

基于个性特征仿真邮件分析系统挖掘犯罪网络核心

乔少杰¹⁾ 唐常杰¹⁾ 彭 京²⁾ 刘 威¹⁾ 温粉莲¹⁾ 邱江涛¹⁾

¹⁾(四川大学计算机学院 成都 610065)

²⁾(北京大学信息科学技术学院 北京 100871)

摘 要 数据挖掘应用于犯罪集团或恐怖组织社会网络分析是一种新兴的研究方法,国内外在分析犯罪和恐怖组织之间通信行为方面的研究工作亟待深入.为了模拟社会网络中个体利用电子邮件进行通信的规律,设计了一种基于个性特征的仿真邮件分析系统 MEP,提出一种利用个性特征判别矩阵计算个性特征矢量各个维度权重的新方法,借助符合用户个性特征的正态分布模型模拟真实的邮件通信行为.为了挖掘犯罪网络的核心成员,提出了一种基于社会网络分析挖掘犯罪组织核心成员的算法 CNKM(Crime Network Key Member mining),并利用时间序列分析方法对邮件的收发规律进行深入分析,发现异常通信事件.实验证明了该文提出的仿真邮件分析系统的有效性和实用性,模拟邮件通信的平均误差小于 10%,并验证了 CNKM 算法的有效性.

关键词 数据挖掘;个性特征;仿真;邮件分析系统;社会网络分析

中图法分类号 TP311

Mining Key Members of Crime Networks Based on Personality Trait Simulation Email Analysis System

QIAO Shao-Jie¹⁾ TANG Chang-Jie¹⁾ PENG Jing²⁾ LIU Wei¹⁾ WEN Fen-Lian¹⁾ QIU Jiang-Tao¹⁾

¹⁾(School of Computer Science, Sichuan University, Chengdu 610065)

²⁾(School of Electronics Engineering and Computer Science, Peking University, Beijing 100871)

Abstract It is a new paradigm to apply data mining technologies to analyze the crime groups and terrorist social networks, there is little work being done on analyzing the communication behavior of criminal and terrorist groups. This paper designed a simulation email system based on personality trait dimensions, called MEP, to model the email users' traffic behavior, proposed a new approach of computing the weight of each dimension in a personality trait vector by using personality trait judge matrix, and simulated the real-world email communication behavior based on normal distribution model satisfying users' personality trait. This paper proposed a social network analysis based algorithm called CNKM (Crime Network Key Member mining) to mine key members of a crime group, and employed time-series analysis techniques to discover the email sending and receiving rules in order to detect the abnormal communication cases. The experimental results show the efficiency and usability of the simulation email analysis system, the average simulation error is less than 10%, and demonstrate that CNKM is efficient.

Keywords data mining; personality trait; simulation; email analysis system; social network analysis

收稿日期:2006-09-26;最终修改稿收到日期:2008-09-07. 本课题得到国家自然科学基金(60773169)、国家“十一五”科技支撑计划项目基金(2006BAI05A01)、四川省青年软件创新工程(2007AA0032, 2007AA0028)和四川省青年科技基金(08ZG026-16)资助. 乔少杰,男,1981年生,博士研究生,主要研究方向为数据挖掘、数据库与知识发现. E-mail: qiaoshaojie@cs.scu.edu.cn. 唐常杰,男,1946年生,教授,博士生导师,主要研究领域为数据挖掘、数据库与知识发现. 彭 京,男,1973年生,博士,主要研究方向为数据挖掘与自然语言处理. 刘 威,男,1975年生,硕士,主要研究方向为数据库与数据挖掘. 温粉莲,女,1982年生,硕士,主要研究方向为数据库与数据挖掘. 邱江涛,男,1972年生,博士,主要研究方向为数据库与数据挖掘.

1 引言

智能化提取和分析技术是执法机关借以捕获犯罪集团和恐怖组织的重要方法^[1]. 目前智能分析专家面临的主要问题不是所需信息的缺乏, 而是如何找到合适的工具和方法从大量变化的数据, 如邮件、电话记录、Internet 网页和金融数据中挖掘出有用的知识用于监控和破坏犯罪/恐怖组织. 数据挖掘是一种强大的工具, 借助它可以使没有接受过特殊训练的犯罪专家迅速而高效地从有限的数据中挖掘大量重要的信息^[2].

由于通过无线电或者电话联络方式极易被执法机关监听, 犯罪/恐怖组织普遍利用电子邮件传达命令和进行信息收集. 对邮件数据的研究集中在邮件内容分析上, 如敏感信息挖掘、垃圾邮件检测与过滤等. 目前尚没有借助犯罪心理学知识分析恐怖组织成员个性特征与通信行为之间的内在联系的工作. 此外, 对于犯罪/恐怖社会网络的分析主要集中在分析稳定或者变化的社会网络, 开发智能化的工具辅助分析犯罪/恐怖组织成员通信行为的研究较少^[3].

本文通过分析个性特征对通信行为模式的影响, 借助个性特征判别矩阵评价不同个性特征维对个体收发邮件的影响, 开发了一个可应用于真实环境下的基于用户个性特征的仿真邮件分析系统 MEP. 通过使用合适的通信数据流分析工具, 模拟和分析社会网络中具有不同个性特征个体的通信规律, 并检测异常通信行为, 进而达到预测犯罪/恐怖活动的目的. 此外, 本文提出了一种基于社会网络分析挖掘犯罪组织核心成员的有效方法.

2 相关工作

目前, 国外的计算机专家所作的研究工作主要集中在开发能够辅助分析犯罪/恐怖组织社会网络的技术和工具. 文献[4]提出了通过分析犯罪事件的相关报道, 将犯罪组织的社会网络可视化显示的方法. 文献[5]提出了一种通过搜集和整理有关“9.11”恐怖袭击事件主要成员之间的社会联系抽取 19 个参与劫机事件的恐怖分子的信息, 并构建恐怖网络. 上述方法在挖掘犯罪网络方面取得了可喜的成果, 但并没有揭示犯罪网络的动态变化规律.

将数据挖掘技术应用于反恐是当前的研究热点. Carley 及其同事成功地开发出了一种基于层次贝叶

斯推理构建恐怖组织网络结构的工具 NETEST^[6], 侦察人员利用该软件可以准确地预测犯罪网络的规模, 确定犯罪组织结构成员之间的关系. Xu 设计了犯罪网络知识发现系统 CrimeNet Explorer^[4], 该工具能够帮助执法机关高效准确地挖掘犯罪网络, 并对其结构可视化显示. 其中集成了多种数据挖掘技术, 包括: 聚类分析查找用相同方法作案的嫌疑犯或区分不同的犯罪组织; 信息抽取技术从警察犯罪记录中自动鉴别罪犯身份和地址; 异常点检测用于发现欺诈犯罪行为、网络入侵和其它的犯罪事件.

国内在犯罪数据挖掘方面的研究刚刚起步, 文献[7-8]中分别提出一种基于基因表达式编程(gene expression programming)和支持向量机对社团成员进行分类的算法, 用于预测潜在的恐怖分子. 此外, 设计了一种查找虚拟社团核心成员的算法.

目前对犯罪/恐怖组织的通信联络方式的研究集中在对其成员电子邮件通信内容的分析. 在与邮件挖掘相关的反犯罪/反恐怖领域中, 代表性工作包括电子邮件挖掘系统 EMT^[9]和 MET^[10]. 其主要思想是通过分析邮件通信流, 挖掘相关邮件主题和获取个人相关信息, 为执法机关的侦察提供决策支持. 文献[11]利用支持向量机学习法则挖掘邮件主题和敏感内容, 鉴定邮件用户的身份. 但上述系统并没有提供一个可以对犯罪组织成员之间通信规律进行模拟的环境. 由于对犯罪组织成员通信内容的获取相当困难, 文献[3]提出了一种基于个性特征的邮件模拟系统, 该系统可以不同用户的性格特征鉴定其对邮件收发的影响, 但随机性较大.

国内应用数据挖掘技术进行反恐的研究尚未深入展开, 笔者对在实验环境下模拟仿真邮件的通信进行了探索. 与已有方法的不同在于, 本文从犯罪心理学^[12]角度分析了个性特征对通信行为的影响, 提出了用个性特征判别矩阵计算个性特征维度权重的方法. 实验结果证明该仿真邮件分析系统模拟的实验结果与真实环境相符, 准确性较高. 此外, 提出了一种利用中心性挖掘犯罪社会网络核心成员的新方法.

3 基于个性特征的仿真邮件分析系统

开发仿真邮件分析系统基于以下 3 点考虑: (1) 从用户中获取含有隐私信息的邮件数据是相当困难的; (2) 如果没有足够的数据量, 挖掘出的犯罪组织社会网络结构不够准确, 从而导致部分网络结

点的通信信息丢失;(3)使用智能技术能够更加准确地检测“异常”通信行为.基于上述原因,我们开发了仿真邮件分析系统 MEP,借助其可以根据现有的犯罪组织成员信息生成与实际情况相符的邮件使用者社会网络,并根据行为学理论和犯罪心理学知识^[12]来模拟犯罪通信行为,用于辅助执法机关进行智能化反恐.

3.1 仿真邮件分析系统模型

本文开发的仿真邮件分析系统对数据信息的处理分为两个主要步骤.

(1)信息获取及过滤,包括数据清理、ETL(数据提取、转换和装载)及过滤.数据清理和过滤阶段消除噪声和不一致数据,经过数据提取、转换与装载将邮件数据装载到数据仓库中;

(2)数据挖掘与知识发现,包括参数设置、社会网络生成器、时间序列预测器、异常通信行为检测器及数据挖掘工具等功能模块.

用户的行为特征可以通过社会网络生成器、时间序列预测器、异常通信行为检测器获取,并可以对异常事件进行报警处理.通过系统提供的数据挖掘工具可以挖掘用户感兴趣的邮件通信行为模式.

基于个性特征的仿真邮件分析系统模型由邮件客户端和个性特征模型两部分组成.其中邮件客户端属性包括邮件地址、邮件服务器名称、社会联系列表,个性特征模型由用户的个性特征维度及每一维对应的取值构成.

通过对犯罪心理学的研究和咨询犯罪心理学专家,我们将仿真邮件系统的个性特征模型用 6 个特征维描述,即乐群性、稳定性、兴奋性、有恒性、敏感性及独立性.每一个个性特征维用来刻划每个个体特殊层面的性格特征,揭示了内在的性格特征与其外在行为之间的联系.例如,高乐群性的人比低乐群性的人表现的外向、热情、善于与人打交道、适应环境能力较强,但容易冲动.高稳定性的人比低稳定性的人表现的情绪稳定,具有更好的心理素质,能很好地面对现实.有恒性高的人,做事持久、有恒心即责任感,相反有恒性低的人表现为做事缺乏恒心和毅力.高敏感性主要表现为对周围事物敏感,易受周围人和事的影响.敏感性低的人,做事理智,重事实,他们能够从客观的角度处理遇到的问题和困难.

为了便于描述邮件用户的个性特征,下面根据实际情况给出本文所用到的几个重要概念.

例 1. 将用户个性属性乐群性、稳定性、兴奋性、有恒性、敏感性及独立性进行量化.它描述了不同个性特征维对个体行为的影响程度.接近 1 的值

表明个性特征维对外在行为影响很强烈,接近 0 的值表明个性特征维对外在行为影响很小.例如,一个人个性特征矢量的取值为 $T=(0.8, 0.5, 0.4, 0.7, 0.0, 0.5)$,即乐群性为 0.8,稳定性为 0.5,兴奋性为 0.4,有恒性为 0.7,敏感性为 0,独立性为 0.5,说明这个人善于与人打交道,遇事比较沉着,做事理智,具有独立性.

将例 1 的观察形式化,引入个性特征矢量概念.

定义 1(个性特征矢量). 个性特征矢量 T 是一个用于描述个体性格特征和行为倾向的一维属性元组, $T=(t_1, \cdots, t_i, \cdots, t_n)$, 其中 $t_i \in [0, 1]$ 表示第 i 个个性特征的取值.

为了更好地度量个性特征维度,根据文献^[3],每个个性维度的取值介于 0 和 1 之间.

通过对每个个性特征维赋予不同的值,可以在邮件分析系统模型中建立代表不同个体的唯一行为模式,为每个个体在不同个性特征维设置不同的值区别于其他人.因为不同的个性特征维对用户收发邮件的影响程度不同,即其所占的权重比例不同,这里引入个性特征判别矩阵的定义,用于计算每个个性特征维度的权重.

定义 2(个性特征判别矩阵). 设 $T=(t_1, t_2, \cdots, t_n)$ 是已知的 n 维个性特征属性矢量, $M=\{m_{i,j} | (i,j=1,2,\cdots,n)\}$ 表示一个 $n \times n$ 的矩阵,其中 $m_{i,j}$ 表示属性 t_i 相对于属性 t_j 的重要程度,将满足表 1 所示判别表的 M 称为个性特征判别矩阵^[13].

表 1 个性特征判别矩阵中各元素的确定		
m_{ij}	两个个性特征维相比	解释
1	同等重要	i 和 j 同样重要
3	稍微重要	i 比 j 略微重要
5	明显重要	i 比 j 重要
7	重要得多	i 比 j 明显重要
9	极端重要	i 和 j 绝对重要
2, 4, 6, 8	介于两相邻重要程度之间	
以上各数倒数	两目标反过来比较	

例 2. 已知邮件系统所采用的个性特征属性分别为乐群性、稳定性、兴奋性、有恒性、敏感性及独立性,通过多方讨论及专家咨询得到本文采用的由上述 6 个属性构成的个性特征判别矩阵.

表 2 个性特征判别矩阵及权重							
M	t_1	t_2	t_3	t_4	t_5	t_6	W
t_1	1	2	3	4	5	5	0.36
t_2	1/2	1	2	4	4	6	0.26
t_3	1/3	1/3	1	3	3	4	0.15
t_4	1/4	1/4	1/3	1	3	4	0.11
t_5	1/5	1/4	1/3	1/3	1	5	0.08
t_6	1/5	1/6	1/4	1/4	1/5	1	0.04

表 2 中,每个具体个性特征维度的权重是我们所关心的不同个性特征对用户收发邮件的影响程度指标,其具体计算方法请参见文献[13].

通过对邮件进行分析和咨询犯罪心里学专家可以发现:乐群性决定个体发信行为;稳定性和兴奋性描述在回复信件时表现的行为特征,影响邮件客户端发送邮件的频率和回复接收邮件的速率;有恒性用于刻画用户收到邮件时有多大可能会回;敏感性将会影响用户发送邮件的延迟变化.

3.2 模拟邮件通信的正态分布算法

本文利用文献[3]提出的已被学术界广泛采用的正态分布模型,结合文献[14]设计的仿真邮件系统生成器,计算每个邮件客户端收发邮件的时延,进而统计在一段时间内用户收发邮件的数量.该正态分布模型受邮件行为模式中个性特征维取值的约束.由前面讨论我们知道:发送时延受乐群性和敏感性的约束;回复时延受稳定性、兴奋性和有恒性的约束.仿真邮件系统模拟邮件的发送时延和回复时延均符合正态分布规律,其分布曲线参见文献[14].

计算发送和回复时延时,时间间隔设置为一周,因为一般超过一周时间大多数用户都会对邮件的收发产生遗忘^[3].本文对文献[3]使用的正态分布模型进行改进,将每个用户个性特征维度的影响赋予不同的权重,而不仅仅采用文中提到的利用外向度计算发送时延,利用责任度计算回复时延.

发送时延正态分布模型采用正态分布函数 $N(Send, \delta)$ 表示^[14],其中 $Send$ 表示发送均值, δ 表示标准偏差.正态分布的均值计算如下:

$$Send = N_E - D_{per} \times (N_E - N_S) \quad (1)$$

$$D_{per} = \sum_{i=1}^6 \omega_i \times D_i \quad (2)$$

其中, $N_E = 1$, $N_E = 7$ 表示一周的开始和结束时间, D_{per} 为个性特征维度在 $[0, 1]$ 之间的取值.由于每个个性特征维对用户收发邮件均产生影响,所以利用式(2)计算 D_{per} ,其中 D_i 表示乐群性、稳定性、兴奋性、有恒性、敏感性及独立性的取值, ω_i 为利用个性特征判别矩阵计算得到的每一维个性特征的权重,

其中 $\sum_{i=1}^6 \omega_i = 1$.

标准偏差 δ 由以下公式计算得到

$$\delta_{Max} = (C_{Max} - C_{Min}) / 2 - \delta_{Gap} \quad (3)$$

$$\delta = \delta_{Max} - D_{per} \times (\delta_{Max} - \delta_{Min}) \quad (4)$$

其中 δ_{Max} , δ_{Min} 表示标准偏差中的最大和最小值, $C_{Max} = 0$ 和 $C_{Min} = 7$ 为一周中天数间隔的最大和最

小值, δ_{Gap} 为标准差 δ 与 7 天分布拐点之间距离的宽限值.

这里发送时延取概率分布密度函数为 0.7 (该值可以根据用户的需要做相应的调整,但必须满足概率密度值大于 0.5) 的值 T_s ,表示前一封邮件与下一封邮件之间的发送时间间隔,且 $T_s \in (0, 7)$.

回复正态分布模型较发送模型稍复杂,因为其标准偏差是在发送时延正态分布的标准偏差的基础上计算得到的.其用正态分布函数 $N(Receive, \delta)$ 表示,其中 $Receive$ 为回复均值,由式(5)~式(7)计算得到; δ 为标准偏差,与发送时延方差计算方法相同^[14].

$$N_M = N_S / 2 + N_E / 2 \quad (5)$$

$$M_{Dep} = \begin{cases} \frac{D_{per} - 0.5}{0.5} \times \frac{N_M - Send}{N_M - N_S}, & Send < N_M \\ -\frac{D_{per} - 0.5}{0.5} \times \frac{Send - N_M}{N_E - N_M}, & Send \geq N_M \end{cases} \quad (6)$$

$$Receive = Send + M_{Dep} \quad (7)$$

其中, N_M 为 N_S 和 N_E 的中值, M_{Dep} 表示个性特征维在正态分布均值上的附加因子, $Receive$ 在发送时延 $Send$ 的基础上加上了个性特征维的附加因子.

正态分布模型中回复时延 T_R 的计算方法与发送时延计算方法类似,根据正态分布函数计算其概率分布密度函数为 0.7 的值 T_R ,且 $T_R \in (0, 7)$.

4 社会网络分析挖掘犯罪组织核心

社会网络分析(Social Network Analysis, SNA)在社会学研究中用来分析社会中个体之间的关系和交互模式,目的是发现和理解社会结构^[4,15].已经有大量的 SNA 方法被用在研究组织行为、组织内部结构关系、文献引用、计算机通信以及其它领域中.

本文提出一种新的用于度量社会网络中结点之间关系的标准——中心性,用于寻找犯罪网络的核心成员,达到辅助反恐部门打击恐怖犯罪活动的目的.为了便于理解,这里给出犯罪网络的定义.

定义 3(犯罪网络). 犯罪网络定义为一个 5 元组 $CN = (V, E, N, \mathbf{M}, a)$,其中 $V = \{v_i | i \in N\}$,表示结点序号是顶点集合, $E = \{(v_i, v_j) | i, j \in N\}$ 是网络中边的集合, N 表示网络中结点个数, \mathbf{M} 为用于表示结点间关系的邻接矩阵, a 表示社会网络的核心结点.

例 3. 假设存在一个犯罪网络 CN , 结点集合

$V = \{v_1, v_2, v_3, v_4\}$ 含有 4 个结点, M 为社会网络结点间的邻接矩阵, 如式(8)所示.

$$M = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \quad (8)$$

其中, $M[v_i][v_j]$ 表示结点 v_i 与结点 v_j 之间邻接关系, $M[v_i][v_j] = 1$ 表示结点相邻, 否则结点不相邻; 边集合 $E = \{(v_i, v_j) | M[v_i][v_j] = 1, \text{其中 } i \neq j\}$; 因为每个结点均与 v_3 相连, 所以 $a = \{v_3\}$.

为了度量犯罪网络每个结点的重要性, 参照文献[16]给出 3 种度量方法及中心性的定义.

定义 4(中心性). 已知犯罪网络 CN 和结点 h ,

(1) 与 h 直接相连的结点数目称为 h 的联系度. 由式(9)定义的 $D_d(h)$ 称为 h 的相对联系度, $D_d(h) \in [0, 1]$, 其中 $a(i, h)$ 是一个布尔型变量, 如果为 1 说明结点 i 和 h 直接相连, 如果为 0 说明不相连, n 表示结点数目.

$$D_d(h) = \frac{\sum_{i=1}^n a(i, h)}{\sum_{h=1}^n \sum_{i=1}^n a(i, h)} \quad (9)$$

(2) h 到其它结点的最短路径数称为结点 h 的居间度. 由式(10)定义的 $D_b(h)$ 为结点 h 的相对居间度, $D_b(h) \in [0, 1]$. 其中 $g(i, j)$ 是一个布尔型变量, 表示结点 i 和 j 之间的最短路径是否通过结点 h , 通过 h , 则为 1; 否则为 0.

$$D_b = \frac{\sum_i \sum_j g_{ij}(h)}{\sum_{h=1}^n \sum_i \sum_j g_{ij}(h)} \quad (10)$$

(3) h 和网络中所有结点之间的最短路径之和称为 h 的紧密度, 由式(11)定义的 $D_c(h)$ 称为结点 h 的相对紧密度, $D_c(h) \in [0, 1]$. 其中 $s(i, h)$ 为结点 i 和 h 之间最短路径长度.

$$D_c = \frac{\sum_{i=1}^n s(i, h)}{\sum_{h=1}^n \sum_{i=1}^n s(i, h)} \quad (11)$$

(4) $C = (D_d(h), D_b(h), D_c(h))$ 称为 h 的中心性.

参照文献[6]对上述定义直观地解释, 联系度反映了结点在网络中的活跃程度. 一个结点的联系度越高, 与该结点相联系的结点数目越多, 表示该结点越有可能操控网络中的其它结点. 居间度用于衡量

某个特殊结点影响其它结点间交互关系的程度. 一个结点的中介度越大, 说明网络中的结点越有可能通过该结点与其它结点联系. 紧密度反映了一个结点到达其它结点的速率. 一个结点的紧密度越小, 说明该结点到其它结点距离越短, 其越可能是结点间通信的必经之路. 我们用中心性集成了上述三者, 刻画某个结点在社会网络中的地位.

算法 1. 犯罪网络核心挖掘算法——CNKM.

输入: 犯罪网络 CN

输出: 核心结点 a

```

1. for ( $CN$  中的每个结点  $i$ ) do
2.    $D_d(i) = CN[i].degree$ ; // 计算结点  $i$  的联系度
3.    $D_b(i) = CN[i].betweenness$ ; // 计算  $i$  的居间度
4.    $D_c(i) = CN[i].closeness$ ; // 计算  $i$  的紧密度
5.    $score = D_d(i) + D_b(i) - D_c(i)$ ;
6.   if  $score \geq max\_score$  then
7.      $max\_score = score$ ;
8.      $a = i$ ;
9.   end if
10. end for
11. return  $a$ ;
```

命题 1. 设社会网络结点数目为 n , 则算法 1 的时间复杂度为 $O(n^3)$.

证明. 算法计算网络中每对结点的最短路径采用的是 Dijkstra 算法, 该算法的时间复杂度为 $O(n^2)$. 经过 n 次循环可以找到其中得分最高的结点, 即犯罪网络的核心结点, 所以该算法的时间复杂度为 $O(n^3)$.

5 实验

一个基于个性特征的仿真邮件分析系统 MEP 在本文所述方法基础上实现, 该模型在 Borland Delphi 7 平台下开发. 实验环境为 Pentium IV 1.0GHz 处理器, 256MB 内存, 运行在 Windows XP Professional 的 PC 机环境下. 为了验证 MEP 系统的有效性, 本文利用某单位邮件服务器获取的邮件通信记录分析成员间的邮件通信关系进而建立社会网络, 采用抽样和问卷的方式获取部分用户在本文提出的 6 个个性特征维度上的取值, 然后利用 MEP 系统模拟这些邮件用户的通信行为, 最后将实验结果与真实数据比较, 验证系统的有效性.

5.1 实验数据

实验中, 仿真邮件系统 MEP 设置了 19 个用户及与其对应的 19 种不同个性特征模式. 每个邮件客

户端用“EM<标号>@edu.cn”标识,其中个性特征模式用“PersonalityMode<标号>”表示,<标号>用不同的数字或字母区分.图1显示了每个邮件用户之间的社会联系.

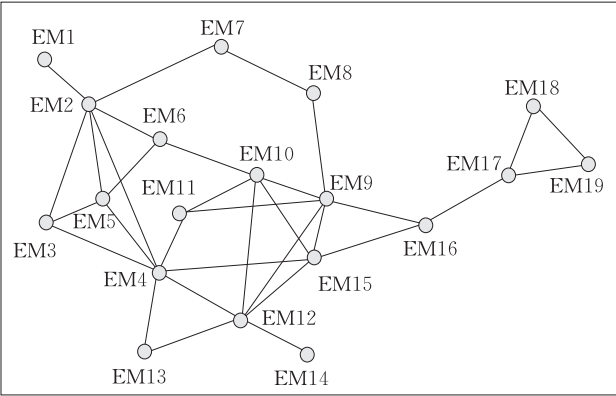


图1 仿真邮件系统客户端社会网络图

5.2 验证仿真邮件分析系统的有效性

为了计算每个邮件用户收发邮件的数量,首先利用个性特征判别矩阵计算6个个性特征维度权重的取值,结果为 $W=(0.36,0.26,0.15,0.11,0.08,0.04)$ (参见表2).为了验证本文提出的个性特征判别矩阵的有效性,我们随机生成了一组权重矢量,即 $W'=(0.3,0.25,0.25,0,0.15,0.05)$.然后利用第3节提出的计算邮件发送和回复时延的方法计算每个邮件用户100天内收发邮件的数量,最后与真实数据比较得到如图2和图3所示的实验结果.其中,横坐标表示不同邮件用户,纵坐标分别表示发送和接收邮件数量.

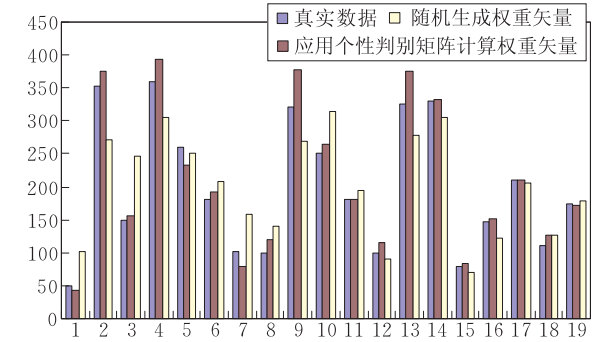


图2 仿真邮件分析系统100天内客户端发送邮件数量

通过分析可以发现利用本文提出的个性特征判别矩阵计算得到的结果与真实数据极其相符,收发邮件的平均误差均小于10%,而随机生成的权重矢量得到的结果与真实数据相差较大,收发邮件的平均误差分别为23.9%和22.3%.

为了证明 MEP 系统的准确性和可用性,我们进一步对真实邮件数据进行模拟,实验数据集来源

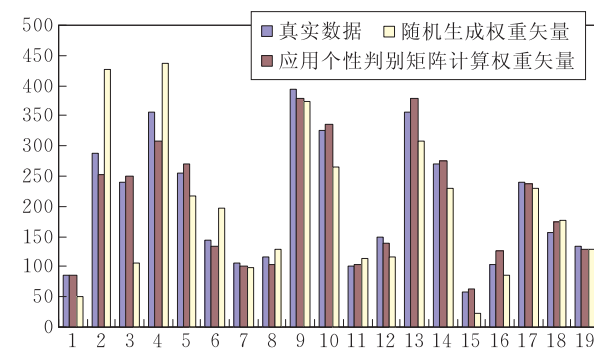


图3 仿真邮件分析系统100天内客户端接收邮件数量

于某单位邮件服务器部分用户一周内收发邮件的数据,采用网上调查问卷的方式获取用户在6个个性特征维度上的取值.我们对每组实验得到的收发邮件的准确率取平均值,实验结果如图4所示.其中,横坐标表示结点的个数,纵坐标表示预测准确率.

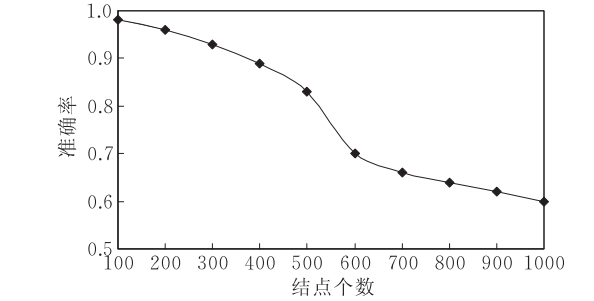


图4 MEP模拟收发邮件的准确性

图4表明,随着网络结点数目的增加,MEP系统模拟收发邮件的准确率虽然有所下降,但是仍然比较高,原因在于本文提出的利用个性特征判别矩阵计算个性特征矢量维度权重的方法能够比较客观地反映用户收发邮件的规律.当网络中结点数目超过500时,准确率下降较快,主要是因为用户之间关系错综复杂及邮件规模庞大,但准确率一般在60%以上.

5.3 时序分析及异常通信行为检测

时间序列分析和异常通信行为检测两大功能模块主要用于分析每个邮件客户端在一段时间内产生的邮件通信量,监测该仿真邮件系统客户端收发邮件的数据流,一旦检测到异常的通信行为,异常点检测器就会启用报警系统将异常的通信行为发送给邮件服务器,可以辅助执法机关进行决策分析.

本文借助时序分析软件 TimeSearcher 2 图形化分析19个邮件用户在100天内进行邮件通信的数据流量.

通过观察图1中19个不同用户的个性特征及其在仿真过程中收发邮件的数量,我们发现个性特

征维中乐群性对邮件发信行为影响最大,而稳定性和有恒性主要影响回信行为.这一结论恰好与 3.1 节给出的经验结论相吻合.例如:已知 EM14 用户的乐群性为 0.928,稳定性为 0.428,兴奋性为 0.811,有恒性为 0.033,敏感性为 0.279,独立性为 0.193,而且社会网络中只有 EM12 与其相连,恰好解释了其每周只回复少量邮件并在某一周突然中止回复.

借助时序分析图也能够很好地检测到邮件通信

中的异常通信行为.例如,人为插入一组数据,模拟 EM10 和 EM15 策划恐怖活动.从时序图 5 中可以观察到,EM10 和 EM15 在第 8 周都出现了异常的通信量,即通信量突然增加,这与 EM10 和 EM15 分别对应的行为模式 PersonalityModeJ 和 PersonalityModeO 产生的通信量不符.然而,从社会网络图中可以发现两者联系紧密,因此可以推测 EM10 和 EM15 可能在策划恐怖行动.

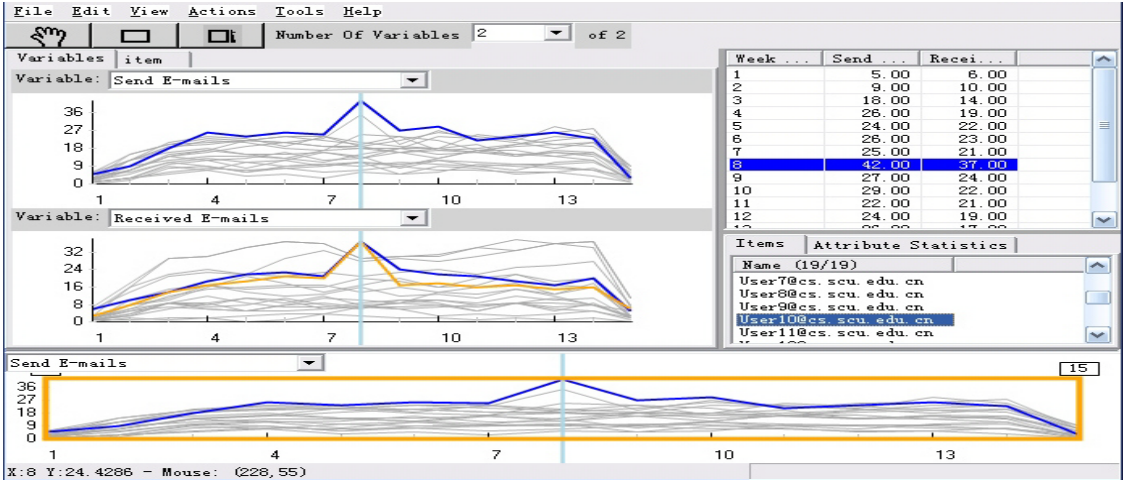


图 5 User10 的每周邮件通信时序图

5.4 利用 CNKM 算法挖掘犯罪网络核心

本实验利用图 1 所示的仿真邮件系统客户端社会网络图计算每个结点的中心性,得到如表 3 所示的结果.利用 CNKM 算法得到 EM2 是该网络的核心结点.而在真实环境中 EM2 恰好是该邮件服务器管理员邮箱对应的结点,该用户确实作为整个网络的核心通过邮件服务器进行信息的交流与传达.

表 3 社会网络结点中心性各个度量标准计算结果

邮件地址	联系度	居间度	紧密度
EM1@edu.cn	0.0151	0.000	0.0585
EM2@edu.cn	0.0909	0.131	0.0442
EM3@edu.cn	0.0455	0.000	0.0471
EM4@edu.cn	0.1060	0.119	0.0414
EM5@edu.cn	0.0606	0.000	0.0485
EM6@edu.cn	0.0455	0.0617	0.0585
EM7@edu.cn	0.0303	0.0329	0.0556
EM8@edu.cn	0.0303	0.0247	0.0656
EM9@edu.cn	0.0909	0.165	0.0528
EM10@edu.cn	0.0909	0.0741	0.0499
EM11@edu.cn	0.0455	0.0247	0.0571
EM12@edu.cn	0.0303	0.00	0.0528
EM13@edu.cn	0.0606	0.0576	0.0528
EM14@edu.cn	0.0606	0.132	0.0571
EM15@edu.cn	0.0303	0.00	0.0528
EM16@edu.cn	0.0606	0.111	0.0599
EM17@edu.cn	0.0455	0.0658	0.0642
EM18@edu.cn	0.0303	0.00	0.0414
EM19@edu.cn	0.0303	0.00	0.0399

为了更好地验证 CNKM 算法的有效性,我们采用“9.11”恐怖分子的网络结构作为算法的输入数据,该数据来源于文献[5].

通过实验我们发现在劫机事件中每个恐怖分子扮演着不同角色,他们之间的联系主要是通过 Nawaf Alhazmi 进行的,因此其在网络中的中心度最高,是网络的核心成员,通过他可以获得更多有用的信息.要破坏整个恐怖组织可以从核心成员开始,这样可以更容易地瓦解整个网络.

5.5 验证 CNKM 算法的有效性

本实验的目的是为了证明当犯罪网络中用户数量增大到海量时,CNKM 算法仍然有效.实验数据集来源于某单位近两年邮件服务器获得的真实数据,总数据量为 3700000 条.实验中将准确率作为算法性能的衡量标准,定义为

Precision =
$$\frac{\text{系统查找的核心结点数}}{\text{正确的核心结点数}}$$
(12)

利用式(12)对不同规模的网络应用 CNKM 算法得到如表 4 所示的实验结果.

通过表 4 可以发现,随着网络规模的增大,CNKM 算法的准确率有所下降,但是保持在一定水平,算法的平均准确率为 83.9%,进而证明了 CNKM 算法查找网络核心结点的准确性较高.

表 4 不同网络规模下算法准确率比较

结点数	边数	准确率/%
10	19	100
20	35	100
50	77	100
100	195	83.3
200	354	80
500	1053	63.3
1000	1560	60.8

为了分析 CNKM 算法中网络规模对运行时间的影响,即观察不同结点数目下算法的运行时间.实验结果如图 6 所示,其中网络中结点的个数分别为 10, 20, 50, 100, 200, 500, 1000. 通过图 6 可以发现,随着网络中结点数目的增加,算法运行时间近似呈线型增长的趋势,进而证明了该算法的有效性.

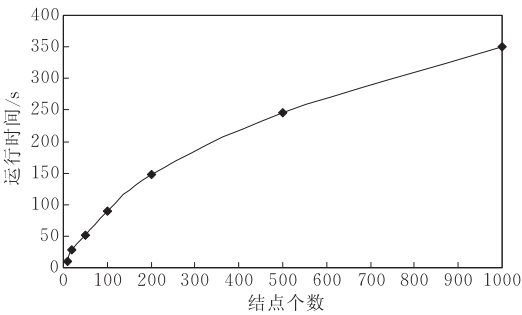


图 6 结点数与运行时间关系

6 结论及未来工作

本文提出了一种基于个性特征的仿真邮件分析系统,并且利用该系统成功地模拟了社会网络中具有不同个性实体的通信行为,对于模拟犯罪组织网络成员间的通信具有指导意义.通过将时间序列分析工具与该系统集成起来,可以有效地检测异常的通信事件,起到预警的作用.在该系统基础上,本文提出一种基于社会网络分析挖掘犯罪组织核心的算法,可以为执法机关提供决策支持.

未来的工作包括:

- (1)使用其它数据挖掘及人工智能方法,如决策树、神经网络、遗传算法和关联规则挖掘等方法对邮件通信情况进行分析,挖掘犯罪组织的活动规律;
- (2)进一步完善 MEP 系统,增加更多的个性特征,模拟更加复杂的通信行为;
- (3)在该仿真邮件系统的基础上应用层次聚类方法挖掘犯罪网络的核心成员;
- (4)设计其它计算个性特征判别矩阵的方法,

如基于粗糙集理论的方法.

参 考 文 献

[1] Berkowitz B D, Goodman A E. Best Truth: Intelligence in the Information Age. New Haven: Yale University Press, 2000

[2] Fayyad U M, Uthurusamy R. Evolving data mining into solutions for insights. Communications of the ACM, 2002, 45 (8): 28-31

[3] Lim M J, Negnevitsky M, Hartnett J. Personality trait based simulation model of the e-mail system. International Journal of Network Security, 2006, 3(2): 164-182

[4] Xu J J, Chen Hsinchun. CrimeNet Explorer: A framework for criminal network knowledge discovery. ACM Transactions on Information Systems, 2005, 23(2): 202-226

[5] Krebs V E. Mapping networks of terrorist cells. Connections, 2002, 24(3): 43-52

[6] Dombroski Matthew J, Carley Kathleen M. NETEST: Estimating a terrorist network's structure. Computational and Mathematical Organization Theory, 2002, 8(3): 235-241

[7] Qiao Shao-Jie, Tang Chang-Jie, Peng Jing, Fan Hong-Jian, Xiang Yong. VCCM Mining: mining virtual community core members based on gene expression programming//Proceedings of the Workshop on Intelligence and Security Informatics 2006. Singapore, 2006: 133-138

[8] Qiao Shao-Jie, Tang Chang-Jie, Yu Zhong-Hua, Wei Jian-Peng, Li Hong-Jun, Wu Luo-Bin. Mining virtual community structure based on SVM. Computer Science (Supplement A), 2005, 32(7): 208-212(in Chinese)

(乔少杰,唐常杰,于中华,韦健鹏,李红军,伍洛宾.基于属性筛选支持向量机挖掘虚拟社团结构. 计算机科学(增 A), 2005, 32(7): 208-212)

[9] Stolfo S J, Hershkop S, Wang Ke, Nimeskern O, Hu Chia-Wei. A behavior-based approach to securing email systems. Computer Networks Security, Lecture Notes in Computer Science 2776, 2003: 57-81

[10] Stolfo S J, Hershkop S, Wang Ke, Nimeskern O, Hu Chia-Wei. Behavior profiling of email. Intelligence and Security Informatics, Lecture Notes in Computer Science 2665, 2003: 74-90

[11] de Vel O, Anderson A, Corney M, Mohay G. Mining e-mail content for author identification forensics. SIGMOD Record, 2001, 30(4): 55-64

[12] Borum R. Psychology of terrorism. Tampa: University of South Florida, 2004: 22-63

[13] Huang Yue-Jun, Li Shu-Cheng. Application of analytical hierarchy process and fuzzy evaluation in recruitment. Modern Management Science, 2006, (4): 6-8(in Chinese)

(黄岳钧,李树丞.层次分析法与模糊评价在企业招聘中的应用. 现代科学管理, 2006, (4): 6-8)

[14] Liu Wei, Tang Chang-Jie, Qiao Shao-Jie, Wen Fen-Lian, Zuo Jie. A new method for crime data mining based on conceptual e-mail system. Computer Science, 2007, 34(2): 213-215(in Chinese)
(刘威,唐常杰,乔少杰,温粉莲,左劼. 基于概念邮件系统的犯罪数据挖掘新方法. 计算机科学, 2007, 34(2): 213-215)

[15] Berkowitz S D. An introduction to structural analysis; The

network approach to social research. Toronto; Butterworth, 1982

[16] Wen Feng-Lian. The research of the key techniques in crime data mining[Ph. D. dissertation]. Chengdu; Sichuan University, 2007(in Chinese)
(温粉莲. 基于犯罪数据挖掘系统的关键技术研究[博士学位论文]. 成都: 四川大学, 2007)



QIAO Shao-Jie, born in 1981, Ph. D. candidate. His research interests include data mining, database and knowledge discovery.

TANG Chang-Jie, born in 1946, professor, Ph. D. supervisor. His main research interests include data mining,

database and knowledge discovery.

PENG Jing, born in 1973, Ph. D. . His research interests include data mining and natural language processing.

LIU Wei, born in 1975, M. S. . His research interests include database and data mining.

WEN Fen-Lian, born in 1982, M. S. . Her research interests include database and data mining.

QIU Jiang-Tao, born in 1972, Ph. D. , lecturer. His research interests include database and data mining.

Background

This work is supported by the National Natural Science Foundation of China under grant No.60773169, the 11th Five Years Key Programs for Sci. & Tech. Development of China under grant No.2006BAI05A01, the Foundation of Innovation Software Engineering for Young People in Sichuan under grant Nos.2007AA0032 and 2007AA0028, and is also supported by Sichuan Youth Science and Technology Foundation under grant No.08ZG026-16. The project aims to help the law enforcement and intelligence agencies discover knowledge from crime networks in an efficient and effective manner. This project proposes a framework for crime data mining, which includes four main stages: Crime and terrorist prediction, crime and terrorist network creation, structure

analysis, and network visualization.

The authors have done research on crime and terrorist network analysis, mining key members of crime and terrorist networks and developing applications related to crime data mining. Currently, they have made some progress in analyzing the communication behavior among terrorist groups, and mining key members of crime networks by Gene Expression Programming (GEP). The existing modules introduced in this paper has been integrated into the Crime Miner system that is funded by the Foundation of Innovation Software Engineering for Young People in Sichuan, and the state-of-art work introduced in this paper is useful and practical in the crime data mining research area.