情報処理学会論文誌 Vol. 50 No. 3 992–1001 (Mar. 2009)

ネットワーク管理におけるイベント発生状況の効率的な把握を実現するイベント分析価値評価手法の提案と評価

佐藤彰洋[†1] 長尾真宏[†1] 小出和秀[†2]
木下哲男[†3] 白鳥則郎[†2]

本論文では，既存の異常検出手法により検出された多量のイベントに対し，管理者によるイベントの発生状況の効率的な把握を支援する「イベント分析価値評価手法」を提案する．本手法の特徴は，イベント発生時のトラフィックデータが持つ情報量に基づき，イベントの分析価値を与え，管理者が分析を行うべきか否かの判断基準とすることである．これにより，多量のイベントが検出され，管理者がすべてのイベントを分析しきれない場合に，管理者が分析するイベントの数を大幅に抑制しつつ，多様な種類のイベントの分析を実現することで，ネットワーク上でのイベントの発生状況の把握が可能となり，管理者の作業負担の軽減に大きく貢献できる．さらに実データに基づく評価実験を通じて，提案手法は従来手法と比較して，希少なイベントを見逃すことや，類似したイベントを何度も分析することなく，分析対象とするイベントを7割以上削減できることを示す．

A Novel Method to Grasp The Detected Events in Event-based Management and Its Evaluation

Akihiro Satoh,[†1] Masahiro Nagao,[†1]
Kazuhide Koide,[†2] Tetsuo Kinoshita[†3]
and Norio Shiratori[†2]

The analysis of events is indispensable to grasp the detected events in an event-based management. However, when large amount of events are generated, overlooking of minor events or repetition of analyzing the same events is a serious concern. In this paper, we propose a method to evaluate the degree of importance of the events. The degree of importance is based on the amount of information that can be obtained by analyzing the traffic data of the events generated by this method. Thus, an efficient way to detect the insights of the events is possible and as a result improvement in network management can be achieved. Our experiment results show that the proposed method is capable of reducing more than 70% of unnecessary events compared to the traditional method without losing any necessary information.

1. はじめに

ネットワーク管理の一般的な方法として，イベントに基づくネットワーク管理が広く行われている[1)]．管理者はネットワーク上の変化をイベントとして検出し，そのイベントを分析することで，原因を特定し対処する．イベントに基づくネットワーク管理の運用上の課題の1つに，イベント検出条件を適切に調整することがあげられる．その目的は，イベント数の増加による，イベントの対処にかかる管理負担を軽減し，また重要なイベントを確実に検出できるようにすることである．そのため，検出されたイベントを分析することで，イベントの数や種類，すなわち発生状況を定期的に把握し，必要に応じて検出条件を調整することが重要である．

近年，高度なイベント検出の方法として異常検出手法が注目され，高い成果をあげている[2)–4)]．しかしながら，異常検出手法により検出されたイベントは，分析が難しく管理者の負担が大きい．特に，多量のイベントが検出された場合は，イベントの発生状況の把握に支障をきたし，イベント検出条件の適切な調整が困難となる[1)]．そのため，イベントの発生状況の把握を支援する技術の開発が期待される．

そこで本論文では，既存の異常検出手法により検出された多量のイベントに対し，管理者によるイベントの発生状況の効率的な把握を支援する「イベント分析価値評価手法」を提案する．本手法の特徴は，イベント発生時のトラフィックデータが持つ情報量に基づき，イベントの分析価値を与え，管理者が分析を行うべきか否かの判断基準とすることである．これにより，多量のイベントが検出され，管理者がすべてのイベントを分析しきれない場合に，管理者が分析するイベントの数を大幅に抑制しつつ，多様な種類のイベントの分析を実現することで，ネットワーク上でのイベントの発生状況の把握が可能となる．

†1 東北大学大学院情報科学研究科
Graduate School of Information Science, Tohoku University
†2 東北大学電気通信研究所
Research Institute of Electrical Communication, Tohoku University
†3 東北大学サイバーサイエンスセンター
Cyberscience Center, Tohoku University

© 2009 Information Processing Society of Japan

提案手法が，イベントの発生状況の把握を効果的に支援し，管理者の作業負担の軽減に大きく貢献することを示すため，プロトタイプシステムを用いた実運用ネットワークでの実験を通じ，有効性の評価を行った．実験の結果，提案手法により導出された分析価値に基づいてイベントを選択することで，管理者が希少なイベントを見逃すことや，類似したイベントを何度も分析することなく，ネットワーク上でのイベントの発生状況を効率良く把握できることを確認した．さらに，従来手法と比較して，管理者が分析すべきイベントの数を 7 割以上削減可能であることを示した．

本論文の構成を以下に示す．2 章でイベントの検出と分析における従来研究と課題について整理し，3 章でその課題を解決するイベント分析価値評価手法を提案し，4 章で具体的なシステムの設計を述べる．5 章で提案手法の評価実験と，その結果に基づいた提案手法の評価を行い，最後に 6 章で結論を述べる．

2. 関 連 研 究

2.1 イベントに基づくネットワーク管理

イベントに基づくネットワーク管理は，ネットワーク管理のうち，障害管理，性能管理，セキュリティ管理の分野でよく用いられる管理方法である[1]．たとえば，障害発生時に観測される可能性のあるトラフィックデータの特徴やログなどをイベントとして定義することで，そのイベントが検出されるごとに，管理者が障害の有無を分析して確認した後，必要に応じてその問題に対処することができる．すなわち，イベントを機械的に検出した後，原因を予測・発見し，対処する管理方法である．本研究ではトラフィックデータに基づいたイベントの検出を対象とする．ここで，イベントの検出手法は，不正検出手法と異常検出手法に大別される．

2.2 不正検出手法によるイベントの検出と分析

不正検出手法は，検出したいイベントの特徴を明確化して，シグネチャとして検出条件を事前に記述し，シグネチャと一致するトラフィックデータの特徴が観測された場合に，イベントとして検出する手法である．この手法により検出されたイベントは，あらかじめシグネチャとしてイベントの特徴が明確化されているため，管理者に対するイベントの分析の負担は小さく，容易に原因の特定と迅速な対処が可能である．

2.3 異常検出手法によるイベントの検出と分析

一方，異常検出手法は，トラフィックデータを抽象化し，その正常/異常を統計的に判定し，異常と判断されたときにイベントとして検出する手法である．この手法は，不正検出手法に比べ，あらかじめシグネチャとして特徴を明記することが困難なイベントに加え，未知のイベントを検出可能なため，ネットワークの複雑化，サービスの多様化にともない，その有効性が期待されている[2)–4)]．しかし，その反面，トラフィックデータを抽象化し統計的に処理するため，不正検出手法と比較して具体的な手がかりに乏しく，加えて，未知のイベントが含まれる可能性があるため，管理者の高度な知識・経験に基づいたイベントの分析が必要となり，管理者にとって大きな負担となる．

イベントの分析における管理者の負担を軽減するため，ネットワークの構成やネットワークから計測した情報の可視化[5),6)]，イベントの分析に有用な付加情報の提供を自動化する手法[7)]，特定のイベント分析を自動化する手法[8),9)] などが提案されている．しかしながら，これらの手法は個々のイベント分析の効率化に着目しており，本研究が対象とする，多量のイベントが検出された場合におけるイベントの発生状況の把握に対しては有効ではない．イベントの数の過多により把握しきれないという問題を解決するためには，単純にイベントの数が少なくなるように正常/異常の判定基準を調整することや，高い異常度が付与されたイベントを優先的に分析することが行われている[10)]．しかしながら，このような方法では検出されるイベントの種類が制限されるため，重要なイベントまで検出することができなくなるという新たな問題が発生する．

上述の問題点により，異常検出手法においては，イベントの数が多く，イベントの発生状況を把握できない場合に，異常検出手法の閾値を変更してイベント数を減少させるのではなく，多数のイベントの効率的な分析とイベントの発生状況の把握を可能とする手法が必要であるといえる．

3. イベント分析価値評価手法の提案

3.1 提案手法の概要

本論文で扱うイベントを，異常検出手法によりトラフィックデータを抽象化し統計的に分析することで検出した「時間軸上の異常点」と定義する．上述のイベントに対して，2.3 節で示した問題を解決し，イベントの発生状況の把握に要する管理者の作業負担を軽減するため，多量のイベントの発生状況の効率的な把握を支援する「イベント分析価値評価手法」を提案する．

本手法の特徴は，イベント発生時のトラフィックデータが持つ情報量に基づき，イベントの分析価値を与え，管理者が分析を行うべきか否かの判断基準とすることである．イベントの分析価値は，障害，性能，セキュリティなどのイベントの分類によらず，管理者がそのイ

© 2009 Information Processing Society of Japan

ベントを把握しているか否か，という観点で評価する．その理由は，本手法が異常検出手法により検出されたイベントを対象としており，イベントの発生状況の把握において，未知のイベントの分析が不可避であるので，それら未知のイベントが，どのような分類のイベントであっても，共通の指標で評価できるようにするためである．

本手法において，トラフィックデータの情報量が担うべき役割は，イベント発生時に観測されたトラフィックデータと，管理者がすでに発生原因を把握している，分析済みのイベント発生時に観測されたトラフィックデータとの相違点の，定量的な評価指標を与えることである．イベントの分析価値をトラフィックデータが持つ情報量に基づき決定する理由は，イベントがトラフィックデータに基づいて検出され，そのトラフィックデータはイベントの発生原因を特定する有力な情報源となることに起因する．すなわち，情報量が大きいトラフィックデータによるイベントは，管理者が把握していない原因により発生したイベントである可能性が高く，イベントの発生状況を把握するために，管理者はそのようなイベントを優先して分析するべきである．逆に，情報量が小さいトラフィックデータは，管理者により分析されすでに原因が把握されているイベントである可能性が高いため，そのイベントの分析を行わないこととする．これにより，多量のイベントが検出され，管理者がすべてのイベントを分析しきれない場合に，管理者が分析するイベントの数を大幅に抑制しつつ，多様な種類のイベントの分析を実現することで，ネットワーク上でのイベントの発生状況の把握が可能となる．

本手法は，以下の 5 つの段階からなる．

(1)　トラフィックデータの注目属性の決定
(2)　分析対象イベントの抽象化と特徴抽出
(3)　参照する既知のイベントの選択
(4)　トラフィックデータの情報量の導出
(5)　イベントの分析価値の導出

次節より，これら 5 つの段階を実現するにあたっての具体的なアイデアを説明する．ここで，本提案において取り扱う 2 種類のイベントを下記のように定義する．

分析対象イベント：異常検出手法によって検出され，これから管理者により分析されるイベント．分析対象イベントの数を N とし，個々のイベントを $e_1, e_2, \cdots, e_N$，その集合を $\mathbb{E}$ と表記する．

既知のイベント：管理者によって分析され，その発生原因についてすでに把握されているイベント．既知のイベントの数を N' とし，個々のイベントを $e'_1, e'_2, \cdots, e'_N$，その集合を $\mathbb{E}'$ と表記する．

3.2　第 (1) 段階：トラフィックデータの注目属性の決定

トラフィックデータを構成する個々のパケットから得られる多次元の情報を属性と呼び，その具体的な値を属性値と呼ぶ．たとえば，送信元アドレスは属性であり，その具体的な値である “192.168.11.120” は属性値である．

イベントの特徴を適切にとらえ，分析価値の正確性を向上させるため，第 (2) 段階におけるイベントの抽象化に用いる属性を選択する．その際，(i) イベントの検出に用いる情報，(ii) ネットワークの構成，について考慮すべきである．(i) の例として，イベントの検出の際に頻繁に用いられる，送受信アドレスおよびポート番号に加え，パケットヘッダから取得可能な TCP フラグ，パケット長，プロトコル番号などの情報が重要な属性であると考えられる．また，(ii) の例として，タグ VLAN で構成されたネットワークや，WEB ホスティングサービス専用のネットワークについては，VLAN タグ，また HTTP プロトコルヘッダにおけるリクエストメソッド，レスポンスコードなどの情報が重要な属性と考えられる．このように，2 つの観点に基づいて決定したイベントの抽象化に用いる属性を，注目属性と呼ぶ．

3.3　第 (2) 段階：分析対象イベントの抽象化と特徴抽出

第 (1) 段階で決定した注目属性を用いて，個々のイベントを抽象化する．イベントの特徴は，イベント発生時のトラフィックデータにおける注目属性の属性値に表れる．そのため，イベント e_i をその発生時のトラフィックデータにおける属性値 r_l の生起確率 $\Pr(r_l|e_i)$ を用いて抽象化できる．しかしながら，イベントの抽象化に用いた属性値の数が膨大になりうるため，すべての属性値を用いたイベント間の比較が困難となる．

そこで，本論文では，属性値を適切に集約することにより，イベント間の比較を実現するため，コンポーネントの概念を導入する．コンポーネントとは，複数の分析対象イベント間に共通，または各イベントで特有に発生しているトラフィックデータである．イベント発生時のトラフィックデータは複数のコンポーネントにより構成される．すなわち，コンポーネントにおける属性値を 1 つの指標に集約することが可能となるため，イベントの比較という観点において，適切な属性値の集約を実現できる．

ここで，コンポーネントの数を M とし，個々のコンポーネントを $c_1, c_2, \cdots, c_M$，その集合を $\mathbb{C}$ と表記する．これにより，イベント e_i 発生時のトラフィックデータにおける属性値 r_l の生起確率 $\Pr(r_l|e_i)$ は式 (1) となる．

$$\Pr(r_l|e_i) = \sum_{k=1}^{M} \Pr(r_l|c_k) \Pr(c_k|e_i) \tag{1}$$

© 2009 Information Processing Society of Japan

3.4　第 (3) 段階：参照する既知のイベントの選択

既知のイベントの数は，分析対象のイベントと比較してきわめて多い．よって，イベントの分析価値の評価のために参照する既知のイベントとして，分析対象イベントと類似したものを適切に選択する必要がある．

まず，第 (2) 段階で述べたコンポーネント集合 $\mathbb{C}$ を用いて，既知のイベント $e'_j \in \mathbb{E}'$ を抽象化する．すなわち，イベント e'_j 発生時のトラフィックデータにおける属性値 r_l の生起確率 $\Pr(r_l|e'_j)$ は式 (2) となる．

$$\Pr(r_l|e'_j) = \sum_{k=1}^{M} \Pr(r_l|c_k) \Pr(c_k|e'_j) + \Pr(r_l|c_\varepsilon^j) \tag{2}$$

c_ε^j は，分析対象イベント N 件を構成するコンポーネント集合 $\mathbb{C}$ には含まれていない，既知のイベント e'_j に特有のコンポーネントである．すなわち，e'_j を構成するすべてのコンポーネントに占める，e'_j に特有のコンポーネントの割合により，e'_j と分析対象イベント N 件との類似性が判断できる．したがって，式 (3) の関係を満たす既知のイベント e'_j を，分析価値の評価の際に参照する既知のイベントとして選択する．ここで，参照する既知のイベントの数は閾値 th_{res} に依存し，そのイベントの数が分析価値の精度や計算時間に大きく影響するため，それらの点を考慮して th_{res} を決定する必要がある．また，L は属性値の数とする．

$$\frac{\sum_{l=1}^{L} \Pr(r_l|c_\varepsilon^j)}{\sum_{l=1}^{L} \Pr(r_l|e'_j)} \leq th_{res} \tag{3}$$

3.5　第 (4) 段階：トラフィックデータの情報量の導出

第 (3) 段階で選択した既知のイベントを参照し，対象イベント e_i 発生時のトラフィックデータが持つ情報量 $IG(e_i|\mathbb{E}')$ を導出する．式 (1)，式 (2) における $\Pr(r_l|c_k)$ がすべてのイベントで共通であり，また，$\Pr(r_l|e_\varepsilon^j)$ は既知のイベント e'_j における特有のコンポーネントである．よって，対象イベント e_i 発生時のトラフィックデータが持つ情報量 $IG(e_i|\mathbb{E}')$ は，一般的な情報量の概念に基づいて式 (4) となる．

$$IG(e_i|\mathbb{E}') = \min\left\{ \sum_{k=1}^{M} \Pr(c_k|e_i) \log \frac{\Pr(c_k|e_i)}{\Pr(c_k|e'_j)} \;\middle|\; \forall e'_j \in \mathbb{E}' \right\} \tag{4}$$

3.6　第 (5) 段階：イベントの分析価値の導出

第 (4) 段階で求めたイベント e_i 発生時のトラフィックデータが持つ情報量 $IG(e_i|\mathbb{E}')$ を，イベント e_i の分析価値とする．管理者はこの分析価値を分析を行うべきか否かの判断基準とする．これにより，多量のイベントが検出され，管理者がすべてのイベントを分析しきれない場合に，管理者が分析するイベントの数を大幅に抑制しつつ，多様な種類のイベントの分析を実現することで，ネットワーク上でのイベントの発生状況の把握が可能となる．

4.　提案手法に基づくイベント分析価値評価システムの設計

4.1　提案手法に基づくイベント分析価値評価システムの概要

図 1 に，提案手法に基づくイベント分析価値評価システムの概要を示す．イベント検出モジュールは，ネットワークトラフィックに基づいてイベントを検出することにより，イベント発生情報を出力する．イベント評価モジュールは，イベント検出モジュールによって出力されたイベント発生情報を用いて，3 章で提案したイベント分析価値の評価を行い，その結果を管理者に通知する．イベント評価モジュールは，(1) トラフィックデータの注目属性の決定，(2) 分析対象イベントの抽象化と特徴抽出，(3) 参照する既知のイベントの選択，(4) トラフィックデータの情報量の導出，(5) イベント分析価値の導出，の 5 つの機能を持つ．これらの 5 つの機能は，3 章で述べた 5 つの段階と対応する．以降，それぞれの機能の設計について述べる．

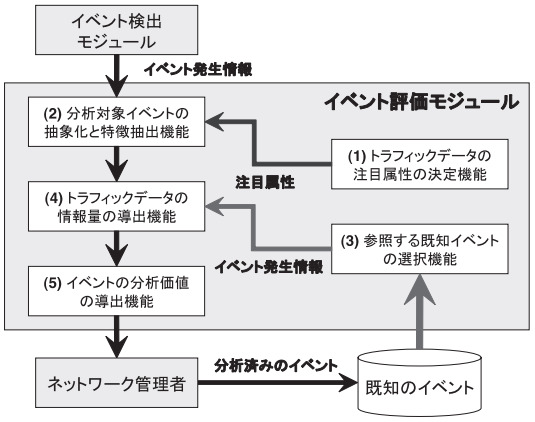


図 1　提案手法に基づくイベント分析価値評価システムの概要
Fig. 1　An overview of the event evaluation system based on proposed method.

© 2009 Information Processing Society of Japan

4.2　機能 (1)：トラフィックデータの注目属性の決定機能の設計

3.2 節で述べた観点に基づいて，管理者がイベント発生時のトラフィックから，イベントの抽象化に用いる注目属性を決定する．

4.3　機能 (2)：分析対象イベントの抽象化と特徴抽出機能の設計

4.3.1　イベントベクトルによるトラフィックデータの抽象化

イベントの特徴は，ネットワークトラフィックにおける複数の属性値の生起確率として抽象化される．よって，複数の属性値およびそれらの生起確率という，多次元の情報を統一的に扱うことが必要となる．このため，イベント発生時のネットワークトラフィックの抽象化に，式 (5) で示すイベントベクトル $\boldsymbol{e_i}$ を用いる．ここで，w^e_{ij} はイベント e_i の属性値 r_j の生起確率とする．

$$\boldsymbol{e_i} = (w^e_{i1}, w^e_{i2}, \cdots, w^e_{iL}) \tag{5}$$

これにより，イベントベクトル $\boldsymbol{e_1}, \boldsymbol{e_2}, \cdots, \boldsymbol{e_N}$ を用いて，分析対象イベント集合 $\mathbb{E}$ に含まれるすべてのイベントを，式 (6) のイベント行列 $\boldsymbol{E}$ で表記できる．ここで，$[\cdot]^T$ は行列の転置を表す．

$$\boldsymbol{E} = [\boldsymbol{e_1}, \boldsymbol{e_2}, \cdots, \boldsymbol{e_N}]^T \tag{6}$$

4.3.2　コンポーネントベクトルによるトラフィックデータの抽象化

分析対象イベントから，3.3 節で述べたコンポーネントを導出するにあたり，独立成分分析（ICA: Independent Component Analysis）[11] を用いる．ICA は複数の観測された混合信号を統計的に独立な信号に分離する手法である．すなわち，ここで観測された信号はイベントベクトルに相当し，独立な信号はコンポーネントによる各属性値の生起確率となる．この変化をコンポーネントベクトル $\boldsymbol{c_k}$ として式 (7) のように示す．

$$\boldsymbol{c_k} = (w^c_{k1}, w^c_{k2}, \cdots, w^c_{kL}) \tag{7}$$

これにより，コンポーネントベクトル $\boldsymbol{c_1}, \boldsymbol{c_2}, \cdots, \boldsymbol{c_M}$ を用いて，コンポーネント集合 $\mathbb{C}$ に含まれるすべてのコンポーネントを，式 (8) のコンポーネント行列 $\boldsymbol{C}$ で表記できる．

$$\boldsymbol{C} = [\boldsymbol{c_1}, \boldsymbol{c_2}, \cdots, \boldsymbol{c_M}]^T \tag{8}$$

式 (1) に基づいて，ICA によりイベント行列 $\boldsymbol{E}$ やコンポーネント行列 $\boldsymbol{C}$ の関係を定式化すると，次の式 (9) となる．

$$\boldsymbol{E}^T = \boldsymbol{X}\boldsymbol{C}^T \tag{9}$$

ここで，混合行列 $\boldsymbol{X}$ は個々のイベントにおけるコンポーネントの影響の大きさである係数となる．ICA を適用することで，コンポーネント行列 $\boldsymbol{C}$ やその係数を表す混合行列 $\boldsymbol{X}$ が未知の場合に，イベント行列 $\boldsymbol{E}$ のみを用いて，コンポーネント行列 $\boldsymbol{C}$ と混合行列 $\boldsymbol{X}$ を推定することが可能である．これにより，イベント行列 $\boldsymbol{E}$ を，コンポーネント行列 $\boldsymbol{C}$ と混合行列 $\boldsymbol{X}$ に分解すること，すなわち，イベントベクトル $\boldsymbol{e_i}$ を式 (10) のように，コンポーネントとその係数に分解することができる．

$$\begin{aligned}\boldsymbol{e_i} &= (w^e_{i1}, w^e_{i2}, \cdots, w^e_{iL}) \\ &= \sum_{k=1}^{M} x_{ik}\boldsymbol{c_k}\end{aligned} \tag{10}$$

ICA はその処理の性質上，あらかじめ求めるコンポーネントの数 M を決定する必要がある．そこで，式 (11) のように，無相関化時の第 M 主成分までの累積寄与率 CP_M が閾値 th_{cp} 以上の場合，コンポーネントの数を M とする．

$$CP_M \geq th_{cp} \tag{11}$$

4.4　機能 (3)：参照する既知のイベントの選択機能の設計

既知のイベント e'_j を式 (12) のように，イベントベクトル $\boldsymbol{e'_j}$ に変換し，前節で導出したコンポーネントベクトル $\boldsymbol{c_1}, \boldsymbol{c_2}, \cdots, \boldsymbol{c_M}$ とその係数に分解する．

$$\begin{aligned}\boldsymbol{e'_j} &= (w'^e_{j1}, w'^e_{j2}, \cdots, w'^e_{jL}) \\ &= \sum_{k=1}^{M} x'_{jk}\boldsymbol{c_k} + \boldsymbol{\varepsilon_j}\end{aligned} \tag{12}$$

$\boldsymbol{\varepsilon_j}$ は，分析対象イベント N 件を構成するコンポーネント集合 $\mathbb{C}$ には含まれていない，既知のイベント e'_j に特有のコンポーネントベクトルと，その係数の積である．すなわち，e'_j を構成するすべてのコンポーネントに占める，e'_j に特有のコンポーネントの割合により，e'_j と分析対象イベント N 件との類似性が判断できる．したがって，式 (13) の関係を満たす既知のイベント e'_j を，分析価値の評価の際に参照する既知のイベントとして選択する．ここで th_{res} は，3.4 節で述べた閾値，$\|\cdot\|$ はベクトルのノルムを表す．

$$\frac{\|\boldsymbol{\varepsilon_j}\|}{\|\boldsymbol{e'_j}\|} \leq th_{res} \tag{13}$$

4.5　機能 (4)：トラフィックデータの情報量の導出機能の設計

3.5 節で述べたように，イベント発生時のトラフィックデータが持つ情報量を導出する．情報量とは分析対象イベントと既知のイベントとの違いを定量的に評価した値である．すなわち，イベントの分析価値はイベント間の違いであり，それらは線形空間上ではベクトル間の距離によって導出される．よって，イベント e_i，e'_j のイベントベクトル間の距離を

© 2009 Information Processing Society of Japan

導出する関数を $dist(e_i, e'_j)$ とすると，イベント e_i におけるトラフィックデータの情報量は式 (14) で求められる．

$$IG(e_i|\mathbb{E}') = \min\left\{ dist(e_i, e'_j) \;\middle|\; {}^{\forall} e'_j \in \mathbb{E}' \right\} \tag{14}$$

本設計では，ベクトル間の距離を導出する関数としてコサイン尺度を用いた．イベント e_i，e'_j のイベントベクトル間のコサイン尺度を用いた距離は，式 (15) で表される．

$$dist(e_i, e'_j) = 1 - \left| \frac{\boldsymbol{e_i} \cdot \boldsymbol{e'_j}}{\|\boldsymbol{e_i}\| \, \|\boldsymbol{e'_j}\|} \right| \tag{15}$$

4.6　機能 (5)：イベント分析価値の導出機能の設計

前節で導出したイベント e_i 発生時のトラフィックデータが持つ情報量 $IG(e_i|\mathbb{E}')$ を，イベント e_i の分析価値とする．管理者はこの分析価値を分析を行うべきか否かの判断基準とする．

5.　実験および評価

5.1　実験の目的と概要

5.1.1　目　　的

提案手法が管理負担の軽減に大きく貢献できることを示すため，実運用ネットワークにおいて実験を行った．すなわち，ネットワーク上でのイベントの発生状況を効率良く把握できること，それにより管理者の作業負担の軽減に大きく貢献できることを検証した．

5.1.2　諸　　元

図 2 に実験環境を示す．LAN 内に 21 台のサーバを設置し，それぞれ HTTP，SMTP，DNS，SSH や PPTP などのサービスを提供している．これらは実運用されているサーバ群である．イベント検出モジュールでは，ルータで観測されるトラフィックデータに基づいてイベントを検出し，イベント発生情報を出力する．イベント評価モジュールでは，イベント発生情報を用いてイベントの分析価値を求め，その値に基づいて，管理者に通知するイベントを削減する．

表 1 に，評価に用いた提案手法および従来手法を示す．従来手法では，管理者がイベントの分析を行うべきか否か，また，どのイベントを優先的に分析するかの判断基準として，2.3 節で述べたように，異常検出に使われる異常度を用いた．また，提案手法では，既知のイベントを参照することで分析価値を調整した場合と，していない場合の 2 種類で評価を行うこととし，前者を提案手法 (A)，後者を提案手法 (B) とする．

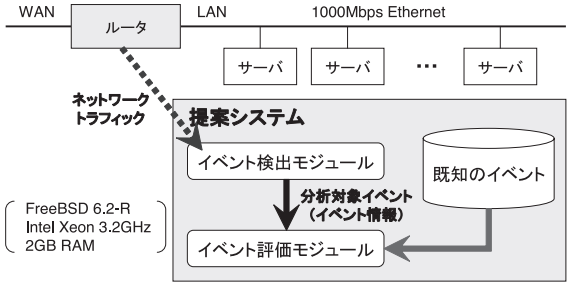


図 2　実験環境
Fig. 2　Experimental environment.

表 1　評価に用いた提案手法および従来手法
Table 1　Proposed method and traditional method for evaluation.

	判断基準	既知のイベント
従来手法	異常度	参照しない
提案手法 (A)	分析価値	参照しない
提案手法 (B)	分析価値	参照する

実験では，3.2 節で述べた注目属性として，送信アドレスとポート番号，および受信アドレスとポート番号を用いた．4.3.2 項で述べたコンポーネントの導出に用いる独立成分分析のアルゴリズムとして FastICA [12] を用い，FastICA における Neg-Entropy 推定式を $G(u) = 1/\alpha \log\cosh(u\alpha)$，その初期値を $\alpha = 1$，コンポーネント数 M の決定に用いる累積寄与率を $th_{cp} = 0.99$ とした．これは，属性値の集約によって失われる情報を最小限に抑制するためである．また，4.4 節で述べた，既知のイベントの選択基準である閾値は，経験的に $th_{res} = 0.7$ とした．このパラメータの最適化については今後の課題とする．

5.1.3　イベントデータ

実験には次の 2 種類のイベントデータ，すなわちイベントとトラフィックデータの組を用いた．

イベントデータ (1)：図 2 の環境で観測されたトラフィックデータに，仮想的に生成したイベント，および他のネットワークで観測されたイベント[13),14)] 時のトラフィックデータを加え生成した．また，個々のイベントには，イベントの種類を示す (1)-1 から (1)-5 のイベントラベルを付与した．Backscatter，Witty Worm，W32/Blaster，Host Scan は，それぞれ注目属性として用いた送信アドレス，送信ポート，受信ポート，受信アドレスに，また

© 2009 Information Processing Society of Japan

表 2　イベントデータ (1) の分析対象イベントの分類
Table 2　Classification of unknown events of event data (1).

イベントの発生原因	Backscatter	Witty Worm	W32/Blaster	Host Scan	Video Streaming	—
イベントラベル	(1)-1	(1)-2	(1)-3	(1)-4	(1)-5	total
分析対象イベントの数	10	10	10	10	10	50

表 3　イベントデータ (2) の分析対象イベントと既知のイベントの分類
Table 3　Classification of known and unknown events of event data (2).

イベントラベル	(2)-1	(2)-2	(2)-3	(2)-4	(2)-5	(2)-6	(2)-7	other	total
分析対象イベントの数	5	2	8	18	15	32	2	—	82
既知のイベントの数	1	0	3	18	32	44	29	132	259

Video Streaming は，それらすべてに特徴が現れる発生原因として選択した．**表 2** にイベントデータ (1) に含まれる，イベントの分類を示す．

イベントデータ (2)：図 2 の環境で観測されたトラフィックデータに対し，文献 2) で提案されているアルゴリズムを適用することで検出した．この手法は，送信アドレスとポート番号，および受信アドレスとポート番号に対するパケット数の分布を主要成分と残差成分，すなわち異常成分に分離し，その異常成分の大きさに基づいてイベントを検出する．したがって，この手法で検出されたイベントの特徴は，本実験で用いた注目属性に現れる．

分析対象のイベントとして，2007 年 1 月 16 日に検出されたイベント 82 件（$N = 82$），既知のイベントとして，2007 年 1 月 6 から同年 1 月 15 日の 10 日間に検出されたイベント 259 件（$N' = 259$）を用いた．本論文では，多量のイベントが検出される状況を想定しているため，あらかじめ 2007 年 1 月の 1 カ月間について，1 日ごとのイベントの検出数を調査し，分析対象イベントと既知のイベントの数が最も多くなる日を選択した．

これらの分析対象イベントを，あらかじめ手作業で分析し，その発生原因ごとに分類した結果，7 種類に分類できた．次に，これらのイベントの分類を一意に識別するため，その発生原因ごとにイベントラベル (2)-1 から (2)-7 を付与した．たとえば，イベントラベルの (2)-6 の発生原因は，WEB ログ解析による多量の DNS アクセスの発生であり，注目属性である送受信ポート番号に変化が現れているものであった．加えて，既知のイベントを同様に分析することで，前述の 7 種類とその他に分類した．**表 3** に，イベントデータ (2) に含まれる，分析対象イベントと既知のイベントの分類を示す．

5.2　分析価値の正確性の評価

あらかじめイベントの発生原因が分かっているイベントデータ (1) を分析対象イベントとして，提案手法によりイベントの分析価値が正確に付与できるか確認する．**表 4** に，提案手法 (A) によるイベントデータ (1) の分析価値とその順位を示す．分析価値の大きいイベントを，降順で 5 位まで示している．

表 4　提案手法 (A) によるイベントデータ (1) の分析順
Table 4　Order of the events of event data (1) to analyze in proposed method (A).

順位	分析価値	イベントラベル
1	1.000	(1)-1
2	0.741	(1)-3
3	0.713	(1)-5
4	0.680	(1)-4
5	0.669	(1)-2

順位が 5 位までのイベントには，同じイベントラベルのイベントが存在せず，イベントの違いを正確に判別し，各ラベルごとに 1 つのイベントのみに高い分析価値を与えている．すなわち，提案手法により，同様の原因により検出されたイベントと，そうでないイベントの違いを，正確に判断可能であることを確認した．

5.3　分析価値の有効性および正確性の評価

実環境において検出されたイベントデータ (2) を分析対象イベントとして，イベントの発生状況を効率良く把握できるか確認する実験を行った．イベントの発生状況の把握とは，管理者が表 3 のイベントラベル (2)-1 から (2)-7 について，それぞれ発生原因が未知である場合に，少なくとも 1 つのイベントを分析し，それらすべての発生原因を把握することである．管理者が，ある発生原因を把握済みの場合，それと同一の原因で発生したイベント，すなわち，同一のイベントラベルが付与されたイベントの分析は不要となる．本実験では，提案手法が既存手法に比べて，このイベントの発生状況の把握をより効率的に，少ない数のイベントの分析で実現できることを確認する．

表 5，**表 6**，**表 7** に，それぞれ従来手法，提案手法 (A)，提案手法 (B) によるイベントデータ (2) の評価値とその順位を示す．各手法における判断基準の値，すなわち評価値の，降順の順位で 10 位までのイベント，および各イベントラベルごとに最も評価値が高いイベントについて，順位，発生時間，イベントラベル，異常度または分析価値の評価値を示している．また，**図 3** に，各手法におけるイベントの評価値とその順位を示す．

表 5 より，従来手法では，順位が 1 位から 5 位に (2)-1 と (2)-3 のイベントが，7 位から 10 位に (2)-4 のイベントが集中している．これは，同一ラベルのイベントに対して，同等の異常度が付与されたためである．また，(2)-5 や (2)-2，(2)-7 のイベントは，最高で 30 位

© 2009 Information Processing Society of Japan

表 5　従来手法によるイベントデータ (2) の評価値とその順位
Table 5　Order of the events of event data (2) to analyze in traditional method.

順位	発生時間	評価値（異常度）	イベントラベル
1	Jan 16 01:15:00	3.172	(2)-1
2	Jan 16 01:20:00	3.132	(2)-1
3	Jan 16 01:05:00	2.972	(2)-3
4	Jan 16 20:05:00	2.954	(2)-1
5	Jan 16 10:35:00	2.933	(2)-3
6	Jan 16 19:35:00	2.882	(2)-6
7	Jan 16 18:05:00	2.823	(2)-4
8	Jan 16 03:30:00	2.813	(2)-4
9	Jan 16 17:55:00	2.758	(2)-4
10	Jan 16 17:50:00	2.732	(2)-4
28	Jan 16 05:05:00	2.351	(2)-5
30	Jan 16 05:15:00	2.322	(2)-2
35	Jan 16 10:40:00	2.037	(2)-7

表 6　提案手法 (A) によるイベントデータ (2) の評価値とその順位
Table 6　Order of the events of event data (2) to analyze in proposed method (A).

順位	発生時間	評価値（分析価値）	イベントラベル
1	Jan 16 05:15:00	1.000	(2)-2
2	Jan 16 22:20:00	0.834	(2)-1
3	Jan 16 10:35:00	0.821	(2)-3
4	Jan 16 10:40:00	0.797	(2)-7
5	Jan 16 17:55:00	0.784	(2)-4
6	Jan 16 17:00:00	0.756	(2)-5
7	Jan 16 12:20:00	0.683	(2)-3
8	Jan 16 02:55:00	0.557	(2)-6
9	Jan 16 20:05:00	0.121	(2)-1
10	Jan 16 02:50:00	0.114	(2)-6

付近に存在する．この結果より，イベントの分析を行うか否かの判定基準として，従来手法によるイベントの異常度の大きさを使用する場合，管理者が，適切にイベントの発生状況を把握するために，1 位から 35 位までのイベントの分析を行う必要がある．

一方，表 6 より，提案手法 (A) では (2)-1 から (2)-7 までの 7 種類のイベントが，8 位までに含まれている．よって，イベントの分析を行うか否かの判断基準として，提案手法 (A) によるイベントの分析価値の大きさを使用する場合，管理者が，適切にイベント発生状況を把握するために，1 位から 8 位までのイベントの分析を行う必要がある．この結果から，提案手法 (A) は従来手法と比較して，イベントの発生状況を把握するために，分析するイベントの数を 7 割以上削減可能であることを確認した．これは，提案手法 (A) は既知のイベントを参照しないこと，すなわち管理者によってすでに把握されているイベントが存在しないことを考慮すると，不必要なイベントの分析が 1 件のみであり，十分に効率的なイベントの発生状況の把握が実現できたといえる．

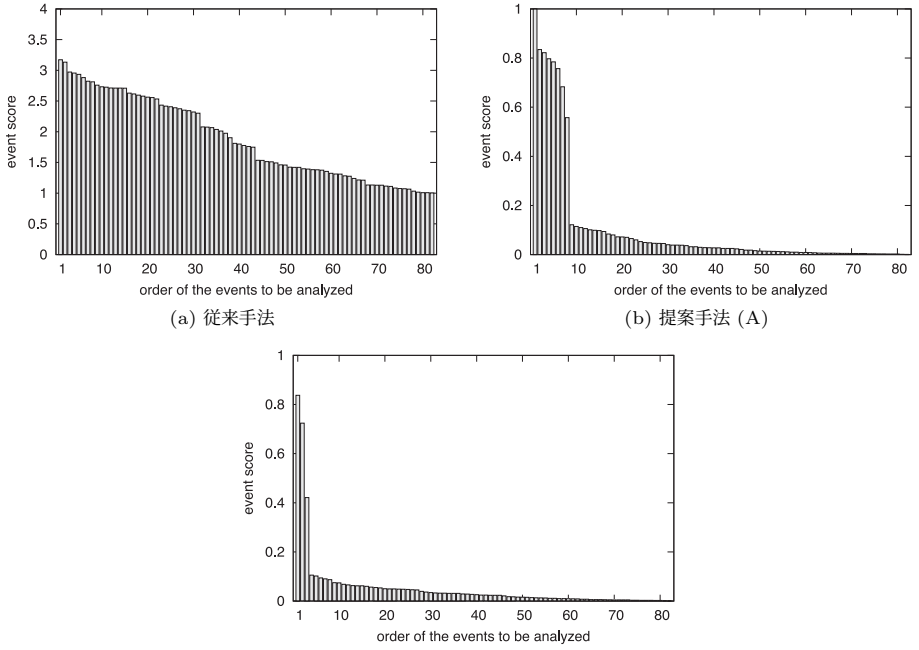
表 7　提案手法 (B) によるイベントデータ (2) の評価値とその順位
Table 7　Order of the events of event data (2) to analyze in proposed method (B).

順位	発生時間	評価値（分析価値）	イベントラベル
1	Jan 16 05:15:00	0.837	(2)-2
2	Jan 16 22:20:00	0.723	(2)-1
3	Jan 16 10:35:00	0.421	(2)-3
4	Jan 16 10:40:00	0.105	(2)-7
5	Jan 16 17:55:00	0.101	(2)-4
6	Jan 16 17:00:00	0.093	(2)-5
7	Jan 16 12:20:00	0.090	(2)-3
8	Jan 16 02:55:00	0.087	(2)-6
9	Jan 16 20:05:00	0.074	(2)-1
10	Jan 16 02:50:00	0.073	(2)-6

表 6 と表 7 を比較すると，提案手法 (B) では，既知のイベント中にもあまり存在しない，イベントラベル (2)-1，(2)-2，(2)-3 の各 1 件のイベントは分析価値が高いまま維持されているが，既知のイベント中に多数存在している他のイベントラベルのイベントは，分析価値が大幅に減少している．この結果は，既知のイベントが多数存在することで，そのイベントラベルが付与されたイベントの発生原因が，すでに管理者によって把握されているためである．すなわち，管理者が，数種類のイベントの発生原因を把握している場合においても，未知のイベントのみの分析を可能とし，効率的なイベント発生状況の把握が実現できたといえる．

図 3 (a) より，従来手法では順位が下がるほどに，評価値が線形に減少している．それに対して，図 3 (b) より，提案手法 (A) では 8 位までの少ないイベントのみが高い分析価値を持っており，9 位以下のイベントの分析価値が大幅に減少している．また，図 3 (c) より，提案手法 (B) では，提案手法 (A) の結果から，さらに既知のイベントの少ないイベントのみが高い分析価値を持っており，4 位以下のイベントの分析価値が大幅に減少している．これにより，提案手法では従来手法と比較して，イベントの発生状況を把握するために，分析の必要があるイベントの判定が容易である．

© 2009 Information Processing Society of Japan



(a) 従来手法　(b) 提案手法 (A)

(c) 提案手法 (B)

図 3　各手法におけるイベントの分析順とその分析価値
Fig. 3　Order and evaluation value of events analyze in each method.

以上の結果より，提案手法により求めた分析価値の高い順に，イベントを分析することで，管理者が分析する対象イベントの数を大幅に抑制しつつ，多様な種類のイベントを分析し，イベントの発生状況の効率的な把握が可能であることを確認した．これにより，提案手法が管理負担の軽減に大きく貢献できることを確認した．

6. おわりに

本論文では，既存の異常検出手法により検出された多量のイベントに対し，管理者によるイベントの発生状況の効率的な把握を支援する「イベント分析価値評価手法」を提案した．また，実験を通じ，提案手法により導出された分析価値に基づいてイベントを選択することで，管理者が希少なイベントを見逃すことや，類似したイベントを何度も分析することなく，ネットワーク上でのイベントの発生状況を効率良く把握できること，それにより管理者の作業負担の軽減に大きく貢献できることを示した．

今後の課題として，イベントの分析価値を評価するうえで，最適なトラフィックデータの注目属性の決定方法，およびその自動化に関して検討を行う予定である．

謝辞　本研究の一部は，科学研究費補助金（19200005）の援助を受けて実施した．

参 考 文 献

1) Martin-Flatin, J.P., Jakobson, G. and Lewis, L.: Event Correlation in Integrated Management: Lessons Learned and Outlook, *Journal of Network and Systems Management*, Vol.15, No.4 (2007).
2) Lakhina, A., Crovella, M. and Diot, C.: Mining Anomalies Using Traffic Feature Distributions, *Proc. ACM SIGCOMM*, pp.169–180 (2005).
3) Barford, P., Kline, J., Plonka, D. and Ron, A.: A Signal Analysis of Network Traffic Anomalies, *Proc. ACM SIGCOMM*, pp.71–82 (2002).
4) 石黒正揮，鈴木裕信，村瀬一郎，篠田陽一：インターネット上の脅威分析を支援する空間および時間的な特徴量に基づく分析手法，情報処理学会論文誌，Vol.48, No.9, pp.3148–3162 (2007).
5) 向坂真一，小池英樹：内部ネットワーク監視を目的とした時間・論理・地理情報の統合的視覚化システム，情報処理学会論文誌，Vol.49, No.1, pp.503–512 (2008).
6) 高田哲司，小池英樹：見えログ：情報視覚化とテキストマイニングを用いたログ情報ブラウザ，情報処理学会論文誌，Vol.41, No.12, pp.3265–3275 (2000).
7) 長尾真宏，北形　元，菅沼拓夫，白鳥則郎：ネットワーク管理におけるイベントのリアルタイム識別の実現のためのログ要約手法の提案と評価，情報処理学会論文誌，Vol.48, No.4, pp.1606–1615 (2007).
8) Estan, C., Savage, S. and Varghese, G.: Automatically Inferring Patterns of Resource Consumption in Network Traffic, *Proc. ACM SIGCOMM*, pp.137–148 (2003).
9) 磯部隆史，渡辺義則，樋口秀光，相本　毅，吉田健一：広域ネットワーク網向け異常通信の探知機能の検討，電子情報通信学会技術研究報告 IN，情報ネットワーク，Vol.105, No.178, pp.109–114 (2005).
10) Ertöz, L., Eilertson, E., Lazarevic, A., Tan, P.-N., Dokas, P., Kumar, V. and Srivastava, J.: Detection and Summarization of Novel Network Attacks Using Data Mining, Army High Performance Computing Research Center Technical Report (2003).
11) Hyvärinen, A. and Oja, E.: Independent Component Analysis: Algorithms and Applications, *Neural Networks*, Vol.13, pp.411–430 (2000).
12) Bingham, E. and Hyvärinen, A.: A Fast Fixed-Point Algorithm for Independent

© 2009 Information Processing Society of Japan

Component Analysis of Complex Valued Signals, *Neural Systems*, Vol.10, No.1, pp.1–8 (2000).
13) Shannon, C., Moore, D. and Aben, E.: The CAIDA Backscatter-2007 Dataset – January 2007 – November 2007. http://www.caida.org/data/passive/backscatter_2007_dataset.xml
14) Shannon, C. and Moore, D.: The CAIDA Dataset on the Witty Worm – March 19–24, 2004. http://www.caida.org/data/passive/witty_worm_dataset.xml. Support for the Witty Worm Dataset and the UCSD Network Telescope are provided by Cisco Systems, Limelight Networks, the US Department of Homeland Security, the National Science Foundation, DARPA, Digital Envoy, and CAIDA Members.

(平成 20 年 6 月 10 日受付)
(平成 20 年 12 月 5 日採録)

佐藤　彰洋（学生会員）
2008 年東北大学大学院情報科学研究科博士前期課程修了．現在，同大学院同研究科博士後期課程在学中．ネットワーク運用管理に関する研究に従事．

長尾　真宏（学生会員）
2006 年東北大学大学院情報科学研究科博士前期課程修了．現在，同大学院同研究科博士後期課程在学中．ネットワーク運用管理に関する研究に従事．

小出　和秀
2006 年東北大学大学院情報科学研究科博士後期課程修了．同年（独）情報通信研究機構研究員．2007 年東北大学電気通信研究所助教．ネットワーク管理，トラフィック計測，モバイルネットワークに関する研究に従事．博士（情報科学）．電子情報通信学会員．

木下　哲男（正会員）
1979 年東北大学大学院修士課程修了．同年沖電気工業（株）入社．1996 年東北大学電気通信研究所助教授，2001 年同大学情報シナジーセンター教授，現在，サイバーサイエンスセンター教授．知識工学，エージェント工学，知識型設計支援，エージェント応用システム等の研究開発に従事．情報処理学会平成元年度研究賞および平成 8 年度論文賞，電子情報通信学会平成 13 年度業績賞等．工学博士．電子情報通信学会，人工知能学会，日本認知科学会，IEEE，ACM，AAAI 各会員．

白鳥　則郎（フェロー）
1977 年東北大学大学院博士課程修了．1984 年同大学助教授（電気通信研究所）．1990 年同大学教授（工学部情報工学科）．1993 年同大学教授（電気通信研究所）．人と IT 環境の共生の研究に従事．本会 25 周年記念論文賞受賞．本会マルチメディア通信と分散処理研究会主査，本会理事，本会副会長，本会フェロー，IEEE フェロー．

© 2009 Information Processing Society of Japan