有很大的提升空间。特别是针对案件文本的分类问题，现有的分类方法都没有考虑案件类别分布的层次性，无法满足二级案件类别的精细化分类要求。此外，串并案件的特征相似度也并没有相关的学者进行定量的分析与研究，导致案件串并分析的效果还有待提升。因此，本文一方面提出了基于朴素贝叶斯和关键词共现图谱的两级分类方法 TLC-NBK，相比于传统的案件分类方法，更能满足刑侦人员对案件精细化分类的要求。另一方面将精细化分类的结果与特征抽取得到的案件时空信息结合后应用在串并案分析工作中，采用基于特征密度的聚类算法 OPTICS-FD 挖掘系列案件的密集簇，取得了较理想的结果。

1.3 本文的主要内容

本文针对刑事案件的文本挖掘技术，包括案件分类、特征抽取和聚类分析等进行了深入的研究和探讨。首先，根据刑事案件文本的特征提出了一种基于朴素贝叶斯和

关键词共现图谱的两级分类算法 TLC-NBK，该算法针对一级案件类别采用朴素贝叶斯方法进行粗分类，由于案件文本的长度短、词频低，在文档频率 DF 的基础上引入词性特征，提出了双因子综合评估特征选择算法，此外，考虑到案件文本分布的不均衡性，选择了面向不均衡类别的改进贝努利模型进行特征向量的表示和计算；在一级案件类别分类完成后，对其所属的二级案件类别，本文以类别文档集作为基本单元构建关键词共现图谱，将关键词共现关系代替词频作为度量指标，解决了二级案件类别间文本差异过小的问题，满足了精细化分类的要求。然后，介绍了一种基于规则和字典相结合的案件特征抽取方法，并以案件的时空维度为例进行了详细的阐述。最后，将精细化分类的结果和案件的时空维度相结合，采用基于特征密度的聚类算法 OPTICS-FD，对系列案件进行串并分析，有效的提高了刑侦人员的办案效率。

为了评估算法性能和效果，实验阶段首先测试了双因子综合评估算法对一级分类结果的影响；然后对比了两级分类算法与单纯的朴素贝叶斯算法对二级案件类别的分类效果；接下来对案件特征抽取的实验结果进行了分析；最后对基于特征密度的串并案聚类结果进行了测试与分析。实验结果表明，本文提出的两级分类方法和基于特征密度的聚类算法相比于其他传统方法在准确率和召回率等指标上均有所提升。下面将概述本文的研究内容组织。

第一章介绍了本文的研究背景和国内外相关研究现状，并以案件精细分类为切入点，结合案件特征抽取引出了系列案件的串并分析，研究和探讨了上述问题中涉及到的文本挖掘技术。

第二章介绍了本文方法所涉及到的各项技术和方法，主要包括特征选择文本分类算法和评价决策方法。

第三章详细介绍了本文方法的设计思路和具体实现，首先从整体上概述了系统的架构和整体流程，然后详细介绍了各个关键模块的实现。

第四章对本文提出的方法进行了详细的性能测试及结果分析，包括双因子评估算法、两级分类方法、特征抽取方法和特征密度聚类算法。

第五章总结了本文的研究内容，并对未来的工作进行了展望。

1.4 本章小结

本章首先介绍了课题研究的背景，包括当前公安领域信息化办公的现状和面临的主要问题，并以案件分类、特征抽取和串并案分析为切入点，介绍了文本挖掘技术应用在刑侦工作中的作用和难点；然后介绍了国外内学者对数据挖掘技术在公安领域的研究与应用现状；最后对本文的主要内容和组织进行了简单的介绍。

2 文本挖掘相关技术与方法

本章主要介绍本文所涉及到的相关技术与方法，主要包括特征选择、文本分类算法和评价决策方法。

2.1 特征选择

特征选择是指从文档中删除信息含量小的项，只选取部分最能反映类别统计特征的项，以提高分类的准确性和效率并降低计算的复杂度。在文本分类中常用的几种特征选择方法[25-27]包括：文档频率[28]、信息增益[29]、互信息[30]和χ^2统计量[31]等。

1. 文档频率（Document Frequency, DF）。是指在文档集中包含某个特征词的文档总数。当特征词的文档频率低于一定的阈值时，我们就认为该词对分类的结果影响较小，可以将其移除，其余的特征词就构成了文档的特征词向量，其计算公式如式 2-1 所示。

$$DF(t)=\frac{\text{含特征词}t\text{的文档数}}{\text{训练集中的文档总数}} \tag{2-1}$$

2. 信息增益（Information Gain, IG）。该方法通过统计特征词在文档中是否出现的次数来预测文档的类别，如式 2-2 所示。

$$G(t)=P(t)\sum_{i=1}^{m}P(C_i|t)\log P(C_i|t)-\sum_{i=1}^{m}P(C_i)\log P(C_i)+P(\bar{t})\sum_{i=1}^{m}P(C_i|\bar{t})\log P(C_i|\bar{t}) \tag{2-2}$$

可以根据特征词的信息增益值降序排列，选择前 k 个词；也可以给定一个信息增益阈值，去除小于阈值的特征词。

信息增益反映了特征词 t 对分类的信息量，但考虑了特征词在类中出现和不出现的两种情况，计算过于复杂。

3. 互信息（Mutual Information, MI）。用于表征两个变量的相关性，如式 2-3 所示。

$$MI(t,C)=\log\frac{P(t|C)}{P(t)} \tag{2-3}$$

互信息的特点是能够体现特征词与类别间的相关性，提取的特征词具有较强的类别区分能力，并能从高频词中移除噪声词，比较结果偏向于低频词。

4. χ^2 统计（Chi-square Statistic, CHI）。其基本思想与 MI 方法类似，都是用于度量特征词与类别间的相关性，但 MI 方法只考虑正相关对特征词重要程度的影响；如果特征词 t 与类别 C_i 反相关，则含有特征词 t 的文档不属于类别 C_i 的概率要大一些，这对于分类的指导也十分具有意义，χ^2 统计对此进行了改进，如式 2-4 所示。

$$\chi^2\left(t_j,c_i\right)=\frac{N\times\left(AD-BC\right)}{\left(A+C\right)\times\left(B+D\right)\times\left(A+B\right)\times\left(C+D\right)} \tag{2-4}$$

根据刑事案件文本的特点，本文在文档频率 DF 方法的基础上，引入了词性特征，提出了双因子综合评估方法作为第一级分类器的特征选择方法，可以有效减少低频词的干扰，降低计算复杂度，同时提高分类的准确率和效率。

2.2 文本分类算法

文本分类算法[32]是文本分类系统中最重要的组成部分，常用的文本分类算法包括 Rocchio 算法[33]、朴素贝叶斯[34]、k-近邻[35]、支持向量机[36]等。

1. Rocchio 算法。该方法的分类思想非常简单，首先为每个类别计算其文本集的算术平均，获得一个能够代表该类别的中心向量；然后将每个类别的中心向量与测试文本进行比较，计算它们之间的距离（相似度）；最后选择距离最近的类别作为测试文本所属分类，具体步骤如下所示：

（1）计算各类别文本集的算术平均，获得中心向量集合。

（2）计算测试文本的特征向量与各类别中心向量间的距离，如式 2-5 所示。

$$sim\left(t,d\right)=\frac{\sum_{k=1}^{M}W_{tk}\times W_{dk}}{\sqrt{\left(\sum_{k=1}^{M}W_{tk}^2\right)\left(\sum_{k=1}^{M}W_{dk}^2\right)}} \tag{2-5}$$

其中，t 为测试文本的特征向量，d 为某个类别的中心向量，M 为特征向量的维数，W_{tk} 为测试文本向量的第 k 维权值，W_{dk} 为中心向量 d 的第 k 维权值。

（3）从上一步的计算结果中选取一个最大值，测试文本将被分到该值对应的类别中。

本文在构建第二级分类器时，对 Rocchio 算法进行了改进，引入了关键词共现向量作为二级案件类别的中心向量，以关键词间的共现关系代替词频作为二级案件类别

特征的表述，进一步提升了精细分类的准确率。

2. 朴素贝叶斯（Navie Bayes, NB）算法。它是一种简单的线性分类器，遵守“贝叶斯假设”：样本的特征是相互独立的。它的基本思想是计算给定测试文本 d 属于不同类别的条件概率，条件概率值最高的类别即为文本 d 所属的类别，如式 2-6 和式 2-7 所示。

$$p\left(C_j \mid d\right)=\frac{p\left(C_j\right)p\left(d \mid C_j\right)}{\sum_{j=1}^{|C|}p\left(C_j\right)p\left(d \mid C_j\right)} \tag{2-6}$$

$$V_{\max}=\arg\max_{C_j \in C}p\left(C_j \mid d\right) \tag{2-7}$$

虽然，实际应用中的数据常常不能满足“贝叶斯假设”，一定程度上降低了分类的准确性。但是在语料库规模足够充分、文本类别间差异较大时，朴素贝叶斯分类方法仍达到了令人满意的准确率和效率。本文将朴素贝叶斯应用在第一级分类器中，对一级案件类别进行粗分类，取得了较理想的效果。

3. k-近邻（k-Nearest Neighbor, kNN）算法。该算法是由 Cover 和 Hart 于 1968 年提出的，通过计算文本间的相似度，找出训练文本集中与测试文本最相似的 k 个文本，然后根据这 k 个文本的类别来判断测试文本的类别，详细的算法流程如下所示：

（1）对训练样本集进行分词和预处理，转换成相应的特征向量集合；

（2）同样地对测试文本进行处理后，提取特征向量；

（3）从训练样本集中选取 k 个与测试文本最相似的训练文本；

（4）上一步中得到的 k 个训练文本所属的类别概率最大的即为测试文本的类别。

4. 支持向量机（Support Vector Machine, SVM）算法。该算法是由 VVap 领导的 AT&Bell 实验室提出的，基本思想是利用简单的线性分类器划分样本空间，如果是在前特征空间不可分的模式，则利用一个核函数将样本映射到高维空间中去，使得样本能够线性可分，再构造一个超平面，称之为决策平面，使得正负模式的间距最大。

2.3 评价决策方法

本文在计算案件特征相似度阶段需要综合考虑案件类别、案发时间和案发地点三个维度对结果的影响，因此需要一种评价决策方法来定量地分析上述三个维度在特征

相似度计算中所占的权值比重。

层次分析法（Analytic Hierarchy Process）简称 AHP[37-39]，是美国运筹学家 T.L.Saaty 教授于 20 世纪 70 年代正式提出，常被运用于复杂决策问题，主要包括以下几个步骤。

1. 建立层次结构模型。一般分为方案层、准则层和目标层，层与层之间的诸因素相互影响。

2. 构造判断矩阵。定量分析各层次各因素间的权重，按照九分位比率判定各个准则的优劣顺序，构造出判断矩阵。判定矩阵 $A=(a_{ij})_{n\times n}$ 具有如下特征：$\forall$ i, j$\in$N, 有 $a_{ij}>0$; $a_{ii}=1$; $a_{ij}=1/a_{ji}$，其中，i=j=1,2,...,n。称判断矩阵 $A=(a_{ij})_{n\times n}$ 为正互反矩阵。

3. 计算权向量。求解矩阵 A 的最大特征根 λ_{max} 及其特征向量 w。如式 2-8 所示。

$$Aw=\lambda_{max}w \tag{2-8}$$

4. 进行一致性检验。上述各要素权值的合理性依赖于一致性检验的结果。如式 2-9 所示。

$$CR=\frac{CI}{RI} \tag{2-9}$$

上式中 *CR* 为随机一致性比率，*CI* 为一般一致性指标，计算公式如式 2-10 所示。

$$CI=\frac{\lambda_{max}-n}{n-1} \tag{2-10}$$

RI 为平均随机一致性指标，其值如表 2-1 所示。

表 2-1 平均随机一致性指标 RI

阶数	1	2	3	4	5	6	7	8	9
RI 值	0	0	0.58	0.90	1.12	1.26	1.36	1.41	1.46

当 *CR* < 0.1 时，认为一致性可接受，否则调整矩阵 *A*。

本文采用层次分析法决策案件类别、案发时间和案发地点在特征相似度计算中所占的权重值。

2.4 本章小结

本章主要介绍了刑事案件文本挖掘所涉及到的相关技术和方法。其中，特征选择是为了不影响文本挖掘质量的前提下，去除非关键词，降低计算复杂度，比较了常见

的特征选择方法，并在文档频率 DF 的基础上引入了词性特征，改进为双因子综合评估算法；两级分类架构采用了朴素贝叶斯和简单向量距离分类方法作为一、二级分类算法的基础，并针对刑事案件文本特征进行了改进；最后，简要介绍了层次分析法的相关概念及应用。

3 案件精细分类与串并案分析方法的设计与实现

本课题是在与某地区公安局的合作项目基础上，深入了解一线刑侦人员的业务需求，以案件分类、特征抽取和串并案分析为切入点，研究并探讨文本挖掘技术在公安领域的应用。本课题总结了刑事案件文本挖掘的一般流程，设计并实现了一个完整的文本挖掘系统。本章主要介绍该系统的总体架构、工作流程以及关键模块的具体实现。

3.1 总体设计

3.1.1 系统架构

本课题的文本挖掘系统从层次上可分为存储层、数据处理层、分析挖掘层和应用接口层，如图 3-1 所示。

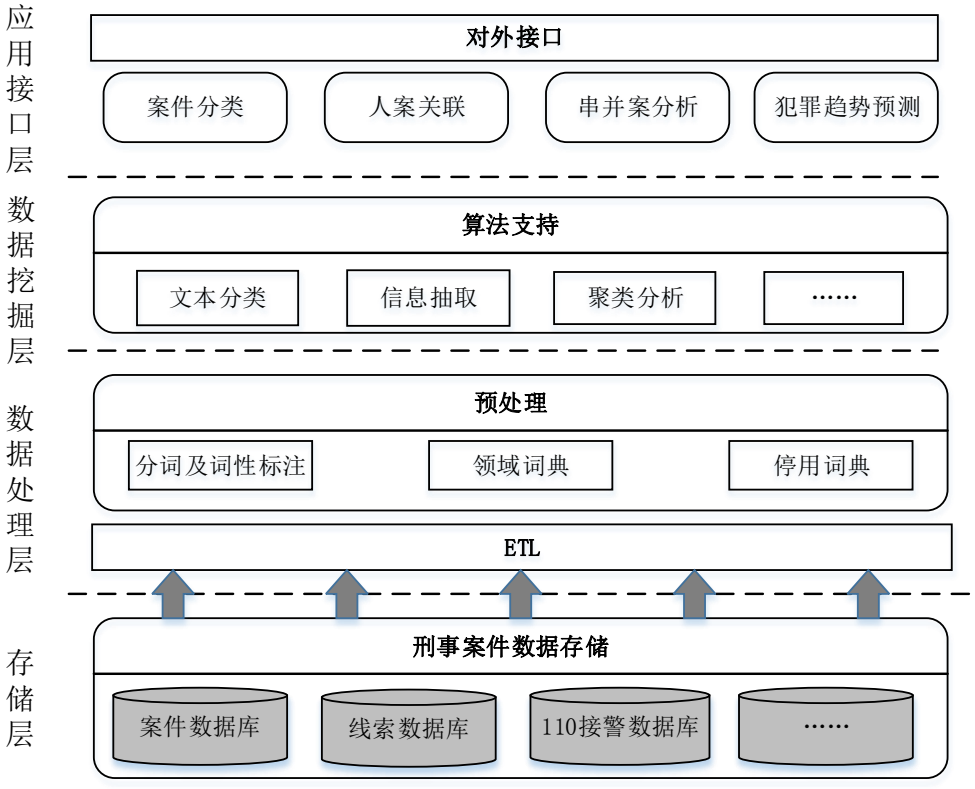

图 3-1 刑事案件文本挖掘系统架构

存储层主要负责刑事案件数据的可靠存储，根据公安业务的需求，又可分为案件数据库、线索数据库和 110 接处警数据库等。

数据处理层主要负责对数据进行预处理和加工。其中，ETL 模块用于处理不同数

据库规范带来的数据格式差异，向上层屏蔽底层存储细节。预处理模块负责中文切分词及词性标注，结合公安领域词典和停用词典更好的实现公安领域词汇的识别。

分析挖掘层是通过文本分类、信息抽取、聚类分析等各种算法和技术挖掘案件文本的内在规律和特征，智能化的分析和处理数据，提供决策支持。

应用接口层就是面向业务和用户提供的一系列功能接口，比如案件分类、人案关联、串并案分析、犯罪趋势预测等等。

3.1.2 整体流程

刑事案件文本挖掘的整体流程主要分为四个部分，分别是数据选择与预处理、文本结构化特征抽取、案件精细分类和串并案分析，如图 3-2 所示。

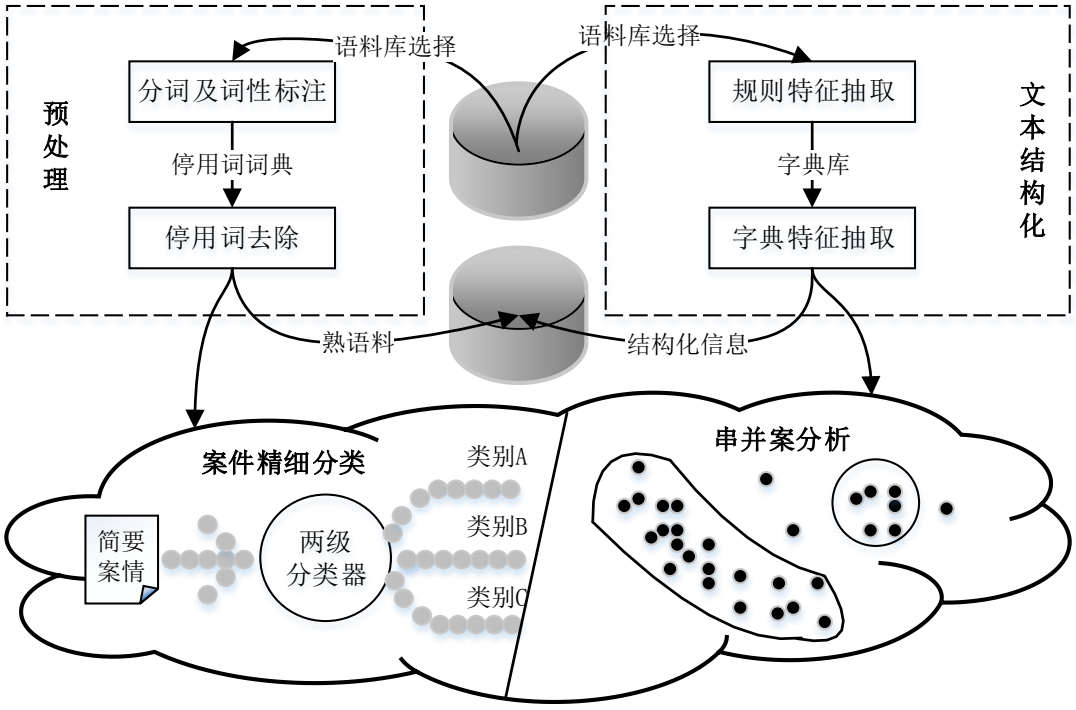

图 3-2 刑事案件文本挖掘流程示意图

第一部分的数据选择与预处理，主要分为语料库选择、中文分词及词性标注和停用词去除；第二部分的文本结构化主要是结合规则和字典进行特征抽取；第三部分是对刑事案件进行精细化分类，采用基于朴素贝叶斯和关键词共现图谱的两级分类方法；第四部分就是将前面得到的分类结果和案件时空特征，采用基于特征密度的聚类算法分析系列案件的密集簇。

3.2 数据选择与预处理

公安情报信息系统由于业务的需求，划分成多个相互交叉的数据子系统，不同子系统间的数据格式与规范存在着较大差异，复杂的数据环境给刑事案件的文本挖掘工作带来了巨大的挑战。

本文首先对案件数据进行预处理与加工，减少因数据不完整性和异常性带来的干扰，以提高文本挖掘的性能和效果。本文的数据预处理过程可分为三个阶段，分别是：语料库选择、中文分词及词性标注和停用词去除，详细的预处理流程如图 3-3 所示。

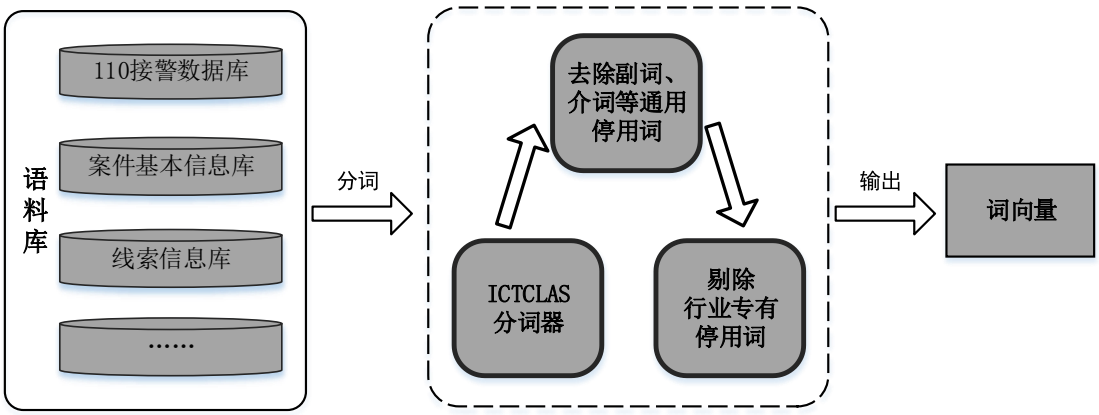

图 3-3 预处理流程

3.2.1 语料库选择

刑事案件的语料库可分为训练语料和测试语料。训练语料主要用于计算先验概率等，对标注后的训练语料进行学习可以得到相应的模型，然后利用测试语料进行验证。因此，语料库的选择恰当与否对案件文本挖掘的性能和效果具有较大影响。理想的语料库应能模拟真实的应用场景，广泛的代表公安系统所要处理的、实际存在的各个类别的文本。本文的语料库主要来源于某地区公安局提供的真实的案件文本信息，包括 110 接处警信息库、案件基本信息库、线索信息库等等。

为了保证文本挖掘的准确性和可靠性，本文共选取 8 种一级案件类别及其所属的 52 种二级案件类别，共计一万余条真实案件信息，并按照 1:2 的原则划分为训练语料和测试语料，每一个案件文本中几乎都包含了案发时间、案发地点、作案手法等案件特征信息。

3.2.2 中文分词及词性标注

英文语料以空格作为单词间的自然分隔符，而中文语料中的词与词之间不存在明显的分隔符，不同的分词结果会对文本挖掘的结果产生一定的影响。因此，本文一方面采用中科院的分词系统 ICTCLAS 进行基本词汇切分和词性标注，并一方面结合刑事案件领域词库对公安领域的专有词汇，例如“故意伤害”“持刀抢劫”“擂肥”等进行更科学的划分。

刑事案件的领域词库汇总了公安部门多个子系统的领域知识积累，主要包括重点人员姓名字典、作案手法字典、作案工具字典等等。本文对刑事案件文本的分词进行了改进，通过建立刑事案件的领域词库，并将其应用在分词环节，有效提高了刑事案件文本的分词效果。

3.2.3 停用词去除

停用词是指对文本表示几乎没有意义，反而会降低文本挖掘准确率和效率的噪声词汇。去除这些词可以降低特征词的维度，提高文本挖掘的效果。本文主要从两方面对刑事案件文本进行停用词过滤：去除无用词性和去除专有停用词。

1. 去除无用词性。刑事案件文本中包含了一些不代表任何案件特征的词性，如拟声词、副词、介词和连词等。本文根据 ICTCLAS 标注的词性信息，去除这些与案件特征无关的词性。

2. 去除专有停用词。刑事案件文本中包含了许多频次高，却与案件特征无关的词，并且它们的词性构成非常复杂，无法轻易的通过词性进行剔除，例如“据称”、“价值”、“报案”等。这类词汇属于公安领域的专有停用词，无法采用传统的方法进行筛选。因此，本文针对公安领域建立了专有停用词表，有效的解决了该类词汇对案件文本挖掘结果的干扰。

3.3 基于朴素贝叶斯和关键词共现图谱的两级分类方法实现

刑事案件数据不同于一般的文本数据，它具有如下特征。

1. 文本篇幅短小，词频较低。刑事案件的描述信息要求简洁精炼，导致文本的长度通常在 30 到 150 个字之间，属于短文本类型。由于文本篇幅较短，词语出现的频率

也相应较低，即使是非常重要的关键性词语通常也只出现一次，使得传统的文本分类方法难以适用。

2. 案件类别分布具有层次性。刑事案件类别存在着天然的层次关系，例如，“盗窃”、“抢劫”和“诈骗”相互独立，属于同一层次；“诈骗”和“电信诈骗”、“丢包诈骗”之间则具有从属关系。本文选取了 8 大类 52 种子案件类别，从属于同一种大类别下的案件描述信息差异较小；不同大类间的案件描述信息差异较大。为了方便区分，本文将大的案件类别称之为一级案件类别，子类别则称为二级案件类别。

3. 案件类别分布具有不均衡性。例如，现实生活中发生的刑事案件主要以“两抢一盗”和诈骗案为主，其他类别的刑事案件相对较少，这使得案件文本的类别分布具备了不均衡性。

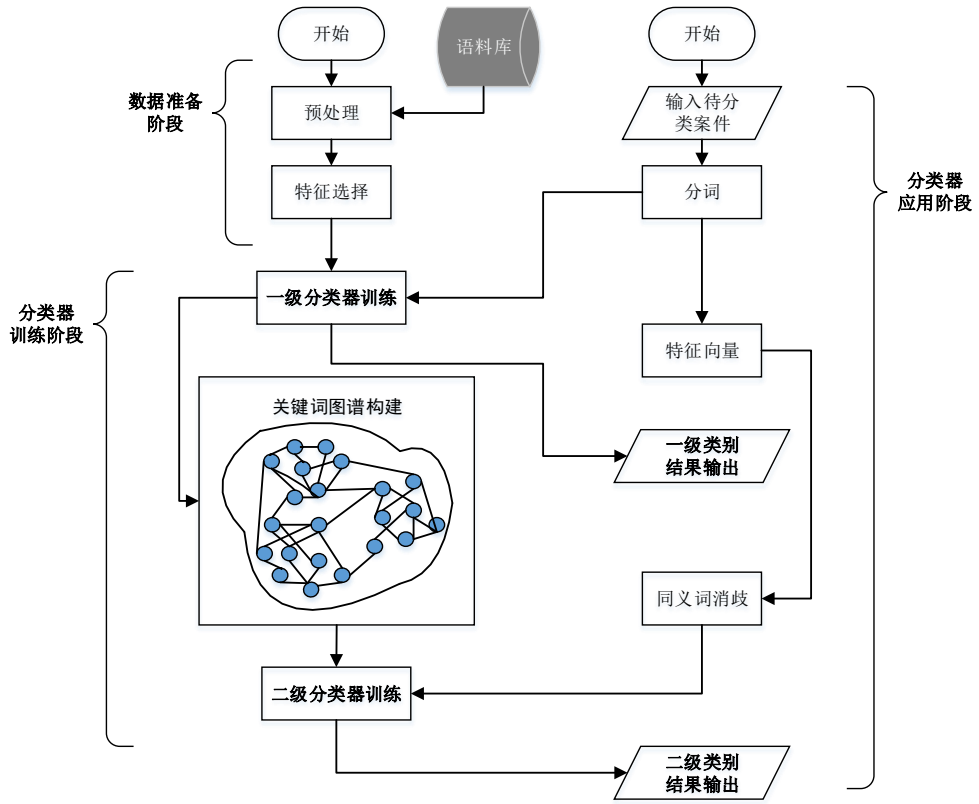

图 3-4 刑事案件文本的两级分类算法流程

鉴于上述分析，本文提出了结合朴素贝叶斯和关键词共现图谱的两级分类方法 TLC-NBK。基本思想是首先利用朴素贝叶斯对一级案件类别进行分类，考虑到刑事案件属于短文本且分布不均衡，提出了双因子评估算法进行特征选择，并采用了面向不

均衡类别的改进多变量贝努利模型，实验表明朴素贝叶斯对粗分类表现出了良好的性能和准确率；其次，借鉴 KeyGraph 思想[40,41]对二级案件类别构建关键词共现图谱，利用词语间的共现关系挖掘案件类别关键词，并借助同义词网屏蔽领域同义词的干扰，最后利用简单向量距离算法进行分类，有效解决了案件文本词频低和子类别差异小的问题。详细的分类流程如图 3-4 所示。

3.3.1 基于朴素贝叶斯的第一级分类器实现

朴素贝叶斯分类在训练语料足够充分且类别特征差异较大的情况下，表现出了良好的准确性和效率优势，非常适用于粗分类。因此，本文根据刑事案件文本的特点和公安业务的需求，提出了基于朴素贝叶斯的第一级分类器，实现了对一级案件文本的准确分类，如图 3-5 所示。

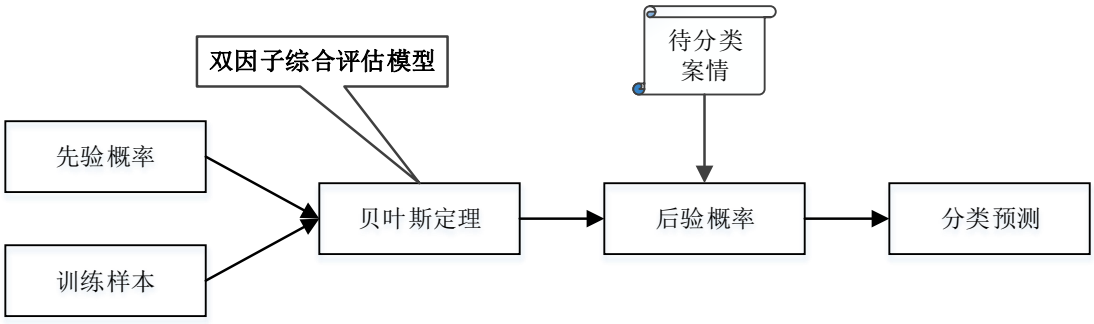

图 3-5 基于朴素贝叶斯的第一级分类器

3.3.1.1 双因子综合评估特征选择

案件文本 D 经过预处理后，被切分为由多个词语组成的特征向量$T(t_1, t_2, \dots, t_n)$。此时，特征向量 T 的维度 n 通常比较大，其中包含了许多跟案件类别无关的通用词汇，可以通过特征选择方法，在不损失分类准确性的同时，减少特征向量 T 的维度，降低系统的计算开销。

目前，常用的特征选择方法主要有：文档频率、互信息、信息增益和χ^2统计量等。然而，刑事案件文本特征稀疏、高度冗余的特点使得上述特征选择算法都难以有效的提取出特征词项，因此，本文在文档频率 DF 方法的基础上引入了词性因子，根据案件关键特征词的词性分布概率放大对应词性的权重，构造了双因子评估函数$Score(t)$，能够有效弥补短文本难以精确抽取特征词项的困难，双因子综合评估函数计算公式如式

3-1 所示。

$$Score(t)=\log W_{DF}(t)\times\left(1+W_{Pos}(t)\right) \tag{3-1}$$

其中，t 表示某个待评估词语，$W_{DF}(t)$ 表示词 t 对应的 DF 权值，如式 3-2 所示。

$$W_{DF}(t)=\frac{\text{包含词}t\text{的文档数}}{\text{文档总数}} \tag{3-2}$$

$W_{Pos}(t)$表示词 t 的词性权值，它相当于一个比例系数，放大重要词性所对应的权重值。例如，大部分案件特征词如“盗窃”、“诈骗”、“抢劫”等都属于动词，因此，在选择特征词时，应适当加大动词的权值，避免方位词、语气词等其他词性的干扰。通过统计案件文本中关键特征词项的词性分布概率，得到了最重要的四种特征词词性系数，如表 3-1 所示。

表 3-1 词性系数对照表

词性	比例系数
v	52.27%
n	34.68%
p	5.71%
a	3.94%

不属于上述四种词性的词项，在计算特征权重时，$W_{Pos}(t)=0$，相当于权值不变；而属于上述四种词性的词项，其权值会相应增大，对应的词性系数如表 3-1 所示，这样就提高了案件文本关键特征选择的准确率。双因子综合评估算法的具体流程如图 3-6 所示。

双因子综合评估算法

```
1: TermList ← 词项集合
2: for all term ∈ TermList do
3:     n ← Nature(term); df ← Df(term)
4:     score(term) ← calScore(n,df)
5:     scoreList ← score(term)
6: end for
7: Sort(scoreList);//排序
8: SelectedFeaturs ← topK(scoreList)//取前 k 个作为特征项
```

图 3-6 双因子综合评估算法流程

3.3.1.2 朴素贝叶斯分类

朴素贝叶斯文本分类[17]的任务就是将待分类文本归类到与其关联最紧密的类别集合的某一类别中，实际的算法过程可分为训练过程和测试过程两个阶段。

1. 训练阶段。首先计算每个类别的先验概率$p(C_j)$，如式 3-3 所示。

$$p(C_j)=\frac{N_{C_j}}{N} \tag{3-3}$$

其中，N_{C_j}是 C_j 类的训练文本数，N是总的训练文本数。

然后，计算特征词 w_i 属于类别 C_j 的概率 $p(w_i|C_j)$。$p(w_i|C_j)$ 的计算方法取决于采用何种模型，在第 3.3.1.3 节中将会详细介绍。

2. 分类阶段。根据朴素贝叶斯算法的独立性假设，计算测试文本 d 属于类别 C_j 的概率，如式 3-4 所示。

$$p(C_j|d)=\frac{p(C_j)\prod_{i=1}^{n}p(w_i|C_j)}{\sum_{k=1}^{|C|}p(C_k)\prod_{i=1}^{n}p(w_i|C_j)} \tag{3-4}$$

对于不同的类别来说，后验概率 $p(C_j|d)$ 的分母$\sum_{k=1}^{|C|}p(C_k)\prod_{i=1}^{n}p(w_i|C_j)$的值都是相同的，所以求最大后验概率等效于求$V_{max}$，如式 3-5 所示。

$$V_{\max}=\arg\max_{C_j\in C}p(C_j)\prod_{i=1}^{n}p(w_i|C_j) \tag{3-5}$$

综上所述，概率 $p(C_j|d)$ 最大的类别即为测试文本 d 所属类别。

3.3.1.3 面向不均衡类别的改进多变量贝努利模型

贝叶斯算法包括多项式和多变量贝努利两种模型。

1. 多项式模型。文本在该模型中被表示为特征词词频的向量模型。给定一个类别 C_j，计算特征词 w_i 属于类别 C_j 的概率如式 3-6 所示。

$$p(w_i|C_j)=\frac{1+\sum_{k=1}^{|D|}N(w_i,d_k)}{|V|+\sum_{s=1}^{|V|}\sum_{k=1}^{|D|}N(w_s,d_k)} \tag{3-6}$$

其中，$|D|$是类别 C_j 所包含的文本总数，$|V|$ 是特征词的总数，$N(w,d_k)$是特征词 w 在文档 d_k 中的词频。

2. 多变量贝努利模型。文本以二进制向量模型表示特征词是否出现。计算特征词

w_i 属于给定类别 C_j 的概率如式 3-7 所示。

$$p(w_i|C_j)=\frac{1+\sum_{k=1}^{k=|D|}B(w_i,d_k)}{|C|+|D|} \tag{3-7}$$

其中，$|C|$ 表示类别总数，$B(w_i,d_k)\in\{0,1\}$，当特征词 w_i 在文本 d_k 中出现时为 1，否则为 0。

由于刑事案件文本的文本长度短、词频低，适用于多变量贝努利模型进行表示。此外，刑事案件的文本分布还具有类别不均衡的特点，例如盗窃类别的案件数相对较多，绑架类别的案件相对较少。而训练语料的不均衡性会直接影响多变量贝努利模型的分类准确率，主要源自两方面原因。

1. 假设类别间的文本数目差异较大，当特征词 w_i 在所有类别中均未出现时，$p(w_i|C)=1/(|C|+|D|)$，也就是说概率会偏向所含文本数较小的类别。

2. 当特征词 w_i 在类别间出现的文档频率相同时，也会导致最终的概率偏向含文本数较少的类别。

上述两种情况都会降低多变量贝努利模型的分类准确性，因此，本文采用了面向不均衡类别的改进多变量贝努利模型[17]，如式 3-8 所示。

$$p(w_i|C_j)=\frac{\frac{|D|}{|D_{max}|}+\sum_{k=1}^{k=|D|}B(w_i,d_k)}{|C|+|D|} \tag{3-8}$$

其中，$|D_{max}|$表示最大的类别文档总数，$|D_{\max}|=\arg\max_{C_j\in C}|D_{C_j}|$，$|D_{C_j}|$表示类别$C_j$ 的文档总数。改进的多变量贝努利模型有效的改善了上述两种情况对分类准确性的干扰。

3.3.2 基于关键词共现图谱的第二级分类器实现

根据公安领域的实际业务需要，对刑事案件类别的划分提出了极为严格的要求：在保证一级案件类别能够准确划分的前提下，实现对二级案件类别进行精细分类。而上文已经介绍了刑事案件文本的特征，从属于相同一级案件类别的二级案件类别间文本差异非常小，例如“撬门入室盗窃案”和“翻窗入室盗窃案”之间的文本差异可能只有一至两个词。传统的面向文档层面的分类方法已经难以适用。因此，本文在构建

第二级分类器时，以二级案件类别的文档集，而不是文档作为特征选择的基本单元。然后，为所有二级案件类别文档集分别构建关键字共现图谱，以关键字间的共现关系来代替词频作为特征选择的度量指标，结合公安领域同义词网提高分类准确率，最后利用简单向量距离算法构建第二级分类器。案件的关键词向量构建流程如图 3-7 所示。

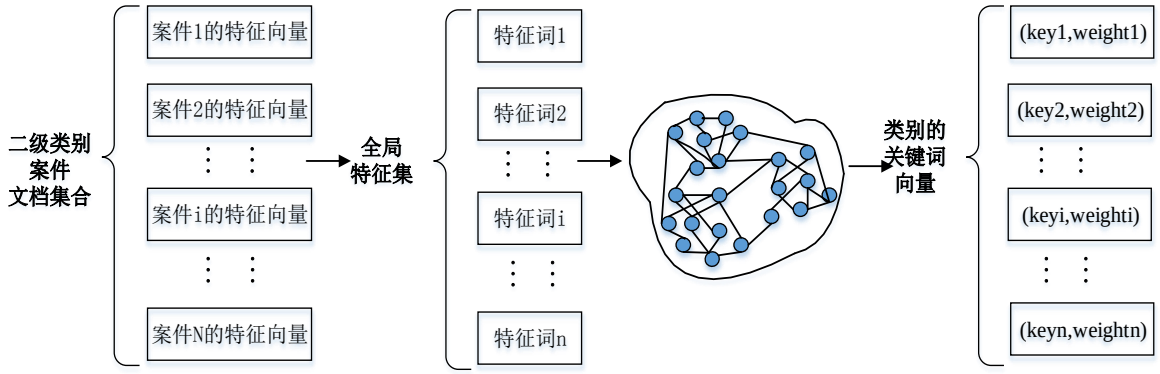

图 3-7 案件的关键词向量构建

3.3.2.1 关键词共现图谱

关键词共现图谱是一种利用词语间共现关系提取文档主题关键词的方法[40,41]，它不仅能够提取文档中的高频关键词，还可以提取能够表达主题意义的低频词。它的基本思想是首先提取文档中的频繁项作为实体顶点，计算频繁项之间的共现度，在共现度超过指定阈值的实体顶点间建立一条实边；其次，找出所有的连通子图，每一个连通子图视为一个岛群；然后，在图中添加能够将岛群连接起来的新项，该项必须满足与岛群之间共现度的要求，这样的项称为虚心顶点，虚心顶点与岛群之间的连接被称之为桥。

由于案件文本的长度特别短，在计算词语共现度时不以句子为单位，而是以文本 D 为基本单位，为每一个二级案件类别文档集构建一个关键词共现图谱 G。然后根据 G 提取文档集的关键词共现向量 $T(t_1,w_1;t_2,w_2;\ldots;t_n,w_n)$，其中 t_i 为G中的顶点，w_i为 t_i 与其他顶点的平均共现度频率，详细流程主要分为以下几步：

1. 提取高频词。定义 $N\times 1$ 的频繁度矩阵 H_f，N 为文本集D_s中词的总数，如式 3-9 所示。

$$H_f[i]=|i|,i\in D_s \tag{3-9}$$

扫描文本集D_s，计算 H_f。将超过指定阈值的词汇加入到高频词集合 HF中。利用HF中的元素建立图 G 的顶点集合，记图 G的顶点数为 N_f。

2. 计算高频词的关联度。定义 $N_f \times N_f$矩阵 $assoc$，如式 3-10 所示。

$$assoc[i,j] = \sum_{D \in D_s} min(|i|,|j|) \quad i,j \in HF \tag{3-10}$$

扫描文本集D_s，计算矩阵$assoc[i,j]$。选取排序靠前的少于 $N_f - 1$个的元素，在图 G 中将元素对应的两个顶点间增加一条边。此时，图 G 中的边数小于 $N_f - 1$，必为非连通图，每一个连通子图构成一个岛群g。岛群的总数即为 N_g，每一个群 g_i 所含的顶点个数为N_{g_i}。

3. 计算非高频词的关键度。定义$N_t \times N_f$矩阵$basic$，N_t表示非高频词的个数，如式 3-11 所示。

$$basic[i,j] = \begin{cases} \sum_{D \in D_s} |i||j| & i \neq j \\ 0 & \text{其他} \end{cases} \tag{3-11}$$

其中，i 表示非高频词，j 表示高频词。$basic[i,j]$可以根据岛群划分为 N_g个矩阵，每个矩阵$basic_j$为$N_t \times N_{g_j}$的矩阵。

定义$N_t \times N_g$矩阵$based$，记录非高频词与所有岛群中高频词的共现度，如式 3-12 所示。

$$based[i,j] = \sum_{k=1}^{N_{g_j}} basic_j[i,k] \tag{3-12}$$

定义 $1 \times N_g$矩阵$neighbors$，记录每一个岛群与非高频词之间的共现度之和，如式 3-13 所示。

$$neighbors[j] = \sum_{i=1}^{N_t} based[i,j] \tag{3-13}$$

定义$N_t \times 1$矩阵key，记录每一个非高频词 i 的关键度，如式 3-14 所示。

$$key[i] = 1 - \prod_{j=1}^{N_g} \left(1 - \frac{based[i,j]}{neighbors[j]}\right) \tag{3-14}$$

根据关键度矩阵key将文本集D_s中的非高频词汇降序排列，取前 r 个关键度最高的词，形成高关键度词集合KF。将KF中的词作为虚心顶点加入到图G中。

4. 构建关键词共现图谱。定义$r \times N_f$矩阵$column$，记录关键度词集KF中的词与高频词集HF中的词之间的共现度，如式 3-15 所示。

$$column[i,j] = \sum_{D \in D_s} min(|i|,|j|) \ i \in KF, j \in HF \tag{3-15}$$

若$column[i,j] \neq 0$，且图G中的顶点和之间没有边，则将边补全。图G中边的权值为两边顶点的共现次数，顶点的权值为与其他顶点的平均共现度频率。关键词共现图谱的实例如图 3-8 所示。

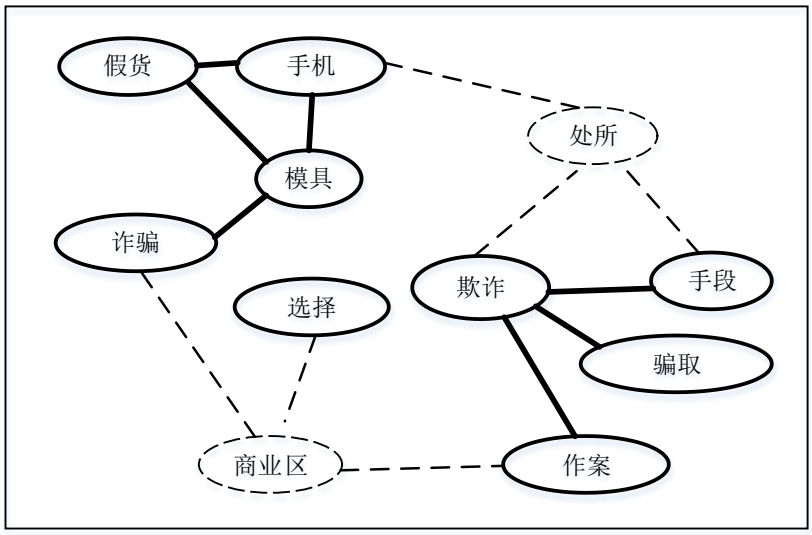

图 3-8 关键词共现图谱实例

5. 提取关键词共现向量。从图G提取关键词共现向量$T(t_1, w_1; t_2, w_2; \ldots; t_n, w_n)$，其中 t_i 为 G 中的顶点，w_i为 t_i 的权值。

6. 利用逆类别频率系数放大领域关键词权值。由于关键词共现向量 T 中包含了一部分出现频率很高的通用词汇，如“XX 区”、“发现”等。这些通用词汇与案件类别特征基本无关，应予以去除。而通用词汇的特点是在所有文档集中均频繁出现，因此，本文对关键词向量T中的每一个关键词权重 w_i 乘上一个逆类别频率系数 λ，放大领域词汇的权值，减少通用词汇的干扰，关键词t_i 的新权值 $Weight(t_i)$ 如式 3-16 所示。

$$Weight(t_i) = w_i * \lambda = w_i * \frac{\text{关键词共现向量总数}}{\text{含关键词}t_i\text{的关键词共现向量个数}} \tag{3-16}$$

3.3.2.2 同义词网

刑事案件文本中包含了许多表示相同含义的领域同义词，例如，“欺诈”、“欺

骗”和“骗取”表达的含义其实都是指诈骗，但却具有不同的表述形式，这在一定程度上影响了分类的准确性。因此，本文通过构建公安领域的同义词网实现刑事案件文本描述的规范化，减少领域同义词对分类结果的干扰。在 3.3.2.2 节计算逆类别频率系数和 3.3.2.4 节计算向量余弦相似度时均利用了同义词网提高计算的准确性。如图 3-9 所示为一个同义词子网示例。

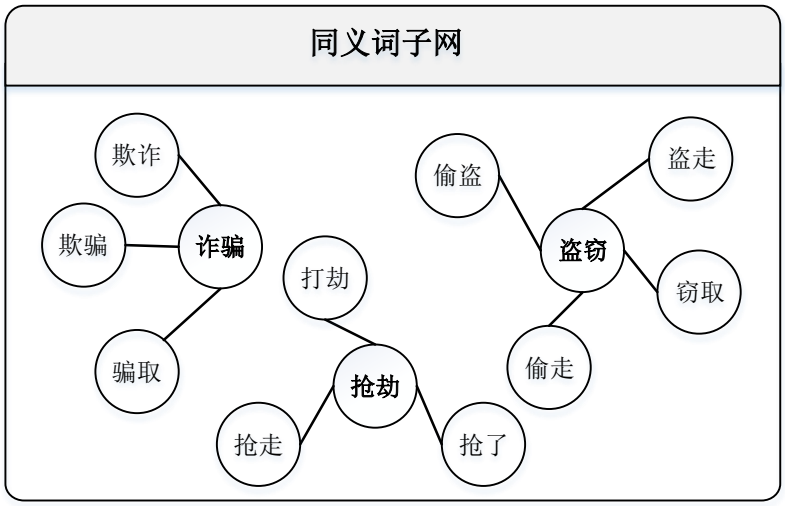

图 3-9 同义词网示例

3.3.2.3 简单向量距离分类算法

二级案件类别的关键词共现向量构建完成后，利用简单向量距离算法进行第二级分类，主要分为以下几步：

1. 计算权重向量。假设测试文本的词向量 $D=\{t_1', t_2', \ldots, t_X'\}$，给定某个二级案件类别的关键词向量 $T=\{t_1, t_2, \ldots, t_Y\}$，为了计算它们之间的距离，需要将其放置在同一个向量空间中，因此，首先需要对 D 和 T 进行“并”操作，如式 3-17 所示。

$$Term = D \cup T = \{term_1, term_2, \ldots, term_N\} \tag{3-17}$$

接下来计算测试文本和案件类别的权重向量DW和TW，如式 3-18 和式 3-19 所示。

$$DW = \{w_1', w_2', \ldots, w_N'\} \tag{3-18}$$

$$TW = \{w_1, w_2, \ldots, w_N\} \tag{3-19}$$

如果词向量 D 包含 $term_i$，则DW对应的权值 w_i' 为 1，否则为 0；如果关键词向量 T 包含$term_i$，则TW对应的权值w_i为 3.3.2.1 节中经过逆类别频率系数计算后的平均共

现频率，否则为 0。

2. 向量空间模型表示。将上一步得到的两个权重向量放置在每一个词为一个维度的空间向量中，如图 3-10 所示。

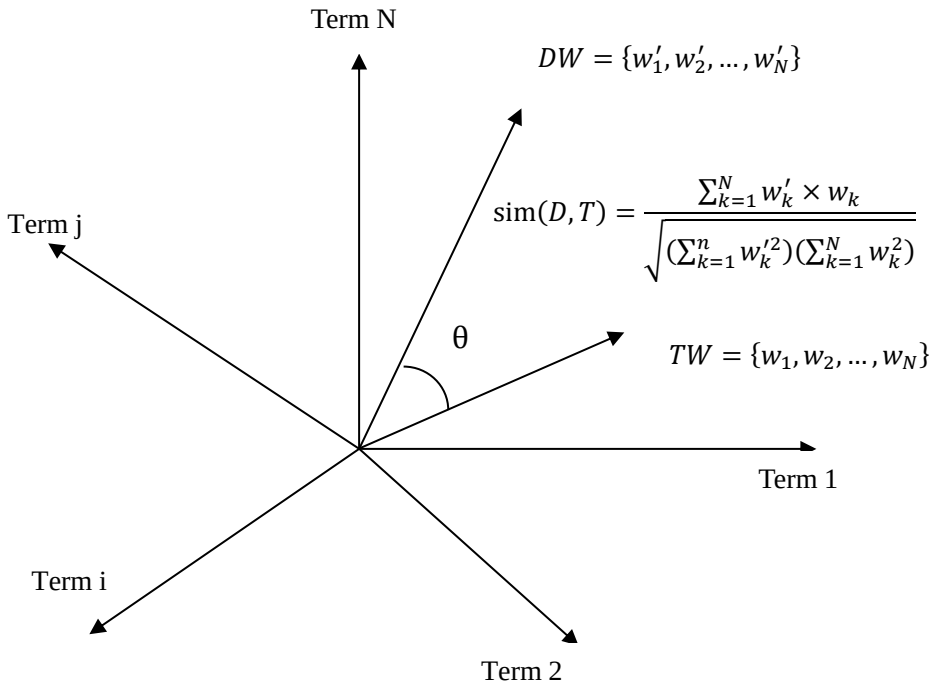

图 3-10 向量空间模型

DW与TW之间的夹角越小，表明测试文本与给定的案件类别越相似，因此，采用余弦相似度计算文本与类别间的相似度，夹角越小，余弦值越大，测试文本属于给定类别的可能性就越大。

3. 计算余弦相似度。可以直接根据测试文本和案件类别的权重向量DW和TW计算余弦相似度，如式 3-20 所示。

$$\mathrm{sim}(D,T) = \frac{\sum_{k=1}^{N} w'_k \times w_k}{\sqrt{\left(\sum_{k=1}^{n} w'^2_k\right)\left(\sum_{k=1}^{N} w_k^2\right)}} \tag{3-20}$$

4. 分类。计算测试文本与所有二级案件类别关键词共现向量间的余弦相似度，选择结果集中的最大值所属的类别作为测试文本的案件类别。

最终，经过一步步的分析处理后，我们就实现了刑事案件文本的精细化分类，其结果可以用于多方面的工作中，如人案关联、串并案分析、公安领域知识库构建等等。

3.4 基于特征密度的系列案件串并分析

串案和并案（简称串并案）是侦破系列案件的常用方法，通过把不同地域、不同时间发生的多起案件，根据其案情特征进行合并分析，便于警方发现犯罪主体并予以打击抓捕。其中，案件的类别信息作为案情的核心特征，是串并案分析过程中需要考虑的最关键的维度，案件类别的准确性和精确性直接影响着串并案分析的最终效果。

本文以案件的精细类别为核心，结合案件发生的时空维度，采用基于特征密度的聚类算法对系列案件进行串并分析，一方面减少了传统串并案分析中维度过多带来的计算复杂度和无用维度的干扰；另一方面细化了案件类别特征，并定量分析了案件的特征相似度，有效的提高了聚类的准确性。

3.4.1 基于规则和字典相结合的案件特征抽取方法

串并案分析需要结合案件的多个特征维度综合考量。其中，案件特征是描述案情的关键性信息，包括案发时间、案发地点、涉案人员、作案工具等等。上文介绍的案件类别更是代表了案情综合信息的一个关键特征。从非结构化的案件文本中快速、准确的抽取出结构化的案情特征，是串并案分析工作的基础。

本文通过对刑事案件文本特征的深入研究和大量实践检验的基础上，总结了三种最常用的案情特征抽取方法，即基于规则的抽取方法，如涉案人身份证号码、联系方式、车牌号等；基于字典的抽取方法，如重点人员姓名、作案手法、作案工具、网络暗语等；基于规则和字典相结合的混合抽取方法，如案发时间、案发地点等。文本重点介绍下第三种混合方法，并以案发时间和案发地点这两个特征作为实例展开描述。

基于规则和字典相结合的混合抽取方法，主要是为了弥补前两种方法在特定场景中的局限性，从而达到互补的效果。其基本思想就是首先根据规则进行特征抽取，然后利用字典来查漏补缺，两种方法抽取的信息可能是一个整体的两部分，也可能是并列关系，互为补充。例如，刑侦人员在判断案发时间时，不仅需要知道确切的日期，更是要精确到案件发生的时段，如“2015 年 3 月 2 日傍晚”或“2015 年 5 月 17 日凌晨”。针对上述情况，首先制定日期抽取规则，如“XX 年 XX 月 XX 日”，利用正则表达式对日期信息进行匹配；然后构建时段字典，包括“上午、傍晚、凌晨……”，

再利用时段字典进行抽取；最后将两者结合就得到了最终的案发时间。同理，案发地点的抽取流程也大体相同：首先制定地址规则“XX 区 XX 路 XX 号”，如果地点特征不满足上述规则，则结合该地区的街道地址字典和机构名称字典进行匹配，后者是对前者的补充。需要特别说明的是，同一种案件特征的抽取可能需要制定多条规则，也可能需要借助多个字典。基于规则和字典相结合的案件特征抽取方法的一般流程如图 3-11 所示。

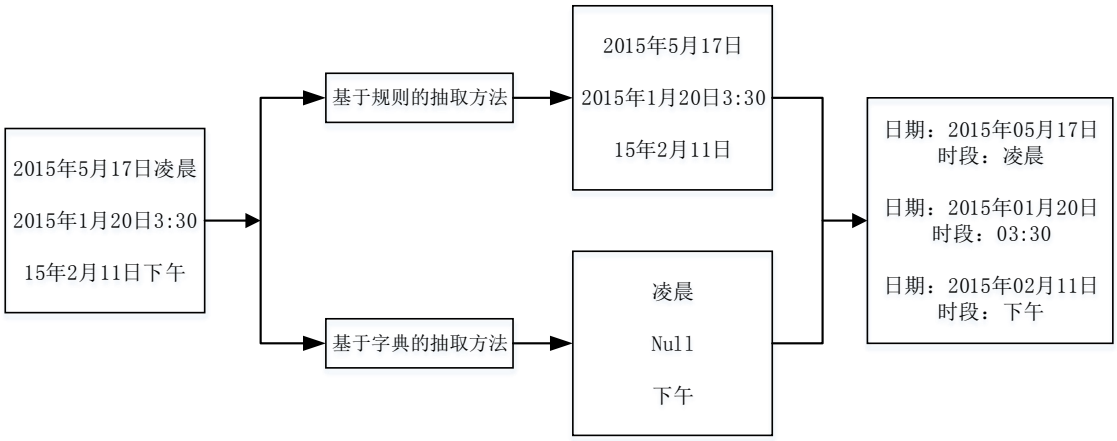

图 3-11 基于规则和字典相结合的案件特征抽取示例

3.4.2 基于特征密度的聚类算法

基于概率密度的聚类算法能够识别任意形状的类簇，弥补了其他基于距离的聚类算法仅能发现“类圆形”簇的缺陷，符合刑事案件分布的不规则性特征。本文在经典密度聚类算法 OPTICS[42]的基础上，结合刑事案件文本的特征，定义了串并案件特征相似度计算公式，提出了基于特征密度的聚类算法 OPTICS-FD，能够更好的挖掘串并案件的密集簇。下面详细介绍下 OPTICS-FD 所涉及到的基本概念。

1. 给定文本集 S，文本 $P_0 \in S$，称$\{P_t | sim(P_t, P_0) > 1-\varepsilon, P_t \in S\}$ 为以P_0为中心，ε 为半径的领域是 P_0 的ε 领域。其中 $sim(P_t, P_0)$为文本 P_t 和 P_0 的特征相似度，具体计算公式将在 4.3.3 详细介绍。

2. 如果给定文本P 的ε 邻域中的文本数量超过预定义的阈值 $MinPts$，则P 为核心对象。

3. 如果文本P 是核心对象，文本Q 在P 的 ε 领域内，则称P 到Q直接密度可达。

4. 如果存在文本序列$P_1, P_2, \ldots, P_n$，其中，$P_1 = P, P_n = Q$，且对于任意$1 \leq \mathrm{i} \leq \mathrm{n}$，

P_i 到P_{i+1} 直接密度可达，则称P 到Q 密度可达。

5. 如果给定文本 O 到 P 密度可达，且 O 到 Q 密度可达，则称 P 和 Q 密度相连。

6. 文本 P 的核心相似度是指 P 成为核心对象的最小邻域半径；文本 Q 关于文本 P 的可达相似度是指 P 的核心相似度和 P 与 Q 的特征相似度之间的较小值。

给定一个核心对象 P，从 P 出发找到密度相连对象的最大集合即为一个聚类簇，如图 3-12 所示。

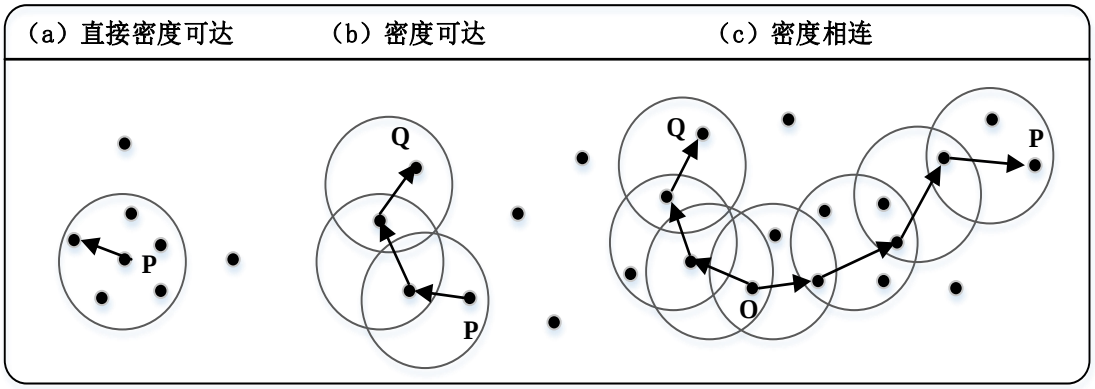

图 3-12 直接密度可达、密度可达和密度相连示例

核心相似度和可达相似度如图 3-13 所示。给定一个领域半径 $\varepsilon = 0.6$，密度阈值 $MinPts = 5$。P 的核心相似度是 P 到第 4 个最近文本之间的特征相似度ε'。Q_1关于 P 的可达相似度是 P 的核心相似度 ε'，而Q_2 关于 P 的可达相似度是 Q_2 与 P 的特征相似度 $sim(Q_2, P)$。

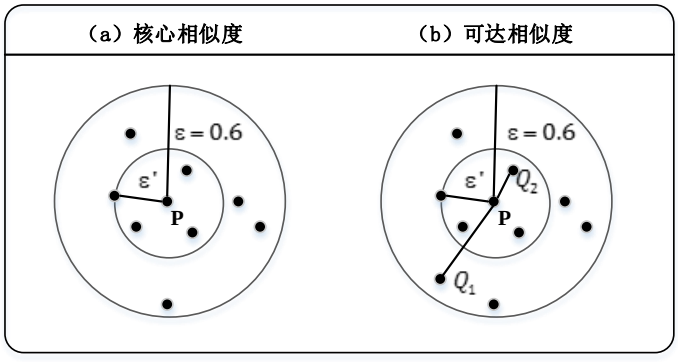

图 3-13 核心相似度和可达相似度示例

OPTICS-FD 的基本思想就是从核心对象出发，找到所有密度相连的对象的最大集合，具体算法描述如图 3-14 所示。

```
   OPTICS-FD
1: Input: CaseDocs← 案件集合
2: Output:   AscSimQue & ResultQue //按相似度递增序列以及结果队列
3: do while  caseDocs ≠ ∅
4:     取文本数据 M ∈ CaseDocs,
5:                AscSimQue=AscSimQue ∪ {M}  && CaseDocs=CaseDocs-{M}
6:     do while  AscSimQue ≠ ∅
7:        取  P ∈ AscSimQue  进行扩张
8:        if P  是核心点
9:           对 P 的 ϵ  邻域内任一未扩张的邻居 Q 进行如下处理：
10:             if  Q ∈ AscSimQue
11:                //这里的相似度为定义的特征相似度
12:                if  从 P 到 Q 的可达相似度大于旧值
13:                   更新 Q 的可达相似度，并调整 Q 到相应位置以保证队列的递增
14:                end if
15:             else
16:             根据 P 到 Q 的可达相似度将其插入到 AscSimQue
17:            end if
18:       else
19:          AscSimQue=AscSimQue-{P},P 存入队列 ResultQue
20:       end if
21:     end do
22:end do
```

图 3-14 OPTICS-FD 算法

3.4.3 案件特征相似度

本文在串并案分析中综合考虑了案件的精细类别、案发时间和案发地点三个维度的特征，因此，在计算案件文本的特征相似度（距离）时也包含了这三个维度因子。案件文本的特征相似度计算公式如式 3-21 所示。

$$sim(P,Q)=\alpha\cdot sim\_type(P,Q)+\beta\cdot sim\_time(P,Q)+\gamma\cdot sim\_place(P,Q) \quad (3\text{-}21)$$

其中α、β 和 γ 分别代表案件类别、时间和地点在特征相似度计算中所占的比例系数。$sim\_type(P,Q)$、$sim_time(P,Q)$和$sim_place(P,Q)$分别表示案件类别相似度、案件时间相似度和案件地点相似度。

3.4.3.1 案件类别相似度

给定两个案件文本 P 和 Q，它们之间的案件类别相似度计算需要考虑两种情况，第一种情况是 P 和 Q 分属于不同的一级案件类别，例如“盗窃案”和“绑架案”，此时它们之间的相似度直接为 0；第二种情况是 P 和 Q 同属于一种一级案件类别，此时，

计算 P 和 Q 的二级案件类别名称的 Jaccard 相似度，如式 3-22 所示。

$$Jaccard(P,Q)=\frac{\left|SubTpye_P \cap SubTpye_Q\right|}{\left|SubTpye_P \cup SubTpye_Q\right|} \tag{3-22}$$

其中，$\left|SubTpye_P \cap SubTpye_Q\right|$表示$P$和$Q$的二级案件类别名称的交集中字的个数，$\left|SubTpye_P \cup SubTpye_Q\right|$表示 P 和 Q 的二级案件类别名称的并集中字的个数。例如，P 和 Q 的二级案件类别分别为“撬门入室盗窃案”和“翻窗入室盗窃案”，则$\left|SubTpye_P \cap SubTpye_Q\right|$值为5，$\left|SubTpye_P \cup SubTpye_Q\right|$值为9，$Jaccard(P,Q)$值为$\frac{5}{9}$；假设 R 的二级案件类别为“拎包盗窃”，则$|SubTpye_P \cap SubTpye_R|$值为 2，$|SubTpye_P \cup SubTpye_R|$值为 7，$Jaccard(P,R)$的值为$\frac{2}{7}$，所以，$P$ 和 Q 的案件类别相似度要大于 P 和 R 的相似度。综上所述，给出案件类别相似度计算公式如式 3-23 所示。

$$sim\_type(P,Q)=\begin{cases}\dfrac{\left|SubTpye_P \cap SubTpye_Q\right|}{\left|SubTpye_P \cup SubTpye_Q\right|} & P和Q同属于一种一级案件类别 \\ 0 & P和Q分属于不同的一级案件类别\end{cases} \tag{3-23}$$

3.4.3.2 案发时间相似度

案发时间相似度的计算需要考虑日期和时段两个因素，案发日期相距越近说明是系列案件的可能性越大。根据刑侦人员的办案经验，一般系列案件的日期跨度不超过 3 个月（90 天），因此，当案件日期跨度大于 90 天时，时间相似度定义为 0。此外，系列案件的作案人员经常选择同一时段作案，例如，某团伙经常在凌晨时分入室抢劫作案。因此，案发时段可以作为案发时间相似度计算的系数因子λ，当案发时段相同时，λ = 1；当案发时段不同时λ = 0.5。案发时段字典如表 3-2 所示。由上述分析可以得到案发时间相似度计算公式如式 3-24 所示。

$$sim\_time(P,Q)=\begin{cases}\dfrac{1}{1+\left|P和Q案发日期相差天数\right|}*\lambda & 日期相差不超过90天 \\ 0 & 日期相差超过90天\end{cases} \tag{3-24}$$

表 3-2 案发时段字典

时段	时钟时间
凌晨	1:00~4:00
早晨	5:00~7:00
上午	8:00~10:00
中午	11:00~13:00
下午	14:00~16:00
傍晚	17:00~19:00
晚上	20:00~22:00
午夜	23:00~0:00

3.4.3.3 案发地点相似度

案发地点的相似度计算也需要考虑两个维度信息，即案件发生的地理位置信息和案发场所信息。由于系列案件的作案人员通常都在某一范围活动，因此，如果两个案件发生地点之间的地表距离（根据经纬度计算）越近，表明它们是系列案件的可能性越大；同理，如果两个案件发生的场所相同，比如 “小区”、“学校”、“网吧”等，也表明是系列案件的可能性较大。案发地点相似度计算公式如式 3-25 所示。

$$sim\_place(P,Q)=\frac{1}{1+|P和Q地表距离公里数|}*\mu \tag{3-25}$$

上式中地表距离的基本单位是公里，μ 表示案件场所系数，当案发场所相同时$\mu=1$，否则$\mu=0.5$。

3.4.3.4 权重值决策

上述三小节分别介绍了案件类别相似度、案发时间相似度和案发地点相似度的计算方法，接下来需要采用评价决策方法来权衡这三个维度在案件特征相似度计算中的权值。本文采用层次分析法[37-39]来决策案件类别、案发时间和案发地点的权值，主要分为以下几步。

1. 构建层次模型。案件特征相似度的度量依赖于案件类别、案发时间和案发地点三个属性值，并且三者之中案件类别占主导。因此，为这三个维度分别赋予不同的权

值，并最终计算得到一个特征相似度距离，案件特征相似度的层次分析模型如图 3-15 所示。

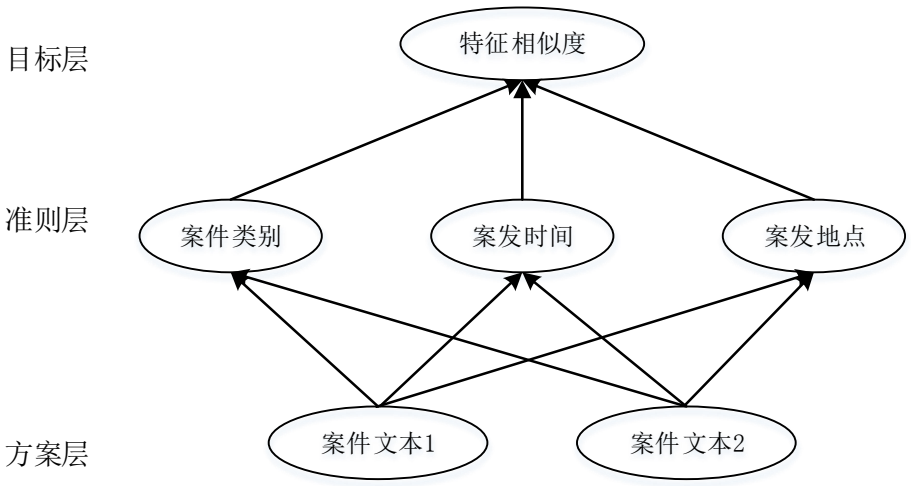

图 3-15 案件特征相似度层次分析模型

2. 创建判断矩阵 $A=(a_{ij})_{n\times n}$，使其满足正互反矩阵的如下特征：$\forall$ i, j∈N, 有 $a_{ij}>0$; $a_{ii}=1$; $a_{ij}=1/a_{ji}$，其中，i=j=1,2,...,n。a_{ij} 表示要素 i 对要素 j 的重要程度值，该值可以取定为 1 到 9 的数字或者其倒数，值越大表示要素 i 比要素 j 的重要程度越高。设 cls 代表案件类别，time 代表案发时间，plc 代表案发地点。则判断矩阵如式 3-26 所示。

$$A=\begin{pmatrix} a_{cls,cls} & a_{cls,time} & a_{cls,plc} \\ a_{time,cls} & a_{time,time} & a_{time,plc} \\ a_{plc,cls} & a_{plc,time} & a_{plc,plc} \end{pmatrix} \tag{3-26}$$

其中，对角线的位置都是要素与自身的重要程度对比，因此值为 1。根据三要素的两两比较，可设置$a_{cls,time}=5$，$a_{cls,plc}=7$，$a_{time,plc}=2$，其他值可根据倒数得出，最终的判断矩阵如式 3-27 所示。

$$A=\begin{pmatrix} 1 & 5 & 7 \\ 1/5 & 1 & 2 \\ 1/7 & 1/2 & 1 \end{pmatrix} \tag{3-27}$$

上式中的标度初始值是通过分析样本数据得到的最优值，然而，随着测试语料的变化，不同维度间的侧重点可能会有所区别，因此，仍然需要一种能够自适应调整机制保证标度值的准确性。

3. 计算权向量 W。首先对判断矩阵 $A=(a_{ij})_{n\times n}$ 按列进行规范化操作，如式 3-28 所示。

$$A=\begin{pmatrix}0.7447 & 0.7692 & 0.7\\ 0.1489 & 0.1538 & 0.2\\ 0.1064 & 0.0769 & 0.1\end{pmatrix} \tag{3-28}$$

然后对判断矩阵按行相加，再进行归一化，得到向量 W，如式 3-29 所示。

$$W=\begin{pmatrix}0.7380\\ 0.1676\\ 0.0944\end{pmatrix} \tag{3-29}$$

上式表明案件特征相似度计算中案件类别维度所占权值为 0.7380，案发时间维度所占权值为 0.1676，案发地点维度所占权值为 0.0944。

4. 一致性检验。以上得出的各要素权值是否合理，还需要进行判断矩阵的一致性检验，如式 3-30 所示。

$$(AW)_i=\begin{pmatrix}1*0.7380+5*0.1676+7*0.0944\\ \frac{1}{5}*0.7380+1*0.1676+2*0.0944\\ \frac{1}{7}*0.7380+\frac{1}{2}*0.1676+1*0.0944\end{pmatrix}=\begin{pmatrix}2.2368\\ 0.5040\\ 0.2836\end{pmatrix} \tag{3-30}$$

将式 3-30 的结果代入式 2-8，最终计算得到 λ_{max} 如式 3-31 所示。

$$\lambda_{\max}=\sum_{i=1}^{n}\frac{(AW)_i}{nw_i}=\frac{1}{3}*\left(\frac{2.2368}{0.7380}+\frac{0.5040}{0.1676}+\frac{0.2836}{0.0944}\right)=3.0141 \tag{3-31}$$

将式 3-31 的结果代入式 2-10 和 2-9，最终计算得到 CI 和 CR 如式 3-32 和 3-33 所示。

$$CI=\frac{\lambda_{max}-n}{n-1}=\frac{3.0141-3}{3-1}=0.0071 \tag{3-32}$$

$$CR=\frac{CI}{RI}=\frac{0.0071}{0.58}=0.0122 \tag{3-33}$$

其中，RI=0.58 是从表 2-1 中根据阶数 3 查询所得。当判断矩阵的 CR 值小于 0.1 或者$\lambda_{\max}=n, CI=0$时，则认为 A 具有良好的一致性，否则需要调整 A 中元素的值。

根据式 3-33 得出 CR=0.0122，一致性检验合格，即权值是合理的。

最后，将案件类别、案发时间和案发地点三个维度的权值代入到式 3-21 中，得到最终的案件特征相似度计算公式如式 3-34 所示。

$$sim(P,Q)=0.7380*sim\_type(P,Q)+0.1676*sim\_time(P,Q)+0.0944*sim\_place(P,Q) \tag{3-34}$$

其中，$sim\_type(P,Q)$、$sim_time(P,Q)$和$sim_place(P,Q)$分别表示案件类别相似度、案发时间相似度和案发地点相似度，如式 3-23、3-24 和 3-25 所示。通过定量分析案件特征相似度，可以更加准确的划分系列案件的类簇，提高串并案分析的准确率。

3.5 本章小结

本章详细介绍了公安领域文本挖掘系统的系统架构、整体流程和关键模块的设计与实现。其中，基于朴素贝叶斯和关键词共现图谱的两级分类器实现以及基于特征密度的串并案聚类分析方法是本章的两个重点。

面对刑事案件文本长度短、词频低、类别分布不均衡等特点，为满足刑侦人员提出的严格的精细化分类要求，本文采用了两级分类器架构，第一级分类器面向海量文本做粗分类，以朴素贝叶斯算法为基础，结合双因子综合评估特征选择方法，能够快速、准确、有效的实现一级案件类别的粗分类要求；第二级分类器面向二级案件类别文档集分别构建关键词构建图谱，以词语间的共现关系代替词频进行特征选择，不仅能够发现高频词，还能发现重要的低频关键词，有效的解决了二级分类面临的子类别差异小、词频低等挑战。此外，结合同义词网，有效消除了领域同义词对分类结果的干扰，进一步提升了分类的准确性。

串并案分析是打击系列案件的重要手段，本章详细介绍了基于特征密度的串并案聚类分析方法，它首先通过基于规则和字典相结合的案件特征抽取方法从案件文本中抽取出案发时间和案发地点信息。再结合上一步得到的案件精细化分类结果，在经典密度聚类算法 OPTICS 的基础上，重新定义了案件特征相似度计算方法，提出了基于特征密度的聚类算法 OPTICS-FD，该方法能够有效的分析出串并案件的密集簇，辅助刑侦人员破案。

4 测试与分析

本章通过实验对提出的方法和改进措施进行了测试，主要包括四部分，分别是双因子评估算法测试、两级分类器测试、案件特征抽取结果分析和串并案聚类测试。

4.1 测试语料

本文的测试语料主要来源于某地区公安局提供的真实的案件文本信息，包括 110 接处警信息、案件基本信息和线索信息等。为了保证测试结果的准确性和可靠性，共选取 8 种一级案件类别及其所属的 52 种二级案件类别，共计 11798 条真实案件信息，并按照 1:2 的原则划分为训练语料和测试语料。一级案件类别的语料数量分布如表 4-1 所示。

表 4-1 一级案件类别语料数量分布

编号	一级案件类别	训练语料数	测试语料数	语料总数
1	盗窃	656	1200	1856
2	诈骗	517	1004	1521
3	抢劫	578	1187	1765
4	抢夺	493	1050	1543
5	涉毒	477	946	1423
6	伤害	489	984	1473
7	杀人	321	659	980
8	强奸	414	823	1237
合计		3945	7853	11798

4.2 双因子评估算法测试

在第一级分类器的特征选择阶段，本文在文档频率 DF 的基础上引入了词性因子，提出了双因子评估算法。下面通过实验结果对比双因子评估算法与 DF 方法在刑事案件文本分类领域对朴素贝叶斯分类效果的影响。主要从准确率、召回率和 F-measure 值三个指标来衡量，如图 4-1、图 4-2 和图 4-3 所示。

图 4-1 表明采用双因子评估算法进行特征选择后的一级案件类别分类准确率要高于传统的文档频率 DF 方法。并且，在 8 种一级案件类别中，涉毒案件的分类准确率最

高，此时，双因子评估算法的准确率为 99.26%，DF 方法为 98.21%；对抢劫案的分类准确率最低，此时，双因子评估算法的准确率为 97.16%，DF 方法为 95.41%。

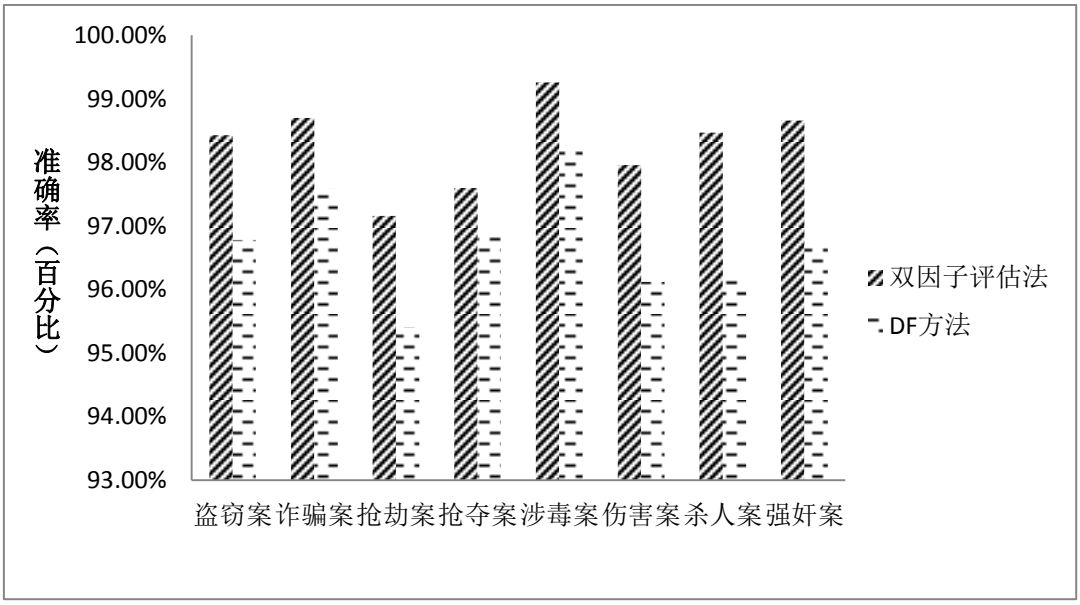

图 4-1 两种特征选择算法的准确率对比

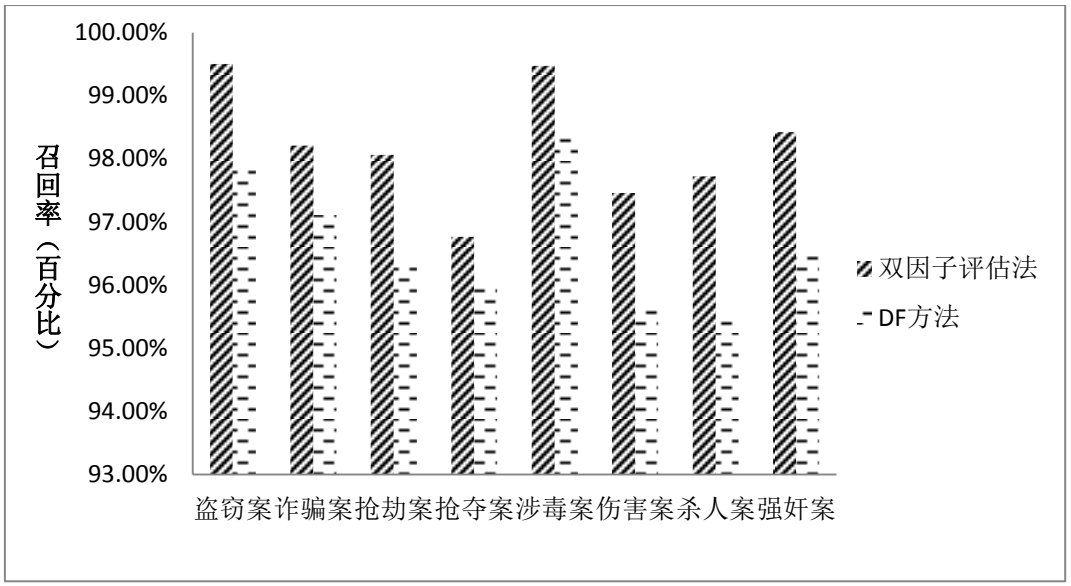

图 4-2 两种特征选择算法的召回率对比

图 4-2 表明双因子评估算法的召回率也明显优于文档频率 DF 方法。在 8 类一级案件类别中，对盗窃案的分类召回率最高，此时，双因子评估算法的召回率为 99.50%，而 DF 方法为 97.83%；双因子评估算法召回率最低的是抢夺案，为 96.76%，而 DF 方法召回率最低的是杀人案，为 95.45%。

图 4-3 表明双因子评估算法的整体性能，即 F-measure 值要高于文档频率 DF 方法。在 8 类一级案件类别中，对涉毒案的分类效果最高，此时，双因子评估算法的 F-measure 值为 99.36%，而 DF 方法为 98.31%；双因子评估算法对抢夺案的分类效果最差，F-measure 值为 96.76%，而 DF 方法对伤害案的分类效果最差，为 94.88%。

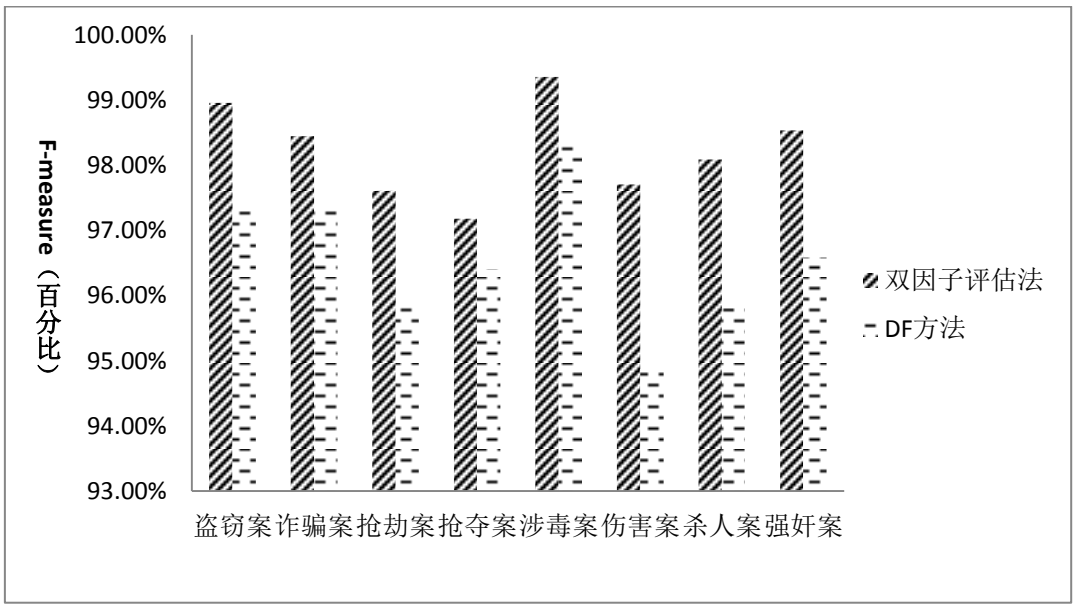

图 4-3 两种特征选择算法的 F-measure 值对比

综上所述，双因子评估算法的宏观 F-measure 值为 98.24%，而 DF 方法为 96.56%。因此，采用双因子评估算法进行特征选择的分类效果要明显优于传统的文档频率 DF 方法。

4.3 两级分类器测试

本文提出的两级分类方法 TLC-NBK 采用朴素贝叶斯作为第一级分类器，然后在一级案件类别分类完成后，对其所属的二级案件子类别再次进行细分，因此，保证了案件文本分类的效果更好稳定、可靠。下面通过实验对比两级分类方法 TLC-NBK 与单纯的朴素贝叶斯算法直接对二级案件类别进行分类的性能对比。

由于测试语料包含了 8 大类 52 种二级案件类别，为了便于分析，随机选取 9 种常见的二级案件类别， 对比 TLC-NBK 方法和 Naïve Bayes 分类算法的准确率、召回率和 F-measure 指标，如图 4-4、图 4-5 和图 4-6 所示。

图 4-4 表明，TLC-NBK 方法虽然在某几类案件的分类准确率略逊于 Naïve Bayes

算法，但是，前者在总体上要明显优于后者，并且 TLC-NBK 方法的准确率稳定性要比 Naïve Bayes 好得多。TLC-NBK 准确率最高的是自行车盗窃案，为 94.33%，而 Naïve Bayes 分类准确率最高的是拎包盗窃案，为 95.71%；TLC-NBK 分类准确率最低的也是迷信诈骗案，为 79.78%，而 Naïve Bayes 分类准确率最低的也是迷信诈骗案，却只有 51.38%。

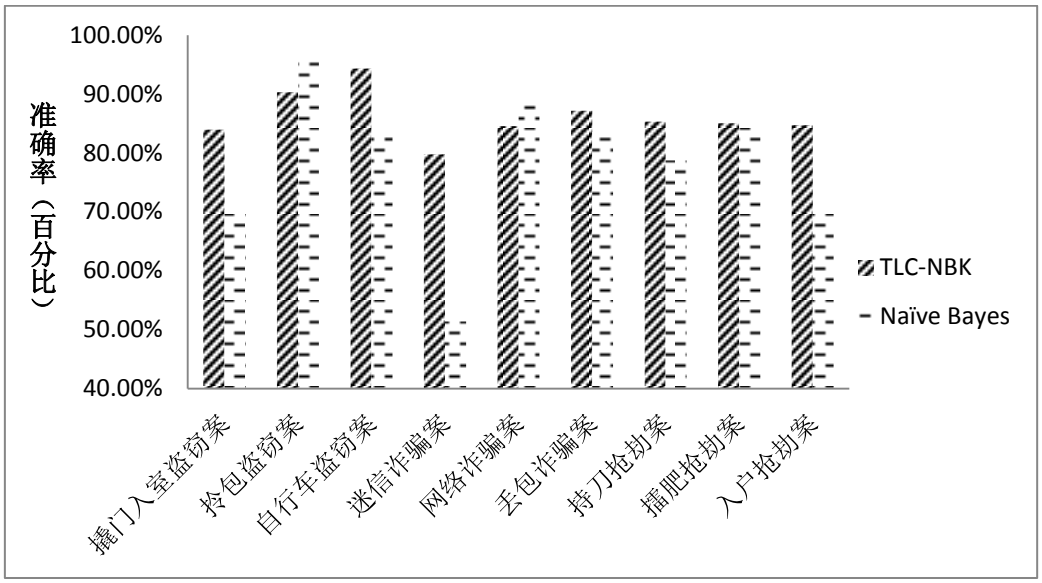

图 4-4 两级分类器与朴素贝叶斯分类的准确率对比

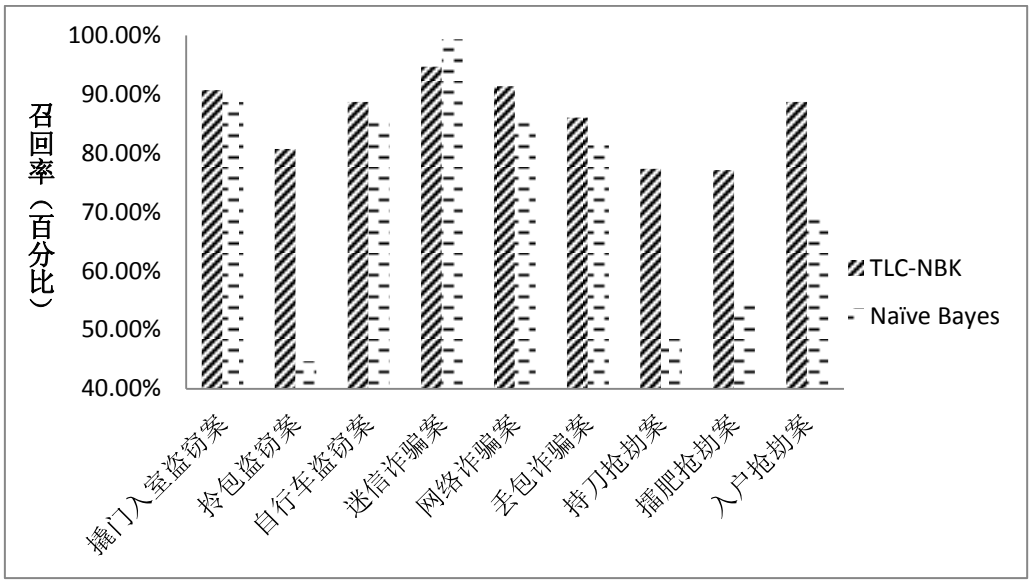

图 4-5 两级分类器与朴素贝叶斯的召回率对比

同样地，图 4-5 表明 TLC-NBK 方法虽然在某几类案件的分类召回率略逊于 Naïve Bayes 算法，但是，前者在总体上要明显优于后者，并且，TLC-NBK 方法的召回率稳

定性也要比 Naïve Bayes 好得多。TLC-NBK 方法分类召回率最高的是迷信诈骗案，为 94.67%，而 Naïve Bayes 召回率最高的也是迷信诈骗案，为 99.33%；TLC-NBK 分类召回率最低的是擂肥抢劫案，为 77.08%，而 Naïve Bayes 召回率最低的是拎包盗窃案，只有 44.67%。

图 4-6 表明，TLC-NBK 方法的整体分类效果要始终优于 Naïve Bayes 算法，主要是由于两方面原因导致的。一方面是因为前者的总体性能本身就比后者高，另一方面是因为前者的稳定性比后者好，因此，对综合指标 F-measure 进行测量时发现，随机挑选的 9 种二级案件类别中，TLC-NBK 方法始终优于 Naïve Bayes 算法，最高差值可达 24.30%，最低差值为 0.75%，平均差值 14.44%。

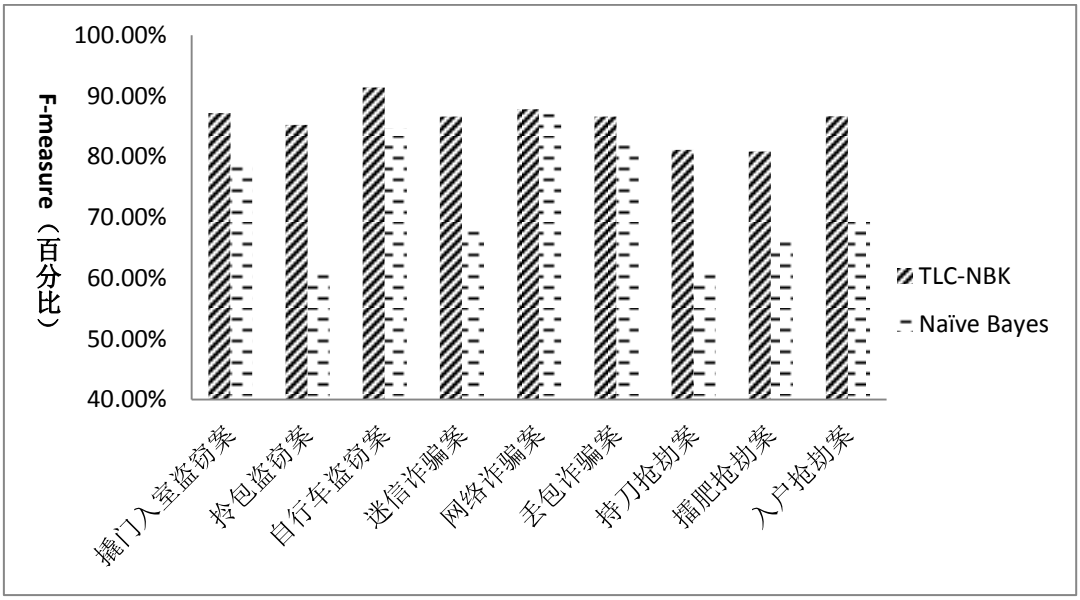

图 4-6 两级分类器与朴素贝叶斯的 F-measure 对比

最后，给出两级分类方法 TLC-NBK 的宏观 F-measure 值为 87.54%，而 Naïve Bayes 的宏观 F-measure 值为 73.10%，因此，本文提出的两级分类方法 TLC-NBK 对刑事案件精细化分类的效果要远优于传统的 Naïve Bayes 分类算法。

4.4 案件特征抽取结果分析

案件特征的抽取采用基于规则和字典相结合的方法，每一种案件特征可能包含多种匹配规则或结合多个字典库。案发时间特征的抽取以基于规则的方法为主，结合时段字典。表 4-2 给出了案发时间特征的日期或时段抽取规则及其示例。

表 4-2 案件时间特征抽取规则及其示例

编号	日期或时段正则表达式规则	抽取示例（以分号分隔）
1	[\d]{2}\.[\d]{2}\.[\d]{2}	06.02.25
2	[\d]{4}-[\d]{2}-[\d]{2}	2008-07-25
3	[\d]{4}年[\d]{2}月([\d]日)?	2011 年 4 月 2 日 ; 2009 年 9 月
4	[\d]{1,2}时([\d]{1,2}分)?	6 时 18 分 ; 14 时
5	[\d]{2}:[\d]{ 2}	07:54 ; 22:08

实际的案发时间特征抽取规则远不止上述 5 种，通过观察抽取结果，不断地反馈和优化，除了极少一部分手工输入的拼写错误和格式不规范引起的歧义错误，最终的案发时间特征抽取几乎能达到百分百的准确率和召回率。

案发地点特征的抽取则是以基于字典的方法为主，辅以少量的地址规则。如表 4-3 所示为案发地点特征涉及的相关字典及其示例。

表 4-3 案发地点特征字典库及其示例

字典库名称	示例
行政区划字典	XX 市 XX 区 ; XX 镇 XX 村
街道详址字典	建设大道 568 号 ； 胜利街 2 号
机构场所字典	XX 大学 ； 美华小区 ; XX 医院

案发地点特征抽取的效果跟地址字典的完备度和精确度息息相关，现有的公安领域字典库能够基本满足应用的要求，案发地点特征抽取的准确率非常高，在随机选择的 500 个测试语料中的综合准确率达到 93.43%，但召回率只有 76.8%，有相当一部分地点特征没有抽取出来，仍有较大进步空间。在地址字典的基础上，我们也构建了一些规则用以查漏补缺，但由于记录人员描述的不规范性，实际提升效果并不明显。此外，本文也尝试了一些机器学习的方法，如隐马尔科夫模型等来识别案发地点，但这类方法需要大量的人工标注，限于人力不足的原因，最终的效果也并不理想。

综上所述，本文对案件特征抽取的实验结果进行分析后发现，基于规则和字典相结合的方法能够快速、准确、有效的抽取出刑事案件文本的相关特征，如案发时间和案发地点等，在实际应用中也表现出了较理想的效果。但是，不足之处在于都需要一定的人工标注和准备工作。

4.5 串并案聚类测试

案件串并的目的是为了筛选出系列案件的密集类簇，避免刑侦人员在海量案件文本中低效率的搜寻。因此，串并案分析的评价指标主要是召回率和缩减率。衡量召回率是为了避免漏判，以防系列案件被作为孤立点遗漏了；而缩减率则是判断聚类结果对刑侦工作效率提升的重要依据。

系列案件大多属于“两抢一盗”和诈骗类型，因此，本文从这四类案件的训练语料库中挑选了已经标注后的案件共计 1413 条，其中包含系列案件 32 簇，共计 146 起。采用基于特征密度的聚类算法 OPTICS-FD 对上述案件进行分析，由于 OPTICS-FD 生成的是增广的簇排序，参数只是起到算法辅助的作用，所以根据经验给定邻域半径 $\varepsilon = 0.4,\ MinPts = 5$，然后观察簇排序结果，综合考虑召回率和缩减率，选取较优情况下的结果，整理后如表 4-4 所示。

表 4-4 OPTICS-FD 聚类分析结果

一级案件类别	案件总数	真实类簇个数	类簇案件数	聚类后类簇数	聚类后案件数	聚类后真实案件数	缩减率	召回率
盗窃案	547	13	67	9	161	59	70.57%	88.06%
诈骗案	483	11	47	7	187	43	61.28%	91.49%
抢夺案	219	5	22	3	92	21	57.99%	95.45%
抢劫案	164	3	10	2	39	9	76.22%	90.00%

从表 4-4 得知，采用 OPTICS-FD 算法聚出的类簇个数普遍小于真实的类簇数，原因是在从簇排序中选取结果时，优先考虑了召回率，导致簇的密度比较低，有一些类簇被融合在了一起，并且在这个过程中，融入了许多孤立点，在聚类的类簇个数降低的同时，类簇包含的案件总数却上升了。

从总体上看，聚类分析后的缩减率与召回率成反比，例如盗窃案的缩减率为 70.57%，抢夺案的缩减率只有 57.99%，而盗窃案的召回率为 88.06%，抢夺案的召回率则高达 95.45%。但是，也不排除异常情况，例如抢劫案的缩减率和召回率都比较高，这说明聚类结果与样本数据的分布也有很大关系。

表 4-5 给出了一个经过 OPTICS-FD 分析得到的盗窃案类簇实例。从表 4-5 可以发

现，聚类的关键因素还是案件类别，同属于一种二级案件类别的特征相似度会非常高，不同二级案件类别之间，只有当案发时间和案发地点的特征非常相近时，才会被聚到同一类簇中。这符合刑侦人员串并案件的规律，证明了基于特征密度的聚类算法 OPTICS-FD 在刑事案件串并分析领域表现出了较理想的效果，能够有效提高刑侦人员的办案效率。

表 4-5 OPTICS-FD 聚类结果实例

案件类别	案发时间	案发地点
撬门入室盗窃案	2012 年 9 月 13 日凌晨	AA 街 49 号，XX 商铺
撬门入室盗窃案	2012 年 9 月 22 日午夜	BB 路 51 号，XX 小区
撬门入室盗窃案	2012 年 10 月 7 日凌晨	CC 街 17 号，XX 酒店
插门入室盗窃案	2012 年 9 月 14 日凌晨	BB 路 188 号，XX 小区
溜门入室盗窃案	2012 年 10 月 8 日凌晨	DD 大道 63 号，XX 小区

4.6 本章小结

本章对提出的方法和改进措施进行了测试与分析，主要包括四个方面。首先对比了双因子评估算法和文档频率 DF 方法对朴素贝叶斯分类结果的影响；然后对比了两级分类方法 TLC-NBK 与 Naïve Bayes 算法在刑事案件精细分类领域的准确率、召回率和 F-measure 指标；再次，分析了基于规则和字典相结合的案件特征抽取的结果，详细阐述了该方法的优势与不足；最后，测试了 OPTICS-FD 算法对系列案件聚类分析的效果，主要是从召回率和缩减率两个指标进行衡量。

实验结果表明，双因子评估算法相比于传统的文档频率 DF 方法，能够更加有效的提高分类效果，前者的准确率、召回率和 F-measure 指标均高于后者；两级分类方法 TLC-NBK 相比于 Naïve Bayes 算法，虽然在某几个小类的准确率或召回率指标略逊于后者，但总体的准确率和召回率要高于后者，并且，前者的稳定性要比后者好的多，因此，TLC-NBK 方法的综合评价指标 F-measure 值均高于 Naïve Bayes，表明前者对刑事案件的精细分类效果要优于后者；基于规则和字典相结合的案件特征抽取方法在准确率和召回率方面均取得了较好的效果，但是需要大量的人工标注和准备工作；特征密度聚类算法 OPTICS-FD 生成的案件类簇通常会比实际类簇个数少，但是包含的案件

总数却会相对较多，这主要是由于在增广簇排序中选取类簇时优先考虑了召回率指标，密度阈值相对较低，类簇之间进行了融合，并在此过程中融入了大量孤立点。因此，聚类后的案件总数约为原始案件总数的三分之一，但平均召回率却在 90%以上，这符合公安人员对串并案分析的指标要求，极大的减少了刑侦人员的工作量，提高了办案效率。

5 总结与展望

5.1 工作总结

本文面向刑事案件文本，重点研究了案件精细分类和串并案分析这两个公安领域文本挖掘的关键问题。根据刑事案件文本长度短、词频低、类别分布具有层次性和不均衡性的特点，提出了基于朴素贝叶斯和关键词共现图谱的两级分类方法 TLC-NBK，显著提高了精细分类的效果；针对刑事案件的文本特点，定义了案件特征相似度计算公式，提出了基于特征密度的聚类算法 OPTICS-FD，能够有效的分析出系列案件的密集簇，辅助刑侦人员破案。具体工作主要包括以下几点：

1 介绍了课题的研究背景和动机，分析了国内外学者对公安领域文本挖掘技术的研究现状。

2 对本文涉及到的相关技术和方法进行了较为详细的介绍，包括特征选择、文本分类算法和评价决策方法。

3 概述了本文系统的总体架构和整体流程，然后详细介绍了各关键模块的具体设计与实现。提出了双因子评估特征选择算法，基于朴素贝叶斯和关键词共现图谱的两级分类方法，基于规则和字典相结合的案件特征抽取方法，定义了案件特征相似度，提出了基于特征密度的串并案聚类分析方法。

4 通过实验对双因子评估算法、两级分类器、案件特征抽取和串并案聚类进行了测试。结果表明，在刑事案件文本挖掘领域，相比于传统方法，上述方法在准确率、召回率或缩减率等指标上均有所提升，更好的支持了刑侦人员进行决策。

5.2 工作展望

在后续工作中，可以继续对本文提出的方法进行优化与扩展，主要包括以下三个方面：

1 将术语抽取方法应用在领域词典的构建过程中，实现公安领域术语的自动识别与抽取，进一步提高分词的准确性，避免领域特征词错误切分带来的分类影响。

2 利用精细分类的结果和特征抽取得到的结构化案件特征，构建公安领域知识库，并在此基础上实现犯罪团伙发现，犯罪趋势预测，活动轨迹分析等功能。

3 可以进一步细化案件特征相似度的计算，测试并分析融入更多特征后对案件相似度的影响，以便进一步提升聚类的效果。

致 谢

时光荏苒，转眼间又到了毕业的季节，三年的研究生生活给我留下了许多的回忆与感叹。从大学时代的懵懂青涩到现在的踏实干练，过程中经历了很多，也学到了很多。特别是研究生期间遇到的许多朋友、同学和老师，他们给予我的关怀、帮助与教诲令我时刻铭感于心。

在这里，我首先要感谢我的导师周可教授，周老师博学睿智、严谨刻苦、乐观随和的作风与性格给我留下了深刻的印象。他为了实验室的发展不辞辛劳、终日奔波，给我们提供了优渥的科研与学习环境。他常常教育我们要脚踏实地，不断进取，做好一件事，有自己的立足之地。同时也要感谢周老师在论文选题和写作过程中对我的帮助与指导。

感谢实验室团队的邹复好老师、李春花老师、王桦老师和程晓燕老师在科研与生活上对我的指导和帮助。

感谢项目组的郑胜老师，他在工作和生活中给了我很多鼓励与关怀，既是良师，更是益友，令我十分感激。

感谢徐涛老师、张胜老师和高路老师在工作中给予的帮助。

感谢金吉祥师兄、何爽师兄和廖正霜师姐，他们教给了我许多宝贵的学习和生活经验。

此外，还要感谢同届的蒋丹、陶灿、饶琦、沈慧羊和杨勇等同学，我们一起学习、进步，我会始终珍惜这份宝贵的友谊。

最后，衷心的感谢我的父母和家人，他们对我的支持、鼓励与理解，是我不断前进的动力，谢谢你们。

祝愿所有的老师、同学和家人，身体健康，工作顺利，幸福平安。

参考文献

[1] S Chou, TP Hsing. Text Mining Technique for Chinese Written Judgment of Criminal Case. Lecture Notes in Computer Science, 2010, 6122: 113~125

[2] M.M.Janeela Theresav, V.Joseph Raj. Modified Fuzzy Neural Network for the Classification of Murder Cases in Criminal Law Using Gaussian Membership Function. International Journal of Computational Intelligence and Applications, 2013,12(2):1~15

[3] CL Liu, TC Chang. Some Case-Refinement Strategies for Case-Based Criminal Summary Judgments. Springer Berlin Heidelberg, 2003, 2871: 285~291

[4] 杨静, 王靖. 基于聚类分析检索团伙多起犯罪的迭代算法. 计算机与现代化, 2013(1): 1~4

[5] 卢睿. 刑事案件的属性约简聚类算法研究. 中国人民公安大学学报:自然科学版, 2015(1): 73~76

[6] 陈龙, Neil Stuart, Williams A.Mackaness. 美国内布拉斯加州林肯市犯罪行为的聚类及热点分布分析. 测绘与空间地理信息, 2015(3): 189~192

[7] CC Yang, KW Li. Cross-Lingual Semantics for Crime Analysis Using Associate Constraint Network. Springer Berlin Heidelberg, 2004, 3073: 449~456

[8] LI Wei, NS Department, B Branch. Analysis and Research on Network Crime Based on Data Mining. Modern Computer, 2013

[9] M Alruily, A Ayesh, H Zedan. Automated dictionary construction from Arabic corpus for meaningful crime information extraction and document classification. in: International Conference on Computer Information Systems & Industrial Management Applications, 2010. 1179~1186

[10] CH Ku, A Iriberri, G Leroy. Crime Information Extraction from Police and Witness Narrative Reports. in: IEEE Conference on Technologies for Homeland Security: 2008. 593~600

[11] CH Ku, A Iriberri, G Leroy. Natural language processing and e-Government: crime information extraction from heterogeneous data sources. in: Proc of the Ninth International Conference on Digital Government Research, 2008. 162~170

[12] JT Bordogna, DE Brown, JH Conklin. Design and Implementation of an Automated Anomaly Detection System for Crime. IEEE Systems & Information Engineering

Design Symposium, 2007: 1~6

[13] 高建强, 谭剑, 崔永发. 一种基于通讯痕迹的社会网络团伙分析模型. 计算机应用与软件, 2012, 29(3): 206~208

[14] 周志涛, 鲍灵佳. 社会网络分析在团伙诈骗犯罪侦查中的应用. 江西警察学院学报, 2014(3): 39~44

[15] 孙沛, 陈世福. 面向公安部门的人脸识别系统的设计和实现. 计算机科学, 2004, 31(9): 183~185

[16] 苏光大, 田青, 徐伟等. 人脸识别技术及其在公共安全领域的应用. 警察技术, 2014(5): 3~7

[17] 程春惠. 公安犯罪案件文本挖掘关键技术研究: [硕士学位论文]. 杭州: 浙江大学, 2010

[18] X Shang, Y Yuan. Social Network Analysis in Multiple Social Networks Data for Criminal Group Discovery. in: International Conference on Cyber-enabled Distributed Computing & Knowledge Discovery, 2012. 27~30

[19] J Hosseinkhani, S Ibrahim, S Chuprat, et al. Web crime mining by means of data mining techniques. Research Journal of Applied Sciences Engineering & Technology, 2014, 7(10): 2027~2032

[20] KR Rahem, N Omar. Drug-related crime information extraction and analysis. in: International Conference on Information Technology & Multimedia, 2014

[21] 杨军. 贝叶斯分类算法在公安犯罪领域的应用研究: [硕士学位论文]. 长沙: 湖南大学, 2014

[22] 韩彦斌. 基于人脸检测和特征提取的移动人像采集系统: [硕士学位论文]. 云南: 云南大学, 2015

[23] 吴文浩, 吴升. 多时间尺度密度聚类算法的案事件分析应用. 地球信息科学学报, 2015(7): 837~845

[24] R Niu, J Zhang, DS Ebert. Classification and Visualization of Crime-Related Tweets. The Summer Undergraduate Research Fellowship (SURF) Symposium, 2015

[25] 胡佳妮, 徐蔚然, 郭军等. 中文文本分类中的特征选择算法研究. 光通信研究, 2005(3): 44~46

[26] 张海龙，王莲芝．自动文本分类特征选择方法研究．计算机工程与设计，2006, 27(20): 3840~3841

[27] 赵世奇，张宇，刘挺等．基于类别特征域的文本分类特征选择方法．中文信息学报, 2005, 19(6): 21~27

[28] 杨凯峰，张毅坤，李燕．基于文档频率的特征选择方法．计算机工程, 2010, 36(17): 33~35

[29] 郭亚维，刘晓霞．文本分类中信息增益特征选择方法的研究．计算机工程与应用，2012, 48(27): 119~122

[30] 范小丽，刘晓霞．文本分类中互信息特征选择方法的研究．计算机工程与应用，2010, 46(34): 123~125

[31] 肖婷，唐雁．改进的 X2 统计文本特征选择方法．计算机工程与应用, 2009, 45(14): 136~137

[32] 卢苇，彭雅．几种常用文本分类算法性能比较与分析．湖南大学学报:自然科学版，2007, 34(6): 67~69

[33] T Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Springer US, 1997: 143~151

[34] D Mladenic, M Grobelnik. Feature Selection for Unbalanced Class Distribution and Naive Bayes. in: Sixteenth International Conference on Machine Learning, 1999. 258~267

[35] WJ Hwang, KW Wen. Fast kNN classification algorithm based on partial distance search. Electronics Letters, 1998, 34(21): 2062~2063

[36] A Mathur, GM Foody. Multiclass and Binary SVM Classification: Implications for Training and Classification Users. IEEE Geoscience & Remote Sensing Letters, 2008, 5(2): 241~245

[37] Saaty T L. The analytic hierarchy process: planning, priority setting, resources allocation. New York: McGraw, 1980

[38] Zhu J, Momoh J A. Optimal VAr pricing and Var placement using analytic hierarchical process. Electric power systems research, 1998, 48(1): 11~17

[39] 莲芬，树柏，层次分析．层次分析法引论．中国人民大学出版社, 1990

[40] 金鑫．基于文本机会发现的共识与非共识标签区分方法: [硕士学位论文]．沈阳:

东北大学, 2011

[41] 陈俊杰, 侯宏旭, 高静. 一种 KeyGraph 的建模思想. 中北大学学报:自然科学版, 2014(2)

[42] 刘金岭. 基于语义密度的文本聚类研究. 计算机工程, 2010, 36(5): 81~83

附 录　攻读学位期间发表论文及申请专利情况

发表论文

[1] 夏明, 金吉祥, 陈起, 蒋金虎, 周可. 海量文本数据的多维度匹配方法研究. 见: 第 20 届全国信息存储技术学术会议. 北京: 2014, 91~98

申请专利

[1] 周可, 王桦, 金吉祥, 夏明. 一种基于多级缓存的混合云存储系统和方法. 申请号: 201310246369.6, 申请日期: 2013.6.20

[2] 郑胜, 张胜, 邹复好, 蒋丹, 夏明, 周可. 一种文本精细分类方法. 申请号: 201510239027.0, 申请日期: 2015.05.12