外れ構造抽出技術を用いた大規模セキュリティログ分析

Large-scale Security Log Analysis based on Outlier Extraction Technology

● 本多聡美　● 丸橋弘治　● 鳥居　悟　● 武仲正彦

あらまし

昨今，サイバー攻撃が激化・巧妙化する中，未知の攻撃やその意図を早期に発見・分析することは喫緊の課題である。しかし，従来のルールベースでの分析やデータ単体の外れ値検知では，大量のネットワーク監視ログに紛れた巧妙な未知攻撃の抽出・発見は困難である。そこで富士通研究所は，この課題解決のために，ビッグデータ分析技術を適用し，セキュリティログのような大規模離散値データから，外れ構造（まれな特徴を有するデータ群）を抽出する「外れ構造抽出技術」を開発した。この外れ構造抽出技術をネットワーク監視ログに適用することで，大量のログすなわちビッグデータに紛れた巧妙な攻撃を抽出できる。実際にサービスを運用している多数のサーバから得られた，IDS（Intrusion Detection System：侵入検知装置）ログに外れ構造抽出技術を適用しセキュリティの観点からその抽出結果を詳細に分析することで，従来知られていなかった巧妙かつ高度な戦略を持つ攻撃を検出することに成功した。

本稿では，実際のIDSログに対して，この外れ構造抽出技術を活用した取組みと事例を紹介する。

Abstract

In recent years, attacks on cyber space have not only increased in number but also become aggressive and sophisticated. To take countermeasures against such attacks, security analysts must detect and analyze unrecognized tactical attack events and their intentions before the start of the main attacks. It is not sufficient to apply conventional methods such as rule-based attack detection and simple outlier detection. The tactical attack events are hidden among numerous known attack events and it is difficult to extract and analyze them with only conventional methods. Therefore, to solve this problem Fujitsu Laboratories has developed Outlier Extraction Technology in order to single out an outlier-structure (a cluster of data shared with rare values) from a large-scale discrete event log such as a security event log. By applying this outlier extraction technology to a network monitoring log, security analysts can detect and analyze a sequence of tactical attacks. In this paper, we present a case study of applying the technology to an intrusion detection system (IDS) log. In this case study, we successfully extract an unrecognized attack sequence with advanced strategies. Detailed analyses on one of the outlier structures conducted by a security analyst led to this finding. Furthermore, we also show a Computer Emergency Response Team (CERT) project that utilizes the outlier extraction technology and is applied to a real IDS log.

ま え が き

サイバー攻撃の激化はとどまるところを知らず，政府や企業を脅かし続けている。その攻撃手法は巧妙化してきており，攻撃者は自身の攻撃対象とする組織に合わせて十分に戦略を練った上で，その対象にカスタマイズした攻撃を仕掛けてくる。そのため，既知の攻撃パターンと照らし合わせて攻撃発生の有無を判断する従来の方式では，巧妙化された最新のサイバー攻撃を十分に検知し防御できない。最新のサイバー攻撃に適切に対処するためには，既知の攻撃だけでなく，未知の攻撃やその意図を早期に発見・分析することが喫緊の課題である。

未知の攻撃やその意図を検知する手段として，ネットワーク機器から出力されたログの分析が重要であると言われている。近年，システム内に設置した様々な機器のログを一元的に管理・分析できるSIEM（Security Information and Event Management）と呼ばれる技術が注目されている。SIEMは，複数の機器のログを横断的に分析できる。しかしその検知ロジックは，ある特定のログ中のレコード件数の急激な変化や，あらかじめ決められたルールに合致する/合致しないレコードの出現を検出するなど，従来の攻撃検知と同様な方式を採用しているに過ぎない。

一方，冒頭に述べた巧妙かつ戦略的な攻撃の検知には，このようなレコード件数の増減やルールベースでの分析だけでは不十分である。なぜならば，攻撃者は従来の検知ロジックでは検知できないようにカスタマイズした手法で攻撃するからである。例えば，あるネットワークサービスに不正にログインするために，よく使われるパスワードを使って何度もログインを試みる攻撃が知られている。従来の攻撃では，短時間に大量のログイン試行を行うものがほとんどであった。そのため，ネットワーク監視装置やSIEMでは，単位時間あたりのログイン試行回数が一定数以上となった場合に，攻撃と判断する検知ロジックが利用されてきた。

これに対し，単位時間あたりのログイン試行回数を攻撃と判断されない程度に少なくした上でログイン試行を長期間続ける，といった特徴を持つ攻撃が確認されている。このような巧妙な攻撃は，従来の検知ロジックを利用したとしても効果的に検知できるとは限らない。そこで，従来のように単純な定常状態を仮定し，それに対する急激な変化や特異点を検出するのではなく，一見正常に見える状態に潜む一連の攻撃を抽出できる技術が必要となる。

この課題に対し，筆者らは，ビッグデータ分析技術の適用による解決を試みた。富士通研究所では，ビッグデータ分析技術の中でも，IPアドレスなどの離散値を持つ大規模データ（**図-1**）から，外れ構造（まれな特徴を有するデータ群）を抽出する技術を開発した。本技術をセキュリティの観点でのログ分析に適用することで，大量のネットワーク監視ログに紛れた巧妙な攻撃を抽出できる。

本稿では，実際にサービスを運用している多数のサーバから得られたIDS（Intrusion Detection System：侵入検知装置）ログに外れ構造抽出技術を適用し，その抽出結果をセキュリティの観点から詳細に分析することで，従来知られていなかった巧妙かつ高度な戦略を持つ攻撃の検出に成功したビッグデータの利活用事例を紹介する。

Webアクセスログ

No.	srcIP	userID	URL
1	a.a.a.a	User1	http://www.f..
2	b.b.b.b	User2	http://login.f...
3	b.b.b.b	User2	https://conf...
4	c.c.c.c	User3	http://rec.pl..
5	d.d.d.d	User1	http://usr-pag..
6	e.e.e.e	User4	https://ad-pho..

⋮

IDS（侵入検知装置）ログ

No.	srcIP	dstIP	Port	Signature
1	a.a.a.a	x.x.x.x	80	TCP Connect DOS
2	b.b.b.b	y.y.y.y	80	TCP Connect DOS
3	c.c.c.c	x.x.x.x	0	TCP Tear Drop
4	b.b.b.b	z.z.z.z	22	Brute Force Attack
5	d.d.d.d	y.y.y.y	0	Finger Print
6	a.a.a.a	z.z.z.z	22	Brute Force Attack

⋮

図-1　離散値データの例

既知攻撃に紛れた未知攻撃の検出

IDSとは，サーバとインターネットの接続部分に設置し，インターネットからの攻撃を検出する装置である。インターネットに接続されたサーバには，世界中にまん延したマルウェアなどから，公開されているサービスの種類を探索するネットワークスキャンや，既知の脆弱性を突くための攻撃通信などが大量に送られてくる。IDSは，これらの攻撃通信を全て記録するため，そのログの量は膨大である。そのため，分析対象となるIDSログは，既知の攻撃を検出した記録が大部分を占めており，その中に紛れたわずかな未知の攻撃を検出することが課題である。以降，IDSが生成する1回の記録（図-1の各行に相当）をレコードと呼ぶ。

上記の課題を解決する手段として，全レコードの値の分布を統計的に学習した上で，大多数のレコードと異なる値を持つレコードを抽出する手法が有効であると考えられる。このような手法は，外れ値検知としてデータマイニングの分野において多数提案されている。[(1)] これにより，特別なルールを記述することなく，未知の攻撃などの検出が期待できる。

ここで，IDSログの分析においては，IPアドレスやポート番号のような離散値の分布が問題となる。既存の外れ値検知手法のほとんどは連続値の分布を対象としたものであるため，これらの手法を直接適用できない。これに対し，離散値データを対象とした外れ値検知手法が近年いくつか提案されている。それらの手法は，基本的には出現頻度の低い値を持つレコードを外れ値とみなす考え方である。例えば，頻出パターンを抽出し，頻出パターンに合わないレコードを外れ値とみなす手法が提案されている。[(2)]

しかし，IDSログは，IPアドレスやポート番号のような数万種類もの値を取り得る変数を持つ大規模離散値データである。この場合，多くのレコードは数回程度しか出現しない値の組合せを持つことになる。そのため，出現頻度の低い値を持つレコードを外れ値とみなす手法では，攻撃に対する誤検知が大量に発生してしまう。更に，出現頻度の高い値の組合せであっても，値を共有するレコード群を集団として見たときに，集団としてまれな特徴を有している可能性もある。これらの問題に対する一つの解決方法は，関連する値を適切にまとめた上で，大多数のレコードと異なる特徴を有するレコードの集団を抽出することである。筆者らは，このようなレコードの集団を「外れ構造」と呼んでいる。

次章では，筆者らが開発した大規模離散値データを対象とした外れ構造抽出技術を紹介する。

外れ構造抽出技術

大規模離散値データから外れ構造を抽出するためには，大きく2種類のアプローチが考えられる。その一つは，少数の主要な値のまとまりで記述できるレコードを取り除き，余ったレコードの中から注目すべきレコードの集団を見つける，言わばトップダウンのアプローチである｛**図-2**（a)｝。筆者らは，トップダウンのアプローチとして，テンソル分解と呼ばれる技術を用いた手法を提案した。[(3), (4)] この手法は，まずテンソル分解を用いて，データの大部分を記述できる主要な値のまとまりをいくつか抽出する。そして，抽出した値のまとまりを用いて記述できないレコードの中から，誤検知とみなせない規模の互いに値を共有するレコードの集団に絞り込むことにより，有意な外れ構造を抽出する。筆者らは，この手法をWebアクセスログやIDSログに適用し，外れ構造の候補を抽出した。[(4)] しかし，この方法は誤検知の低減に成功しているものの，最初に出現頻度の高い値の組合せを持つレコードを除外するため，まれな特徴を有する出現頻度の高い値の組合せのレコード集団を見逃している可能性がある。

もう一つのアプローチは，大小様々な値のまとまりによってデータ全体を多数の集団に分割した上で，ほかと異なる特徴を有するレコードの集団を抽出する，言わばボトムアップのアプローチである｛図-2(b)｝。筆者らは，ボトムアップのアプローチとして最小記述長原理に基づくデータ分割を利用した手法を開発した。最小記述長原理とは，機械学習の分野で採用されることの多い考え方である。これは，データを説明するためのモデルとして，モデルを記述するための情報量と，モデルに基づきデータを記述するための情報量の総和が小さいほど良いモデルとするものである。筆者らが開発

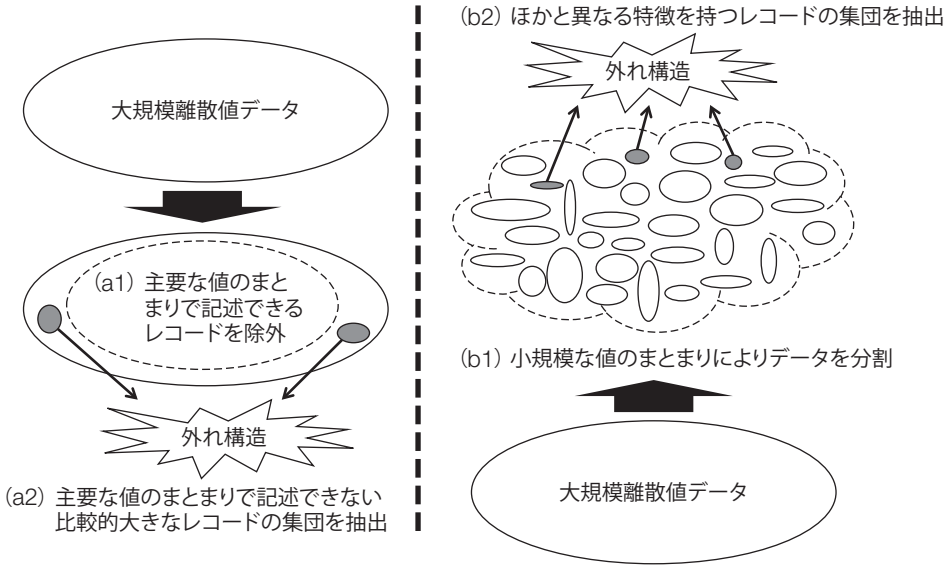

図-2　外れ構造抽出の2種類のアプローチ

した手法では，まず最小記述長原理に基づきデータをレコードの集団に分割する。具体的には，各集団の値の構成（モデル）を記述するための情報量と，モデルに基づき各レコードを集団ごとに記述するために必要な情報量の総和が最小になるようなレコードの集団に分割する。そして，各レコードの記述長に関する特徴量により各レコード集団のアノマリスコア(注)を算出し，スコアの高い集団を外れ構造として抽出する。これにより，集団としてまれな特徴を有するレコード集団を抽出できる。

ここで，大規模離散値データに対するボトムアップのアプローチにおいては，計算量の問題が避けられない。すなわち，大規模離散値データでは，数回しか発生しない数万種類の値の膨大な組合せを探索しなければならないため，現実的な計算量では最適な集団を見つけられない。そのため，限られた範囲の探索により近似解を得ることになる。ここで，こうした探索では，大規模なデータ集団よりも小規模なデータ集団が適切に分離されるほうが望ましい。なぜなら，大量の既知攻撃に紛れた巧妙な攻撃は，小規模なレコード集団から見出される可能性が高いからである。筆者らは，集団の中で低頻度の値の組合せを持つレコード集団を分離し，最小記述長原理に基づく再評価を繰り返す手法を考案した。これにより，小規模のレコード集団を構成する値の組合せを優先的に探索しながら外れ構造を抽出するため，大量の既知攻撃に紛れた巧妙な攻撃を効率的に発見できる。

IDSログでの適用事例

本章では，前章で述べたボトムアップのアプローチによる外れ構造抽出技術を実際にネットワーク監視ログに適用した事例を紹介する。

今回，実際にサービスを運用している多数のサーバから得られたIDSログに外れ構造抽出技術を適用した。適用結果として出力されたIDSログのクラスタ群からアノマリスコアが大きかったクラスタを抽出し，詳細に分析した。

更に，セキュリティの観点での分析を進めるため，抽出したクラスタに対し関連度の観点から可視化した。分析・可視化の例を**図-3**に示す。クラスタに含まれるレコードを円（ノード）で表現し，レコード同士の関連度が強い場合には，該当するノードを線（エッジ）で接続する。図-3では，クラスタに含まれるレコード1，2，3について，1と2，2と3が，それぞれ関連度が高いため，レコード1

（注）　集団内のレコードの平均記述長が，ほかの集団の平均記述長と異なる程度を表すスコア。

1クラスタあたりのIDSログ

No.	srcIP	dstIP	Port	Signature	レコード数
1	a.a.a.a	x.x.x.x	22	Brute Force Attack	10
2	a.a.a.a	y.y.y.y	22	Brute Force Attack	10
3	b.b.b.b	y.y.y.y	80	TCP Connect DOS	50

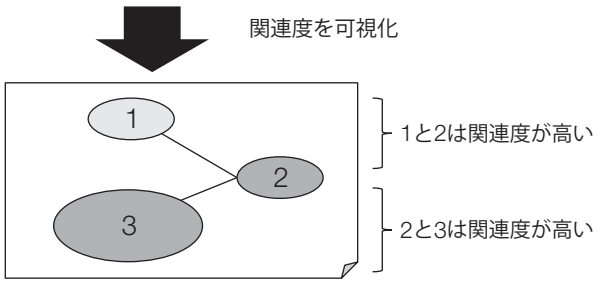

図-3　関連度の観点による可視化

と2を示すノード同士，レコード2と3を示すノード同士がそれぞれエッジで接続されている。ここで，レコード間でまれな値を多く共有するほど，レコード間の関連度が強いと定義する。また，全属性の値が一致するレコードは一つのノードにまとめ，ノードの大きさはまとめたレコードの件数に比例させる。各ノードの色は，含まれるレコードの持つ属性の値に応じて決定する。

IDSログに外れ構造抽出技術を適用し，アノマリスコアが高いと判定されたクラスタの可視化結果を**図-4**に示す。srcIP（攻撃元IPアドレス），dstIP（攻撃対象IPアドレス），Port（攻撃対象ポート番号），Signature（IDSが検知した攻撃名）の4種類の属性ごとに別々に色分けして示している。当該クラスタは，SSH（Secure Shell）サービスに対するブルートフォース攻撃（ログインできるユーザー名とパスワードの組合せを順番に試す攻撃）が検知されたレコードが多く含まれている。このクラスタに含まれるブルートフォース攻撃は，様々なIPアドレスから特定のサービスに対してブルートフォース攻撃を繰り返し行う特徴を持つ「IP使い捨て型ブルートフォース攻撃（Ephemeral Brute Force Attack：EBF）」であることが既に分かっている。[5]

ところで，この図-4中のSignatureにより色分けされたグラフを観察すると，グラフ外周に異なる色のノードがひも付けられた箇所（図中に矢印で示した箇所）が複数存在することが分かる。これ

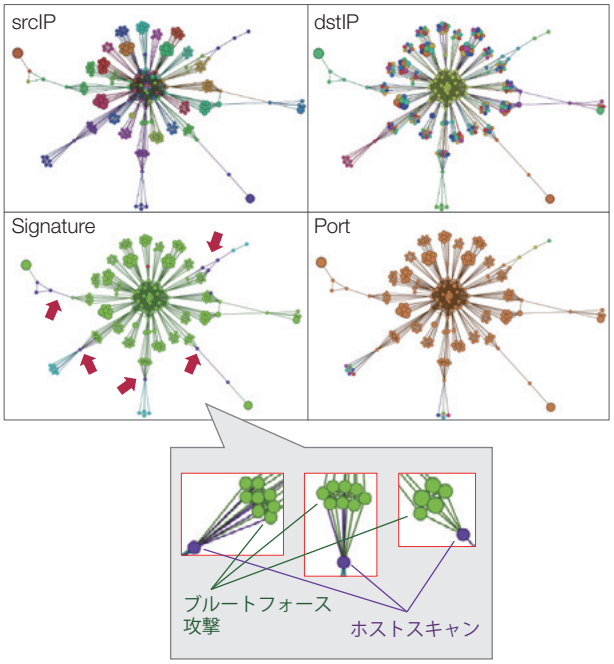

図-4　アノマリスコアが高いと判定されたクラスタの可視化結果

らのノードは，Signatureとしてそれぞれ「ブルートフォース攻撃」と「ホストスキャン」を持つ。この「ホストスキャン」とは，あるネットワークサービスについてサービスを運用しているホストを探索する行為を示す。この観察結果から，EBFとホストスキャンに関連性が存在する可能性があるという知見が得られた。更に，この知見を起点としてログの統計を計算した結果，EBF開始の数分から数十分前にホストスキャンが行われていたことが分かった。[6]

こうした知見は，ホストスキャンとブルートフォース攻撃を実施した攻撃元IPアドレスをEBFに該当すると素早く判断し，ほかのホストへの影響を最小限に抑えるといった対策技術にも活用できる。

クラウドサービス監視での運用

開発技術は現場のセキュリティ監視・分析チームとも連携を図っており，その一環として富士通のクラウドサービス監視においても運用を開始している。

図-5は分析ダッシュボードの画面例である。

本図では，ある一定期間について外れ構造抽出

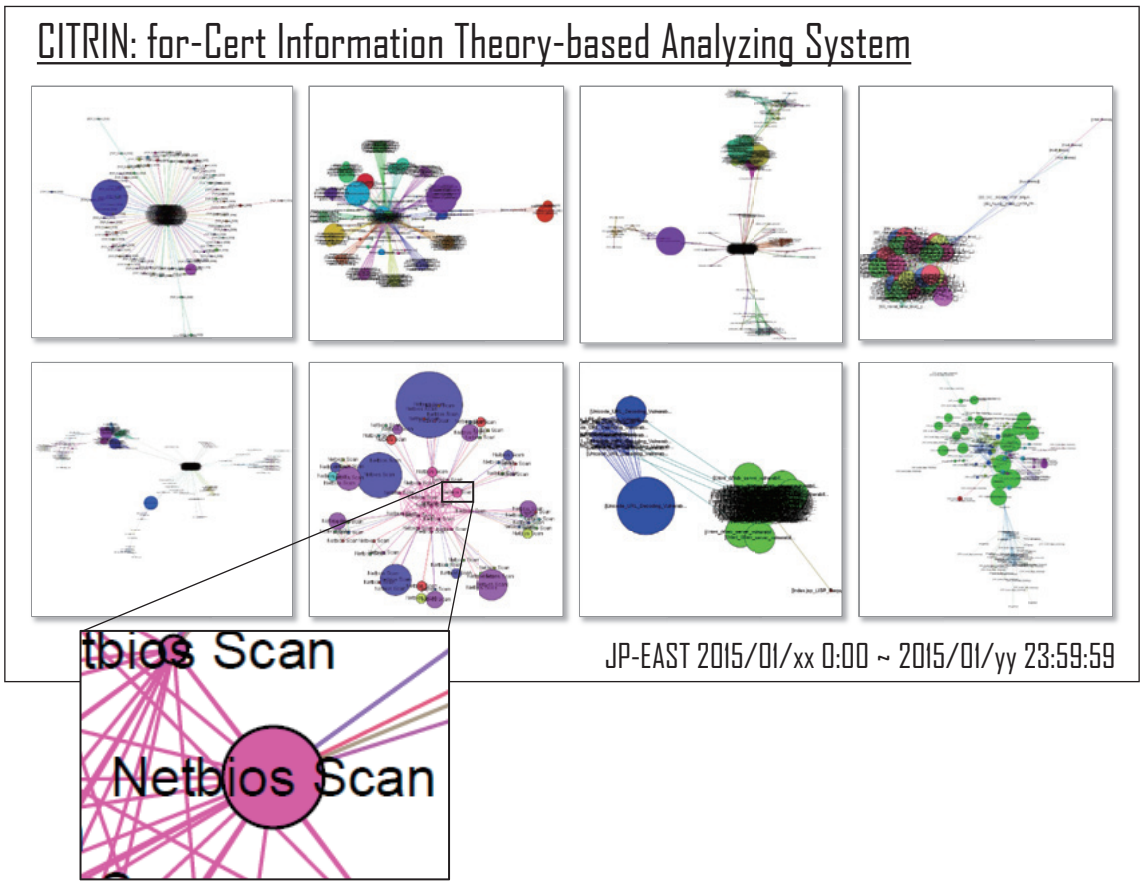

図-5　分析ダッシュボード画面例

技術，および関連度に着目した可視化を適用したIDSログのクラスタをパネル状に並べて表示している。ノード上に表示されている文字列は，各ノードが持つSignatureを表している。監視・分析者は，グラフそのものの形状やカラーバリエーション，Signatureなどを観察し，詳細分析が必要なクラスタを見つけられる。

むすび

本稿では，まれな特徴を有するデータ群「外れ構造」を抽出する「外れ構造抽出技術」をセキュリティログ分析へ適用し，大量のネットワーク監視ログに紛れた巧妙な攻撃を検出する取組みとその分析事例を紹介した。

富士通のクラウドサービスにおける本技術の運用は始まったばかりである。今回の取組みを土台とし，富士通独自のセキュリティ監視技術の実現を目指したい。

参考文献

(1) V. Chandola et al.：Anomaly detection：A survey. ACM Computing, Surveys, Vol.41 (3), (2009).
(2) L. Akoglu et al.：Fast and reliable anomaly detection in categorical data. CIKM '12, AVM, New York, USA,：p415-424, 2012.
(3) K. Maruhashi et al.：MultiAspectForensics：Pattern Mining on Large-Scale Heterogeneous Networks with Tensor Analysis. ASONAM 2011, p.203-210, 2011.
(4) K. Maruhashi et al.：MultiAspectSpotting：Spotting Anomalous Behavior within Count Data Using Tensor. PAKDD 2014, p.474-485, 2014.
(5) 本多聡美ほか：拠点横断分析によるIP使い捨て型ブルートフォース攻撃の検知とその抽出手法．情報処理学会論文誌，Vol.56，No.3，p.911-920，2015.
(6) S. Honda et al.：CITRIN：Extracting Adversaries Strategies Hidden in a Large-Scale Event Log. RAID2014, Presentations and Posters.

著者紹介

本多聡美（ほんだ　さとみ）
知識情報処理研究所
データ・プライバシー保護プロジェクト 所属
現在，サイバーセキュリティに関する研究開発に従事。

鳥居　悟（とりい　さとる）
知識情報処理研究所
データ・プライバシー保護プロジェクト 所属
現在，サイバーセキュリティに関する研究開発に従事。

丸橋弘治（まるはし　こうじ）
知識情報処理研究所
Big Intelligenceプロジェクト 所属
現在，異常検知などのデータマイニングに関する研究開発に従事。

武仲正彦（たけなか　まさひこ）
知識情報処理研究所
サイバー・システムセキュリティプロジェクト 所属
現在，サイバーセキュリティをはじめとするシステムセキュリティに関する研究開発に従事。