基于概念邮件系统的犯罪数据挖掘新方法*)

刘 威 唐常杰 乔少杰 温粉莲 左 劼
(四川大学计算机学院 成都 610065)

摘 要 将数据挖掘技术应用于反犯罪和反恐怖是目前各国安全部门的研究热点。目前国内在分析犯罪和恐怖团伙之间联系行为等方面的研究工作有限。本文主要做了下列探索:(1)建立了一个可用的基于邮件用户个性特征和情报属性的概念仿真邮件系统CEM(Conceptual based EMail system),模拟潜在的犯罪和恐怖组织利用电子邮件进行通信的规律;(2)利用符合个性特征和情报属性上的正态分布,模拟真实的邮件进行数据的收发;(3)使用社会网络分析和时间序列分析方法对邮件通信量进行深层次分析,挖掘有意义的邮件通信模式,进而发现异常通信行为;(4)通过实验证明CEM系统具有很好的鲁棒性和伸缩性,可以准确地模拟大量用户的邮件收发,解决了目前仿真数据不足的缺点,并用于发现不同性格特征群体收发邮件的规律。

关键词 犯罪数据挖掘,邮件,仿真,特征属性,行为模式

A New Method for Crime Data Mining Based on Conceptual E-mail System

LIU Wei TANG Chang-Jie QIAO Shao-Jie WEN Fen-Lian ZOU Jie
(School of Computer Science, Sichuan University, Chengdu 610065)

Abstract It is hot topic to apply data mining techniques to anti-criminal and anti-terrorism research in many law enforcement agencies. The main contributions of this paper include: (1) designed a conceptual e-mail system(CEM) based on personality trait dimensions and intelligence attribute of e-mail users in order to model the e-mail traffic behavior of criminal and terrorist organizations; (2) used normal distribution controlled by the personality trait dimension and intelligence attribute to generate e-mail data; (3) used social network analysis and time-series visualization to search for interesting e-mail behavioral patterns and abnormal communications; (4) demonstrated that CEM has good robust and scalability, exactly simulate the behavior in e-mail's, and solve the problem of lacking simulation model of the e-mail system.

Keywords Crime data mining, E-mail, Simulation, Personality traits, Behavior pattern

1 前言

恐怖组织越来越多地利用网络进行恐怖活动,从网络信息中提炼有利于监控或预测恐怖行动的信息是亟待解决的问题。已有的研究集中在犯罪数据挖掘和恐怖组织社会网络分析,但研究分析恐怖组织间的通信模式的工具和技术方面的工作做得很少。

犯罪数据挖掘得到了很大的关注,各种自动数据挖掘技术的研究为地方法律和国家安全机构提供了帮助[1]。恐怖组织社会网络分析领域中,主要研究集中在分析静态的恐怖组织社会网络的技术和工具上。文[2]中提出一种基于基因表达式编程和支持向量机对社团成员利用属性筛选方法进行分类的算法,并设计了一种查找虚拟社团核心成员的算法。卡内基大学的卡得在文[3]中提出了一种不同的方法,利用多代理技术模拟恐怖组织的进化去分析社会网络动态的变化。

在与邮件挖掘相关的反恐和犯罪领域中,利用邮件挖掘工具[4]分析邮件流,挖掘相关的邮件主题,获取个人相关补充信息,为法律实施的侦察和分析提供支持。通过向量机学习法则挖掘邮件主题和具体内容鉴定邮件客户的身份[5],为犯罪调查提供证据。

在文[6]中也提到了一种基于用户模式的安全邮件保护方法,通过截取海量邮件分析用户收发邮件的模式,但没从用户性格特性的角度分析对通信行为的影响;文[7]中提到了利用邮件用户性格模拟邮件交往行为,通过生活常识选择两个性格特征来推测对邮件收发影响的可能性,随机性较大。本文主要区别在于利用现实存在的恐怖组织社会网络结构,从犯罪心理学的角度分析行为属性影响通信模式,用一组加权值反复验证各个性特征维如何影响通信行为,最终得到一组最优解建立仿真邮件系统,与真实环境下邮件通信相符,准确性较高。

2 真邮件系统生成器

面向概念的邮件系统仿真模型根据邮件使用者的个性特征模拟邮件的收发,然后扩展仿真模型使其包含不同类型的恐怖行为模式,通过可能的人性理论和恐怖分子心理学仿真恐怖组织的行为。面向概念的邮件系统模型分成邮件客户端和行为模式属性两部分,模拟不同邮件客户端的交互,通过定义唯一的基于概念的性格个性去模拟不同的个体性格特征,

*)基金项目:国家自然科学基金(60473071),高等学校博士学科点专项科研基金 SRFDP(20020610007号),四川省青年软件创新工程(350号)。刘 威 硕士研究生,研究方向:数据库与知识工程;唐常杰 博士生导师,教授,研究方向:数据库与知识工程,数据挖掘;乔少杰 硕士研究生;温粉莲 硕士研究生;左 劼 博士。

每个个体生成一个唯一的邮件交易行为模式，这个模式与个体行为特征紧密联系在一起。

对犯罪心理学[8]提到的个性特征形式化有：

定义 1(个性特征元组) 个性特征元组是一个用于描述个体性格特征和行为倾向的属性元组，$f=(x_1,x_2,x_3,\cdots,x_n)$，其中 $x_i\in(0,1)$，表示第 i 个属性的取值。

例 2 个性特征元组的每个个性特征维是一个介于 0 和 1 之间的值。属性有年龄，种族，文化程度，情绪稳定度，外向度，责任度，都可以量化至某个具体值。它描述了每个个性特征维对人的行为影响深度。靠近 1.0 的值表明个性特征维对行为有很强的影响，靠近 0.0 的值表明个性特征维对行为影响很小。例如，一个人年龄为 0.5，外向度为 0.9，责任度为 0.7，说明这个人非常外向，冲动，努力，而且很可靠。相比较的，一个人年龄为 0.2，外向度为 0.2，责任度为 0.1，说明这个人非常内向，懒惰，不可靠。通过对每个个性特征维赋予不同的值，可以在邮件系统模型中建立代表不同人的唯一的行为特征，每个人在 6 个性特征维中指定不同的值区别于别人。

3 生成邮件客户间的通信行为

用文[9]正态分布模型，生成每个邮件客户端发送邮件信息的时延，以及生成收到邮件回复的时延。每个正态分布受邮件行为模式中个性特征维取值的约束。

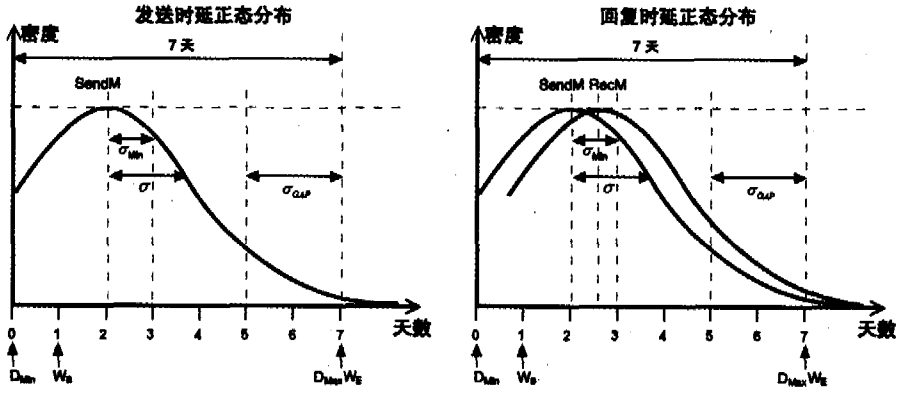

图 1 发送和回复时延正态分布

发送和回复时延的正态分布都用 7 天间隔，用来限制一周内发送和回复时延的可能值。发送时延正态分布用图 1 正态分布函数 N(SendM, σ)表示：

$$SendM = W_B - Ext\,(W_B - W_E) \tag{1}$$

$$Ext=\sum_{i=0}^{5} W_i \times D_i \tag{2}$$

(D_i 个性特征元组取值，W_i 为[0, 1]之间的随机加权值，其中 $\sum W_i=1$)

其中 W_B，W_E 为一星期的开始和结束值[$W_B=1$, $W_E=7$]，Ext 为外向度在[0,1]之间的取值。

$$\sigma_{Max}=(D_{Max}-D_{Min})/2-\sigma_{Gap} \tag{3}$$

$$\sigma=\sigma_{Max}-Es(\sigma_{Max}-\sigma_{Min}) \tag{4}$$

σMAX，σMIN 为方差中的最大和最小值，其中 D_{Max} 和 D_{Min} 为一星期中天数间隔最大值和最小值[$D_{Max}=0$, $D_{Min}=7$]，σ_{Gap} 为标准方差 σ 与 7 天分布拐点之间的取值。Es 为情绪稳定度在[0, 1]之间的取值。

为从正态分布上选择一个发送时延，从正态分布函数上先择一个任意数值 D_S，D_S 为一星期[D_{Max}, $D_{Min}=7$]中的取值，并符合发送时延正态分布条件。一旦 D_S 被选取，则 D_S 为前一封信到下一封信之间的发送时间。每发送一封信时，重新从正态分布中选择一个发送时延值 D_S，以决定下一封邮件的发送时延。

回复时延正态分布用正态分布 N(RecM, σ)表示，σ 与发送时延方差不变，RecM 由上式发送方差及个性属性的责任度值联合决定：

$$W_{Ave}=W_B/2+W_E/2 \tag{5}$$

$$M_{Dep}=\begin{cases}\dfrac{Con-0.5}{0.5}\times\dfrac{W_{Ave}-SendM}{W_{Ave}-W_B}, & m<W_{Ave}\\[2ex] -\dfrac{Con-0.5}{0.5}\times\dfrac{SendM-W_{Ave}}{W_E-W_{Ave}}, & m\geqslant W_{Ave}\end{cases} \tag{6}$$

$$RecM = SendM + M_{Dep} \tag{7}$$

其中，W_{Ave} 为 W_B 和 W_E 的平均值，Con 为责任度在[0,1]间的取值，M_{Dep} 为责任度在正态分布均值上的附加因子，RecM 为在发送时延 SendM 的基础上再加上责任度的附加因子值。

生成邮件系统数据的过程分为如下 6 步：1)模型数据生成，生成邮件系统客户端和行为模式。2) 分配行为模式给客房端，每个行为模式至少分配给一个邮件客户端。3)生成邮件客户端之间的社会联系，由发送和回复正态分布决定邮件交易数据。4)存储生成的数据，用邮件系统模拟器程序去仿真概念性的邮件系统模型。5) 装载邮件系统模型数据到模拟器程序，读出邮件系统模型的结构参数和设置。6)模拟客户端邮件的生成过程，生成邮件交易数据。

仿真器结束条件由设定的模拟天数 T 决定。当模拟结束，从每个邮件客户端邮箱得来的仿真交易数据将存储到文本文件中。这些文本数据过滤和填充到邮件数据交易库，用邮件交易分析系统来视图化。

4 实验结果与分析

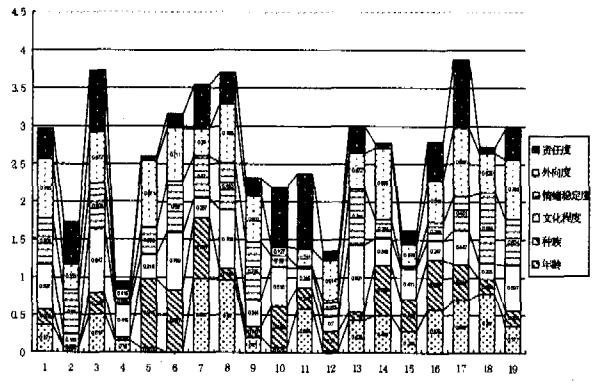

图 2 邮件用户对应的行为模式及取值

为了验证 CEM 系统的有效性，本文采用文[11]中提供的 911 事件中犯罪分子的真实数据作为实验数据来建立社会网络，用于模拟恐怖分子间邮件的收发。根据文献中提到的 19 个对象的具体行为特征，我们得到图 2 所示的结果。

4.1 行为模式权重对实验结果的影响

本实验的目的是观察随着行特征维的随机生成加权值的变化，仿真邮件系统 CEM 发送和接收邮件的变化情况，进而得到一组最优解。参数的设置和实验结果如表 1 所示。试验表明，随着加权值的变化，相应的行为属性值对邮件用户收发邮件产生了比较大的影响。加权值分别取 $W_1=0.05$，$W_2=0.05$，$W_3=0.1$，$W_4=0.2$，$W_5=0.5$ 时，得到一组最优的实验结果，该组加权值比较符合真实邮件收发情况。

表 1 通过修改行为模式权重得到仿真邮件收发结果

	$W_1=0.2$ $W_2=0.2$ $W_3=0.2$ $W_4=0.2$ $W_5=0.2$		$W_1=0$ $W2=0$ $W_3=0$ $W_4=0.2$ $W_5=0.8$		$W_1=0.1$ $W_2=0.1$ $W_3=0.1$ $W_4=0.3$ $W_5=0.4$		$W_1=0.05$ $W_2=0.05$ $W_3=0.1$ $W_4=0.2$ $W_5=0.5$	
	发	收	发	接	发	收	发	收
User1@cs. scu. edu	36	15	85	19	27	27	42	48
User2@cs. scu. edu	159	173	267	345	157	208	233	253
User3@cs. scu. edu	90	77	151	184	122	96	144	98
User4@cs. scu. edu	151	164	324	392	194	230	227	212
User5@cs. scu. edu	108	85	206	167	133	119	186	159
User6@cs. scu. edu	95	69	169	122	100	72	103	118
User7@cs. scu. edu	55	60	111	86	73	52	69	93
User8@cs. scu. edu	69	39	113	93	67	51	114	58
User9@cs. scu. edu	154	148	281	325	214	161	229	311
User10@cs. scu. edu	129	146	253	208	128	136	182	171
User11@cs. scu. edu	80	76	116	160	82	61	120	106
User12@cs. scu. edu	39	58	141	102	65	53	55	88
User13@cs. scu. edu	166	116	236	315	185	167	256	185
User14@cs. scu. edu	138	101	330	234	163	185	235	197
User15@cs. scu. edu	17	26	58	18	37	19	40	41
User16@cs. scu. edu	74	92	155	116	97	101	138	132
User17@cs. scu. edu	93	82	151	133	119	82	123	116
User18@cs. scu. edu	54	51	91	98	67	66	76	77
User19@cs. scu. edu	50	47	151	140	62	74	84	61

4.2 时序分析

邮件交易量分析系统的时间序列图程序的图形化显示采用的是 Time Searcher 2.4 软件，图 3 显示的是邮件交易数据的输出结果。

把社会网络图和时序图结合起来分析，能很好地观察到我们感兴趣的通信行为模式。通过观察分配给不同邮件客户的个性属性特征值和在整个仿真过程中发送的邮件数，我们发现权值重的属性中外向度对邮件发信行为影响最大，而情绪稳定度主要影响回信行为。通过社会网络图和图 3 发现：User15 用户的参数为年龄 0.288，种族 0.428，文化程度 0.411，情绪稳定度 0.033，外向度 0.279，责任度 0.193，而且社会联系只有 User13，这就解释了 G 每周只发出少量邮件和在第 9 周突然中止。而 User1 和 User19 都采用相同的行为模式 A，但由于各自的社会联系不同，每周收发邮件数也各不相同。这表明客户发送的邮件数不仅仅取决于邮件客户的个性属性特征值，也取决于邮件客户经常联系的人和他们如何回复邮件。

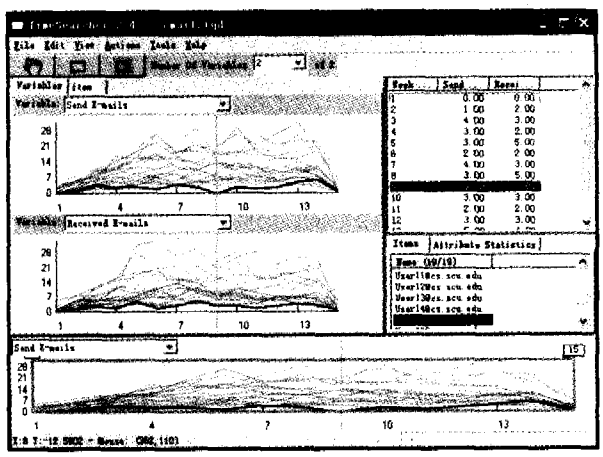

图 3 时间序列视图，时间轴为每星期的邮件收发数量

总结 本文提出了基于用户个性特征和情报属性的面向概念的仿真邮件系统 CEM，证明了邮件用户个性特征和社会网络联系共同决定用户通信行为，通过改变各个属性不同的加权值实验，找出了最符合真实情况的个性特征维。利用符合个性特征和情报属性上的正态分布计算发送时延和回复时延，模拟生成了真实的邮件收发数据。用社会网络和时间序列图分析邮件通信量，挖掘有意义的邮件通信模式，发现异常通信行为。实验表明 CEM 系统有好的鲁棒性和伸缩性，可准确模拟大量用户邮件收发。未来工作包括完善该仿真邮件系统，通过分类和预测的新方法例如决策树，挖掘有趣行为模式，并应用于实践，辅助公安部门进行智能化的反恐怖行动。

参 考 文 献

1 Chen H, Chung W, Jie J, et al. Crime Data Mining: A General Framework and Some Examples. The IEEE Computer Society, 2004

2 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，2005，32(7. 增 A)：208～212

3 Carley K M, Dombroski M, Tsvetovat M, Reminga J, Kamneva N. Destabilizing dynamic covert networks. In: 8th International Command and Control Research and Technology Symposium, National Defense War College, Washington DC, 2003

4 Stolfo S J, Hershkop S, Wang K, Nimeskern O, Hu C W. A Behavior-based Approach to Securing Email Systems. Math. Methods, Models and Architectures for Comp. Networks Security, 2003

5 de Vel O, et al. Mining E-Mail Content for Author Identification Forensics. SIGMOD Record, 2001,30(4):55～64

6 Li Y, Somayaji A. Securing Email Archives through User Modeling. ACSAC, 2005. 547～556

7 Lim M J H, Negnevitsky M, Hartnett J. Personality Trait Based Simulation Model of the E-mail System. International Journal of Network Security, 2006,3(2):164～182

8 Borum R. Psychology of terrorism. Tampa: University of South Florida,2004

9 Lantz A. Does the use of e-mail change over time. International Journal of Human-Computer Interaction, 2003,15(3):419～431

10 Aris A, Khella A, Buono P, et al. Timesearcher 2, Human Computer Interaction Laboratory, Computer Sci nce Department, 2005. http://www.cs.umd.edu/hcil/timesearcher/

11 Krebs V E. Mapping Networks of Terrorist Cells. CONNECTIONS, 2002,24(3):43～52