# 基于分段、聚类和时序关联分析的用户行为分析

常慧君，单 洪，满 毅
（电子工程学院 网络系，合肥 230037）

**摘 要**：分析用户行为对网络用户的管理控制有着重要意义。用户行为实质上是一系列的数据交换过程，最终会体现为业务流，且这些业务流在时间上表现出一定的规律性。通过研究业务流的时序关系来分析用户行为的规律，提出一种用户行为的分析方法。该方法分为三个阶段，分别基于分形模型、改进的最大距离聚类法和 Apriori 算法进行分段、聚类和时序分析，最终从用户的数据交换中获知用户的行为规律。实验表明，该方法在无法获知用户消息的具体内容的前提下，仍能较为准确地区分各类报文序列，并能有效发现用户信息发送行为的规律。

**关键词**：分段；最大距离聚类；Apriori 算法
**中图分类号**：TP393 **文献标志码**：A **文章编号**：1001-3695(2014)02-0526-06
doi:10.3969/j.issn.1001-3695.2014.02.049

# User behavior analysis based on segmentation, clustering and timing relationship analysis

CHANG Hui-jun, SHAN Hong, MAN Yi
(*Dept. of Computer Network, Electronic Engineering Institution, Hefei* 230037, *China*)

**Abstract**: Analyzing the user behavior is of great importance to the management and control of network users. User behavior, which is actually s set of data exchanges, is ultimately displayed as traffic flow, showing regularity along time. This paper explored this regularity through studying the timing relationship between traffic flows, and proposed a novel user behavior analysis method. The proposed method utilized three steps, which were based on fractal model, improved maximum distance clustering method, and Apriori algorithm for segmentation, clustering, and timing analysis, respectively, to obtain the user behavior regularity from user data exchange. Theoretic analysis and simulation results illustrate that this method can satisfactorily classify the packet sequence and discover the regularity of user transmit behavior without decrypting user information.

**Key words**: segmentation; max-distance clustering; Apriori algorithm

## 0 引言

本文提出一种可用于用户异常行为发现的业务流时序分析方法。该方法简单、易实现，且不会产生大量通信，具体实施是在网络中布置监测代理，所有的代理可以监测整个网络的信息交换，并在代理上运行如下算法：首先依据时间间隔和报文长度信息对用户的信息发送报文序列进行分段，然后依据其包长分布、时间间隔分布、段长度等统计特性对这些序列段进行聚类分析，最后用 Apriori 算法对各类序列段进行时序关联分析，时序分析的结果可以用来进行用户异常行为的分析。

用户行为分析是指用某些特征量的统计特征或特征量的关联关系定量或定性地表示网络用户行为的规律。通过掌握用户行为规律，得以控制并预测网络用户行为。用户行为分析还可用于发现异常用户行为，对网络安全管理有着重要意义。

用户行为最终通过承载用户业务的数据流来体现，而数据流的实质是一段时间序列，因此对用户行为的分析实质上是对时间序列的分析。对时间序列的分析方法主要分为三类，即基于时间序列分解的方法、基于聚类的方法和基于关联分析的方法。基于时间序列分解的用户行为分析方法主要是用来分析互联网骨干链路上的多用户查询、访问的规律性[1~3]，为网络资源分配和管理提供依据，这类研究的已知信息是在某个时间点或时间段内的流量、查询量或数据量等统计量，在分析单个用户的行为特征时不能有效发挥作用；基于聚类的方法如文献[4~6]根据用户的访问偏好等特性对用户进行分类，主要用于营销领域，也可能被用来实现针对用户的认知欺骗攻击；基于关联分析的方法如文献[7~9]分析了网络用户的访问偏好与位置和时间等因素的关系，可以用于对单个用户行为的预测。以上方法都假设网络用户的身份、地址和所发的数据内容等信息已知。当用户对自身的数据内容进行加密时，这些方法将不再适用。而为了保证通信内容的完整性和机密性，目前很多用户业务都采用了端到端加密机制[10~12]，使得在通信业务源和目的之间路径上的任何一个第三方无法解析用户的数据内容。

本文提出一种基于分段、聚类和关联分析的用户行为分析方法。首先依据时间间隔和报文长度信息对用户的信息发送报文序列进行分段，然后依据其包长分布、时间间隔分布、段长度等统计特性对这些序列段进行聚类分析，最后用 Apriori 算法[13]对各类序列段进行时序关联分析。该方法在用户对自身数据进行加密时仍能使用，且经实验检验，能够取得较好的分析效果。

**收稿日期**：2013-05-03；**修回日期**：2013-06-13

**作者简介**：常慧君(1986-)，女，博士研究生，主要研究方向为无线网络对抗技术(changhj2417@126.com)；单洪(1965-)，男，教授，博导，主要研究方向为无线网络对抗技术；满毅(1985-)，女，博士，主要研究方向为无线网络对抗技术.

## 1　问题描述与分析

不同的用户行为会产生不同类型的业务流,而根据协议规定,在一段业务流内部,其数据报文长度和时间间隔等流量特征具有区分于其他段的规律性。图 1 和 2 是从林肯实验室公开的局域网数据集[14]中选取的报文子序列。每一行的各属性依次是一个报文的序号、接收时间戳、源节点、目的节点、协议、报文长度等信息。比较两图可见,不同类型的数据流其报文长度和时间间隔有所不同。从这一点出发,本文将根据报文长度和间隔等流量特征信息,对具有不同流特性的相邻数据流进行分段。

| | | | | | |
|---|---|---|---|---|---|
| 71860 | 4668.51 | 172.16.11 | 135.8.60. | TELNET | 60 Telnet Data ... |
| 71862 | 4668.539 | 135.8.60. | 172.16.11 | TELNET | 60 Telnet Data ... |
| 71863 | 4668.54 | 172.16.11 | 135.8.60. | TELNET | 60 Telnet Data ... |
| 71865 | 4668.569 | 135.8.60. | 172.16.11 | TELNET | 60 Telnet Data ... |
| 71866 | 4668.57 | 172.16.11 | 135.8.60. | TELNET | 60 Telnet Data ... |

图1　TELNET数据流段

| | | | | | | |
|---|---|---|---|---|---|---|
| 472 | 46.12216 | 135.8.60. | 172.16.11 | SMTP | 94 | C: RCPT To:<hyacintl@pasca |
| 473 | 46.15052 | 172.16.11 | 135.8.60. | SMTP | 106 | S: 250 <hyacintl@pascal.ey |
| 474 | 46.15081 | 135.8.60. | 172.16.11 | SMTP | 92 | C: RCPT To:<selmam@pascal. |
| 475 | 46.17268 | 172.16.11 | 135.8.60. | SMTP | 104 | S: 250 <selmam@pascal.eyri |
| 476 | 46.17295 | 135.8.60. | 172.16.11 | SMTP | 93 | C: RCPT To:<yvonnea@pascal |

图2　SMTP数据流段

假设用户报文序列表示为 $P=(p_1,p_2,\cdots,p_n)$,其中 $p_i=(t_i,l_i)$分别对应序列中每个报文的包长和包截获时间。则由于用户行为在时间轴上的规律性,$P$ 应能够按照其流量特征分为 $h$ 个连续的段 $Q_1,Q_2,\cdots,Q_h$:

$$Q_1=(p_1,p_2,\cdots,p_{c_1})$$
$$Q_2=(p_{c_1+1},p_{c_1+2},\cdots,p_{c_2})$$
$$\vdots$$
$$Q_h=(p_{c_{h-1}+1},p_{c_{h-1}+2},\cdots,p_{ch}),c_h=n$$

使得每一个 $Q_j$ 都是一个相对独立的子序列,即在 $Q_j$ 内部,其相邻点的时间间隔以及每个点相应的报文长度与 $Q_{j-1}$或 $Q_{j+1}$有所区分。对 $P$ 分段的结果 $Q_1,Q_2,\cdots,Q_h$ 是时间轴上的业务流段。由于用户具体业务类型有限,特定类型的业务流在时间轴上通常会重现。因此,按照流量特性可以将 $Q_1,Q_2,\cdots,Q_h$ 分为 $k$ 类 $V_1,V_2,\cdots,V_k$,其中 $V_i$ 由流量特征相似的报文时间子序列组成,即 $V_i=\{Q_j\},Q_j=p_{j1},p_{j2},\cdots,p_{jx}$。

由网络行为决定,用户的不同业务流之间可能会存在某种时序关联。例如,在 $V_i$ 类业务流段之后总是跟着 $V_j$ 类的业务流段,则 $V_i$ 和 $V_j$ 存在时序关联。业务流之间的时序关联性体现了用户行为的规律性。

## 2　基于分段、聚类和关联分析的用户行为分析方法

### 2.1　数据流分段

现有的分段技术,又称摘要技术,是一种在保持时间序列基本特性的前提下,以减少其维数为目标的时间序列表示技术。时间序列的分段技术主要分为三类:滑动窗口法[16~18],自顶向下法[19~21]和自底向上法[22~24]。其中,滑动窗口法的基本原理是以时间序列中某个点为锚,对时间序列进行抽象或表示,直到与原时间序列的误差超过用户设定的门限值。自顶向下法是一种分而治之的方法,每一次遍历都将时间序列在最可能的点将其分割为不同的子序列,然后再对不满足条件的子序列进行分割,直到超过用户规定的门限。自底向上法则是每一次遍历都将合并代价最低的点或段合并,直到代价超过用户规定的门限。按照目的,时间序列的分段技术可以分为两类:以存储空间最小为目的[20,22,24]和以误差最小为目的[16]。这两种分段技术一般都应用在时间序列的简化存储或表示中,而本文的分段是作为聚类分析的预处理,以每段内的各点具有最大程度的相似性为目标,从本质上与两者不同,因而现有的时间序列分段技术在本文中并不适用。本文以分组长度和间隔为主要标准,提出一种简单的遍历方法对报文时间序列进行分段,将可能属于不同业务流的相邻分组分为不同的段。

1)简单分段方法

分段技术是一种在保持时间序列基本特性的前提下,以减少其维数为目标的时间序列表示技术。现有的时间序列分段技术或以存储空间最小为目的[15~17],或以误差最小为目的[18],一般都应用在时间序列的简化存储或表示中,面向的对象一般都是时间轴上的一维数据。在本文中,由于报文的长度和时间间隔都与业务类型相关,因此对时间序列进行分段时需要考虑两个属性的影响,造成这些分段方法在本文应用中的失效。

一般地,节点之间在通信状态下其报文接收时间序列的时间间隔比较稳定,不会大于一定阈值,且两个节点不会一直保持通信状态。如果两个节点之间的通信报文时间序列时间间隔超过 $T$ 秒,则认为间隔前后的时间序列分属于不同的两次业务流。因此,大于一定阈值的时间间隔可以作为一个分段的标准。而在通信密集时,两节点间的业务流之间时间间隔可能很短。在这种情况下,考虑利用报文长度和时间间隔的共同作用来分段。由于不同的业务采用的协议不同,不同业务流的报文长度很可能也有所区分。但简单地利用报文长度比值变化来进行分段(如算法 1),并不能获取良好的分段效果。例如在图 2 中,报文长度一直在变化,但其业务类型并没有变化。按照算法 1,在每个报文处都要进行分段,这种分段显然并不合理。

算法 1　简单分段方法

```
1   输入:时间序列 P=(p1,p2,…,pn)。
2   k=1;Qk.s=p1,Qk.e=pn//开始和结束点
3   for i=3:n
4     if ti-ti-1>T或max(li+1,li)/min(li+1,li)>Δ
5       k=k+1;Qk-1.e=pi-1;
6       Qk.s=pi;Qk.e=pn;i=i+1;
7     else
8       i=i+1;
9     end if
10  end for
11  输出:分段的时间序列{Qi}。
```

2)基于分形模型的分段

(1)初步分段　相邻的两个报文长度相差超过一定的阈值不代表业务类型的变化。当相邻的报文长度不相等时,报文长度的变化可用公式 $x_i=\frac{l_{i+1}-l_i}{l_i-l_{i-1}}$表示,这种计算方法避免了对图 2 中所示情况的误判。若 $x_i$ 超过一定阈值,则很可能业务类型有所变化。所以,首先经过一次遍历找出所有时间间隔大于 $T$ 和 $x_i$ 大于一定阈值 $\Delta$ 的点,对数据流进行初步划分。

业务数据流可分为控制报文流和数据报文流,表现形式可以为大段的数据报文流,数据报文中夹杂控制报文或者大段的控制报文流。其中,大段数据报文流的特征是报文长度相同,大段控制报文流则表现为不同长度报文的交互,而数据报文中夹杂控制报文则由于一般控制报文长度较短表现出报文长度的突变。前两者接近准确分形的数据,而后者由于报文长度的突变

会造成初步分段时的误分割。如图 3 所示,(1)中的方法会在第 266 个报文处进行分段,而事实上,这种分段是没有必要的。

```
264  43.9689 135.8.60.172.16.11SMTP 1514 C: DATA fragment, 1460
265 43.96971 135.8.60.172.16.11SMTP 1514 C: DATA fragment, 1460
266 43.96981 172.16.11135.8.60.TCP   60 smtp > 1027 [ACK] Seq=1
267 43.97106 135.8.60.172.16.11SMTP 1514 C: DATA fragment, 1460
268 43.97229 135.8.60.172.16.11SMTP 1514 C: DATA fragment, 1460
```

图3 误分段情况

(2)段合并 将每一段数据流中的第 $i$ 个点表示为(时间戳 $t_i$,报文长度 $l_i$),其中,$s<i<e$,$s$ 和 $e$ 分别为该段数据流的开始和结束报文序号。根据分形理论[19]和同一业务流的自相似性,若相邻的两段数据流属于同一业务流,则存在收缩映射 $M_k$,使得两段数据的并集 $P_k$ 可以映射到第一段或第二段 $P'_k$,并且使得 Error($M_k(P_k)-P'_k$)小于一定阈值。其中,$M_k$ 如式(1)所示,Error 是两序列段之间的欧氏距离。

$$M_k\begin{pmatrix} t \\ l \end{pmatrix}=\begin{pmatrix} a_{11}^k & 0 \\ a_{21}^k & a_{22}^k \end{pmatrix}\begin{pmatrix} t \\ l \end{pmatrix}+\begin{pmatrix} b_1^k \\ b_2^k \end{pmatrix} \tag{1}$$

假设原数据序列 $P_k$ 的区间为($t_s^k$,$t_e^k$),所映射到的 $P'_k$序列区间为($t'^k_s$,$t'^k_e$),则 $P_k$ 与 $P'_k$之间存在如式(2)和(3)所示的关系。令 $A_i=l_i-(\frac{t_e-t_i}{t_e-t_s}l_s+\frac{t_i-t_s}{t_e-t_s}l_e)$,$B_i=l_i-(\frac{t_e-t_i}{t_e-t_s}l'_s+\frac{t_i-t_s}{t_e-t_s}l'_e)$,则 Error 可以表示为式(4)所示。使得 Error 最小的 $a_{22}^k$ 可由对 Error 相对于 $a_{22}^k$ 的偏导求得,如式(5)所示,$a_{22}^k=\sum_{i=s}^{e}A_iB_i/\sum_{i=s}^{e}A_i^2$,$M_k$ 的其他参数可通过迭代求得。如果 $M_k$ 使得 Error 小于一定阈值,则认为该段数据序列属于同一业务流。

$$M_k\begin{pmatrix} t_s \\ l_s \end{pmatrix}=\begin{pmatrix} t'_s \\ l'_s \end{pmatrix} \tag{2}$$

$$M_k\begin{pmatrix} t_e \\ l_e \end{pmatrix}=\begin{pmatrix} t'_e \\ l'_e \end{pmatrix} \tag{3}$$

$$\text{Error}(M_k(P_k)-P'_k)=\sum_{i=s}^{e}(A_ia_{22}^k-B_i)^2 \tag{4}$$

$$\frac{\partial \text{Error}}{\partial a_{22}^k}=2\sum_{i=s}^{e}(A_ia_{22}^k-B_i)A_i \tag{5}$$

(3)分段算法与性能分析 综上所述,基于分段模型的分段方法具体操作如算法 2 所示。首先根据时间间隔和报文长度突变进行初步分段,然后基于分形模型对属于同一数据流的相邻段进行合并。

算法 2 基于分形模型的分段

```
1  输入:时间序列 P=(p1,p2,…,pn)。
2  k=1;Qk.s=p1,Qk.e=pn//开始和结束点
3  for i=3:n
4     if ti-ti-1>T
5        k=k+1;Qk-1.e=pi-1;
6        Qk.s=pi;Qk.e=pn;i=i+1;
7     else if li-li-1>1&&((|(li+1-li)/(li-li-1)|>Δ)‖(|(li-li-1)/(li+1-li)|>Δ))
8        k=k+1;Qk-1.e=pi-1;
9        Qk.s=pi;Qk.e=pn;i=i+1;
10    else
11       i=i+1;
12    end if
13 end for
14 for all Qk+1.s.t-Qk.s.t<T
15    if Error(M(Qk,Qk+1),Mk+1)<ζ
16       Qk.e=Qk+1.e,后续所有分段序号-1;
17    end if
18 end for
19 输出:分段的时间序列{Qi}。
```

假设报文总个数为 $n$,则算法的时间复杂度为 $O(n)$。

## 2.2 基于马氏距离的聚类

时间序列分段的结果是业务流段,本节将对这些业务流段进行分类。由于加密技术的采用,类别数目未知,因而考虑聚类技术。聚类是一种无监督的机器学习方法,目的是将一组样本数据按照其特征分组,使得每一组内的样本具有最大程度的相似性,而组间的区别最大。聚类方法中有三个关键点:样本属性集的选择、样本距离计算方法和聚类方法的选择。

在加密条件下,可获取到的业务流流量主要特征如表 1 所示。由于前面在分段时已将分组间隔和分组长度有显著变化的报文分为不同的段,因而这里不考虑分组长度和间隔的最大报文长度和最小报文长度参数。对每一段业务流计算其报文个数、报文间隔和报文长度特征,作为一个样本,将其输入聚类算法,则可以将这些业务流段按照其流量特征分为不同的类。

表 1 MAC 层与业务流相关的特征

| | |
|---|---|
| 数据报文<br>控制报文 | 报文个数 |
| | 报文间隔(平均、最大、最小) |
| | 报文长度(平均、最大、最小) |

考虑到这些特性的度量尺度不同,在本文中采用马氏距离来计算样本距离。马氏距离[20]是一种有效的计算两个未知样本集的相似度的方法,考虑各种特性之间的联系并且尺度无关。设 $X_i$、$X_j$ 是协方差阵为 $\Sigma$ 的样本集中抽取的两个样本,则 $X_i$、$X_j$ 两者之间的马氏距离如式(6)所示。其中,$\Sigma$ 为形如式(7)的矩阵,$\mu$ 是样本 $X$ 的所有属性的均值组成的向量,其中 $m$ 为样本属性个数,在本文中 $m=3$。$\Sigma$ 的物理意义是 $X$ 的各属性值之间的协相关系数,则经过 $\Sigma^{-1}(X_i-X_j)$ 消除了行向量之间的相关性,再用该结果右乘$(X_i-X_j)'$,消除了其列向量上各值的相关性,最终得到一个无量纲的值,消除了 $X$ 的各样本属性之间的度量尺度对不同距离值产生的影响。

$$d_m^2(X_i,X_j)=(X_i-X_j)'\Sigma^{-1}(X_i-X_j) \tag{6}$$

$$\begin{bmatrix} E((X_1-\mu_1)(X_1-\mu_1)) & E((X_1-\mu_1)(X_2-\mu_2)) & \cdots & E((X_1-\mu_1)(X_m-\mu_m)) \\ E((X_2-\mu_2)(X_1-\mu_1)) & E((X_2-\mu_2)(X_2-\mu_2)) & \cdots & E((X_2-\mu_2)(X_m-\mu_m)) \\ \vdots & \vdots & & \vdots \\ E((X_m-\mu_m)(X_1-\mu_1)) & E((X_m-\mu_m)(X_2-\mu_2)) & \cdots & E((X_m-\mu_m)(X_m-\mu_m)) \end{bmatrix} \tag{7}$$

在聚类类别数目未知时,主要有两种聚类方法[13],即自适应样本集构造法和最大距离聚类法。自适应样本集构造法的基本思想是利用距离测度建立各类聚类中心,该方法不仅对样本的紧致性要求较高,也要求异类样本之间彼此尽量远离。由于加密机制的采用,链路层不同业务流的特征之间彼此区分可能并不是很大,该方法在本文中并不适用。

最大距离聚类法则可以保证每次取到的新的聚类中心离已有的聚类中心的距离都比较远,且该算法不需要预先给出聚类个数,可以根据一定的计算规则智能地确定初始聚类中心的个数。由于后续的样本聚类都要依据其与初始聚类中心的距离,且初始聚类中心与其他样本的距离越大,聚类越准确。而在原算法中,随机选取一个样本作为初始聚类中心显然不妥。在本文中,抽取 $n$ 个样本,从中选择距离其他样本最远的样本作为初始聚类中心,保证了两个初始的聚类中心距离其他样本相对最大。改进之后的具体操作过程如下:

a)从样本集中抽出 $n$ 个样本,计算每一个样本与其他样本之间的马氏距离之和,取使得该值最大的样本作为第一个聚类中心。取第一个样本 $X_1$ 为 $Z_1$。

b)从样本集中取出距 $Z_1$ 最远者作为第二个聚类中心 $Z_2$。

c)对剩下的样本,分别计算各自到 $Z_1$ 和 $Z_2$ 的距离。

d)在每一个样本 $X_i$ 至各类 $Z_j$ 的距离值中,取其最小者,令其为 $D_{X_i}$,这样对全部的剩余样本就得到一个最小距离序列 $\{D_{X_i}\}$。

e)从其中挑出最大距离值。如果该值大于或等于 $Z_1$ 与 $Z_2$ 之间距离的 $\alpha$ 倍,则将相应的样本作为新的聚类中心 $Z_{k+1}$;如果这一条件不满足,则确定聚类中心的过程结束,转向 g)。

f)重复 d)和 e),直至找不到符合上述条件的 $X_i$。

g)把剩余样本分配到离它最近的聚类中心所属类别。

假设样本集中有 $l$ 个样本,每个样本有 $m$ 个属性,则改进的最大距离聚类方法的聚类复杂度为 $l^2m^2$。在本文中 $m=3$,所以算法时间复杂度为 $O(l^2)$。

### 2.3　基于 Apriori 的时序关联分析

经过聚类,节点之间的通信报文序列可以表示为如图 4 所示的形式。其中,字母 ABC 等分别代表一类业务流段。如果在 A 类业务流之后总是紧跟着 B 类业务流,则可以据此推断该用户的行为规律,为网络用户的行为管控提供依据。判断两个相继出现的业务流段 $Q_i$ 和 $Q_{i+1}$之间存在衔接的标准是满足 $Q_{i+1}.s-Q_i.e<t$,其中,$t$ 稍大于 $T$。业务流在时间上的这种衔接性实质上是一种关联规则。

Apriori 算法是一种关联规则分析方法,通过项集的支持度和置信度来计算频繁项集,从而产生强关联规则。Apriori 算法一般用来分析样本属性之间的关联。

在本文中,业务类型可看做 Apriori 关联规则分析中的属性,两类业务相继出现可认为是两个属性值同时出现,相应地,两类业务流衔接出现的次数可认为是关联规则发现中的项集支持度,项集支持度大于一定阈值的则认为是频繁项集。本文的业务流时序关联分析即是要发现用户业务在时间上的关联性。

综上所述,发现业务流之间时序关联的方法如算法 3 所示。经过一次遍历,统计各类分段之后出现其他类型分段的次数,然后根据这些次数用 Apriori 算法计算各类型时序关系的支持度和置信度。置信度满足一定阈值的两类分段被认为有时序关联。例如,对分段类型 A 和 B,A 之后 B 出现的置信度为 support(A,B)/support(A),其中 support(A)表示在整个分段序列中 A 出现的次数,而 support(A, B)表示在分段序列中 B 类分段随 A 类分段出现的次数。同理,support(A,B,C)/support(A,B)表示在(A,B)的组合流之后出现 C 类业务流的概率。

算法 3　业务流时序关联分析

1　输入:分段的时间序列$\{Q_i\}$和聚类结果$\{V_j\}$。
2　for $i=1:(|Q|-1)$
3　　$j=\{j|Q_i\in V_j\}$,support$(j)$ ++;
4　　if $Q_{i+1}.s-Q_i.e<t$//衔接
5　　　$k=\{k|Q_{i+1}\in V_k\}$,support$(j,k)$ ++;
6　　end if
7　end for
8　for $j,k=1:|V|$
9　　$M(j,k)=\dfrac{\text{support}(j,k)}{\text{support}(j)}$//时序关联矩阵
10　end for
11　输出:业务流之间的时序关联矩阵 $M(j,k)$。

经过以上分段、聚类和时序关联分析过程,则可以掌握不同类型的业务流之间存在的时序关联,为研究用户的行为规律提供依据。

## 3　仿真实验与结果分析

麻省理工学院林肯实验室搭建了一个局域网[14],对该局域网进行了时间跨度为两周的数据收集工作,生成的数据集主要用于入侵检测,同样也可以用于用户行为分析。该数据集包括网络中每个节点上信息发送的源、目的节点、信息截获时间、协议等信息。选定节点 172.16.114.50 的报文序列为输入,在 MATLAB 平台上运行本文提出的算法。

判断分段准确性的方法是将分段的结果与通信节点上记录的业务类型相比对,计算将同一次业务交互所产生的通信报文分为同一个段的百分比。其中,建立控制连接所产生的控制报文段也作为一类特殊的业务类型。

判断聚类准确性的方法是将聚类的结果与通信节点上记录的业务类型相比对,计算将同一类型的业务流聚为一类的百分比。

判断时序分析准确性的方法则是与林肯实验室对该节点上信息序列的介绍相比对,如果本文分析的结果符合网站的介绍,则认为时序分析结果可靠。

仿真实验分为三部分,分别检验本文提出的分段、聚类和时序分析方法。

### 3.1　基于分形模型的分段

分段是决定聚类和业务时序分析性能的重要步骤,其性能决定着整个算法的性能。首先通过与简单分段方法的性能对比来说明基于分形模型的优越性。图 5 是 $\Delta=6$ 时算法 1 的分段准确率 $P$ 随 $T$ 的变化情况,图 6 是在 $T=10$ 时算法性能随 $\Delta$ 的变化情况。由图可见,算法的分段准确率最高只能达到将近 0.63,其分段性能并不好。

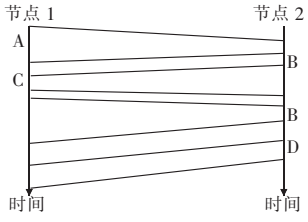

图 4　节点之间的业务流交互示意图

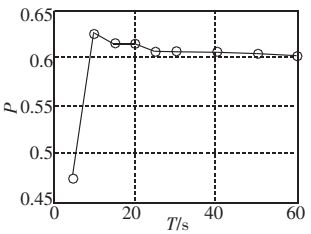

图 5　准确率与 $T$ 的关系

算法 2 的识别准确率情况如图 7 ~ 9 所示。图 7 是 $\Delta=3$、$\zeta=1.5$ 时识别准确率随 $T$ 的变化情况。由图可见,随着 $T$ 的增加,准确率逐渐升高,并趋于稳定。这是因为时间间隔大于一定 $T$ 值的两段数据流一般属于不同的业务流,时间间隔的门限高于一定值时,对于初步分段的影响越来越小。

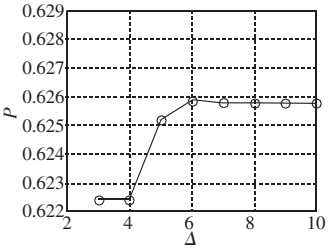

图 6　准确率与 $\Delta$ 的关系

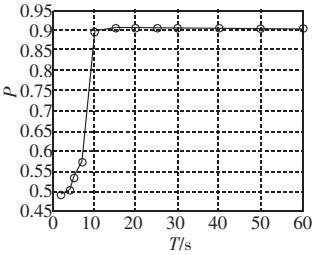

图 7　识别准确率随 $T$ 的变化情况

图 8 是 $T=10$ s、$\zeta=1.5$ 时识别准确率随 $\Delta$ 的变化情况。由图可见,随着 $\Delta$ 的增加,算法识别准确率逐渐降低并趋于稳定。这是因为在 $\Delta$ 比较小时,会将大部分报文长度有变化的点分段,而在利用分形模型的段合并中算法又将属于同一流的段进行合并。而随着 $\Delta$ 的增加,越来越多的报文长度突变点被选择不分段,造成了准确率的降低。

图 9 是 $T=10$ s、$\Delta=3$ 时识别准确率随 $\zeta$ 的变化情况。由

图可见,随着 $\zeta$ 的增加,识别准确率先增加后降低。这是因为 $\zeta$ 太小时造成合并规则的苛刻,本该合并的同一业务流的段不被合并,而 $\zeta$ 太大,则会造成段合并部分算法的失效。

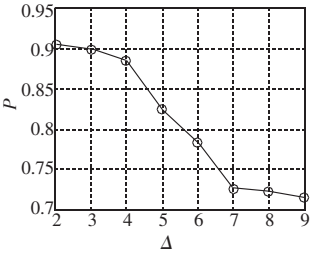


图 8 识别准确率随 $\Delta$ 的变化情况

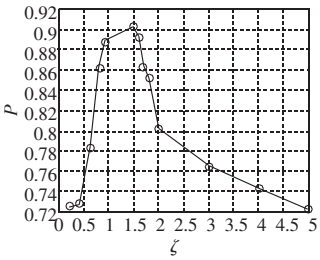


图 9 识别准确率随 $\zeta$ 的变化情况

将图 7 ~9 与图 5、6 相比较,基于分形模型的分段方法比简单分段方法的准确率至少提高了 20% 。

### 3.2 基于马氏距离的聚类

基于马氏距离的聚类将节点的业务流段分为五类。以 $T=10$ s、$\Delta=3$、$\zeta=1.5$、$n=15$、$\alpha=0.5$ 时的分段结果为输入,在 MATLAB 平台上运行以上基于马氏距离的聚类方法,用分段长度、平均分组间隔和平均分组长度三个参数表示的聚类中心如表 2 所示。经与已知的业务流信息相比对,各类业务流段的聚类中心分别对应于 FTP 流、SMTP 流、TELNET 流、HTTP 流、以及 SSH 流。

表 2 业务流分类结果

| 类别 | 分组个数 | 分组间隔 | 分组长度 | 对应的业务类型 |
|---|---|---|---|---|
| 1 | 18 | 0.0095282 | 575.623 | SMTP 流 |
| 2 | 142 | 0.079918 | 67.226 | TELNET 流 |
| 3 | 11 | 0.13527 | 342.359 | HTTP 流 |
| 4 | 72 | 0.0032541 | 78.567 | FTP 流 |
| 5 | 451 | 0.14152 | 74.2 | SSH 流 |

以 $T=10$ s、$\Delta=3$、$\zeta=1.5$ 时的分段结果为输入,在 MATLAB 平台上运行以上基于马氏距离的聚类方法,聚类的结果如图 10 和 11 所示。如图 10 所示,当 $n=15$ 时,聚类的准确率在 $\alpha=0.5$ 左右时急剧下降,在此之前和之后都比较稳定。这是因为 $\alpha>0.5$ 时,很可能将与 $Z_1$ 和 $Z_2$ 属于不同类别的剩余样本归入 $Z_1$ 或 $Z_2$。图 11 是 $\alpha=0.5$ 时准确率随 $n$ 的变化情况。第一个聚类中心要选择与剩余样本尽可能最远的样本,$n$ 越大,此条件越容易满足,因而准确率越高。而随着 $n$ 的增加,准确率趋于平缓,则是因为抽取样本的类型已经涵盖样本集中所有的类。

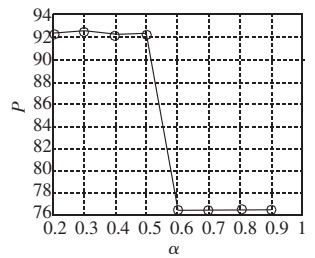


图 10 准确率与 $\alpha$ 的关系

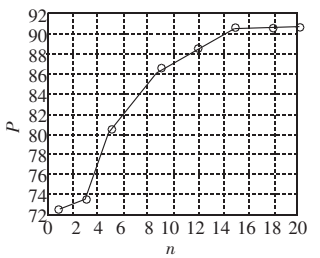


图 11 准确率与 $n$ 的关系

### 3.3 基于 Apriori 的时序关联分析

对 3.2 节中聚类的结果运行算法 3,具体操作过程为:a)第一次遍历统计各类业务流出现的次数;b)第二次遍历统计两类业务流衔接出现的次数。基于两次遍历的统计结果,可以计算两类业务流衔接出现的置信度,即两类业务流的时序关联强度。基于上述的分段和聚类,具体参数设置为 $T=10$ s,$\Delta=3$,$\zeta=1.5$,$n=15$,$\alpha=0.5$,$t=13$,各类业务流关联分析的结果如表 3 所示。其中,1 ~6 分别对应于表 2 中给出的业务流聚类结果,第 $i$ 行 $j$ 列的值代表在 $i$ 类业务流之后出现 $j$ 类业务流的概率,第 $i$ 行、“ – ”列则是指在 $i$ 类业务流之后较长时间没有数据发送的概率。

表 3 业务流时序关系

| 业务流类型 | 1 | 2 | 3 | 4 | 5 | – |
|---|---|---|---|---|---|---|
| 1 | 0.01 | 0.0985 | 0.4142 | 0 | 0.1333 | 0.3438 |
| 2 | 0.0946 | **0.5627** | 0.0806 | 0.2625 | 0 | 0 |
| 3 | **0.6405** | 0.2333 | 0 | 0 | 0.1262 | 0 |
| 4 | 0 | 0.2964 | 0 | 0.4357 | 0 | 0.2679 |
| 5 | 0.0064 | 0.2793 | 0.0802 | 0 | **0.63333** | 0.0008 |

由表 3 可见,在第 2 类流后以较高概率出现第 2 类流,第 5 类流后以较高概率出现第 5 类流。这是因为第 2、5 类流分别对应 TELNET 流和 SSH 流,这两类流的报文长度较短,传送与其他类型的流相同的数据量需要占用更长的时间,且需要进行大量信息交换来建立应用层连接,造成同一流的数据报文之间出现较长时间间隔的情况;另外也从侧面反映了用户的行为规律,大量连续的 TELNET 和 SSH 信息交换。在第 3 类流后以较高概率出现第 1 类流,这与该用户的操作习惯(即用户行为)有关,即通过 HTTP 页面来使用 SMTP。

对另一个节点 135.8.60.182 上的报文序列进行以上的分段、聚类和时序关联分析操作,参数设置同上,业务流之间的时序关联如表 4 所示。

表 4 业务流时序关系

| 业务流类型 | 1 | 2 | 3 | 4 | 5 | – |
|---|---|---|---|---|---|---|
| 1 | 0.1066 | 0.2753 | 0.1247 | 0.1438 | 0.0934 | 0.2562 |
| 2 | 0.1963 | 0.2037 | 0.01 | 0.1135 | 0.2124 | 0.2876 |
| 3 | 0.3625 | **0.6375** | 0 | 0 | 0 | 0 |
| 4 | 0.1544 | 0.1522 | 0 | 0.2377 | 0 | 0.4527 |
| 5 | 0.0987 | 0.3013 | 0 | 0 | 0.1 | 0.4 |

由表 4 可见,第 3 类业务流之后以较强的概率出现第 2 类流。这是因为 135.8.60.182 在网络中扮演攻击节点的角色,通过发送大量特定设置的 TELNET 数据包,远程改变目标节点上的用户权限。而 HTTP 信息交换则相对较少,且都在 TELNET 信息发送的间隙,因此会形成 HTTP 流与 TELNET 流之间的关联。

## 4 结束语

本文提出了一种业务流时序分析方法。首先依据时间间隔和报文长度信息,基于分形模型对节点的信息发送时间序列进行分段;再依据报文长度、报文间隔和段长度等统计特性,基于马氏距离对业务流段进行聚类分析;最后用 Apriori 算法对各类业务流段进行时序关联分析。实验结果表明,这一系列方法可以有效地将用户通信中的不同业务流进行分类,并分析出其时序关系。时序分析的结果可以为用户行为的分析与管控提供技术支持。

**参考文献:**

[1] VLACHOS M, MEEK C, VAGENA Z. Identifying similarities, periodicities and bursts for online search queries[C]//Proc of ACM SIGMOD International Conference on Management of Data. 2004: 131-142.

[2] RADINSKY K, SVORE K, DUMAIS S, *et al*. Modeling and predicting behavioral dynamics on the Web[C]//Proc of the 21st International Conference on World Wide Web. 2012: 599-608.

[3] SHAFER I, REN Kai, BODDETI V N, *et al*. Rainmon: an integrated approach to mining bursty timeseries monitoring data[C]//Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012:1158-1166.

[4] TAL L, MUNTEAN G. User-oriented cluster-based solution for multimedia content delivery over VANETs [C]//Proc of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting. 2012:1-5.

[5] LU Zheng, YANG Xiao-kang, LIN Wei-yao, *et al*. Inferring user image-search goals by mining query logs with semi-supervised spectral clustering[C]//Proc of IEEE Visual Communications and Image Processing. 2012:1-6.

[6] ZHENG Kai, XIONG Hai-ling, CUI Yu-jie, *et al*. User clustering-based Web service discovery[C] //Proc of the 6th International Conference on Internet Computing for Science and Engineering. 2012: 276- 279.

[7] WU Liang, CHIN A, ZHOU Yuan-chun, *et al*. Context-aware prediction of user's first click[C]//Proc of the 1st International Workshop on Context Discovery and Data Mining. 2012:7.

[8] SPIEGEL S, GAEBLER J, LOMMATZSCH A, *et al*. Pattern recognition and classification for multivariate time series[C]//Proc of the 5th International Workshop on Knowledge Discovery from Sensor Data. 2011:34-42.

[9] BO Han, HAO Xing-wei, LIU Chao-feng. The design and implementation of user behavior mining in E-learning system [C]//Proc of International Conference on Automatic Control and Artificial Intelligence. 2012:2078-2081.

[10] CASTIGLIONE A, CATTANEO G, MAIO G D, *et al*. SECR3T: secure end-to-end communication over 3G telecommunication networks [C]//Proc of the 5th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. 2011:520-526.

[11] CHENEAU T, SAMBRA A V, LAURENT M. A trustful authentication and key exchange scheme (TAKES) for Ad hoc networks[C]//Proc of the 5th International Conference on Network and System Security. 2011:249-253.

[12] ZHANG Wu-jun, ZHANG Yue-yu, CHEN Jie, *et al*. End-to-end security scheme for machine type communication based on generic authentication architecture[C]//Proc of the 4th International Conference on Intelligent Networking and Collaborative Systems. 2012:353-359.

[13] 蒋艳凰,赵强利. 机器学习方法[M]. 北京:电子工业出版社, 2009:249-251.

[14] http://www. ll. mit. edu/mission/communications/cyber/CSTcorpora/ideval/docs/index. html[EB/OL].

[15] LI Jiang-ping, YANG S X, GREGORI S. Combining top-down and Ncut methods for figure-ground segmentation[C]//Proc of International Conference on Apperceiving Computing and Intelligence Analysis. 2008:216-219.

[16] ALPERT S, GALUN M, BRANDT A, *et al*. Image segmentation by probabilistic bottom-up aggregation and cue integration [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 34 (2):315-327 .

[17] BIRDWELL J D, DJOUADI M S, PAN Yong-sheng. Efficient bottom-up image segmentation using region competition and the Mumford-Shah model for color and textured images[C]//Proc of the 8th IEEE International Symposium on Multimedia. 2006:376-390.

[18] GATTAL A, CHIBANI Y. Segmentation and recognition strategy of handwritten connected digits based on the oriented sliding window [C]//Proc of International Conference on Frontiers in Handwriting Recognition. 2012: 297-301.

[19] 朱华. 分形理论及其应用[M]. 北京:科学出版社,2011.

[20] 郭帅,马书根,李斌,等. VorSLAM 算法中基于多规则的数据关联方法[J]. 自动化学报,2012,38(1):1-12.

(上接第 517 页)条件下航空自组网具有一定的组网可行性。

## 5　结束语

本文结合一维航线的高度分层、双向飞行等实际情况,建立适当的飞机分布模型,并考虑反向飞机对信息转发的作用,推导出一维航线连通概率表达式。通过实验模拟航线上飞机速度、分布等环境,证明了本文连通概率表达式的正确性,并对北京与三亚之间部分实际航线飞行场景下的连通性进行了仿真分析。仿真结果表明:组建航空自组网在一维航线双向通信情况下是客观可行的。本文对航空自组网的研究和应用具有一定的参考价值,下一步主要研究飞机的多维航线及不固定航线的连通性以及将其应用在路由协议等技术上。

**参考文献:**

[1] TU H D, SHIMAMOTO S. A proposal of relaying data in aeronautical communication for oceanic flight routes employing mobile Ad-hoc network[C]//Proc of the 1st Asian Conference on Intelligent Information and Database Systems. 2009:436-441.

[2] SCHNELL M, NEWSKY S S . A concept for networking the sky for civil aeronautical communications [J]. Space Communications, 2008, 21(3/4):157-166.

[3] 郑博,张衡阳,黄国策,等. 航空自组网的现状与发展[J]. 电信科学,2011,27(5):38-47.

[4] 党亚茹,丁飞雅,高峰. 我国航班流网络搞毁性实证分析[J]. 交通运输系统工程与信息,2012,12(6):177-185.

[5] 黄后彪,罗长远,宋玉龙. 航空自组网漫游接入认证方案[J]. 计算机应用研究,2013,30(2):500-502.

[6] BETTSTETTER C. On the minimum node degree and connectivity of a wireless multihop network[C]//Proc of the 3rd ACM International Symposium on Mobile Ad hoc Networking & Computing. New York: ACM Press,2002:80-91.

[7] MEDINA D, HOFFMANN F, AYAZ S, *et al*. Feasibility of an aeronautical mobile Ad hoc network over the North Atlantic Corridor[C]// Proc of the 5th Annual IEEE Communication Society Conference on Sensor. San Francisco: IEEE Press, 2008:109-l16.

[8] DESAI M, MANJUNATH D. On the connectivity in finite Ad hoc networks[J]. Communications Letters, IEEE, 2006, 6 (10): 437-439.

[9] GHASEMI A, NADER-ESFAHANI S. Exact probability of connectivity one-dimensional Ad hoc wireless networks [J]. Communications Letters, IEEE, 2006, 10(4): 251-253.

[10] CHENG M X, ZHAO Y Y. Connectivity of Ad hoc networks for advanced air traffic management[J]. Journal of Aerospace Computing , Information, and Communication, 2004, 1(5):225-238.

[11] 郑博,黄国策,张衡阳,等. 甚高频航空自组网的组网概率及连通性研究[J]. 西安交通大学学报,2011,45(8):24-29.

[12] YAN Jian-shu, SONG Ge, LI Hua, *et al*. Critical transmission range for connectivity in aeronautical Ad hoc networks [C]//Proc of the 10th World Congress on Intelligent Control and Automation. 2012:4446-4451.

[13] 郑博,张衡阳,孙鹏,等. 航空自组网单、双向航路连通性研究[J]. 上海交通大学学报,2012,46(4):624-629.

[14] LI Hua, YANG Bo, CHEN Cai-lian, *et al*. Connectivity of aeronautical Ad hoc networks[C]//Proc of IEEE GLOBECOM Workshops Aerial Vehicles. 2010:1788-1792.