doi :10.3969/j.issn.1671-1122.2013.11.017

数据挖掘在防范和打击计算机犯罪中的应用研究

叶明[1,2]，叶猛[2]

（1. 武汉邮电科学研究院，湖北武汉 430074；2. 武汉虹旭信息技术有限责任公司，湖北武汉 430074）

摘　要：面对海量的互联网数据，挖掘计算机犯罪数据是目前需要研究和解决的问题。文章提出一个方案，即挖掘已有的海量案件，寻找海量案件中的规律性，建立一个案件知识库模型，从而达到预防和打击计算机犯罪的目的。

关键词：数据挖掘；大数据；知识库

中图分类号：TP309　　**文献标识码：**A　　**文章编号：**1671-1122（2013）11-0067-04

The Research on Data Mining in the Preventing and Combating Computer Crime

YE Ming[1,2],YE Meng[2]

(1.*Wuhan Research Institute of Posts and Telecommunications, Wuhan Hubei* 430074, *China;* 2.*Wuhan Hongxu Information Technology Co., Ltd., Wuhan Hubei* 430074, *China*)

Abstract: Facing the massive Internet data, digging out computer crime data is a problem to be studied and solved. This paper presents a scheme ,which mines the existing mass cases,finds the regularity in massive case,and establishs a case knowledge base model, thus preventing and combating computer crime.

Key words: data mining;big data; knowledge base

0 引言

计算机和互联网在给人们的生活带来极大便利的同时，它们带来的负面影响也随着它们的迅速发展而急速增加。计算机犯罪日益猖獗，呈现的形式越来越多样化。计算机犯罪不仅影响个人和企业的经济利益，而且严重危害社会的发展和稳定。在此背景下，计算机取证研究受到司法部门的高度重视。计算机取证中收集的数据是海量的，并且格式各异、来源复杂，想要在海量的数据中提取有用的数据，提前判断可疑对象和可疑区域，做到打击和预防计算机犯罪，是当前面临的问题之一。

公安机关在“科教兴警”的方针指导下，大力发展信息化建设，建立起横向到边、纵向到底的公安信息化网络[1]。当前，各个警种的业务全面实现了信息化管理，积累了大量的基础业务，但这些通过工作经验积累起来的业务只是用来进行简单的数据查询、统计、更新。目前，在大数据背景下很多行业，包括生物界、金融界、医学界、电信业、保险业、零售业等[2]，为获取有效信息各自利用数据挖掘技术取得了一定的实效。因此，如何利用海量案件数据，通过数据挖掘技术，发现大量杂乱无章的数据背后的规律性信息，为各个警种提供技术支持和参考，值得我们进一步思考。本文提出建立一个案件知识库模型，在这个模型中存放挖掘出的规律，让这些规律来帮助判断海量网络数据中哪些信息对警务人员是有价值的。

1 计算机犯罪及数据挖掘技术

计算机犯罪和传统案件相比有着显著的不同：1）犯罪人员一般具有专业技术知识，很多还都有计算机工作背景，相比传统犯罪人员更具有智能性。2）犯罪行为表现出复杂性和多样性，相比传统犯罪人员的犯罪行为更加多变，规律并不是非常清晰。3）犯罪手段具有隐蔽性，相比传统犯罪人员的犯罪痕迹往往不易被发现且更加容易被销毁。以上几个特点给打击计算机犯罪带来

收稿日期：2013-08-23

基金项目：“十二五”国家科技支撑计划 [2012BAH38B05]

作者简介：叶明（1989-），女，湖北，硕士，主要研究方向：网络信息安全；叶猛（1975-），男，湖北，高级工程师，硕士生导师，博士，主要研究方向：网络信息安全。

很大困扰。但是通过仔细分析多种犯罪数据发现，很多计算机犯罪手法是相似的或者在某一点上有一定的关联性。由此，深入挖掘海量犯罪数据，寻找数据背后的规律并建立知识库，对后期在海量网络数据中寻找犯罪线索很有帮助。美国的很多城市都采用“预测警务”来决定哪些街道、群体或个人需要更严密的监控[3]。“预测警务”说白了就是在海量数据中挖掘出可以预见的犯罪，即用当前已知的数据来预测未来[4-6]。

数据挖掘只是一个工具，它又称为数据库中的知识发现（KDD）[7-10]。这个技术是由数据驱动的，而不是由用户驱动的。当算法或规则确定时，只要给出数据，不用告诉算法程序怎么做和用户期待什么结果，算法自动在大型数据库、数据仓库或其他大量信息存储设备中提取尽可能多的隐藏知识[11]。而我们需要做的就是“听”数据在“讲”什么。这种能够收集和分析海量数据的技术将帮助我们更好地理解世界[12]。

2 建立案件知识库

2.1 案件知识库的初始值

建立案件知识库最开始就需要采集初始值。初始值可以人为随意设定吗？当然是不行的，建立案件知识库的宗旨就是：一切听从数据的。根据对大数据相关资料的研究得知，数据量越大，挖掘出来的规律就越准确、越值得信赖。所以在知识库建立之初，需要使用以前积累的所有和计算机相关的案件数据，这就是前期的采集工作。办案过程中人工提取数据项的过程等同于将案件数据进行了预处理，即去除其中部分无用数据，同时把使用人类语言抽象描述的案件数据转化成比较具体的数据项，如案件类型、姓名、作案时间、作案地点、作案目标、作案动机等。所有数据项中的数据都需要手动输入或导入到知识库中。表1给出了知识库中部分数据项的说明：

表1 知识库中部分数据项说明

序号	数据项	类型	备注
1	案件类型（代号）	int	按照案件类型表中的编号填写
2	姓名	char	
3	作案时间	Time	
4	籍贯	char	
5	作案地点	char	
6	性别	char	1.男；2.女
7	年龄	char	
8	作案动机	char	按照动机类型表中的编号填写
9	文化程度	char	1.高中及以下；2.大专；3.本科及以上
10	敏感词语	char	按照敏感词表中的编号填写

具体每条案件导入格式如表2所示：

表2 案件格式示例

案件序号	案件类型	姓名	绰号	性别	籍贯	文化程度	年龄	作案动机	作案途径	敏感词语	...	作案地点
1	1003	张三	张麻子	1	湖北黄冈	3	32	13	QQ聊天	98、136	...	武汉市武昌区

表2中，案件类型1003代表网络钓鱼类案件，作案动机13代表对目前的生活水平极度不满意，敏感词语98和136分别表示“银行账号”和“中奖”这两个词语。一些数据项没有数据时可以不予填写。海量案件数据全部收集和预处理完毕之后，下面就需要对这些数据进行数据挖掘，以获取其中的深层规律。

1）关联分析

1993年，Agrawal等人设计出Apriori关联规则频繁项集算法，主要用来在大型数据库上快速挖掘关联规则。即从数据源中找出形如“由于某些事情的发生而引起另外一些事情的发生”这样的规则[13-15]，推出每个数据项之间的相关性，从而找出属性集中不同属性之间的联系。

设定知识库的数据项为犯罪类型、姓名、性别、籍贯、文化程度、家庭住址、出生背景、出生年月、身高、工作经历、绰号、体型、脸型、作案动机、作案时间、作案地点、作案工具、流窜类别以及流窜范围。我们提取犯罪人员的年龄、文化程度、工作经历和作案动机这几个特定项，对这几个特定项使用Apriori关联频繁项集算法进行关联分析，发现在网络钓鱼、造谣诽谤、黑客攻击这3类常发生的计算机犯罪中，犯罪人员的特点有如下几个相似之处：（1）年龄。大部分处于20到35岁这个阶段。（2）文化程度。基本为大专以上学历。（3）工作经历。一半以上有过计算机相关背景的工作。（4）作案动机。对目前工作和生活极其不满。设定每一个关联条件为W_i，将同时满足n个条件的集合称作一个相关点（$d=(W_1,W_2,...,W_n)$）。利用Apriori关联频繁项集算法得到这个相关点在案件发生过程中的可信度。通常得到的可信度的值越大，满足相关点中所有条件的人员的犯罪概率就越高。

2）聚类分析

将数据对象划分成若干个类，同一类中的对象具有较高的相似度，而不同类中的对象差异较大[16-18]，聚类就是把具体的或者抽象的对象按照相似程度分类的一个过程。可以根据事先设定好的规则，合理划分数据集合，这个规则是根据系统的自身需求设定的。具体到本文的案件知识库，根据警务人员想要获取的不同内容设定不同的规则。例如，若警务人员想知道犯罪人员喜欢浏览哪些网页，那么针对Web浏览数据进行模式分析[19-20]，将不同人员浏览的同类网页归到一个浏览模式中，则这个具体的浏览模式即为该规则下的聚合点。又如，若警务人员想知道不同犯罪人员平时对社会的关注点分别在什么地方，可以对社交网站、微博等社交型应用进行聚类分析，挖掘出不同犯罪人员同时对哪类人群或哪类文章有强烈兴趣，这时关注该类人员或该类文章就成为这个规则下的聚合点。

3）敏感词统计

各类警务人员根据办案经验和近期的敏感话题、热点话题，建立一个符合本地区情况的敏感词表。敏感词表中的每个敏感词都有一个权值，这个权值直接反映了该敏感词的敏

感程度。

例如，某地区提供的敏感词表如表 3 所示。

表3 敏感词表示例

序号	敏感词	权值	所属类型
53	法轮功	0.544	313
64	AV	0.694	412
76	游行	0.362	06、123
98	银行账号	0.336	915、1006
136	枪支走私	0.756	01
190	购买器官	0.675	01

表 3 中，“所属类型”中的每一个代号对应一种案件类别。其中，313 代表“宗教迷信”类、412 代表“黄色淫秽”类、06 代表“政治活动”类、123 代表“非法游行”类、915 代表“网络诈骗”类、1006 代表“网络钓鱼”类、01 代表“刑事案件”类。从表 3 中可以看到有些词所属的案件类别可能不止一个，这表示该词在多类案件中均为敏感词语。表 3 中的“序号”和表 2 中的“敏感词语”是对应关系。

根据警务人员提供的敏感词表，定期对知识库的所有数据统计其中敏感词的词频，计算每个敏感词在各类案件中的出现频率。系统最终会根据加权计算方法，算出每个敏感词的权值和词频的综合值，从而推选出每类案件综合值最高的前 5 个敏感词作为该类案件的重点敏感词。

通过数据挖掘得到了相关点和其可信度、聚合点和其聚合程度以及重点敏感词后，知识库的初始值才算真正建立成功。

2.2 利用知识库预防和打击计算机犯罪

2.2.1 评估值的确定

利用 2.1 节中的挖掘方法，不同案件类型会得到不同的初始值，这个初始值也称为评估点。评估点来源于以下几点：

1）由知识库的关联分析得到的相关点。2）由知识库的聚合分析得到的聚合点。3）各类案件中的重点敏感词。

针对每一个评估点系统也会给予相应的评估值。评估值的确定需要综合考虑以下几点因素：

1）知识库中可信度越高，对应相关点的评估值就应设置越高。例如，假设买牛奶是一个犯罪行为 h，分析知识库中购买牛奶的所有数据，得出购买面包和购买黄油这两个行为形成一个相关点 d，并且在海量网络数据中满足 d 的人 90% 都进行 h 这个行为。很显然满足相关点 d 的人对于警务人员来说就是“危险人物”，该评估点的值必然设置较高。

2）知识库中聚合度在案件库和海量网络数据中的差值越高，对应聚合点的评估值就应设置越高。例如，频繁攻击某一端口，这个聚合点在黑客类案件中聚合度为 0.76，但在海量网络数据中聚合度很低，为 0.03，那么两者的差值很大。又如，关注雅安地震这条新闻，这个聚合点在知识库中的聚合度很高，为 0.81，同时在海量网络数据中的聚合度也很高，为 0.84（已经超过在知识库中的聚合度），这时两者的差值就很小。那么，虽然后一个例子的聚合点相较于第一个例子的聚合点在知识库中的聚合度更高，但是最终得到的评估值应远远小于第一个例子的聚合点的评估值。

3）敏感词表是由办案人员人工输入或导入的，表中的词语一般为法律敏感词汇，如“枪支买卖”、“毒品走私”等。按照系统推出的 5 个重点敏感词的综合值从大到小的顺序，应从高到低赋值。

4）警务人员特别关注某类数据时，可以人工将一些评估点的评估值调高。但需要注意的是，当不再需要重点关注时，应当恢复这些评估点的评估值到正常值。

2.2.2 告警阈值的设定

各个评估点的评估值确认完成后，需要根据不同案件设定不同的告警阈值。告警阈值可以根据以下几点来综合评定：

1）不同案件种类对社会的危害程度。

2）近期各类案件在该地区发生的频繁程度。

3）案件种类是否需要重点关注。

本文按照案件种类对社会的危害程度分为极度危害、一般危害和轻微危害 3 种，给予的分值分别为 0.2、0.5 和 0.8。根据近 3 个月警务人员的办案情况，算出不同种类案件发生的概率 f，再通过公式 $\log(0.15/f)$ 得到每类案件发生的频繁程度。当案件类别不属于重点关注类别时，案件的告警阈值还要再增加 0.5。例如，网络钓鱼的危害程度为一般危害，赋值 0.5；近 3 个月网络钓鱼类案件发生概率为 3.2%，算出的对应频繁程度为 0.67；同时网络钓鱼属于一般案件，告警阈值再增加 0.5。最终计算得到网络钓鱼的告警阈值为 0.67+0.5+0.5=1.67。

告警阈值的设定还应遵循如下原则：关注度越高的案件，告警阈值应设置得越低。系统会根据事先设定的算法给出一个告警阈值，并同时给出一个建议可控范围。这个阈值可以根据具体警种人为调控，但最好控制在系统给出的建议可控范围内。当某个个体或某个区域的网络数据的评估值之和超过某类案件的阈值的时候，警方需要重点关注一下该人员或该区域的活动情况，观察是否已有犯罪事实或犯罪倾向，加强防控并尽可能消除犯罪诱因，最终达到打击和预防犯罪的目的。

3 知识库模型

互联网数据每天都会有很大的更新和变化，系统不能以一个不变的标准看待变化的数据。前面讲述一个静态的知识库是如何收集数据、清洗数据和分析数据的，要想创建出来的知识库一直保持有效性，必须随时注入新的数据，更新知识库中各类参数的值。

图 1 所示是一个不断更新的案件知识库模型。在这个模型中必须有与案件知识库形成对比的海量网络数据。本文的

目的也就是利用知识库中的规律知识找到海量网络数据中的异常情况。

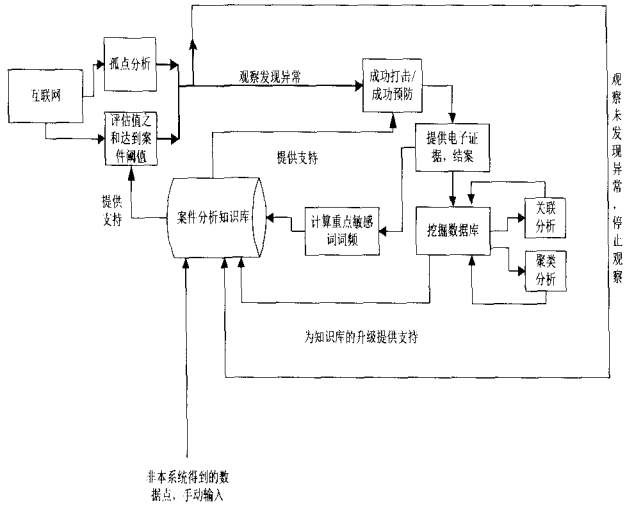

图1 案件知识库模型图

该案件知识库模型操作流程如下：

1）利用数据挖掘技术挖掘出互联网数据中的孤立点。所谓孤立点是指在互联网海量数据中和一般模型数据不一样的数据对象。大部分系统的数据挖掘都是对这类数据冠以“噪声”而予以剔除，但是对本文案件知识库模型而言，互联网中的孤立点意味着可能有案情发生，这对警务人员来说是非常有价值的。所以对互联网中的孤立点数据进行分析和跟踪观察，可以起到一定的预防计算机犯罪的作用。

2）将海量互联网数据和案件库中的评估点进行对比，算出该数据在每类案件中的评估值之和，如果某个个体或某个区域的所有评估值的总和达到案件的阈值，该人员或该区域需要重点分析和加强防范。分析和观察一段时间后，再结合知识库的数据和办案人员的经验判断是继续防范还是撤销警戒。

3）如果按照这套系统成功预防或打击了某个计算机犯罪案件，则说明这个案件所使用的评估点和评估值比较有效，可以在可调范围内酌情提高这些评估点的评估值，同时将这个案件的所有数据项输入案件知识库，给案件知识库注入新鲜血液。反之，如果分析和观察一段时间后，发现系统给出了错误的判断，则说明该案件的评估点和评估值设定有些偏差。这类数据也需要作为“反面数据”返回知识库，并对相应的评估点减值。

4）对于已经确定案情发生，但暂时还找不到突破口侦破的计算机犯罪案件，知识库中同类型案件的评估点可以为警务人员提供一种思考方向。

5）任何一个系统都不能保证考虑到所有因素，为了让系统更具完备性，可以适当综合考虑其他系统数据，手动调控系统中的评估点和评估值以及其他数据。所有返回值都是为了完善系统，使系统数据更具实时性和有效性。

4 结束语

综上所述，案件知识库在网络信息监控体系中起到了集中信息、挖掘信息、分析信息和排查信息的作用，可用于全面搜索网络中可能出现的涉警信息。虽然数据量越大，我们的判断越准确，但目前还没有一个大数据系统能100%地准确预估未来。所以本文系统只能作为一种参考和思路，指导需要重点对哪些区域和个人进行防御，但不能作为判定一个人犯罪的真实证据。本系统在综合治安方面，在加强重点场所和重点人员的防控工作，消除犯罪诱因等方面，起到了预防和打击计算机犯罪的作用。（责编 马珂）

参考文献

[1] 徐腾．计算机数据挖掘技术在犯罪规律分析中的应用研究[J]．自动化与仪器仪表，2012，(04)：80-81.
[2] 钟秀玉，凌捷．计算机动态取证的数据分析技术研究[J]．计算机应用与软件，2004，(9)：26-27
[3] 维克托·迈尔－舍恩伯格，肯尼思·库克耶著．盛杨燕，周涛译．大数据时代[M]．杭州：浙江人民出版社，2013.
[4] 刘军，吕俊峰．大数据时代及数据挖掘的应用[N]．国家电网报，2012-05-12(010)．
[5] 李广建，杨林．大数据视角下的情报研究与情报研究技术[J]．图书与情报，2012(06)：1-8.
[6] 西安美林电子有限责任公司．大话数据挖掘[M]．北京：清华大学出版社，2013.
[7] 毛国君．数据挖掘技术与关联规则挖掘算法研究[D]．北京：北京工业大学，2003
[8] 李强．数据挖掘中关联分析算法研究[D]．哈尔滨工程大学，2010.
[9] 周纪．数据挖掘技术在安全技术防范管理中的应用研究[J]．中国公共安全(公共版)，2009(11)：144-149
[10] 陈宏．浅谈数据仓库与数据挖掘技术及应用[J]．科技广场，2011(09)：90-93.
[11] Jiawei Han,Micheline Kamber著．范明，孟小峰译．数据挖掘一概念与技术[M]．北京：机械工业出版社，2001.
[12] 甄彤．基于层次与划分方法的聚类算法研究[J]．计算机工程与应用，2006，(08)：178-180.
[13] 李菁菁，邵培基，黄亦潇．数据挖掘在中国的现状和发展研究[J]．管理工程学报，2004，(03)：10-15.
[14] 钱雪忠，孔芳．关联规则挖掘中对Apriori算法的研究[J]．计算机工程与应用，2008，(17)：138-140.
[15] 董振华，李喜艳，张开便．基于关联规则的经典Apriori算法研究[J]，2012(06)：148-149.
[16] 陈宇．聚类算法研究[J]．福建电脑，2007，(07)：27-29.
[17] 王鑫，王洪国，王珺，王金枝．数据挖掘中聚类方法比较研究[J]．计算机技术与发展，2006，(10)：20-22.
[18] 马利．基于数据挖掘的聚类分析和传统聚类分析的对比研究[J]．数理医药学杂志，2008，(5)：530-531
[19] 平金珍，王茜，于莉莉．聚类分析在Web数据挖掘中的应用研究[J]．科技信息，2013，(19)：75-75.
[20] 贾丙静．聚类分析在Web文本挖掘中的应用研究[D]．阜新：辽宁工程技术大学，2007.