社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DPS－130 (17)
2007－CSEC－ 36 (17)
2007／3／1

不正アクセスのトラフィックによるセンサの独立性

福野 直弥† 菊池 浩明† 寺田 真敏†† 土居 範久¶

†東海大学大学院
259-1292 平塚市北金目 1117
fukuno,kikn@ep.u-tokai.ac.jp
†† 日立製作所 Hitachi Incident Response Team (HIRT)
212-8567 神奈川県川崎市幸区鹿島田 890 日立システムプラザ新川崎
¶中央大学理工学部情報工学科
112-8551 東京都文京区春日 1-13-27

あらまし インターネット上で分散観測したパケットからインシデントの傾向や特徴を解析，予測する分散観測システムが広く試みられている．この解析の精度はセンサ間の独立性に左右する．そこで本研究では，センサのポートごとのトラフィックに，TF-IDF や主成分分析等の指標を適用し，識別に重要なポート，送信元アドレス，センサを抽出する方法を研究する．

Research of Sensor Independently Using Malicious Access Traffic

Naoya Fukuno† Hiroaki Kikuchi † Masato Terada†† Norihisa Doi¶

†Couse of Information Engineering, Graduate School of Engineering Tokai University
1117 Kitakaname, Hiratsuka, Kanagawa 259-1292
††Hitachi, Ltd. Hitachi Incident Response Team (HIRT)
890 Kashimada, Kawasaki, Kanagawa 212-8567
¶Dept. of Info. and System Engineering, Faculity of Scienece and Engineering, Chuo University
1-13-27 Kasuga, Bunkyo, Tokyo 112-8551

Abstract Observations of packets sent over the Internet with distributed sensors have been widely attempted in order to estimate and analyze the characteristics of intrusions. The accuracy of estimation, however, depends on the distribution sensors. In this paper, we propose a new method for analyzing independence of sensors given statistics of port, source-IPaddress and sensors based on the scan logs observed by JPCERT/CC and our own sensors using the TF-IDF (Term Frequency - Inverse Document Frequency) and PCA (Principal Component Anaysis) and evaluate the propose method.

1 はじめに

インターネット上にセンサを複数設置することで，パケットを分散観測して，インシデントの傾向や特徴を解析，予測する定点観測システムがある．この解析の精度はセンサ間の独立性に左右する．[1] によれば，ワームやウィルスに感染した不正ホストが放つパケットの宛先は，TCP では IP アドレスの上位 16 ビットが一致している比率は 58.9% であり，上位 8 ビットが一致する比率は 82.4%，という報告がある．つまり，センサのインターネット上での設置場所によってはインターネット全体から考えた場合，一部のインシデントの傾向や特徴しか採取できないことになる．しかし，インターネット上に無数のセンサを配置するのは難しい．そこで，センサが設置されていないネットワークについては，そのネットワークへのトラフィックを推定する [2] などが行われている．

また，実際に不正ホストが感染活動を行う際のア

クセス先ポートは不正ホストによってそれぞれ異なる．それゆえに，ネットワーク管理者は代表的な特定のポートを閉じることでインシデントの軽減を図っている．センサといっても管理は全て同一ではなく，ポートがフィルタリングされていたり，センサのIPアドレスが固定のものや変動してしまう等，各種特徴を持っている．

そこで本研究では．各センサやポート別パケット数や送信元アドレスのパケット数にTF-IDFや主成分分析等の指標を適用し，解析を行う上で重要なポート，送信元アドレス，センサを識別する方法を研究する．

2 提案方法

2.1 基本定義

不正ホストとは，ウィルスやネットワークワームなどにより他のホストへのスキャン（ポートスキャン等）を仕掛けるホストである．センサとは，不正ホストからのスキャンを観測する正規ホストであり，決して感染しない．n 台のセンサからなるセンサの集合を $S = \{s_1, s_2, ..., s_n\}$ とする．

センサが観測するパケットには，あて先のポート番号と送信元のIPアドレスが格納されている．これらのポートの集合を $P = \{p_1, p_2, ..., p_m\}$，送信元のIPアドレスの集合を $A = \{a_1, a_2, ..., a_\ell\}$ とする．P と A はの空間は $IPv4$ の場合で各々最大で 2^{16}，2^{32} であるが，ここでは観測出来たものだけを扱う．すなわち，$m \ll 2^{16}$, $\ell \ll 2^{32}$ である．IPアドレスは4つのオクテットから成り立つ．また，本研究では，上位の2つのオクテットが同じ場合，下位の2つのオクテットが異なってもその不正ホストからのパケット数としてカウントする．例えば，$a = 221.10$ は 221.10.0.0 から 221.10.255.255 までの65336個のアドレスを含む．

2.2 TF-IDFによる重要度

ある期間に T において，センサ s_i で観測されたパケットのうち，あて先ポートが p_j であるものの数を c_{ij}，送信元アドレスが a_k であるものの数を b_{ik} で表す．従って，センサ s_i の観測データは，ポートと送信元アドレスの空間上の2つの列ベクトル

$$\boldsymbol{c_i} = \begin{pmatrix} c_{i1} \\ \vdots \\ c_{im} \end{pmatrix}, \boldsymbol{b_i} = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{im} \end{pmatrix}$$

で特徴づけられる．P や A はの空間は大きすぎて，そのほんの一部にしかスキャンパケットは届かない．そこで，ここでは自然言語の手法を用いて，P や A の要素を順序付けし，意味のある部分集合を求める方法を提案する．TF-IDF(Term Frequency - Inverse Document Frequency)[4] とはある単語が複数の文書から重みを調べるもので，その単語がもつ特徴量となる．本研究では，分類する対象のポートと送信元アドレスの出現頻度を索引語頻度TFとみなし，センサの集合を文書と解釈して，文書頻度DFを求める．n 台のセンサ全体では $C = (\boldsymbol{c_1}, ..., \boldsymbol{c_n})$, $B = (\boldsymbol{b_1}, ..., \boldsymbol{b_n})$ の $m \times n$ と $b \times n$ の行列で表すことが出来る．ポート p_j についての出現頻度 $TF(p_j)$ を

$$TF(p_j) = \frac{1}{n} \sum_{i=1}^{n} c_{ij}$$

対するセンサ頻度 $DF(p_j)$ を

$$DF(p_j) = |\{c_{ij} \in C | c_{ij} > 0, i \in [n]\}|$$

と定める．ここで，$[n] = \{1, .., n\}$ とする．TFが高いポートは，よくスキャンされる代表的かつ特徴的なポートであり，観測対象として重要と考える．一方，DFが高いポートは，どのセンサでも普通に観測される「ありふれた」ポートであり，それ故に重要でないと解釈する．この2つの指標を合わせて

$$TF\text{-}IDF(p_j) = TF(p_j) \times \log_2(\frac{n}{DF(p_j)} + 1)$$

を定義する．DFの逆数の対数を取っているのは，値の変動を小さくするためである．定数1は，全センサで観測されたポートを0にしないために用いている．TF-IDFの値を高くするポートは，センサを特徴づけるポートと言うことができる．以後の実験では，TF-IDF値の高い20件を用いて，$m = 20$ の C を用いる．

ポートと同様にして，送信元アドレス a_k についても

$$TF(a_k) = \frac{1}{n} \sum_{i=1}^{n} c_{ik}$$

$$DF(a_k) = |\{b_{ik} \in B | b_{ik} > 0, i \in [n]\}|$$

を定めることで，センサの特徴付けに利用することができる．

2.3 主成分分析による分類

TF-IDFによってセンサを識別するのに主要なポートや送信元アドレスを絞り込むことが出来るが，20の次元はまだ大きすぎる．そこで，多次元データ解析の手法の一つである主成分分析を導入する．観測された基礎データを「それぞれの意味を考えて組み直す」ということで，方法論として重要な機能を持つ[3]．データから互いに無関係の因子(主成分)を取り出し，観測データをそれらの因子の線形結合で表す．例えば，センサ s_i における各ポートのパケット数が

$$\boldsymbol{c_i} = \begin{pmatrix} c_{i1} \\ \vdots \\ c_{im} \end{pmatrix}$$

で与えている時，

$$v_1 = \boldsymbol{u_i^T} \cdot \boldsymbol{c_i} = u_{11}c_{i1} + ... + u_{1m}c_{im}$$

で求まる新たな指標を v_1 を定義する．ここで，$u_{11}, ..., u_{1m}$ は係数である．主成分分析はこの u_1 の分散を最大化することで，$\boldsymbol{u_1}$ を定める．次に，$\boldsymbol{u_1}$ に直交して分散を最大化する $\boldsymbol{u_2}$ を求め，同様にして，$\boldsymbol{u_3}$, $\boldsymbol{u_4}$ の直交ベクトルを求めていく．$\boldsymbol{u_1}$, $\boldsymbol{u_2}$... は，$\boldsymbol{C}$ についての相関行列 $\boldsymbol{R}$

$$\boldsymbol{R} = \frac{1}{n}\sum_{i}^{n} \tilde{c}_i \tilde{c}_i^{\top}$$

についての固有値 $\boldsymbol{\lambda_1}$,$\boldsymbol{\lambda_2}$,...($\boldsymbol{\lambda_1 > \lambda_2 >}$ ...) と固有ベクトル $\boldsymbol{u_1}$, $\boldsymbol{u_2}$... により求まる．ただし，$\tilde{c_i}$ は $\boldsymbol{c_i}$ を平均0，分散1に標準化したベクトル，すなわち，$\boldsymbol{j = 1, ..., m}$ について

$$\tilde{c}_{ij} = \frac{c_{ij} - \mu(c_i)}{\sigma(c_i)}, \mu(c_i) = \frac{1}{m}\sum_{j=1}^{m} c_{ij},$$

$$\sigma(c_i)^2 = \frac{1}{m}\sum_{j=1}^{m}(c_{ij} - \mu(c_i))^2$$

とする．

ポートとセンサについてのパケット数 $\boldsymbol{C}$ から相関行列の固有ベクトルから定まる主軸(m行の列ベクトル) $\boldsymbol{u_1(C)}$,$\boldsymbol{u_2(C)}$，及び主成分を

$$\boldsymbol{V(C)} = \begin{pmatrix} \boldsymbol{v_1(C)} \\ \boldsymbol{v_2(C)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{u_1(C)}^{\top} \\ \boldsymbol{u_2(C)}^{\top} \end{pmatrix} \boldsymbol{C}$$

で表す．(2行n列の行列)．同様にして，送信元アドレスについてのパケット数 $\boldsymbol{B}$ から定まる主軸を $\boldsymbol{u_1(B)}$,$\boldsymbol{u_2(B)}$，その主成分を

$$\boldsymbol{V(B)} = \begin{pmatrix} \boldsymbol{v_1(B)} \\ \boldsymbol{v_2(B)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{u_1(B)}^{\top} \\ \boldsymbol{u_2(B)}^{\top} \end{pmatrix} \boldsymbol{B}$$

とする．こうして，n個のセンサ集合は，ポートと送信元アドレスについて各々2次元のベクトルで表される．

ここで，$\boldsymbol{C}$ と $\boldsymbol{B}$ を転置して同様に主成分分析を行うと，今度はセンサの集合から，2通りの因子を抽出できる．こうして各々，主軸 $\boldsymbol{u_1(C^{\top})}$, $\boldsymbol{u_2(C^{\top})}$, $\boldsymbol{u_1(B^{\top})}$, $\boldsymbol{u_2(B^{\top})}$ と主成分 $\boldsymbol{V(C^{\top})}$, $\boldsymbol{V(B^{\top})}$ を得る．結局，パケット数についてのデータ $\boldsymbol{C}$ と $\boldsymbol{B}$ から4通りの主成分分析が実行出来る．それぞれの解釈について，次章で考察する．

3 実験方法

3.1 2種類の解析データ郡

提案解析法の妥当性をみるために，次の2種類の観測データを用いた．

1. 既知センサ．東海大学のサーバで分散観測のログを管理しているセンサである．センサ数 $\boldsymbol{n}$ は8台あり，観測期間は2006年11月30日- 2007年1月12日 ($\boldsymbol{T_1 = 44}$) で，ポートのフィルタリングやDHCPでアドレスが変動するセンサが混在している．詳細なデータを表1に示す．

2. ISDAS．ISDAS(Internet Scan Data Acquisition System) は，JPCERT/CCが運用しているセキュリティに関するトラフィックの分散観測システムである[7]．単位時間当たりの主要なポート (135, 445, 8080, 1026, ICMP) のトラフィックなどを定期的にWebで公開している[7]．

   本実験では，2005年10月1日-2006年3月30日 ($\boldsymbol{T_2 = 182}$) の $\boldsymbol{n = 30}$ 台のセンサのログデータを解析する．観測センサのパケット数を表12に示す．ただし，本センサの接続形態やアドレス，動的割当等については非公開である．各センサごとに観測するデータに偏りがある．例えば，パケット数では $\boldsymbol{s_{15}}$ と $\boldsymbol{s_{26}}$ で約300倍の開きがあり，それぞれのセンサの特徴の一つを示している．

3.2 TF-IDF によるポートとアドレスの重要度

ポートと送信元アドレスの TF-IDF 値を求めた．上位の 20 件について，表 2,3,4,5 に示す．

表 2 における $\boldsymbol{DF(p_j)}$ に着目すると，フィルタされているポートがわかる．代表的でないポート (例えば 5 位の **23556** など) が上位に上がっているものがある．$\boldsymbol{TF(23556) = 122}$ なので観測頻度はそれおほど高くないが，$\boldsymbol{DF(23566) = 1}$ よりセンサの特徴をとらえていると解釈されて，重要度を上げている．表 2 と表 4 を比較すると，全般的に $\boldsymbol{TF(p_j) \geqq TF(a_k)}$ である．ポートの方が空間が小さいためであろう．表 2 と表 3 を比較すると第 1 位の 135 は共通で，20 位までで共通しているのは，135,1433 などの 10 ポート (**50%**) であった．445 は ISDAS では 2 位だが既知センサではフィルタリングされている．DF の低いものはバックドアに使用されている可能性が高いと考えられる．表 3,5 を見ると，既知のセンサと比べ全体的に値が高い結果となった．$\boldsymbol{p_j}$ や $\boldsymbol{a_k}$ の $\boldsymbol{DF(p_j)}$ が高く，どのセンサでも観測するポートや送信元アドレスであることがわかる．

3.3 既知センサの主成分分析

TF-IDF を求めたポート別のパケット数と送信元アドレスを主成分分析し，$\boldsymbol{C}$ と $\boldsymbol{B}$ の主軸，すなわち得られた因子の大きさを表 6，表 8 に各々示す．各因子から因子から求めた第一主成分と第二主成分によるセンサの分布を図 1，2 に示す．図 1 のセンサの分布を見ると $\boldsymbol{v_1(C)}$ と $\boldsymbol{v_2(C)}$ の両方が負の値でセンサ $\boldsymbol{s_{103} - s_{106}}$ は大学 2 で観測していることがわかる．更に，ADSL によるセンサ群 $\boldsymbol{s_{107}}$，$\boldsymbol{s_{108}}$，CATV によるセンサ $\boldsymbol{s_{102}}$ が各々分離していることが観察できる．この ADSL 郡に共通している特徴はポート 135 へのパケット数であり，これは，表 6 の因子の大きさを見ても **0.33** であり，第一主成分 $\boldsymbol{v_1(C)}$ を支配していることが推論できる．135 は Windows RPC の脆弱性を利用するワームの対象であり代表的な「Blaster 軸」と考える．一方，第二主成分 $\boldsymbol{v_2(C)}$ を決めていたのは，10421 などのポートだが，表 2 を見ると $\boldsymbol{DF(10421) = 1}$ なので $\boldsymbol{s_{102}}$ に固有なポートである．従って，「ランダムのバックドア軸」と解釈する．

同様にして，図 1 の送信元アドレスの主成分上のセンサの分布を分析すると，やはり，大学内のセンサと ISP のセンサ ($\boldsymbol{s_{107}, s_{108}}$) に分離できる．ただし，このときの主軸は，表 4 より第一主成分 $\boldsymbol{v_1(B)}$ は 219.97 のアドレス，第二主成分は $\boldsymbol{v_2(B)}$ は 59.147 の送信元アドレスのパケット数が効いていることがわかった．従って，第一は「クラス C 軸」，第二は「クラス A 軸」と考えることもできよう．図 1,2 を比べると $\boldsymbol{s_{107}}$,$\boldsymbol{s_{108}}$ の分布が大きく食い違っている．あて先ポート ($\boldsymbol{C}$) を見ると，2 つのセンサは近いと判断したが，送信元 ($\boldsymbol{B}$) で主成分分析を行うと離れている．これはポートだけではなく，送信元アドレスのばらつきが影響を与えていると考えられる．

表 2: 既知センサが観測したポートの $\boldsymbol{TF\text{-}IDF(p_j)}$

ポート $\boldsymbol{p_j}$	$\boldsymbol{TF(p_j)}$	$\boldsymbol{DF(p_j)}$	$\boldsymbol{TF\text{-}IDF(p_j)}$
135	4785.75	2	11420.21
1026	1269.88	8	1269.88
1433	502.88	8	502.88
ICMP	307.12	6	395.48
23566	122.12	1	376.08
1434	373.62	8	373.62
6812	120.75	1	371.84
137	206.12	5	303.00
30010	95.38	1	293.70
60618	85.88	1	264.45
1027	196.00	8	196.00
2967	166.25	8	166.25
80	85.25	4	144.34
5900	128.75	8	128.75
6881	52.00	2	124.09
15651	37.25	1	114.71
8080	107.12	8	107.12
10421	32.25	1	99.31
10426	32.12	1	98.93
22	96.75	8	96.75

表 3: ISDAS の観測するポートの $\boldsymbol{TF\text{-}IDF(p_j)}$

ポート $\boldsymbol{p_j}$	$\boldsymbol{TF(p_j)}$	$\boldsymbol{DF(p_j)}$	$\boldsymbol{TF\text{-}IDF(p_j)}$
135	19499.73	29	20160.80
445	15326.47	27	16941.27
ICMP	6537.40	29	6759.03
139	5778.23	27	6387.03
80	3865.90	30	3865.90
1026	3705.97	30	3705.97
23310	789.57	2	2927.75
1433	2423.33	30	2423.33
631	552.17	3	1823.58
1027	1268.73	30	1268.73
1434	1130.90	27	1250.05
137	989.53	26	1131.14
4899	1007.90	30	1007.90
1025	713.13	29	737.31
4795	150.67	1	663.11
22	470.47	30	470.47
32656	119.17	2	441.88
12592	92.47	1	406.96
113	174.57	8	405.30
1352	108.37	2	401.83

3.4 ISDAS センサの主成分分析

ポート別のパケット数と送信元アドレスを主成分分析を行った．得られた因子の大きさを表 7,9 に示す．$\boldsymbol{C}$ から求めた第一主成分と第二主成分によるセ

表 1: 既知センサ実験環境

	s_{101}	s_{102}	s_{103}	s_{104}	s_{105}	s_{106}	s_{107}	s_{108}
観測期間	2006 年 11 月 30 日-2007 年 1 月 12 日							
クラス	B	C	B	B	B	B	C	C
帯域 [bps]	100M	8M	100M	100M	100M	100M	12M	8M
ネットワーク	大学 1	ISP1	大学 2	大学 2	大学 2	大学 2	ISP2	ISP3
接続携帯	LAN	CATV	LAN	LAN	LAN	LAN	ADSL	ADSL
アドレスの変動	×	○	×	×	×	×	○	○
総パケット数	796	5456	2152	1289	2306	2483	31443	27191

表 4: 既知センサの送信元アドレスの *TF-IDF*(a_k)

アドレス a_k	$TF(a_k)$	$DF(a_k)$	$TF-IDF(a_k)$
219.97	1605.38	2	3830.90
220.211	1153.12	2	2751.70
218.221	431.75	1	1329.55
219.116	519.75	2	1240.28
124.144	205.25	1	632.06
219.95	260.38	3	515.76
204.16	424.38	8	424.38
218.22	91.38	2	218.05
80.24	51.12	1	157.44
221.208	38.50	2	91.87
221.3	37.38	2	89.19
219.94	37.75	3	74.78
59.147	29.88	2	71.29
219.104	29.12	2	69.50
219.96	25.88	2	61.75
219.132	34.38	4	58.20
219.82	23.75	2	56.67
192.168	22.50	2	53.69
211.2	22.12	2	52.80
220.21	29.25	4	49.52

表 5: ISDAS の送信元アドレスの *TF-IDF*(a_k)

アドレス a_k	$TF(a_k)$	$DF(a_k)$	$TF-IDF(a_k)$
219.111	4668.60	23	5909.06
58.93	4490.57	24	5492.61
203.205	2939.13	16	4786.70
222.148	3055.33	25	3612.39
61.252	2159.63	21	2929.92
61.193	1994.30	21	2705.62
61.205	1858.40	21	2521.24
220.221	2035.27	26	2326.52
61.199	1810.27	25	2140.32
222.13	1504.80	20	2114.94
219.2	561.77	12	1076.51
218.255	676.33	17	1060.48
222.159	774.90	23	980.79
220.109	722.17	22	946.15
221.208	861.07	29	890.26
219.114	750.70	25	887.57
203.174	408.50	12	782.80
221.188	600.40	25	709.87
221.16	245.23	6	639.92
219.165	533.77	25	631.08

表 6: 既知センサのポートの主軸$\boldsymbol{u(C)}$

p_i	$u_1(C)$	$u_2(C)$
137	-0.15	-0.20
ICMP	-0.06	0.44
10421	-0.05	0.44
10426	-0.05	0.44
80	0.06	0.44
1434	0.14	0.24
15651	0.19	0.01
23566	0.19	0.01
30010	0.19	0.01
1027	0.21	-0.20
6812	0.24	-0.04
60618	0.24	-0.04
6881	0.24	-0.04
8080	0.24	-0.15
22	0.26	0.07
2967	0.29	0.20
1026	0.29	-0.06
1433	0.31	0.09
135	0.33	-0.03
5900	0.33	0.00
固有値	9.00	4.85

表 7: ISDAS センサのポートの主軸$\boldsymbol{u(C)}$

p_i	$u_1(C)$	$u_2(C)$
445	-0.37	0.01
135	-0.36	0.01
137	-0.34	-0.07
1433	-0.33	0.17
4899	-0.30	0.27
1434	-0.30	0.16
1026	-0.28	-0.27
1025	-0.28	-0.01
1027	-0.25	-0.28
22	-0.23	0.08
32656	-0.13	-0.27
12592	-0.13	-0.27
139	-0.10	0.18
23310	-0.09	-0.03
80	-0.02	0.45
ICMP	-0.02	0.44
113	0.00	0.25
4795	0.00	0.25
631	0.05	-0.04
1352	0.09	-0.08
固有値	6.19	2.49

ンサの分布を図 3, $\boldsymbol{B}$ から得られた第一主成分と第二主成分による分布を図 4 にそれぞれ示す. 図 3 の第一主成分に特徴的なのは, センサ s_1 であり表 7 の因子を見ると, ポート 445 と 135 が係数が $\boldsymbol{-0.37}$ で支配的である. 従って, これらのポートをよく用いる, 例えば「Bot 軸」と考えられる. ただし, 第二主成分の特異な点 s_{26} と s_{11} は各々, ポート 4795 と 32656 のパケットが多く, 表 7 より, これが $v_2(C)$ の因子である. これを主に狙うワームは報告されておらず, プライベートなポートと予想される.

図 4 のアドレス $\boldsymbol{B}$ による主成分によると, センサ s_1 とアドレス 221.188, センサ s_3 とアドレス 58.188 に大きな相関があることが分かる. 従って既知センサの図 2 と同様に各々, クラス C と A の軸と考えるのが妥当であろう. 図 4 では, 図 3 より強い偏りを見せている. その原因は因子の大きさで, 表 9 の因子の絶対値は表 7 のものよりも高い値を示していたことより裏付けられる. 観測ポートの基準を因子の絶対値の大きさを **0.3** とすると, 観測に必要なポートは 445, 135 , 137 , 1433 , 4899 , 1434 , 80 , ICMP であり, 送信元アドレスは 221.188, 222.148, 219.114, 219.165, 221.208, 220.221 である. TF-IDF 値と因子の大きさはほぼ一致しているが, TF-IDF では 12

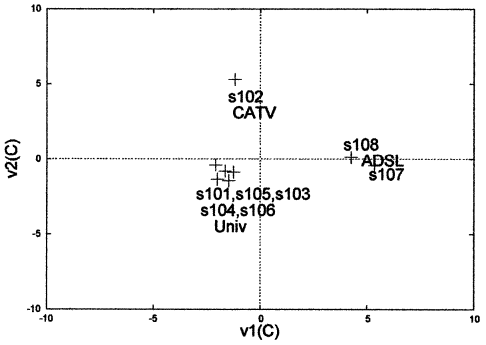

図 1: ポートの主成分 $v_1(C) \times v_2(C)$ 上の既知センサの分布

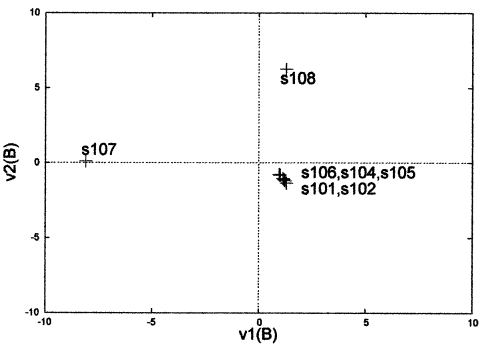

図 2: 送信元の主成分 $v_1(B) \times v_2(B)$ 上の既知センサの分布

位のポートは 137 が因子では第三位になるという矛盾した例をいくつか残す.

3.5 ISDAS センサからのポートとアドレスの識別

ポートについてのパケット数 C と送信元アドレスについての B を各々転置して $C^\top$, $B^\top$ とし, 主成分分析を行うとセンサの主成分の空間上にポート集合 P とアドレス集合 A を分布することができる. 各々の結果を図 5,6 と固有ベクトルを表 10,11 に示す.

図 5 からは, よく知られたポート 135, 445, 139, 1434 が独立したクラスタに現れており, 理解しやすい. 表 10 から, 第一主成分の因子は s_1, s_2, s_3, 第二主成分は s_7, s_{20} である. 一方, B^T からの分析結果の 6 は, 中心に大きなクラスタができてしまっており, 送信元アドレスを分類するのは困難であった.

表 8: 既知センサの送信元アドレスの主軸 $u(B)$

a_k	$u_1(B)$	$u_2(B)$
219.97	-0.31	0.01
220.211	-0.31	0.01
218.221	-0.31	0.01
219.116	-0.31	0.01
124.144	-0.30	0.00
219.95	-0.30	0.01
204.16	-0.30	0.01
218.22	-0.30	0.01
80.24	-0.30	0.01
221.208	-0.30	0.02
221.3	-0.21	0.28
219.94	-0.09	0.14
59.147	0.04	0.39
219.104	0.04	0.39
219.96	0.05	0.38
219.132	0.05	0.38
219.82	0.05	0.38
192.168	0.05	0.38
211.2	0.05	-0.08
220.21	0.05	-0.08
固有値	10.73	6.56

表 9: ISDAS センサの送信元アドレスの主軸 $u(B)$

a_k	$u_1(B)$	$u_2(B)$
221.188	-0.54	0.20
222.148	-0.54	0.20
219.114	-0.53	0.20
219.165	-0.28	-0.52
221.208	-0.17	-0.41
220.221	-0.14	-0.59
58.93	-0.01	-0.20
222.13	0.00	-0.09
222.159	0.01	-0.06
61.199	0.03	0.03
219.111	0.03	0.02
220.109	0.03	0.03
61.205	0.03	0.03
221.16	0.03	0.03
61.252	0.03	0.04
203.174	0.03	0.04
61.193	0.03	0.04
203.205	0.04	0.04
219.2	0.06	0.14
218.255	0.06	0.14
固有値	3.16	2.29

4 結論

TF-IDF 値によって重要なポートを特徴づける方式を提案した. 更に, 集約されたデータに主成分分析を行い, 意味のあるセンサを分類する方式を示した. 提案方式を評価するために ISDAS のデータで分析したところ, 意味のあるセンサ群を分類できることが明らかになった. 観測に必要なポートと送信元アドレスは因子の絶対値で 0.3 を閾値にすると, ポートは 445, 135, 137, 1433, 4899, 1434, 80 ,ICMP で 8 種類あり, 送信元アドレスは 221.188, 222.148, 219.114, 219.165, 221.208, 220.221 で 6 種類あった. 今後の課題として, 主成分の分布から, センサとポートと送信元アドレスの関係を分析し, 独立のセンサ群にクラスタリングを行うことが挙げられる.

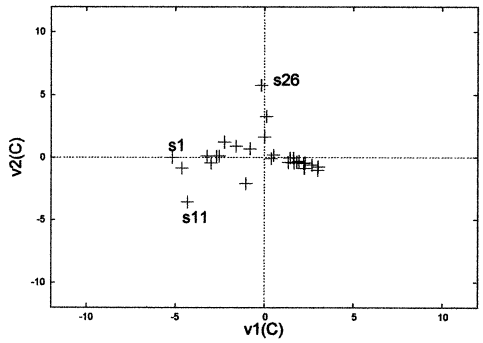

図 3: ポートの主成分 $\boldsymbol{v_1(C)} \times \boldsymbol{v_2(C)}$ 上の ISDAS センサの分布

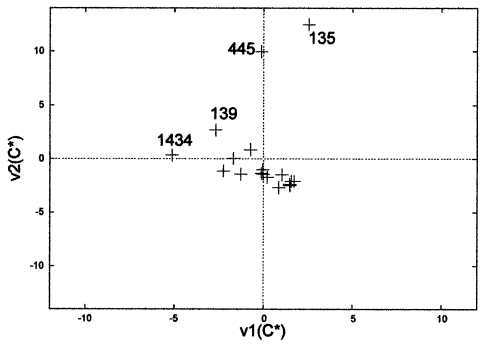

図 5: ISDAS センサによるによる $\boldsymbol{v_1(C^\top)} \times \boldsymbol{v_2(C^\top)}$ 上のポートの分布 $(\boldsymbol{C*} = \boldsymbol{C}^\top)$

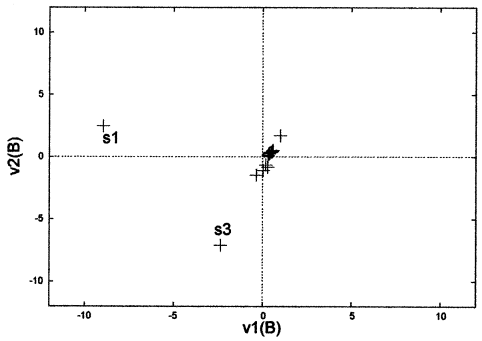

図 4: 送信元の主成分 $\boldsymbol{v_1(B)} \times \boldsymbol{v_2(B)}$ 上の ISDAS センサの分布

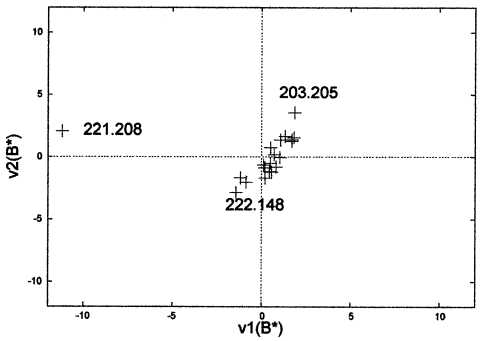

図 6: センサによる $\boldsymbol{v_1(B^\top)} \times \boldsymbol{v_2(B^\top)}$ 上の送信元アドレスの分布 $(\boldsymbol{B*} = \boldsymbol{B}^\top))$

謝辞

本研究を遂行するにあたり，定点観測データ提供いただいた JPCERT/CC の竹田 春樹氏，中谷 昌幸氏，鎌田 敬介氏に感謝する．

参考文献

[1] 石黒，伊藤，戸田，他，インターネット上のポート観測による不整パケットの分布に関する分析，CSS 2005 予稿集，pp 283–288, 2005.

[2] 石黒，松浦，今井，定点観測システム収集データを利用したインターネット空間補間手法の提案と早期異常検知への適用，SCIS 2005, 2005.

[3] 上田　向一，主成分分析，朝倉書店, 2003.

[4] 北，津田，獅々堀，情報検索アルゴリズム，3.2 節索引語の重み付け，pp.33-45，共立出版，2002

[5] 菊池 他，ネットには何台の不正ホストがいるのか？，情報処理学会, CSS 2005, pp.421-426, 2005.

[6] 寺田，高田，土居，ネットワークワーム動作検証システムの提案，情報処理学会論文誌，Vol. 46, No. 8, pp. 2014-2024, 2005.

[7] JPCERT/CC,ISDAS,
(http://www.jpcert.or.jp/isdas, 2007 年 2 月参照)

[8] J. Jung, V. Paxson, A. W. Berger, and H. Balakrishnan, "Fast Portscan Detection Usig Sequential Hypothesis Testing", proc. of the 2004 IEEE Symposium on Security and Privacy (S&P'04), 2004.

[9] "Number of Hosts adrertised in the DNS", Internet Domain Survey, July,2005. (http://www.isc.org/ops/reports/2005-07).

表 10: ポート別パケット数から求めた各 ISDAS センサの主軸$\boldsymbol{u}(\boldsymbol{C}^{\top})$

s_i	$u_1(C^{\top})$	$u_2(C^{\top})$
s_7	-0.04	0.34
s_{20}	-0.03	0.30
s_8	-0.03	0.42
s_{22}	-0.01	0.42
s_{26}	-0.01	0.25
s_{30}	0.03	-0.12
s_{28}	0.05	-0.19
s_{12}	0.06	0.37
s_{15}	0.06	-0.16
s_{29}	0.07	-0.22
s_{25}	0.17	-0.01
s_{23}	0.18	-0.08
s_6	0.18	0.24
s_{24}	0.19	0.04
s_5	0.21	0.02
s_4	0.22	0.08
s_{17}	0.22	-0.12
s_{16}	0.22	-0.09
s_{21}	0.22	-0.02
s_{27}	0.23	-0.06
s_{13}	0.23	0.03
s_{14}	0.24	-0.02
s_{18}	0.24	0.10
s_{11}	0.24	0.07
s_{19}	0.24	0.01
s_3	0.24	0.05
s_1	0.24	0.03
s_2	0.24	0.01
s_{10}	0.24	-0.02
s_9	0.24	0.03
固有値	16.64	3.73

表 11: 送信元アドレス別パケット数から求めた各 ISDAS センサの主軸$\boldsymbol{u}(\boldsymbol{B}^{\top})$

s_i	$u_1(B^{\top})$	$u_2(B^{\top})$
s_{12}	-0.34	0.16
s_{18}	-0.34	0.18
s_6	-0.34	0.18
s_{20}	-0.34	0.02
s_{22}	-0.34	0.18
s_{13}	-0.32	0.21
s_{17}	-0.32	0.01
s_{29}	-0.28	-0.20
s_{28}	-0.21	-0.35
s_{27}	-0.20	-0.11
s_4	-0.17	-0.27
s_{23}	-0.10	-0.33
s_1	-0.05	-0.30
s_3	-0.05	-0.21
s_5	-0.03	-0.03
s_{11}	-0.01	0.03
s_{10}	0.00	-0.15
s_{14}	0.01	-0.08
s_{26}	0.01	-0.05
s_9	0.01	0.07
s_2	0.01	0.06
s_{15}	0.02	-0.11
s_{30}	0.02	-0.07
s_{16}	0.03	-0.00
s_{19}	0.03	0.12
s_{24}	0.04	0.15
s_8	0.04	0.13
s_{25}	0.04	0.32
s_{21}	0.06	0.31
s_7	0.07	0.18
固有値	7.81	2.66

表 12: ISDAS の各センサの観測パケット数

センサ	パケット数
s_1	192953
s_2	132141
s_3	185463
s_4	8297
s_5	139024
s_6	74109
s_7	52707
s_8	23809
s_9	119558
s_{10}	89410
s_{11}	125450
s_{12}	10169
s_{13}	28951
s_{14}	152905
s_{15}	853
s_{16}	99728
s_{17}	8849
s_{18}	12842
s_{19}	109684
s_{20}	2600
s_{21}	102859
s_{22}	22369
s_{23}	3242
s_{24}	28459
s_{25}	14650
s_{26}	244635
s_{27}	14242
s_{28}	6203
s_{29}	6120
s_{30}	88688