计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(4):234−240 [doi: 10.15888/j.cnki.csa.007948]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 人工智能语言处理技术在非结构化案件数据中的应用①

罗冬梅[1], 刘瑞军[2], 林锡平[3]

[1](武夷学院 信息技术与实验室管理中心, 武夷山 354300)
[2](武夷学院 科研处, 武夷山 354300)
[3](福建省南平市公安局 刑警支队, 南平 353000)
通讯作者: 罗冬梅, E-mail: luodmei@wuyiu.edu.cn

**摘　要**: 随着信息技术的快速发展, 以文本、音频形式记录在案的非结构化数据急速增长, 传统的案件人工处理方式已经很难满足应用需求, 对公安机关案件侦查带来了重大挑战. 对此, 本文提出了利用人工智能技术领域的自然语言处理技术, 对侵财类案件、电信诈骗类案件、团伙类案件等业务类型的信息系统中报警内容、简要案情、询问笔录等特征信息进行提取分析, 实现非结构化文本挖掘分析, 为侦查部门、情报部门提供研判支撑, 再通过发案时空与犯罪轨迹的信息比对碰撞, 并结合作案工具、作案手段等犯罪特点, 从中发现高危犯罪嫌疑人进行主动推荐, 可极大地缩小侦查范围, 提高侦破效率.

**关键词**: 人工智能; 自然语言处理; 案件串并; 实体识别

引用格式: 罗冬梅,刘瑞军,林锡平.人工智能语言处理技术在非结构化案件数据中的应用.计算机系统应用,2021,30(4):234–240. http://www.c-s-a.org.cn/1003-3254/7948.html

## Application of Artificial Intelligence Language Processing Technology in Unstructured Data of Cases

LUO Dong-Mei[1], LIU Rui-Jun[2], LIN Xi-Ping[3]

[1](Center for Information Technology and Laboratory Management, Wuyi University, Wuyishan 354300, China)
[2](Scientific Research Division, Wuyi University, Wuyishan 354300, China)
[3](Criminal Police Detachment, Nanping City Public Security Bureau, Nanping 353000, China)

**Abstract**: With the fast development of information technology and the consequent surge in the unstructured text and audio data, traditional manual ways of processing the cases are not suitable for practical applications, which has posed great challenges to the public security organs in case investigation. Thus, this study devises the artificial intelligence-based natural language processing technology to extract and analyze the characteristic information such as reports to the police, brief cases, and records of inquiries from the information system of cases of encroachment, telecom fraud, and gang. In this way, unstructured texts can be mined and analyzed, further supporting the judgment by investigation departments and intelligence departments. Moreover, spatio-temporal information, trajectories of the crime, and the characteristics of tools and means are compared. In this way, the high-risk suspects can be found and actively recommended, greatly reducing the scope of investigation and improving the efficiency of detection.

**Key words**: artificial intelligence; natural language processing; cases concatenation and combination; entity recognition

① 基金项目: 福建省教育厅科技项目 (JAT190779)
Foundation item: Science and Technology Program of Education Bureau, Fujian Province (JAT190779)
收稿时间: 2020-08-27; 修改时间: 2020-09-23, 2020-10-23, 2020-11-09; 采用时间: 2020-11-19; csa 在线出版时间: 2021-03-30

## 1 引言

随着科技的飞速发展, 刑事违法犯罪的手段也变得越来越多样化, 这便要求刑事侦查部门不断提供打击防范能力, 通过以信息化工作方式创新办案思路, 提高办案效率. 2018 年 1 月 24 日, 在全国公安厅局长会议上, 公安部党委书记、部长赵克志提出“建设智慧公安, 打造数据警务”的警务新理念.

当今, 国内外的学者越来越关注公安领域的数据挖掘技术研究, 利用公安部门多年来积累的犯罪信息数据及侦察破案的经验, 对其进行分析挖掘, 发现犯罪行为的规律、趋势, 了解案件之间的关联, 进行串并案分析是当前公安机关分析人员的主要任务. 利用知识图谱技术可以将公安情报部门掌握的琐碎、零散的情报信息相互连接, 以构建自动化、智能化海量文本情报处理业务流程和方法. 针对公安领域的数据挖掘工作在不断的深化, 虽然已取得了不错的进展, 但是仍具有很大的提升空间. 特别是针对案件串并和实体识别问题, 目前的文本挖掘主要解决案件的分类问题, 基于自然语言处理应用到公安案件数据挖掘中, 面向公安系列性刑事案件, 通过中文分词、词性标注、实体识别、文本聚类等方式, 为实现精细化的案件串并提供借鉴与参考, 实现案件串并过程“智能化”、“自动化”, 节省警务资源, 提高侦破效率.

自然语言处理 (Natural Language Processing, NLP)[1–5] 是一门融合了语言学、计算机科学、人工智能为一体的交叉性学科, 研究能实现人与计算机之间用自然语言进行有效通信的理论和方法, 解决“让计算机理解和合成人类的自然语言”. 自然语言处理技术主要包括词法分析、句法分析、命名实体提取、语义分析等, 它主要应用于自动摘要、信息检索、信息抽取、问答系统等领域. 其中, 命名实体提取技术作为自然语言处理的核心技术之一, 能有效提取文本内容中的命名实体信息, 对自然语言处理技术在实践应用有非常重大的意义.

当前, 自然语言处理技术已受到了国家中央政府、大型互联网企业的关注. 自然语言处理技术是机器学习当前最神秘, 最红火, 最具难度, 也最引人关注的分支. 在搜索引擎、情感分析、大批量文档处理、案件分析等各个领域有着前程无可限量的应用.

## 2 总体研究框架

本文从智能化案件串并和高危嫌疑研判两条线出发, 针对系列性案件, 对公安 110 警情、侵财类案件、电信诈骗类案件、团伙类案件等业务类型的信息系统中报警内容、简要案情、现场勘查、案件回访、询问笔录等特征信息进行提取分析, 实现非结构化文本数据自动分析、自动案件特征提取、案情特征聚类等数据挖掘分析, 为侦查部门、情报部门提供实体对象识别、案件串并研判支撑, 再通过发案时空与犯罪轨迹的信息比对、数据碰撞, 并结合作案工具、作案手段等犯罪特点, 通过轨迹数据的时空碰撞最终确定重点嫌疑人. 研究提供了从基于自然语言处理支撑案件串并、实体识别, 到高危嫌疑人智能推荐的一整套解决方法, 实现了沉睡警务数据的深度利用, 充分激发多源异构数据的融合与碰撞, 形成实用性的战法模型, 可极大地缩小侦查范围, 提高破案效率. 智能人案研判方法流程如图 1 所示.

## 3 自然语言处理应用于公安刑事案件串并

### 3.1 案件自然语言要素分析

基于对大量案件研判数据的深入分析, 利用开源基于人工智能系统的自然语言解析模型分析和机器学习技术, 通过中文切块分词、词性标注统计、命名实体提取、语义情感分析、热词推介等方式, 帮助警务人员从结构化和非机构化案件信息中提取其他关键要素.

(1) 中文切块分词

分词是自然语言处理的基础, 特别是中文切换分词[6] 的准确度, 它直接决定了后面的词性标注、句法语义分析、词向量以及文本分析的质量.

1) 基于字符串匹配的字典查找算法

先对语句进行分词, 然后从字典中查找每个词语的词性, 对其进行标注即可.

2) 基于统计的词性标注算法

和分词一样, 可以通过 HMM 隐马尔科夫模型[7] 来进行词性标注. 观测序列即为分词后的语句, 隐藏序列即为经过标注后的词性标注序列. 起始概率发射概率和转移概率和分词中的含义大同小异, 可以通过大规模语料统计得到. 观测序列到隐藏序列的计算, 利用统计得到的起始概率发射概率和转移概率来得到. 得到隐藏序列后, 就完成了词性标注过程.

针对公安案件的简要案情内容, 文本利用 Python 脚本语言封装和调用 jieba 中文分词组件的词性标注算法, 实现对中文分词切片, 如图 2 所示.

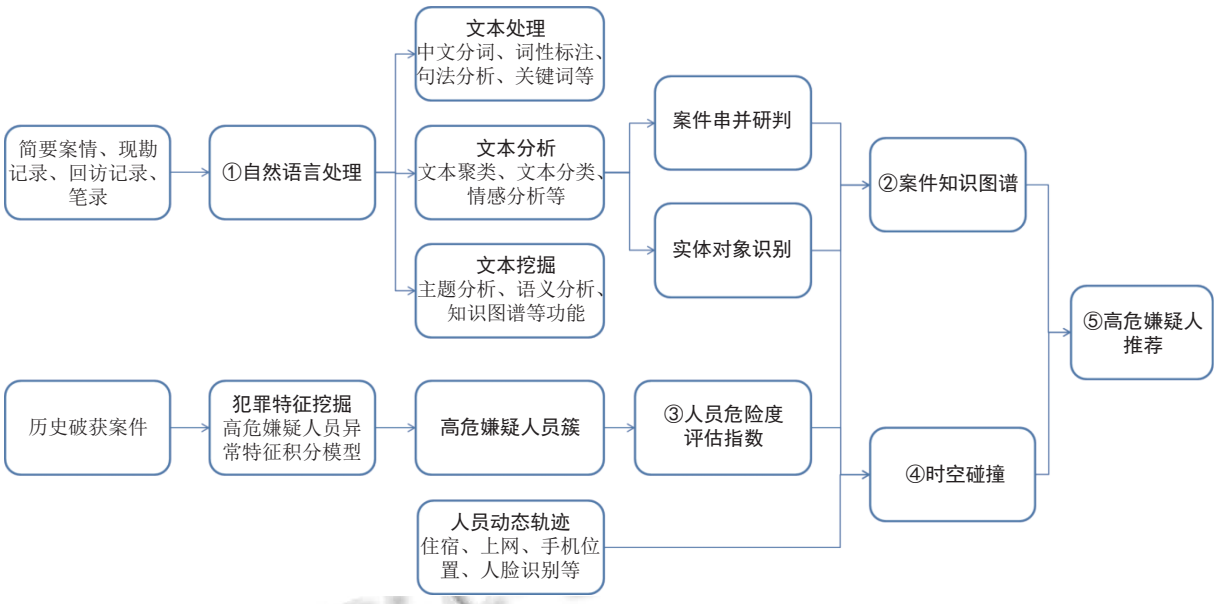


图 1 智能人案研判方法流程图

图 2 中文切换分词展示

(2) 词性标注统计

在中文分词切片基础上, 按照名称、动词等词性进行词频统计分析, 如图 3 所示.

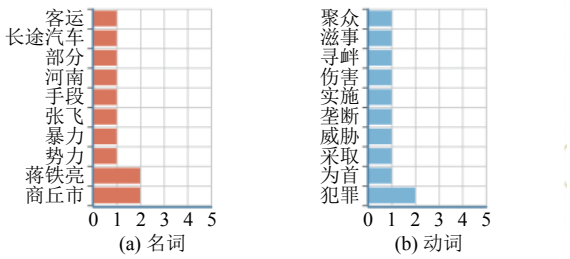


图 3 词性标注统计展示

(3) 命名实体提取

通过定义规则, 自动地对公安文本数据进行重要实体对象的提取, 提取的信息包括命名实体、常用术语等信息. 从公安案件的简要案情内容文本信息中提取出如人名、地名、公司名称、证件号码、时间、手机、QQ、微信、银行卡号等实体及实体间关系、事件等信息.

以人名识别描述其识别过程:

1) 初略实体

将称谓词、句首、前缀词、标点符号等作为分隔触发信息, 如果该触发词的后续词为人名等可用词, 则直到后缀词或连续字符为止, 中间的部分组成粗略人名对象集合.

2) 待选实体

结合实体识别规则, 在粗略人名集合中进一步提取待选人名信息.

3) 实体集合

如果待选人名为并列结构, 则将并列的词语分别加入待选人名集合中; 如果待选人名为正向结构, 且修饰的主语为人际关系指示词, 则将该待选人名的修饰词也加入待选人名集合.

4) 重复过程 3), 直到获得长度最小的待选人名.

通过以上步骤, 利用 Python 开发语言定义命名实体提取规则, 实现人名、地名、公司名称、证件号码、电话、时间等不同实体类型、实体信息的提取和识别, 如图 4 所示.

(4) 语义情感分析

语义情感分析是自然语言处理中常见的语义分析场景, 可以实现对案情的自动分类提供依据. 语义情感

分析可以采用基于情感语料库的典型方法和采用基于机器学习的情感分类方法.

1) 基于情感语料库的情感分类

基于情感语料库的方法, 先对文本进行分词和停用词处理等预处理, 再利用先构建好的情感语料库, 对文本进行字符集匹配, 从而挖掘正面和负面情感信息.

2) 基于机器学习的情感分类

基于机器学习的情感分类, 首先对语句进行分词、停用词、简繁转换等预处理, 然后进行词向量编码, 然后利用 LSTM 或者 GRU 等 RNN 网络进行特征提取, 最后通过全连接层和 Softmax 输出每个分类的概率, 从而得到情感分类.

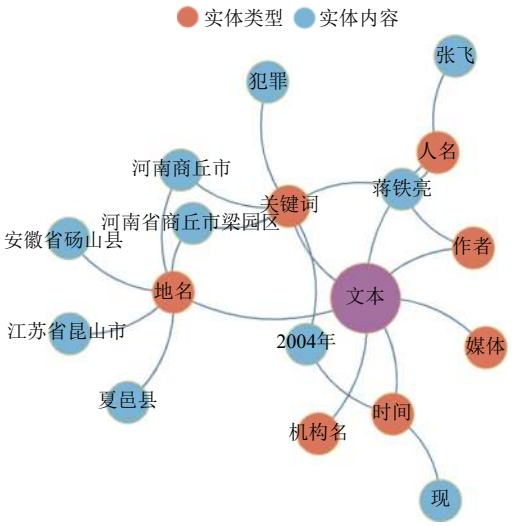


图 4 实体识别统计展示

(5) 热词推介

对历史案件样本数据, 通过以上语义情感分析组件提炼公安专用语料库. 通过得分算法案情描述的词频、命名实体提取对象进行计算, 最终形成案情热词、关键词组, 如图 5 所示.


图 5 关键词展示

### 3.2 案件智能串并

智能研判串并根据机器智能提取要素特征, 通过主题分析、语义分析等智能研判分析, 实现案件串并研判、实体对象识别. 围绕完成结构化处理后的案件信息, 建立基于领域知识库的多维数据模型, 与更多可对接的系统数据碰撞, 进一步挖掘关联价值, 形成案件知识图谱, 为案件侦破推荐特征类似的嫌疑人提供有力支撑.

(1) 案件分类

本文相似性算法的自动聚类分析技术, 自动将公安非结构化文本内容中对无类别的事件、警情信息进行归类, 把案情相近的案事件信息归为一类, 赋予文本内容一个预先设定的案件类别分类库, 实现根据文本内容进行案件类别划分, 从而达到提高分类精度的目的, 不需人工干预.

(2) 相似案件分析

基于以上案件分析对案件数据的提取和分类标记, 系统提供相似案件查询、相似案件基本信息、相似案件分析等. 通过案件类别进一步对相似案件的建进行研判分析, 实现基于案件特征的相似案件挖掘, 同时照新的按人工监督下分类规则进行相似案件学习和分析.

(3) 案件串并分析

基于对案件分类标记和相似案件分析, 系统自动对新发案件进行关联分析、关联值评估、串并分析、串并案可视化分析等.

1) 案件数据关联串并分析是将案件嫌疑人的姓名、身份证号、发案时间、手机号码、虚拟身份等要素进行关联, 并根据关联度进行关联权值评估, 从而找出案件之间的关联关系.

2) 基于案件特征的相似案件挖掘, 利用大数据对案件之间的相似特征进行整理推荐. 在数据整合的过程中, 将案件涉及的人名、地名、电话、虚拟身份、银行卡、体貌特征、身高、作案手段、作案时间等案件特征识别并添加至案件的标签.

3) 案/事件智能串并分析利用资源库数据, 结合可视化关系挖掘工具, 利用大数据技术, 挖掘出案件之间的内在关联, 实现串并案分析. 通过手机号码实现案件与案件串并的关系, 结合车辆、时间、人物、地点、作案工具等要素实现串并案分析.

## 4 利用大数据轨迹碰撞发现嫌疑对象

侦查工作就是利用事实的相关性来捕捉案件线索,

“环环相扣”构建数据证据链条, 而大数据体现的相关关系是立体的、多维度的, 信息范围广, 更有助于侦查工作的开展. 基于大数据的数据关联碰撞、数据挖掘分析出的预警预测方法, 可以为系列性案件侦破提供从“案到人”的犯罪预测, 使得侦查部门能够尽早甚至第一时间发现犯罪嫌疑人, 达到犯罪预测预防的能力.

尤其针对系列性入室盗抢等侵财案件[8], 犯罪嫌疑人习惯与原有作案手法继续作案, 真实办案过程中, 侦查部门会运用案件侦查经验和现场勘查情况, 将同一个或同一犯罪团伙所做的案件串并起来统一侦查. 案件串并之后明确根据案件发案时间、发案地址, 系列性案件的发案时间、空间两个维度就是轨迹数据时空碰撞的主要输入条件选, 根据犯罪对象在案发区域产生的包括旅馆、网吧、手机位置、车辆等数据轨迹, 并结合案件类别、作案手段、作案方式等特点, 通过轨迹数据时空碰撞, 作案信息比对排查嫌疑人范围, 并按评估指数精选排名, 最终确定重点嫌疑人.

### 4.1 高危嫌疑人员评估指数

充分利用历史破获案件通过建立高危嫌疑人[9] 异常特征积分模型实现犯罪特征挖掘, 形成高危嫌疑人员簇, 整理多种数据标签, 从多维度进行人物描绘, 通过机器学习的回归算法提供精准犯罪评估指数.

(1) 本地案件嫌疑人员分析

对本地办案系统中的同类型系列性、团伙性案件, 通过对抓获嫌疑人员的高危地区 (户籍地、籍贯) 进行分析, 按照案件类别、作案手段、作案特点等属性, 归纳出某一类型案件的高危地区人群.

(2) 跨区域案件嫌疑人员分析

利用全国刑侦系统数据及全国前科人员数据, 针对跨区域系列性案件, 重点针对相邻的省市侦办的同类型案件, 通过分析已抓获嫌疑人员高危地区 (户籍地、籍贯) 进行分析, 比对案件类别、作案手段、作案方式等特点, 形成某一类型案件的高危地区人群. 并可重点关注, 越是相邻距离近的高危地区人群作案特点越相似, 相距较远地区的高危人群可作为参考.

(3) 侦查部门归纳总结

对于刑侦、情报等侦查部门已经掌握形成作案专项的高危地区 (如外币诈骗、抛物诈骗、婚姻诈骗、抢劫出租车、麻醉抢劫等) 高危人群, 形成高危地区与案件类别、案件手段经验归纳库. 可通过办案经验不断归纳完善, 或直接与高危地区 (户籍地、籍贯) 的公安机关确认联系, 提高对高危人群与案件类别、作案特点关联的准确性.

(4) 通过分析作案特点分析

对案件信息、案件嫌疑人建立关键字组合检索工具, 以案件的案件类别、作案工具、作案手段、侵入方式、侵害对象、案件状态、简要案件等为条件, 细化高危地区作案特点分析.

(5) 前科特征人员积分

对前科人员、前科侵财人员、同类案件前科人员、多人同时来、多人同住、作案后离开、(多次) 凌晨入住 (上下网)、频繁变更旅馆住宿、案发期间频繁活动、夜间跨区活动、流窜作案有驾驶证、是已破同类案件关系人且同时来本地过 (同住宿)、住宿登记人员的关系人有侵财前科等因素进行自动赋分.

### 4.2 案件时空碰撞推荐

从“地域”和“时域”两个维度洞察案件关联特征, 从而清晰地了解某区域特定的案件类型, 发案位置, 作案时间等规律信息, 在上面要素合并的基础上结合公安各类轨迹数据, 基于公安地理信息系统, 对串并的案件进行时空轨迹碰撞, 达到高危嫌疑人智能推荐.

(1) 案件时空特征提取

本文研究的案件主要是系列性侵财案件, 案件的关键数据主要是发案时间、发案地址 (定位到地图坐标) 是案件时空碰撞的前提条件.

1) 提取案件发案时间

一般入室盗抢、扒窃等侵财类案件发现, 受害人基本上都无法准确提供案件发生的精确时间点, 只能推断出大概的时间段, 所有对时间提取需根据案件发案日期提供按照日期段提取、时间段提取多个维度提取案件发案时间的范围.

2) 提取案件地图坐标

根据受害人报案时提供的案件案发地点描述的抽象地址信息, 通过报案电话地址地图定位范围和描述的地址信息, 利用地图服务坐标转换, 将文字描述的地址信息, 转换为精确的地图坐标.

(2) 轨迹数据时空维度碰撞

利用警用地理信息系统, 通过对公安大数据的综合应用, 以多个串并案件发生地为中心, 可在地图上标注案发地, 在案件发生前后对经过的地图轨迹, 高危地区人员的旅馆住宿、网吧上网、火车票、汽车票、飞机票、以及从互联网公司获取的各类消费信息、活动

轨迹等, 同时接入手机位置轨迹、车辆轨迹, 以及全息感知网建成后设备采集轨迹数据, 与案发地的重合度进行系统自动比对发现高危嫌疑人员, 可查看案件详情及案后侦查情况, 通过算法按积分倒叙推荐可疑对象.

(3) 重点嫌疑人落地查证

以上案件时空提取和轨迹时空碰撞, 通过分类赋分、数据挖掘、综合计算, 自动对特定人群进行立体、综合研判, 从海量数据中自动筛查具有高作案嫌疑指数的对象重点目标, 很大程度上减少了警力排查研判的过程. 但是, 这种系列性案件只是代表了某类案件的高危地区人员在某一时间段的高危嫌疑, 高危地区人员作案特点会随着新型犯罪手法出现发生变化. 因此, 通过案件时空碰撞推荐的高危嫌疑人需推送相关警种及基层一线落地核查, 进一步分析认定或排除其作案嫌疑人, 从而不断检验、修正、完善推荐结果.

## 5 非结构化案件数据分析研判和碰撞挖掘的设计和实现

本文研究的基于自然语言处理的非结构化案件数据分析研判和碰撞挖掘, 技术上利用 Java 开发语言, 基于开源 jieba 自然语言处理组件, 采用主流 Hadoop+Spark 大数据框架体系对大数据进行存储、处理和挖掘, 结合综合预警模式, 从智能化案件串并和高危嫌疑研判两条线出发进行设计和实现.

### 5.1 智能化案件串并方面

首先通过自然语言处理, 利用 Python 脚本语言封装开源 jieba 自然语音处理组件的词性标注算法、命名实体提取规则, 语义情感分析算法, 利用历史案件中简要案情样本数据, 提炼出公安专用语料库; 再通过 Java 调取 Python 实现对案件实体对象的识别和热词、关键词组的提取. 然后采用 SparkMLlib 中 K-means 相似性算法的自动聚类分析技术, 对以上实体对象识别和热词提取的结果, 进行案件分类, 并结合相识案件分析功能按人工监督下分类规则进行相似案件学习和分析, 实现对新发案件的自动串并. 如图 6 所示.

### 5.2 高危嫌疑研判方面

首先基于 SparkMLlib 回归算法之决策树算法, 对本地案件嫌疑人员数据、跨区域案件嫌疑人员数据和侦查部门归纳总结数据进行分析, 形成高危嫌疑人员簇, 实现高危嫌疑人员评估指数建立人员积分. 然后对接公安掌握的网上网下各类轨迹数据, 采用 Hadoop 大数据框架对高危嫌疑人员海量轨迹数据进行分布式存储. 最后对以上新发串并案件的时空特征包括发案时间、发案地址进行提取, 通过人员轨迹数据与串并案件的时空维度采用 Spark 实时计算引擎进行计算碰撞, 利用警用地理信息系统进行直观展示, 并按积分倒叙推荐重点可疑对象, 实现高危嫌疑人挖掘, 大大提高刑侦办案民警办案效率, 极大提高破案率. 如图 7 所示.

图 6 利用自然语言处理进行案件串并分析

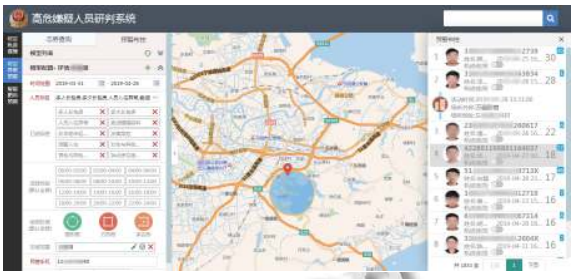

图 7 通过时空碰撞挖掘实现嫌疑人推送预警

## 6 结论及展望

基于自然语言处理的非结构化案件数据分析研判和碰撞挖掘的研究, 旨在大数据、人工智能环境下, 为案件侦查和情报分析的数据挖掘、研判工作提供更好途径, 以解决公安机关案件线索提取的实际困难, 为公安机关侦查的实际工作提供支撑. 利用大数据、人工智能等技术辅助案件侦查应用, 是一个不断学习优化的过程, 后期的价值判断还需要侦查人员的核实反馈, 输入准确的学习样本以提高数据分析挖掘的准确性. 另外, 随着的作案手段和犯罪类型的变化, 需要专业的侦查人员对预警模型不断进行监督、修正和完善. 目前公安机关全面推进“智慧警务”建设, 基于自然语言处理的非结构化案件数据分析研判和碰撞挖掘的研究, 是智慧警务一个实战应用的缩影, 是公安业务实战应用的一个前沿探索和实践, 有助于打造智慧警务新模式.

## 参考文献


1 张凌霄. 基于聚类的串并案分析研究与实现 [ 硕士学位论文 ]. 上海: 东华大学, 2017.

2 白继峰, 张蕾华. 公安文本情报的智能化处理方法与实践. 山西警察学院学报, 2018, 26(3): 90–94. [doi: 10.3969/j.issn.1671-685X.2018.03.017]

3 Allen J. 自然语言理解. 刘群, 张华平, 骆卫华, 等译. 2 版. 北京: 电子工业出版社, 2005.

4 张德. 自然语言处理技术在司法过程中的应用研究. 信息与电脑, 2017, (17): 33–34. [doi: 10.3969/j.issn.1003-9767.2017.17.013]

5 李静, 罗文华, 林鸿飞. 自然语言处理技术在网络案情分析系统中的应用. 计算机工程与应用, 2012, 48(3): 216–220. [doi: 10.3778/j.issn.1002-8331.2012.03.064]

6 梁喜涛, 顾磊. 中文分词与词性标注研究. 计算机技术与发展, 2015, 25(2): 175–180.

7 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别. 通信学报, 2006, 27(2): 87–94. [doi: 10.3321/j.issn:1000-436X.2006.02.013]

8 杨万方. 大数据驱动下多发性侵财犯罪的侦防路径研究. 广州市公安管理干部学院学报, 2018, 28(3): 9–13.

9 井晓龙. 高危分析在多发性侵财案件侦查中的应用. 中国人民公安大学学报 (社会科学版), 2013, 29(4): 67–72.