見えログ: テキストマイニングと情報視覚化を用いたログ情報ブラウザ

高田哲司　小池英樹

電気通信大学大学院 情報システム学研究科

不正侵入対策においてログ情報の調査は必要不可欠である．しかしログ情報は膨大な量の文字情報であるため，ログ情報の調査は手間がかかるとともに知識を必要とする作業であるという問題がある．そこで本研究では，テキストマイニングをログ情報に適用し，頻度情報を抽出して管理者が注目すべきログ情報の抽出を支援するとともに，情報視覚化を用いて得られた情報を図的に表現し，かつそれらの表示を用いて対話的に調査を行なうことを可能にするログ情報ブラウザ“見えログ”を構築した．本システムを用いてログ情報の調査を行なうことにより，事前に注目すべきログ情報に関する規則を定義しなくてもその抽出が可能なる．またログ情報が持つ種々の特徴が図的に表現されることにより，テキストマイニングによって抽出された注目すべき情報を明示することが可能になり，人間によるログ情報の調査作業を支援することが可能になる．

MIELOG: Log Information Browser with Visualization and Text Mining

Tetsuji Takada and Hideki Koike

The Graduate School of Information Systems
University of Electro-Communications

It is necessary for system administrator to investigate log information. The reason is that log-files contain intrusion marks. It is, however, time-consuming task and needs the knowledge of intrusions.

In this research, we built a log information browsing system which is called “MIELOG”. MIELOG extracts a frequency information using text mining, because such information helps administrator to find notable information that is generated by intrusive actions. Our system represents some log characteristics visually using information visualization technique. This function makes it possible to represent notable information clearly. As a result, MIELOG makes it easier for administrator to investigate log information.

1. はじめに

今日，計算機への不正アクセスや組織内部のユーザによる不正行為が多発しており，これらを検出するとともに，その原因追求や犯人追跡のために調査を行う必要がある．その主たる方法として，計算機の管理者はログ情報を閲覧し，前述の事象に関連すると思われる情報を抽出して，今後の対策に必要な情報を得なければならない．

しかし，ログ情報は文字として記録された膨大な量の情報である．したがって調査を行なうためには記録された情報を読んで理解し，さらに個々の情報に対して不正行為との関連性について判断を行なわなければならない．したがって，その作業は非常に単調かつ退屈な作業になる．よってログ情報を人間が手作業で閲覧し，調査を行うのは極めて困難な作業であるといえる．

そこで本研究では，ログ情報の調査を支援するログ情報ブラウザ「見えログ」を構築した．本システムでは，テキストマイニングを用いて管理者が注意を払うべきログ情報の抽出を支援する．さらに，情報視覚化を利用してログ情報を図的に表現することにより，人間によるログ情報の認識負荷を軽減する．

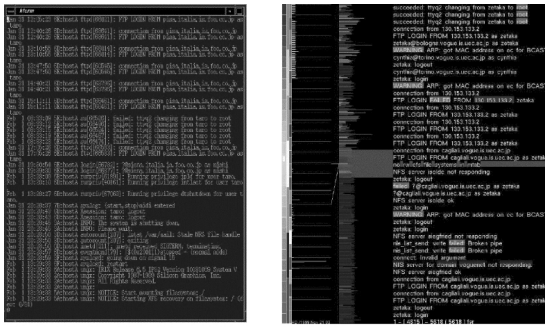
以降，本論文では2章でログ情報の調査における問題点とその改善法について述べ，3章で見えログのシステム概要と視覚化手法について説明する．さらに4で見えログを用いた調査事例について例を挙げてその有用性を解説する．

2. ログ情報調査の問題点

ログ情報は不正侵入に関する唯一の情報源であり，不正侵入対策においてその調査作業は必要不可欠である．しかし現状は様々な要因により作業が敬遠されていると推測され，警視庁が行ったアンケート[1)]でもシステムログを定期的に調査しているのは44%という結果が出ている．そこで本章ではこの原因について考察を行うとともに，本手法で採用した改善手法について述べる．

ログ情報の調査には大きく二つの問題点が存在する．その一つはログ情報の認識負荷の大きさである．ログ情報は膨大な量の文字情報である．したがってそれらを理解するためには，記録されている情報を読んで理解しなければならず，結果として管理者は極めて単調で退屈な作業を強いられることになる．

この問題に対して，本研究では情報視覚化を用いて文字情報として存在するログ情報を図的に表現し，その認識負荷の軽減を実現する．



現在のログ情報　　見えログ使用時のログ情報

図1　ログ情報提示法の変化

二つめの問題は，管理者が注意を払うべき情報の不明確さである．ログ情報を理解した後には，個々のログ情報が不正侵入に関連したものかを判断しなければならない．しかし，この作業には不正侵入やログ情報に対する広範な知識が必要となる．

なおこの作業を行う方法として，既定の規則を用いる手法が存在する[2),3)]．しかし不正侵入手法は日々新たな手法が出現するため，規則による方法では既知の注目すべき情報しか抽出できないという問題が発生する．

そこで本研究では，注目すべきログ情報の抽出に規則による方法とあわせて頻度情報を用いる．ログには不正侵入に関連する情報だけではなく，多数の正常な状態を表す情報や管理上有用な情報などが多数記録されている．一方不正侵入が常に発生しているということは考えにくく，したがって管理者が注意を払うべき情報もログの中に少数だけ存在すると考えられるからである．

よって見えログでは，ログ情報にテキストマイニングを適用し，種々の要素情報に関して頻度情報を抽出する．この頻度情報を基に出現頻度の低い情報を抽出し，ユーザに提示することで管理者が注目すべきログ情報の抽出を支援する．

3. システムの概要と視覚化法

見えログのシステム構成図は，図2の通りである．図2から，見えログは三種の処理によって構成されていることがわかる．本章ではこれらの処理内容について説明を行なう．

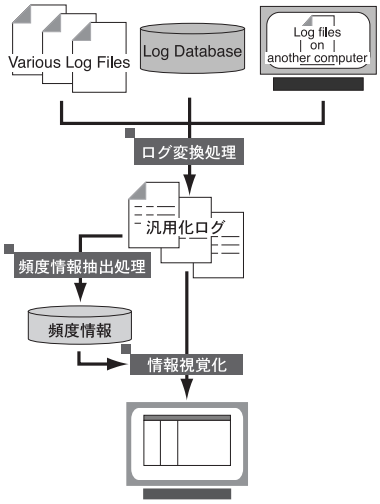


図2　見えログの処理概要

3.1 ログ情報変換処理

見えログでは，種々のログ情報を汎用ログフォーマットとして定義した中間形式に変換する．汎用ログフォーマットの定義を図3に示す．

図3　汎用ログフォーマット

この変換処理の目的は二つある．一つは，ログ情報の記録形式の統一である．ログ情報を本形式に変換することにより，複数のログを時刻を基準として統合することが可能になる．これにより，複数の調査対象の存在による調査負担の増大や広範な知識的の必要性という問題を改善することが可能になる．

もう一つは，後に行なわれるテキストマイニングのための前処理である．

3.2 頻度情報抽出処理

本処理では，汎用ログフォーマット化されたログを対象にテキストマイニングを適用して，種々の頻度情報を抽出する．抽出される情報は次の通りである．

- タグ種別毎の出現頻度
- 時刻の周期的特徴に基づく出現頻度
- 指定時間間隔毎の出現頻度

- 単語の出現頻度
- 連続した二単語に基づく句の出現頻度

ここでは単語と二単語からなる句の出現頻度についての説明を行なう (図 4).

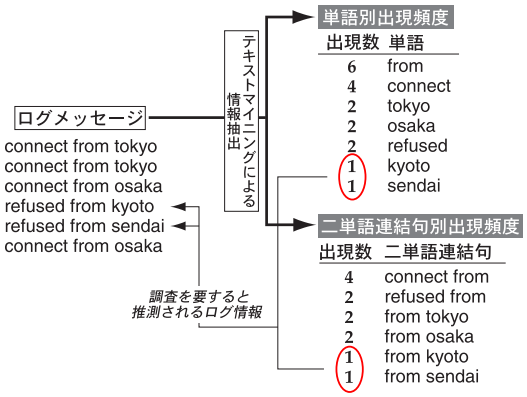


図 4 単語と二単語に基づく句の出現頻度抽出

その抽出方法は，メッセージ部を単語毎に分割し，個々の単語と連続した二単語による句を抽出して，それぞれの出現数を全メッセージ内にわたって数えあげることである．連続した二単語による出現頻度を求めた理由は，出現頻度の高い単語同士が組み合わさり出現頻度の低い句を構成する可能性があるからである．

見えログでは，この処理で得られる出現頻度を基にその値が低い情報を管理者が注目すべき情報として提示する．

3.3 視覚化手法

見えログでは，ログ情報の調査における情報認識負荷の軽減を目的としてログ情報を視覚化する．また前述の頻度情報抽出処理で得られた情報を通常のログ情報と統合して提示するとともに，その関連性を明示することで注目すべきログ情報の認識を支援する．

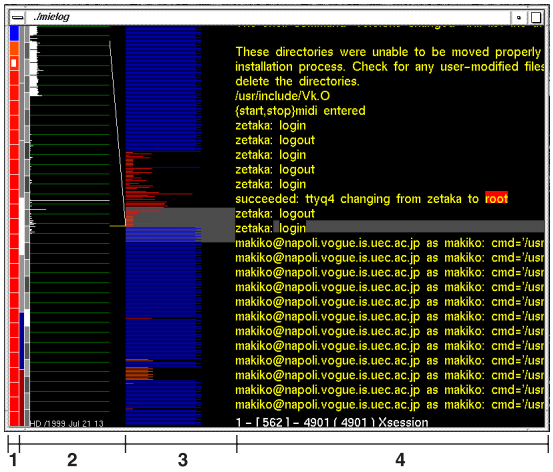
見えログの画面には，4つの表示領域が存在する (図 5)．本節ではそれぞれについて表示法の説明を行なう．

一番左の領域は，ログ種別表示領域である．この領域では汎用ログフォーマットのタグ情報に関して抽出した頻度情報を視覚化している (図 6)．

この領域ではタグ情報を格子として表現し，その色で出現頻度を表す．格子の色は出現数が多いほど青，少ないほど赤となる．またこのタグ情報は出現頻度の高い順にソートされて視覚化されるため，画面上位ほど出現頻度が高く，下位ほど出現頻度が低いことになる．

これにより，調査対象のログに含まれるログ情報の種類とその出力数分布を図的に認識することが可能になる．

次の領域は，時間情報表示領域である．この領域は汎用ログフォーマットにおける時刻に関して抽出された情報を視覚化している (図 7)．



1 ログ種別表示領域
2 時刻情報表示領域
3 アウトライン表示領域
4 ログメッセージ表示領域

図 5 見えログの表示画面

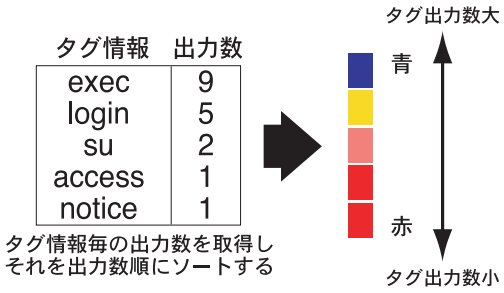


図 6 ログ種別表示領域の視覚化法

この領域には，ログ情報の周期的特徴を格子として表示している部分と，一定時間毎のログ出力数をヒストグラムとして表示している領域がある．周期的特徴は曜日と毎時のログ出力数をログ種別表示同様に格子として表現している．一方，ヒストグラムによる表示は，ログが持つ全時間領域における一定時間毎のログ出力数を表している．縦軸は時間軸で上から下に向かっており，横軸が各時間帯におけるログ出力数を表す．これにより記録されているログ情報の出力傾向を図的に認識することが可能になる．

次の領域は，アウトライン表示領域である．この領域はログメッセージをそのメッセージ長に応じた長さを持つ線として描画するとともに，ログ種別表示領域で決定された色を個々のログ種別に応じて割り当てている (図 8)．

これによりログメッセージをメッセージ長とログ種別の出現頻度に基づくパターンとして認識可能になるとともに，文字による表示よりも広範囲な情報を同時に認識可能になる．

一番右の領域は，ログメッセージ表示領域である．

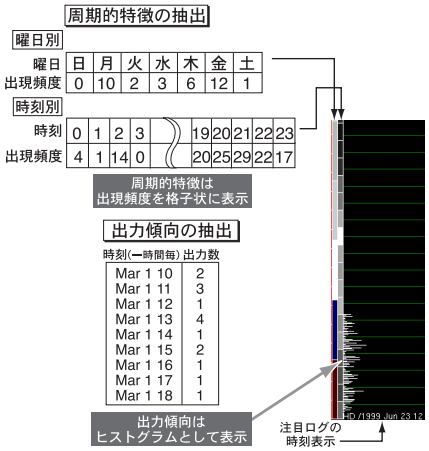


図 7　時刻情報表示領域の視覚化法

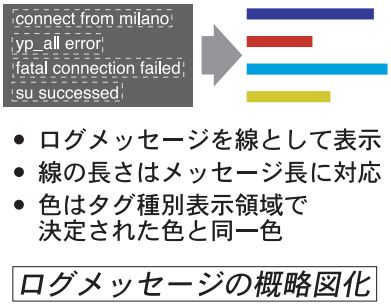


図 8　アウトライン表示領域の視覚化法

この領域は汎用ログフォーマットにおけるメッセージ部を文字として表示する．さらにこの領域では出現頻度抽出処理で得た情報を利用し，出現頻度の低い単語をハイライト表示することで注目すべき情報の存在を管理者へ明確に通知する．

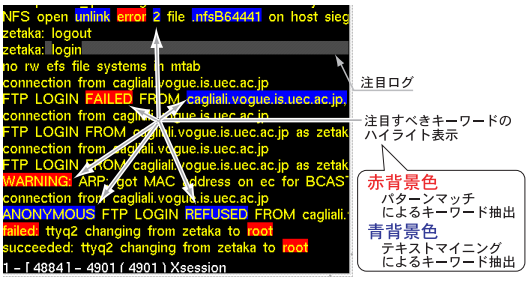


図 9　ログメッセージ表示領域の視覚化法

図9は本領域における視覚化例である．図からもわかる通りこの領域には背景が違う色で描画された単語がいくつか存在する．これは管理者が注目すべき単語を表しており，その基準は二つ存在する．一つは，出現頻度の低い単語であり，背景色が青で描画される．この出現頻度の閾値はユーザが自由に変更可能である．もう一つは既知の異常事象を表すログのキーワードを表すものであり，背景色が赤で描画される．このキーワードは管理者の知識に基づいて事前に定義される．

なお見えログでは管理者が現在注目しているログ情報を注目ログと定義し，アウトライン表示領域やログメッセージ表示領域で常に画面中央に表示される．また注目ログの詳細情報，すなわちログの生成時刻やタグ情報は画面下部に文字で表示される．よって，なんらかの契機によって不正侵入事象に関連すると推測されるログ情報を抽出した場合には，そのログ情報の詳細をログメッセージ表示領域とあわせて即座に得ることが可能である．

3.4　対話処理による調査機能

見えログでは，情報視覚化によって提示される図を用いてフィルタリング機能を主とした様々な対話的処理を行なうことが可能である．その機能を以下に挙げる．

- 特定タグ種別を持つログ情報の抽出
- 指定出現頻度未満のタグ情報を持つログ情報の抽出
- 特定時間帯に出力されたログ情報の抽出
- ログメッセージ長に基づくログ情報の抽出
- 特定単語を持つログ情報の抽出

これにより，視覚化された情報から注目すべきログ情報を発見した場合，種々のフィルタリング手法を用いて即座にそのログへアクセス可能であると同時に，そのログ情報がもつ特徴を用いたフィルタリングにより類似したログ情報の抽出も可能である．

4.　調 査 事 例

本章では，見えログを用いてどのように管理者が注目すべき情報を抽出可能かについて例を挙げて説明する．

時刻情報に注目した調査

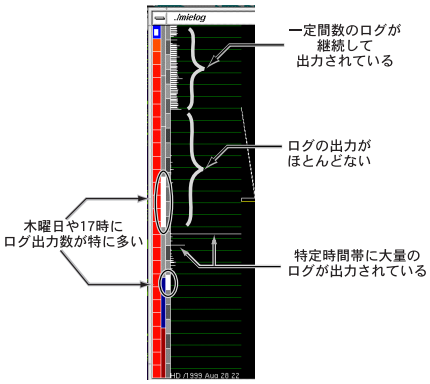


図 10　時刻情報に注目した調査例

図10を見ると，ログの記録開始時刻付近では一定数

のログ情報が出力し続けていることがわかる．その後，ある時刻を境にログ情報は出力されなくなっている．さらに進むと特定の時間帯のみ大量のログ情報が出力されていることを認識することが可能である．

これらの状況から，二つの調査契機を得ることができる．一つはある時刻を境になぜログ情報が出力されなくなったかである．もう一つは特定の時刻に大量に出力されているログ情報が何に起因するものかである．

このように従来までの文字表示による調査では認識の極めて困難なログ情報の時刻に関する情報認識を支援することで，管理者が注目すべき情報の抽出を支援することが可能である．

アウトライン表示に注目した調査例

次にアウトライン表示に注目した調査例について述べる．図11はアウトライン表示領域の表示例を三種類提示したものである．

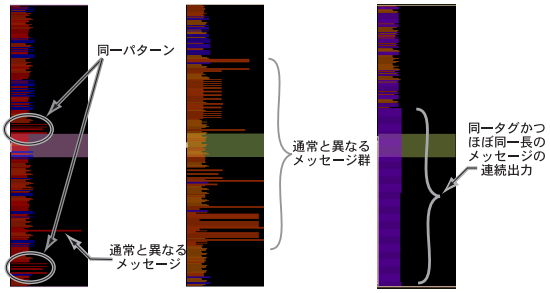


図 11　アウトライン表示に注目した調査例

図11左の表示例が通常時の一表示例である．このようにログ情報の概要を図として認識可能にすることで，その詳細はわからないものの，ログ内に存在する同一パターンのログ情報の出力や，通常と明らかに異なるログ情報を認識することが可能になる．

図11中および右は管理者が注目すべきログ情報の表示例である．図11中の表示では通常とは異なるログメッセージ群が存在することが一目で認識可能である．またこれらの色が黄色であることから，そのログ種別の出現頻度も高くないことが理解できる．したがって，管理者はその詳細を調査する必要がある．

図11右の表示では，青色の線だが同一の長さのメッセージが連続して出力されていることが認識できる．このようにアウトライン表示を用いることで大量に記録されているために出現頻度による方法では抽出が困難だが，同一か類似したログ情報が連続して出力されているといった状況もその抽出が可能である．

ログメッセージに注目した調査例

最後にログメッセージに注目した調査例について述べる．

見えログでは従来通り文字によるログ情報の提示も行なう．さらに管理者が注目すべき情報をキーワードによる規則と頻度情報の双方に基づいて抽出し，ハイライト表示を用いてそれらを明確に提示することでログ情報の調査を支援する．

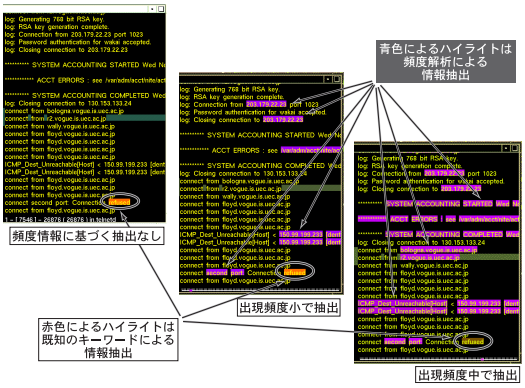


図 12　ログメッセージに注目した調査例

図12は同一ログにおける三つの視覚化例を示している．図12左は，キーワードによる抽出のみを行なっている例であり，画面下部に既定のキーワードに一致した単語が赤い背景色で描画されているのを認識することができる．

図12中および右は，キーワードによる抽出と頻度情報に基づく抽出の双方を用いた例であり，双方の違いは頻度情報による抽出の閾値の差によるものである．図12中は閾値 “小” で，図12右は閾値 “中” で抽出を行なった例であり，閾値に応じて青いハイライト表示で提示される単語の数が増減していることが認識できる．

管理者は，この機能を用いて注目すべきログ情報を抽出した上で，ログ情報を閲覧することにより，従来までは認識することの困難であった調査すべきログ情報を容易に認識することが可能になる．

5. 考　　察

本章では，見えログについての考察を行なう．

5.1 利　　点

見えログにおける利点を以下に述べる．

一つ目は，管理者が注目すべきログ情報の抽出を支援できることである．見えログでは，テキストマイニングを用いてログ情報から頻度情報を抽出し，管理者が注目すべきであると推測されるログ情報の抽出を支援する．これにより文字情報を読んで調査するだけでは認識の困難なものや，キーワードによる方法では抽出の困難な情報の抽出を支援することが可能になる．

また本手法は，不正侵入検知システムにおける異常検出☆手法をログ情報の調査に適用した例として考えることができる．よってキーワードによる手法を不正検出☆☆ 手法と考えると，ログ情報調査作業に双方の

☆ 原語は Anomaly Intrusion Detection
☆☆ 原語は Misuse Intrusion Detection

手法に基づくログ情報の抽出を可能にした対話的ログ情報調査支援システムであるといえる．

二つ目は，情報視覚化による情報認識負荷の軽減である．たとえ頻度情報を用いて注目すべきログ情報を抽出したとしても，それを人間が把握し，適切な対処を行なわなければ不正侵入対策として意味がない．そこで本研究では，テキストマイニングによって得られた情報を情報視覚化を用いて視覚化するとともに，対話的に調査作業を進めることを可能にした．これによりログ情報の把握における認識負荷を軽減することが可能になると同時に，図的表現と従来の文字による表示が連係しているため，概要から詳細情報に基づく調査までシームレスに行なうことが可能である．

5.2 問 題 点

次に見えログにおける問題点について二点ほど述べる．

一つは，ログ情報を“見る”という作業における作業負担の軽減が必要なことである．見えログによってログの調査作業を支援することが可能になるが，それでも調査作業時にはログ情報を全て閲覧しなければならない．この問題に対する対処法として音を用いた通知機能や，ログ情報の圧縮による閲覧すべき情報量の縮小を今後の課題として考えている．

二つ目は抽出情報の妥当性改善である．頻度情報を用いた注目すべきログ情報の抽出と実際の異常事象を表しているログ情報との間にずれが生じることがある．その一例として数値が挙げられる．ログ情報内に記録されている数値は繰り返し出力されない場合が多く，それゆえ頻度情報に基づく抽出ではそれが管理者が注目すべき情報でなくても，注目すべき情報として抽出されるという問題が発生する．したがって，頻度情報を抽出しない単語や句を規則として設け，例外処理とするなどの改善法を考案する必要がある．

5.3 関 連 研 究

ログ情報の調査を支援するシステムの関連研究について例を挙げる．

Xlogmaster[4]は，X-window system上で動作するログ情報監視システムである．しかしログ情報の提示は文字による提示のみであるため，認識負荷の問題が生じる．また色による情報の分類法はあるものの，そのためには規則を定義しなければならないという問題がある．

Swatch[2]やLogsurfer[3]は，管理者によって定義された規則を基に注目すべきログ情報の抽出を行なう．したがって，これらのツールを利用するためにはキーワードやパターンを抽出規則として定義しなければならないという問題が生じる．

なお不正侵入や不正行為の抽出にデータマイニングが有効であるという指摘は[5]や[6]でも行なわれており，今後，その応用が期待されている．

6. おわりに

本論文では，ログ情報の調査を支援するログ情報ブラウザ「見えログ」について述べた．不正侵入対策においてログ情報の調査は必要不可欠である一方で，文字による情報提示のため，情報認識の負荷が大きいことや管理者が注目すべきログ情報が不明確であるという問題がその作業の実施を阻害していた．

そこで本研究では，ログ情報にテキストマイニングを適用して頻度情報を抽出するとともに，それらを情報視覚化を用いて図的に表現した．テキストマイニングを利用して頻度情報を取得することにより，事前に規則を定義しなくても，管理者が注目すべきであると推測されるログ情報を抽出することが可能になる．また情報視覚化を用いることにより，ログ情報に対する人間の認識負荷を軽減するとともに，注目すべきログ情報の存在を明確に提示することが可能になる．

これらの特徴により，システム管理者が見えログを用いてログ情報を調査することにより，従来までは認識の困難であったログ情報の特徴を加味しながら注目すべきログ情報の抽出を行なうことが可能になり，結果として不正侵入に関連すると推測されるログ情報の抽出を支援することが可能になる．

参 考 文 献

1) 警視庁，“ハイテク犯罪に関するアンケート”，Feb-Mar (1998). http://www.npa.go.jp/soumu6/
2) S.E.Hansen and E.T.Atkins, “Automated System Monitoring and Notification With Swatch”, USENIX 7th System Administration Conference, Nov (1993).
3) W.Ley and U.Ellerman, “Logsurfer”, (1995). http://www.cert.dfn.de/eng/logsurf/home.html
4) Georg C.F.Greve, “The Xlogmaster”, (1998). http://www.gnu.org/software/xlogmaster/
5) W.Lee and S.Stolfo, “Data Mining Approaches for Intrusion Detection”, In Proc. of 7th USENIX Security Symposium, Jan (1998).
6) Kenneth C. Cox, Stephen G. Eick, Graham J. Wills and Ronald J. Brachman, “Visual Data Mining: Recognizing Telephone Calling Fraud”, Journal of Data Mining and Knowledge Discovery, Vol.1, No.2, pp.225-231, (1997).
7) 高田 哲司，小池 英樹，“ログファイルの視覚化による不正侵入検知手法の提案”，情報処理学会 コンピュータセキュリティシンポジウム’98，pp.153-158，(1998).
8) Pieter Adriaans and Dolf Zantinge著 山元英子 梅村恭司訳，“データマイニング”，共立出版，(1998).