

基于关联规则的通信行为指纹编码研究

刘萍, 郑彦

(南京邮电大学 计算机学院, 江苏 南京 210003)

摘要:现代社会多种多样的通信方式改变了人们的生活工作方式,各种通信行为普遍存在于社会交往中。但是在目前的研究成果中还没有出现过一个通信个体的通信特征提取的方法,因此文中针对电子邮件这种通信行为提出一种编码方式,应用关联规则的相关理论和方法,深入剖析了电子邮件通信行为的特征与属性,提取每个通信实体的特征编码,并用MD5算法进行加密,最终得到电子邮件通信实体的通信行为指纹编码。得到每个通信实体的“指纹”码后能够组成“指纹”码库,能够给数据挖掘、公共安全等领域提供大量数据信息。

关键词:关联规则;通信行为;指纹;编码

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2012)07-0231-04

Research of Fingerprint Coding of Communication Behavior Based on Association Rules

LIU Ping, ZHENG Yan

(School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Variety of communication in modern society is changing people's lives, a variety of communications exist in social interaction behavior. But there is not a research of the methods of extracting the communication features. So it presents a coded way for the communication of e-mail. In view of this, using the theory of association rules and methods, analyse the characteristics of e-mail communication behavior and attributes in detail, extract the characteristics of each communication entity encoded, and encrypts with the MD5 algorithm. Ultimately get the communication behavior of e-mail communications entities fingerprint coding. These will give important information for data mining, public safety and other areas.

Key words: association rules; communication behavior; fingerprints; coding

0 引言

数据挖掘中最重要的方法是关联规则的挖掘^[1-3],关联规则主要反映了事物之间的关联性,旨在大量的数据中挖掘出关注点之间的关联性^[4,5],文中选取电子邮件这种通信行为,提取电子邮件中出现的发信方、收信方、客户端IP、服务端IP、客户端MAC、服务端MAC、加密类型七个因素作为研究对象,在发信方与其他六个因素之间建立起关联规则,计算每个关联规则的置信度与支持度,从而得到某一个发信方与其他因素的关联度,继而能得到这个发信方的通信特征。将这些特征采用编码方式进行量化从而得到唯一的“指纹”原始码字,将这些“指纹”原始码字通过MD5方式加密之后就得到最后的“指纹”编码序列。

由于每个通信实体一定时间内的通信对象等各方面都是不一样的,所以得到的码字必定是唯一的,这就体现了通信行为的指纹特性。同时MD5加密方式仍然是目前密码学界应用最为广泛的加密算法,将这种优秀的加密算法应用到文中的编码,能够保证所得到信息的安全性。

1 关联规则的确定及编码

1.1 关联规则的概念

关联规则定义:关联规则主要反应了事物之间的关联性。对反映同一事物的一条记录而言,若其具有特征属性A的同时,也具有特征属性B,则称特征属性A和B是关联的,即 $A \Rightarrow B$ 。这种关联性仅表现为“共生现象”,即两者同时存在,但并不一定表现两者之间必然存在前后因果关系。

一般关联规则都要从置信度和支持度两个方面进行研究,置信度表现了规则的强度,而支持度则表现了该项在全部数据中占的比例大小。

收稿日期:2011-11-23;修回日期:2012-02-29

基金项目:国家重点基础研究发展规划(973)课题(2006AA01Z201)

作者简介:刘萍(1986-),女,山东泰安人,硕士研究生,研究方向为数据库与知识库系统;郑彦,教授,硕士生导师,研究方向为数据挖掘、信息安全。

1.2 置信度

规则 $X \Rightarrow Y$ 在事务集中的置信度^[6] 是指支持 X 和 Y 的事务数与支持 X 的事务数之比。

$$\text{Confidence}(X \Rightarrow Y) = \frac{|\{t: X \cup Y \subseteq t, t \in T\}|}{|\{t: X \subseteq t, t \in T\}|} \quad (1)$$

式中: X, Y 都为事务项集, t 是一组数据项, $X \cup Y$ 为包含 X 和 Y 的事务, 也就是项集 X 和 Y 的并集, T 为所有用户会话事务的集合。公式(1)说明, 设 T 中支持全部的数据项集 X 的事务中, 有 $s\%$ 的事务同时也支持数据项 Y , 则 $s\%$ 称为关联规则 $X \Rightarrow Y$ 的置信度。在文中是指在每条规则事务中, 出现规则前件时规则后件也出现的概率, 是对关联规则准确度的描述。

1.3 支持度

一个关联规则是形如 $X \Rightarrow Y$ 的蕴涵式, 这里 $X \cap Y \neq \emptyset$ 。规则 $X \Rightarrow Y$ 在事务 T 中的支持度^[4] 是事务集中支持 X 和 Y 的事务数与所有事务数之比。

$$\text{Support}(X \Rightarrow Y) = \frac{|\{t: X \cup Y \subseteq t, t \in T\}|}{|T|} \quad (2)$$

式中各符号含义同式(1)。公式(2)说明, 设 T 中有 $m\%$ 的事务同时支持数据项集 X 和 Y , $m\%$ 称为关联规则 $X \Rightarrow Y$ 的支持度。支持度的含义是 X 和 Y 这两个数据项集的并集 C 在所有的事务中出现的概率的大小^[7]。即关联规则发信方 \Rightarrow 收信方中发信方与收信方出现的次数在所有的邮箱中出现的次数中所占的比例, 是对关联规则重要性的描述, 说明该规则在所有的事物中有多大的代表性。

2 电子邮件通信行为特征提取

首先提取出电子邮件中涉及到的能够体现出通信行为的特征。选取电子邮件中的几个特征因素: 发信方、收信方、客户端 IP、服务端 IP、客户端 MAC、服务端 MAC、加密类型。将这些通信行为关注点与发信方之间建立起规则。建立发信方与收信方之间的规则: 发信方 \Rightarrow 收信方, 发信方与客户端 IP 之间的规则: 发信方 \Rightarrow 客户端 IP, 发信方与服务端 IP 之间的规则: 发信方 \Rightarrow 服务端 IP, 发信方与客户端 MAC 之间的规则: 发信方 \Rightarrow 客户端 MAC, 发信方与服务端 MAC 之间的规则: 发信方 \Rightarrow 服务端 MAC, 发信方与加密类型之间的规则: 发信方 \Rightarrow 加密类型。将每条规则作为原始指纹编码的一段, 该原始指纹编码共包括六段。每一段对应一条关联规则得出的原始编码信息。

对于每段规则都求出相应的前项与后项之间的置信度与支持度, 统计某一个邮箱一年内的邮件联系人按照置信度和支持度大小逆向排列得到每条规则的数据表格。统计资料来源于某公司数据库, 得到发信方

\Rightarrow 收信方关联规则的数据如表 1 所示:

表 1 发信方 \Rightarrow 收信方关联规则数据表

发信方	对应的收信方	置信度	支持度
fleezom@163.com	huanghuang@dma800.com	50.000%	0.4500%
	13913877624@139.com	16.667%	0.3500%
	jjcat91@163.com	10.000%	0.3300%
	shuangli1112@yahoo.cn	8.000%	0.3240%
	thomson060616@163.com	7.367%	0.3221%
	fiaozhiren@yahoo.com.cn	5.667%	0.3170%
	lzw_090570124@163.com	1.667%	0.3050%
	zhangjie11100@163.com	0.333%	0.3010%
	wanghui6010616@163.com	0.200%	0.3060%
	wenjingxia@tom.com	0.100%	0.3030%

同理统计 fleezom@163.com 这个邮箱对应的客户端 IP 之间的数据得到表 2:

表 2 邮箱对应的客户端 IP 数据表

发信方	对应的客户端 IP	置信度	支持度
fleezom@163.com	10.32.20.168	57.800%	0.4735%
	10.10.126.20	19.200%	0.3576%
	10.10.126.12	5.800%	0.3174%
	10.10.224.18	3.233%	0.3097%
	10.10.224.23	2.967%	0.3089%
	10.10.102.45	2.833%	0.3085%
	196.168.0.57	2.467%	0.3074%
	172.18.226.35	0.300%	0.3009%
	10.10.213.77	0.300%	0.3009%
	192.168.83.67	0.133%	0.3004%

由于篇幅的问题文中就不一一将剩下的四个规则列出, 按照上面介绍的方式统计出剩下的四条规则的置信度和支持度, 为下面的编码提供基本数据。

3 指纹编码

3.1 原始编码方式介绍

第一步: 首先将所有的关注点数据进行编码, 采用等长编码的方式。文中为了提高数据的可靠性与唯一性, 对于每一个研究对象均采集一百万条数据。对这一百万条数据采用等长编码方式需要 20bit。每类数据用 20bit 二进制码表示之后存入到 oracle 数据库中作为基础数据使用。

第二步: 通过第二章中得到的表格数据, 将每个规则中排列前十位的数据编码采用连接的形式组合在一起。这样每条规则便得到 200bit 二进制码。

第三步: 同第二步将六条规则的二进制以连接的形式组合在一起。这样最终得到 1200bit 的原始编码数据信息。

尽管在每个分类里面都是 20bit 二进制信息, 但是由于置信度和支持度的不同, 所以对于每个研究对象

的原始编码方式必定是不同的^[8]。

3.2 通过原始编码方式得到数据结果

通过上述编码方式对统计结果进行编码可以得到
的编码数据如表3所示:

表3 原始编码数据

0000000001	000000100	000000000	000000000	000000000
0110000111	1000010010	0101011000	0010000000	0000000000
0000000000	0000000001	0000000000	0000000000	0000000001
0110000110	0001010000	0001100000	0000000001	0000000000
0000000001	0000000001	0000000001	0000000001	0000000000
0110000001	0110000111	0110000111	0110000111	0000101000
0000000000	0000000000	0000000000	0000000000	0000000000
0010010000	0000001001	0000101000	0000001000	0000000111
0000000000	0000001000	0000000000	0000000000	0000000000
0000110000	0000010000	0000000110	0000110001	0000001111
0000000000	0000000000	0000000000	0000000000	0000000000
0000011000	0000101000	0000100011	0000011000	1001000000
0000000100	0000000000	0000010000	0000000000	0000010000
0000001000	0110000000	0000100001	0000100100	0010000100
0000000000	0000100000	0000000000	0000000000	0000000000
0001010011	0010000000	0001100110	0111010111	1101000000
0000001001	0000010100	0000000000	0000000000	0000000000
0010001111	0110110100	0000000001	0001010101	0000011110
0000000001	0000000001	0000000001	0000000001	0000001100
0110000111	0110000111	0110000111	0110000111	0001010001
0000000000	0000000000	0000000000	0000000000	0000001100
0000011101	0000101111	0010011100	0001010110	0000110101
0000000000	0000000000	0000000000	0000000000	0000000000
0000001111	0000001111	0000001111	0000111000	0001001010

由上面的表格可以得到1200bit编码。由于清晰度的原因将其表示成上述图表形式。最后一步是将上面的原始编码采用MD5方式加密^[9,10]。下面介绍一下MD5算法过程。

4 MD5 加密方法

4.1 MD5 算法描述

MD5 算法具体描述如下^[11,12]:

1. 将字符串信息以512作为基数分组来处理。
2. 每一分组又被划分为16个32位子分组,然后进行一系列的处理。
3. 输出4个32为分组。
4. 将4个32位分组合级联后生成128位散列值。

4.2 MD5 算法具体实现过程

第一步:需要对信息进行填充,使其字节长度对512求余的结果等于448,因此,信息的字节长度(Bits Length)将被扩展至 $N * 512 + 448$,即 $N * 64 + 56$ 个字节(Bytes), N 为一个正整数。填充的方法是在信息的后面填充一个1和无数个0,直到满足上面的条件时才

停止用0对信息的填充^[11]。

第二步:在第一步得到的结果后面附加一个以64位二进制表示的填充前信息长度。经过前两步的处理,现在的信息字节长度 $= N * 512 + 448 + 64 = (N + 1) * 512$,即长度恰好是512的整数倍。这样做的原因是为满足后面处理中对信息长度的要求^[11]。

第三步:算法的四轮循环运算^[11]。

MD5中有4个32位被称作链接变量(Chaining Variable)的整数参数,分别为: $A = 0x01234567$, $B = 0x89abcdef$, $C = 0xfedcba98$, $D = 0x76543210$ 。设置好这四个参数之后就进入循环运算。循环的次数是要加密的信息中512位信息分组的数目。首先将4个链接变量复制到另外四个变量中: A 到 a , B 到 b , C 到 c , D 到 d 。

主循环有四轮,每轮进行相似的循环。第一轮先进行16次操作。每次操作对 a 、 b 、 c 和 d 中的其中3个作一次非线性函数运算,然后将所得结果加上第四个变量,文本的一个子分组和一个常数,再将所得结果向右环移一个不定的数,并加上 a 、 b 、 c 或 d 中之一,最后用该结果取代 a 、 b 、 c 或 d 中之一。每次操作中用到的四个非线性函数(每轮一个)分别为 $F(X, Y, Z) = X \& Y \mid \text{NOT}(X) \& Z$; $G(X, Y, Z) = X \& Z \mid Y \& \text{NOT}(Z)$; $H(X, Y, Z) = X \text{ xor } Y \text{ xor } Z$; $I(X, Y, Z) = Y \text{ xor } (X \mid \text{NOT}(Z))$,其中如果 X 、 Y 和 Z 的对应位是独立和均匀的,那么结果的每一位也应是独立和均匀的。 F 是一个逐位运算的函数。即,如果 X ,那么 Y ,否则 Z 。函数 H 是逐位奇偶操作符。

4.3 通过MD5加密算法实现过程

算法流程图如图1所示。

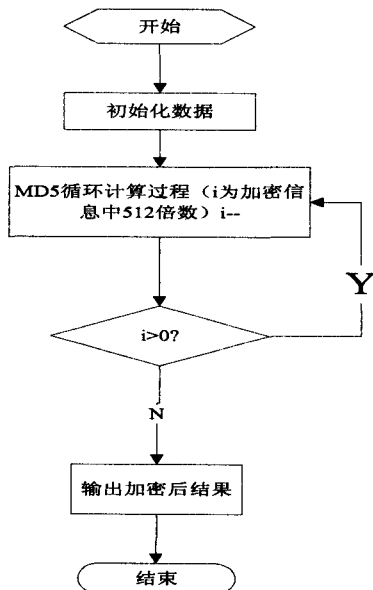


图1 MD5 算法流程图

根据上面介绍的MD5算法编写具体的实现代码,

可以求得上述编码最终得到的加密后的指纹编码为: 7a08b67df7954bfba042f2da852593a4, 将得到的指纹编码存入到数据库中, 为后面数据挖掘、公共安全等方面提供数据支持。

5 结束语

文中通过统计一个邮箱一年内发送的邮件数, 使用的客户端 IP、服务端 IP、客户端服务端 MAC 地址、服务器端 MAC 地址, 以及每封邮件所使用的加密方式, 按照排列顺序把这些使用排名前十位的数据用已经存放在数据库中的对应码字组合起来, 经过加密之后得到的数据便是这个邮箱的通信行为指纹编码, 由于设定的一百万大数据量, 以及编码的原始性避免了出现相同编码的可能性。

通过文中介绍的方法得到这些通信行为的指纹编码, 把这些编码存放到数据库里, 能够给数据挖掘、公共行为安全等领域带来大量有效数据。但同时由于文中还并未把置信度和支持度两个特性体现在编码中, 所以将关联规则的置信度和支持度加入到编码中这是文中课题下一步要研究的方向。

参考文献:

- [1] Han J W, Kamber M. Data Mining: Concepts and Techniques [M]. 2nd ed. Beijing: China Machine Press, 2007: 147-155.

(上接第 230 页)

监控平台。进行了智能终端设备的功能设计及信息管理公共服务系统的构架设计, 解决了智能集装箱监控系统在现代物流行业应用的关键技术问题, 初步形成了行业应用的智能集装箱成熟产品和技术方案。

参考文献:

- [1] Zhou Shouqin, Ling Weiqing, Peng Zhongxiao. An RFID-based remote monitoring system for enterprise internal production management[J]. The International Journal of Advanced Manufacturing Technology, 2007, 33(7-8): 837-844.
- [2] 周受钦. RFID 技术与集装箱追踪管理[J]. RFID 技术与应用, 2007(2): 42-46.
- [3] Wei G F, Yu J, Tang Z A, et al. A Novel Tank Monitor for Transporting Hazardous Chemicals[C]//The 3rd International Conference on Environmental Science and Technology. Houston, Texas, USA: [s. n.], 2007.
- [4] 周受钦, 段战归. 应用传感和网络技术实现危化品运输的智能与安全[J]. 物流技术与应用, 2007(9): 84-87.
- [5] 高飞. 基于 RFID 实时监控系统的通信数据处理方案研究[J]. 计算机技术与发展, 2007, 17(11): 96-98.

- [2] Agrawal R, Imielinski T, Swami A. Mining Association Rules-between Sets of Items in Large Databases[C]//Proceedings of the 1993 ACM SIGMOD Conference. Washington DC: [s. n.], 1993: 207-216.
- [3] Tsay Y, Chiang J. An efficient method for mining association rules[J]. Knowledge-based Systems, 2005, 18(3): 99-105.
- [4] 李杰, 徐勇, 王云峰, 等. 最简关联规则及其挖掘算法[J]. 计算机工程, 2007, 33(13): 46-48.
- [5] 郭有强. 一种高效的关联规则维护算法研究与实现[J]. 计算机技术与发展, 2007, 17(10): 123-126.
- [6] 米娜瓦尔·努拉合买提, 玛依拉·别克强塔依娃, 张太红, 等. 基于 Web 日志文件的关联规则挖掘模块的实现[J]. 计算机技术与发展, 2011, 21(9): 51-54.
- [7] 曹珍富, 薛庆水. 密码学的发展方向与最新进展[J]. 计算机教育, 2005(1): 19-21.
- [8] 徐茂智, 游林. 信息安全与密码学[M]. 北京: 清华大学出版社, 2007: 121-127.
- [9] 杜昌钰. MD5 算法的过程分析及其 C#实现[J]. 通信技术, 2008(8): 71-72.
- [10] 张裔智, 赵毅, 汤小斌. MD5 算法研究[J]. 计算机科学, 2008, 35(7): 295-297.
- [11] 王津涛, 覃尚毅, 王冬梅. 基于 MD5 的迭代冗余加密算法[J]. 计算机工程与设计, 2007, 28(1): 41-42.
- [12] 李霞. MD5 加密算法浅析及应用[J]. 运城学院学报, 2005, 23(5): 36-37.

- [6] 王浩远, 梁昌勇, 俞家文, 等. 基于 RFID 技术的汽车总装 MES 系统研究[J]. 计算机技术与发展, 2010, 20(9): 222-226.
- [7] 肖楠, 郑文岭, 马文丽, 等. 一种基于 RFID 的物流管理系统的设计[J]. 计算机技术与发展, 2008, 18(7): 237-239.
- [8] 张捍东, 朱林. 物联网中的 RFID 技术及物联网的构建[J]. 计算机技术与发展, 2011, 21(5): 56-59.
- [9] Evan W, Leilani B, Garret C, et al. Building the internet of things using RFID: the RFID ecosystem experience[J]. IEEE Internet Computing, 2009, 13(3): 48-50.
- [10] Weber R H. Internet of things - new security and privacy challenges[J]. Computer Law and Security Report, 2010, 26(1): 23-28.
- [11] Liu Hai, Miodrag B, Amiya N, et al. Taxonomy and challenges of the integration of RFID and wireless sensor networks[J]. IEEE Network, 2008, 22(6): 26-35.
- [12] Joshua C, Anne J. Challenges for database management in the internet of things[J]. IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India), 2009, 26(5): 320-324.