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

高田哲司[†] 小池英樹[††]

計算機の運用管理においてログ情報の調査は必要不可欠である．この作業は，計算機への不正侵入の多発にともないその重要性を増しつつある．しかしログ情報の調査は退屈で単調な作業であるため，その作業を敬遠する傾向が高く，結果として種々の問題を見過ごす原因となっている．さらに，異常を表すログ情報の抽出が困難であることも調査作業を困難にする原因としてあげられる．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，異常事象を表すログ情報の抽出を支援する．さらにそれらの情報を情報視覚化を用いて図化してユーザに提示することによりログ情報の認識を支援する．我々はこれらの特徴により人間によるログ情報の調査作業を支援するログ情報ブラウザ“見えログ”を開発した．本システムを利用することにより，システム管理者が異常事象を表すログに関する種々の規則を事前に定義しなくても異常を表すログ情報の抽出を可能にする．また抽出された特徴情報は，情報視覚化によって図化されるとともに文字による表示と連係して提示される．これにより個々のログ情報に関する種々の特徴を容易に把握することが可能となり，人間が様々な観点をもとに注目すべき情報か否かを判断をすることが可能となる．これらの特徴によ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となる．

MieLog: Log Information Browser with Information Visualization and Text Mining

Tetsuji Takada[†] and Hideki Koike[††]

It is necessary for system administrator to investigate some log information. The reason is that log-files contain enormous information generated from an operating system and various programs and these information are useful to solve a variety of troubles on computer. Moreover, an intrusion to the computer becomes serious problem more and more. A system administrator, therefore, has to watch a log information periodically in order to find out the intrusion marks. In this research, we developed log information browsing system, which is called “MieLog”, in order to support such task. MieLog extracts some characteristics from log information. An example of these characteristics is the number of log outputting in fixed time or the length of log text. MieLog, moreover, represents their characteristics visually with textual information. As a result, MieLog makes it easier for system administrator to investigate log information.

1. はじめに

計算機の運用管理においてログ情報を定期的に調査することは重要である．管理している計算機において発生している事象を把握し，異常事象が発生している場合には適切な対処を迅速に行う必要があるためである．また，計算機への不正侵入の多発がこの重要性をますます高めている．なぜならば，不正侵入を検知するため，ならびに不正侵入に対する適切な対処を行うためにログ情報の調査が必要不可欠だからである．

しかしこの作業は単調で時間のかかる作業であるため，重要な作業であることが認識されているにもかかわらず，その実施度は低いといわれている[3)]．

そこで我々は，ログ情報の調査を支援するログ情報ブラウザ“見えログ”を開発した．本システムでは，複数のログ情報を汎用ログフォーマットと呼ばれる形式に変換/統合し，それに対してテキストマイニングを適用してログ情報が持つ特徴情報を抽出する．さらに得られた特徴情報を情報視覚化を用いて図化するとともに，それを文字によるログ情報表示と連係させて調

† 電気通信大学サテライト・ベンチャ・ビジネス・ラボラトリ
Satellite Venture Business Laboratory, University of Electro-communications
†† 電気通信大学大学院情報システム学研究科情報システム運用学専攻
Graduate School of Information Systems, University of Electro-communications

査者に提示する．

これにより従来までのログ情報調査における問題であったログ情報の認識負担を軽減することが可能になるとともに，異常事象として疑わしい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．

以降，本論文では 2 章でログ情報調査の重要性とその作業における問題点について述べ，3 章で本論文で提案するログ情報ブラウザ “見えログ” についてその詳細を述べる．4 章で見えログを用いたログ情報調査の事例について例をあげ，5 章で考察を行う．

2. ログ情報調査における問題点

計算機の運用管理においてログ情報の調査を定期的に行うことが必要不可欠であることに疑いの余地はない．この作業は計算機への不正侵入が多発するにつれてその重要性を増しつつある．

しかしログ情報の調査作業は,「難しい，面倒，空しい」[1]といわれ，作業を定期的に遂行する必要性が認識されているにもかかわらずその作業は敬遠される傾向にあり，実際には遂行されていないといえる．そこで我々はログ情報の調査作業における問題について考察を行う．

人間によるログ情報の調査を阻害する問題として大きく 2 つの問題が考えられる．1 つはログ情報把握の困難さに関する問題であり，もう 1 つは異常事象と推測されるログ情報の抽出に関する問題である．

ログ情報を把握する際の問題として以下の事項が考えられる．

- 文字としての記録
- 膨大な量の情報量
- 情報源の多様性

これらの問題から，ログ情報の調査では膨大な量のログ情報を読まなければならず，その認識負荷は大きい．またログ情報は様々な種類の情報や記録形式が存在するため，調査者にはそれらに対する知識も必要となる．さらに異常事象の調査においては，複数のログ情報を調査し，断片的な情報を組み合わせて状況を判断する必要もある．

次に，異常事象と推測されるログ情報の抽出における問題点を以下にあげる．

- 異常事象を表すログ情報は少数
- 抽出規則の構築は困難

ログ情報には様々な情報が記録されており，そのすべてが異常事象を表すものではない．また，異常事象を表すログ情報は膨大な量のログ情報の中にごく少数存在するといわれている[2]．

またその一方で出現頻度の観点だけではその抽出が困難だが，出現頻度の低いもの以外にも異常事象と推測されるログ情報は考えられる．その例を以下にあげる．

- 通常時に出力されるものとは明らかに異なるログメッセージの存在
- 突発的に大量のログ情報が出力されている場合
- ログが出力されなくなる，または出力数が増減している場合
- 定期的に出力されるべきログ情報が，それ以外の時刻に出力された場合

したがってログ情報の調査においては，注目する必要のない情報が大量に含まれているログ情報から前述のような特徴を持つログ情報を抽出しなければならない．

さらに，これらを自動的に抽出するための規則を構築することが困難であることがあげられる．パターンマッチングによる方法は多数存在するが[8)~11)]，これらの問題点は既知の異常事象を表すログ情報しか抽出できないことである．また，あらゆる異常事象を意味するログ情報のキーワードを規則として構築することも現実的ではない．またこの手法はシステム管理者の知識に依存してしまうため，一般には既知の異常事象であってもシステム管理者の技術レベルによってはそれを見過ごす恐れもある．

そこで本研究では，ログ情報把握の困難さという問題に対して情報視覚化を，異常事象と推測されるログ情報の抽出の困難さという問題に対してテキストマイニングを用いることでログ情報の調査を支援する．

ここでテキストマイニングについて説明を行う．テキストマイニングとは文書データを様々な観点から分析し，有用な知識や情報を取得しようとする技術である．本研究では特に出現頻度に注目してログの分析を行っている．

なお同様の目的を持つ技術にデータマイニングが存在するが，テキストマイニングとデータマイニングの違いは処理対象となるデータにある．データマイニングで扱うデータはデータベーススキーマによって定義されたレコードが処理対象となるのに対し，テキストマイニングは形式化されていない文書を処理対象とするのである[12]．

また見えログでは既存のツールのように自動的にログ情報を調査し，その結果を出力するのではなく，対話的にログ情報の調査が遂行できることを目指した．これによりテキストマイニングだけでは抽出の困難な疑わしいログ情報の抽出を人間が実際に見て調査することで補完することが可能になる．

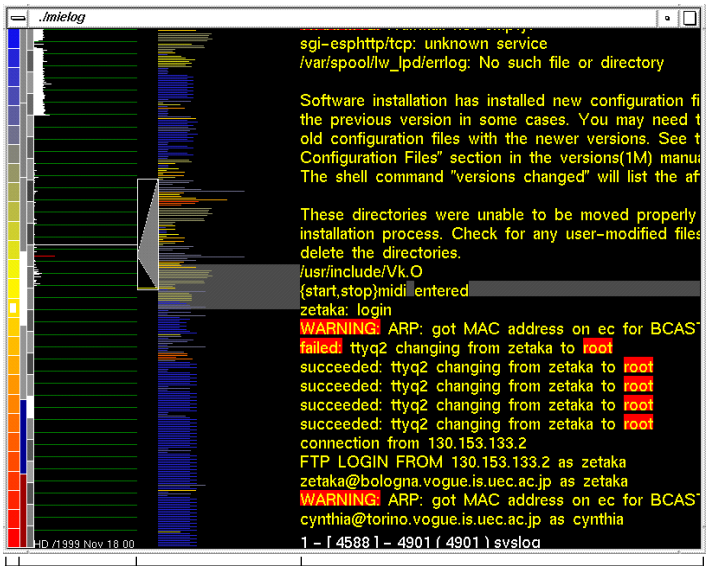

図 1 見えログの画面表示
Fig. 1 A display image of MieLog.

3. 見えログ：システム概要と視覚化手法

本研究ではこれまでの考察をもとに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．本システムの概観を**図 1** に示す．

図からも分かるとおり，見えログでは画面右側のログメッセージ表示領域に文字でログ情報を表示すると同時に，画面左側ではテキストマイニングによって取得したログ情報の特徴情報を提示する．さらに両情報間の関連を明確に提示し，文字情報だけでは把握の困難な各ログ情報の特徴を認識可能にしている．

図 2 は見えログの処理概要である．

見えログは 3 つの処理から構成されている．まずはじめに種々のログ情報をログ情報監視変換処理によって汎用ログフォーマットに変換するとともに，1 つのログ情報として統合する．この変換されたログ情報に対して特徴情報抽出処理を行い，テキストマイニングを用いて特徴情報を抽出する．最後に，情報視覚化処理を用いてログ情報を調査者に図的に提示するとともに対話的な調査作業を可能にする．

本章ではこれらの処理についてその詳細を述べてゆく．

3.1 ログ情報収集と汎用ログフォーマット化 ——ログ情報監視変換処理

ログ情報の調査における問題として，ログ情報の形式や内容の差異，ログ情報の取得方法や存在場所の知識を必要とすること，さらに複数のログ情報を調査する必要があることなどがあげられる．そこで本研究では，調査対象となる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形式に変換する．**図 3** はその形式と変換例である．

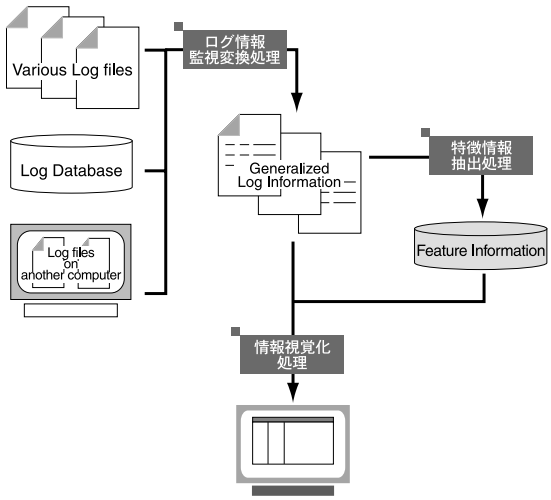

図 2 見えログの処理概要
Fig. 2 A process module overview about MieLog.

汎用ログフォーマットは 4 つの情報から構成され，時刻は秒数として，それ以外は任意の文字列として定

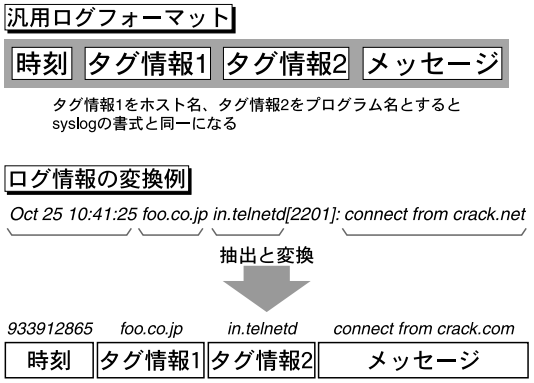

図 3 汎用ログフォーマットと変換例
Fig. 3 General Log Format specification and translation example.

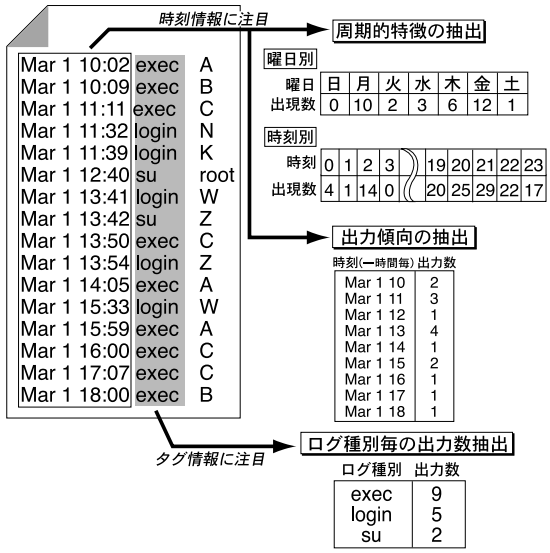

図 4 時刻とタグ情報に注目した特徴情報の抽出
Fig. 4 Feature extraction about time and tags.

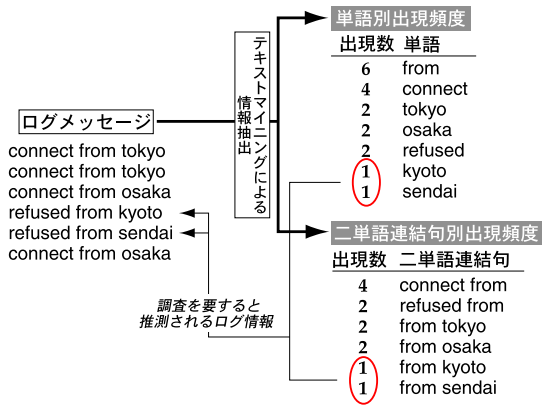

図 5 メッセージに注目した特徴情報の抽出
Fig. 5 Feature extraction about messages.

義される．種々のログ情報をこの形式へ変換することによって一定の形式にするとともに，複数のログ情報を時刻を基準として 1 つに統合することが可能になる．これにより，複数のログ情報に対する調査作業を 1 つのログ情報の調査に集約することが可能になる．また複数のログ情報を 1 つに統合することで，複数のログ情報間に存在する情報間の関連性を明確にすることが可能になるという利点も生じる．なおこの変換処理は，後に行われる特徴情報抽出のための前処理としての役割もある．

3.2 テキストマイニングによる特徴情報の抽出 ——特徴情報抽出処理

ログ情報の調査のおける問題点の 1 つに膨大な量の文字情報から異常事象であると推測されるログ情報を抽出することの困難さがあげられる．

そこで本研究では「異常事象として注目すべきログ情報は，膨大な量のログ情報の中に少数含まれる」という特徴に注目し，出現頻度の低い情報を発見するためにテキストマイニングを用いて頻度情報を抽出する．

以下では見えログで抽出する特徴情報を汎用ログフォーマットの各要素情報ごとに説明する．

- 時刻情報に注目した特徴情報
  ログ情報に含まれる時刻情報の周期的特徴と出力傾向を抽出する（**図 4**）．周期的特徴とは曜日と毎時別のログ情報出力数を指し，出力傾向とはログ情報が記録された全時間帯における一定時間間隔ごとのログ情報出力数である．
- タグ情報に注目した特徴情報
  タグ情報ごとのログ情報出力数を抽出する（図 4）．
- メッセージに注目した特徴情報
  メッセージからはメッセージを構成する単語と 2 単語からなる句ごとの出現頻度を抽出する（**図 5**）．

これらの処理は，異常事象と関連するログ情報が膨大な量のログ情報の中に少数存在するという前提に基づいている[2]．システム管理者によって管理作業が行われている計算機ならば，異常事象を表すログ情報が頻繁にまたは大量に記録されているとは考えにくい．つまりログ情報の調査において抽出すべき情報は，膨大な量のログ情報の中にごく少数埋もれていると考えられる．また異常事象を表すログメッセージの中にはふだん出力されない単語や句が記録されていると考えられるためである．

なお見えログでは，パターンマッチングを用いた既知の異常事象を表すログ情報の抽出も可能である．システム管理者は異常事象を表すキーワードを事前に定義することにより，パターンマッチングによるログ情

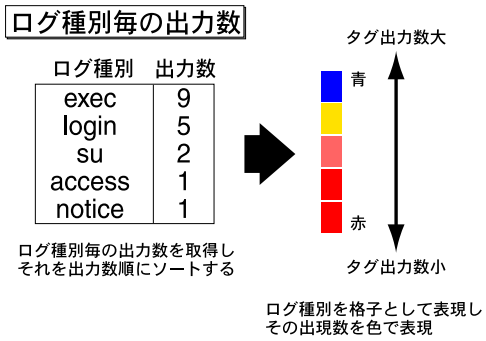

図 6 ログ種別表示領域の視覚化手法
Fig. 6 Visualization method of tags in log.

報の抽出が可能になる．

3.3 情報視覚化による情報の図的表現と対話的機能

ログ情報の調査では，ログ情報の認識負荷が問題としてあげられていた．そこで本研究では，情報視覚化を用いて特徴情報収集処理で得られた情報を図化して提示する．さらに文字によるログ情報提示と図化された特徴情報の提示を連係させることにより，調査者が注目しているログ情報の特徴情報を容易に認識できるようにする．本節ではその視覚化手法について述べる．

見えログの表示画面は 4 つの領域から構成されており，左からログ種別表示領域，時刻情報表示領域，アウトライン表示領域，ログメッセージ表示領域となる．ログ種別表示領域はタグ情報をもとにした表示領域で，タグ情報をその出力数によってソートし，出力数の多い順に上から下へ格子として表示する（**図 6**）．

なお，各格子の色はログ情報の出力数を表しており，値の大小に応じて青から赤へと色が割り当てられる．この表示により，ログ情報が保持しているログ種別数とその出力数分布を迅速に認識することが可能となる．なお，個々のログ種別名はログメッセージ表示領域の画面下部に文字で表示される．

次の領域は時刻情報表示領域である．この領域は左側の格子状表示領域と右側のヒストグラム表示領域の 2 つに分割される．領域左側の格子状領域は時刻情報の周期的特徴を表示している．縦方向に 2 つの格子が存在するが，左側の格子が曜日に注目した周期的特徴の表示であり，上の格子から順に月曜日から日曜日の情報を表している．一方右側の格子は時刻に注目した周期的特徴情報の表示であり，上から順に 0 時から 23 時の情報を表している．また各格子の色の濃淡でログ情報の出力数を表している（**図 7**）．

一方，領域右側はログ情報の出力傾向を表しており，縦軸方向は上から下に向かって時間軸を，横軸は左から右に向かって各時間帯におけるログ情報の出力数を

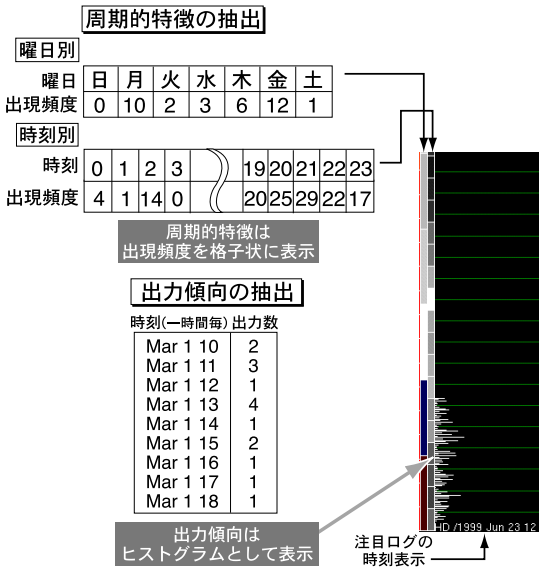

図 7 時刻情報表示領域の視覚化手法
Fig. 7 Visualization method of time in logs.

表している．なお見えログでは，ログメッセージ表示領域で画面中央に表示されているログ情報を調査者が注目しているログとして注目ログと定義しているのだが，この領域の画面下部には注目ログの記録された時刻情報が文字で表示される．これらの視覚的表現により，ログ情報がどのような時間的傾向を保持しているかを迅速に認識すること可能になる．

次はアウトライン表示領域である．この表示領域では各ログメッセージをその長さに応じた線として表現し，線の色にはログ種別表示領域で定義したログ種別ごとの色をそのメッセージのログ種別に応じて割り当てている．この表現により，ログ情報をメッセージ長とログ種別ごとの出力頻度に基づくパターンとして認識することが可能になる．結果としてログ情報の概要を図的に把握することが可能になる（**図 8**）．

なおこの領域の画面中央には背景が灰色の部分が存在するが，これはログメッセージ表示領域に文字で表示されているログ情報の領域を表しており，この領域の中央に存在するログが前述した注目ログとなる．なお注目ログの位置は本表示領域の左側に緑色の線でも示されており，緑色の線を囲むように描画されている白枠は，注目ログと同一時間帯に出力されたログ情報の領域を示している（図 8）．

一番右側の表示領域はログメッセージ表示領域であり，ログ情報が文字で表示される．さらにこの表示領域では，パターンマッチングにより抽出された単語と，出現頻度をもとに抽出された単語をハイライト表示す

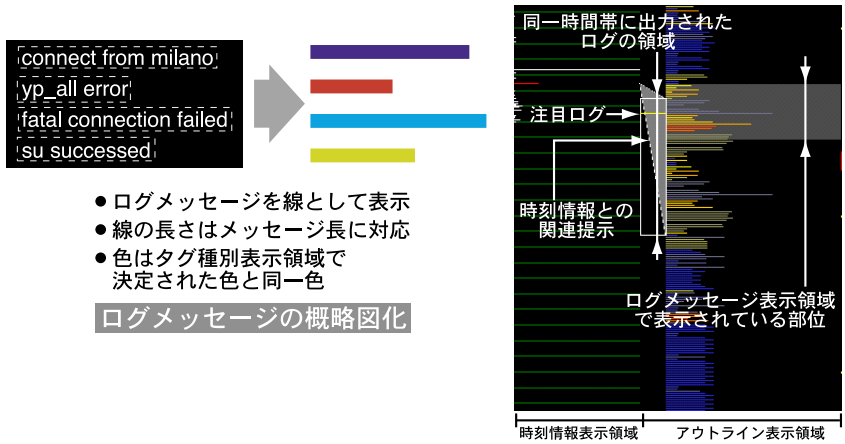

図 8 アウトライン表示領域の視覚化手法
Fig. 8 Visualization method of outlines about log messages.

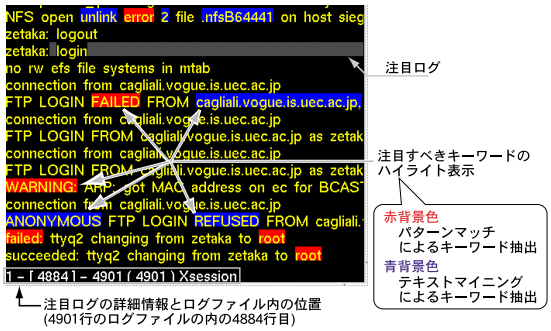

図 9 ログメッセージ表示領域の視覚化手法
Fig. 9 Visualization method of log messages and its features.

る（**図 9**）．これにより，異常事象を表すと推測されるログ情報の抽出および認識を支援する．

ハイライトの背景色は 2 種類存在し，赤色はパターンマッチングにより抽出された単語を，青色は出現頻度に基づいて抽出された単語を表している．なお出現頻度を用いて抽出する際の基準となる出現頻度の基準値は，調査者が対話的に指定および変更可能である．

また画面中央部に 1 行だけ背景色が灰色の行が存在するが，これは注目ログを表しており，注目ログに関する詳細情報が本表示領域や時刻情報表示領域の画面下部に文字で表示される．

次に見えログが提供する対話的機能について述べる．見えログでは情報抽出やフィルタリングを対話的に行うことが可能である．現在実装されている対話的機能を以下にあげる．

(1) 特定ログ種別を持つログ情報の抽出
(2) 指定出現頻度以下のログ種別を持つログ情報の抽出
(3) 特定時間帯に出力されたログ情報の抽出
(4) ログメッセージの長さに基づくログ情報の抽出
(5) 指定した単語を含むログ情報の抽出
(6) 既定キーワードを含むログ情報の抽出

具体例としてタグ情報に基づいたログ情報のフィルタリング例を**図 10** に示す．この図では (1)，(2) の機能によるフィルタリング例が示されている．

4. 調査事例

本章では，見えログを用いてどのように異常事象を表すログ情報を抽出可能かについて述べる．

4.1 時刻情報に注目した抽出

まずはじめに時刻情報に注目したログ情報の抽出例について示す．

図 11 は，見えログの時刻情報表示領域による表示例である．まずはじめにこの図におけるヒストグラム表示に注目する．ヒストグラムの上部における表示から，このログ情報の記録が始まってからある時刻までは，定期的に一定数のログ情報が出力されていたことが容易に認識できる．

一方，画面中央部では，ヒストグラムによる表示がなくなっている．すなわちある時刻を境にログ情報が出力されなくなったことを表している．さらに時刻が進むと，ある特定の時間帯においてのみ多数のログ情報が出力されていることが認識できる．

また画面左側の格子状表示に注目すると，ある特定のタイルだけ明るい白色で描画されているのを認識することができる．これはその表示法から，木曜日と 17 時台にログ情報が多数出力されていることを意味している．

これらの表示から 2 つの調査作業が必要になるとい

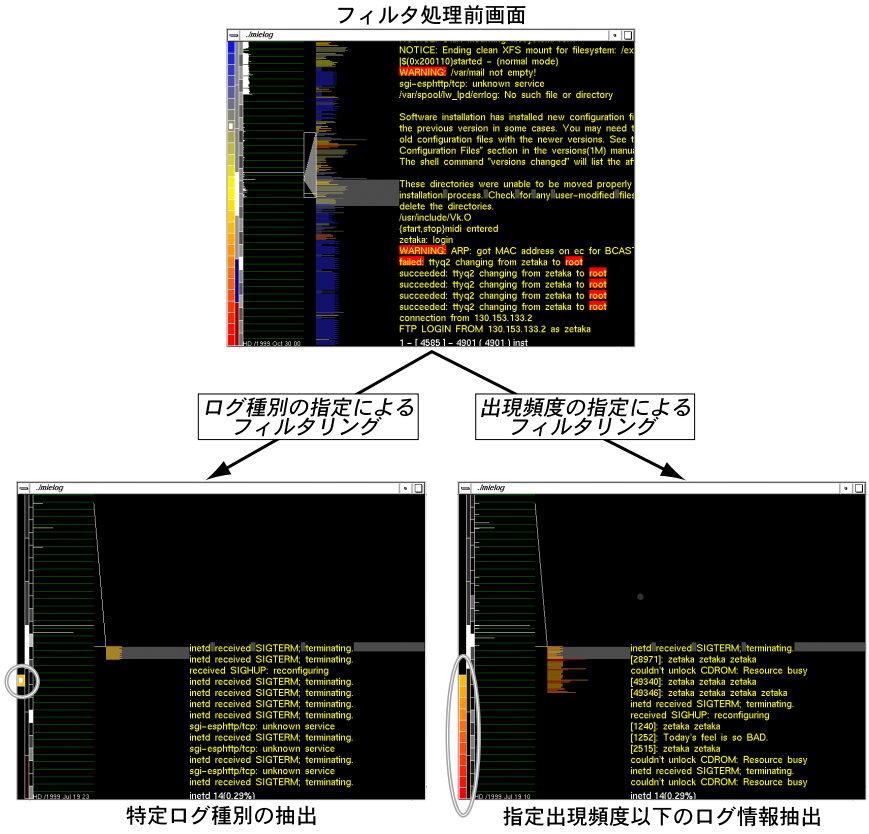

図 10 タグ情報に基づいたフィルタリング処理
Fig. 10 An example of interactive process: log filtering by tags.

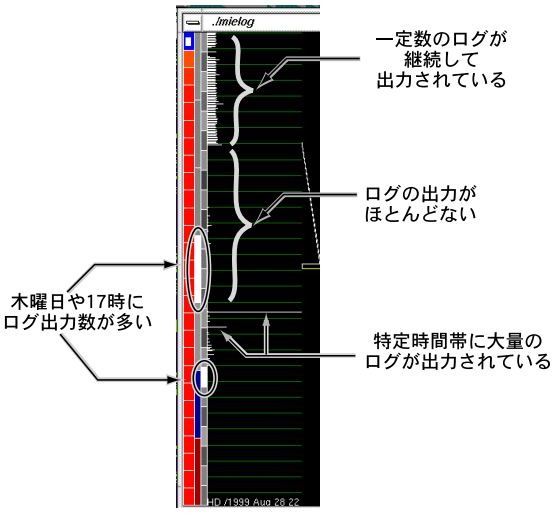

図 11 時刻情報に注目した調査例
Fig. 11 An example of an investigation focusing on time.

うことができる．1 つは特定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 1 つは特定時間帯になぜ大量のログ情報が出力されているかである．

これらはそれぞれ該当時間帯のログ情報をより詳細に調査する必要があることを意味している．この作業も該当時間帯のログ情報を見えログ上で抽出することで容易に実施可能である．

これらの事象に対する調査の結果，ログ情報が出力されなくなった原因はシステムの設定変更によりこれまでログ情報を出力していたプログラムが起動しなくなったことであり，大量のログ情報出力の原因はソフトウェアの導入における稼働試験に起因するログ情報であることが分かった．

時刻に関する情報は，文字によるログ情報調査ではその認識がきわめて困難である．しかし見えログでは，時刻に関する特徴を抽出し図化して提示することで異常事象と推測される事象の抽出を支援することが可能になる．

4.2 アウトラインに注目した抽出

次にログ情報のアウトラインに注目した調査例について述べる．

図 12 は，アウトライン表示領域の表示例を 3 種類提示している．左の図は通常時の表示例，その他の 2

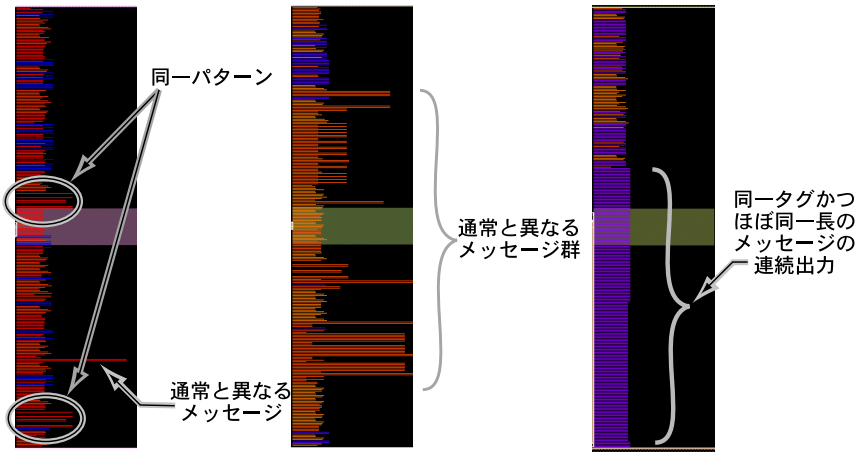

図 12 アウトラインに注目した調査例
Fig. 12 An example of an investigation focusing on outlines about log messages.

つは異常事象であると推測される表示例である．図 12 左を見ると，調査対象のログメッセージはほぼ似たような長さのログ情報が出力されていることが認識できるとともに，画面中央と下部に同一パターンのログ出力があることが認識できる．さらに画面下部には他のメッセージとは明らかに異なる長さのログメッセージが存在することが認識できる．このようにその詳細は分からないものの，ログメッセージの概要をパターンとして認識可能にすることで，通常のログとは異なるログメッセージの抽出が可能となる．

図 12 中には，通常とは明らかに異なるログ情報が連続して出力されていることが一目で認識できる．またそれらの線の色が黄色であることから，該当ログメッセージの出現頻度がそれほど高くないことも認識することができる．したがって，これらのログ情報をより詳細に調査する必要があるといえる．

一方，図 12 右は，線の色が青色であることからその出現頻度は高いものの，まったく同一のログ情報が連続して出力されていることが認識できる．これも通常時のメッセージとはまったく異なるパターンであるとともに，異常事象であってもそれが連続して出力され続けた結果，出現頻度が高くなってしまった可能性もあるためその詳細を調査する必要があるといえる．

このように，ログメッセージをアウトライン表示として図化することによってログ情報をパターンとして認識することが可能になる．これにより，ログ情報の概要を把握することが可能になると同時に，通常時とは異なるログ情報の抽出を支援することが可能になる．

4.1，4.2 節で述べられている調査事例は，見えログが出現頻度の低い情報だけを異常事象と推測して提示するものではないことを示している．ログ情報に含まれる異常事象には出現頻度の低い情報だけではなく，短時間に集中して大量のログ情報を出力する場合や，平常時には周期的に出力されるログ情報が例外として非周期的な時刻に出力される場合も考えられる．さらに前述のような傾向もなく，突発的に平常時とは異なるメッセージを出力する場合も考えられる．これらの異常事象を検出するためには，時刻情報に関する周期的特徴や出力傾向ならびにメッセージ長といった特徴情報を抽出し，それらを図化して調査者に提示することが有効であるといえる．この特徴により見えログは単に出現頻度の低い情報を提示するだけでなく，それ以外の見地からも異常事象として疑わしい情報を抽出することが可能である．

4.3 ログメッセージに注目した抽出

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．見えログではログメッセージを文字で提示しているが，単に文字表示するだけでなく，メッセージを構成している単語が持つ特徴をあわせて提示している．

図 13 は，ログメッセージ表示領域における同一ログ情報の表示例を 3 つ示している．3 つの図の違いは，特徴情報抽出処理で抽出した頻度情報を利用し，単語の出現頻度に基づいた特徴情報をあわせて表示しているか否かである．図 13 左は頻度情報に基づく特徴情報の提示をまったく行っていない例である．赤色でハイライトされている単語は，パターンマッチングによりシステム管理者が事前に定義したキーワードに一致した単語を表している．

その他の 2 つの表示例は，頻度情報に基づき，出現頻度の低い単語を明確に表示している例であり，調査者が指定した閾値より出現頻度の低い単語が青色の背景色でハイライト表示される．図 13 中は閾値を出現

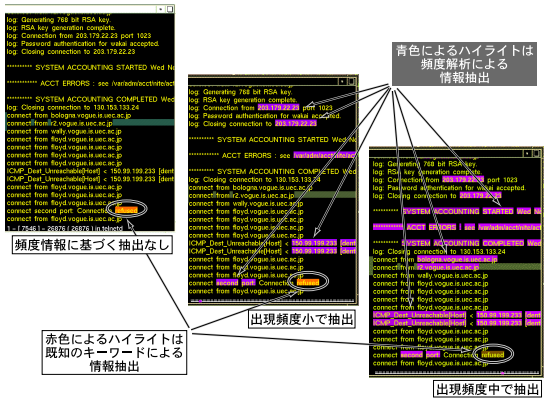

図 13 ログメッセージに注目した調査例
Fig. 13 An example of an investigation focusing on log messages.

頻度 “小” に設定し，図 13 右は閾値を出現頻度 “中” に設定して抽出した例である．閾値を大きくした図 13 右の表示の方が，より多くの単語をハイライト表示していることが認識できる．

このように，出現頻度の低い単語をハイライト表示で明確にした後にログ情報の閲覧を行うことにより，従来の方法による調査では見落としがちな異常事象を表すログ情報の抽出を支援することが可能である．

5. 考　　察

5.1 利　　点

見えログはログ情報の調査を支援する目的でその特徴情報を抽出し，図的な情報提示と対話的な操作によるログ情報の調査作業を可能にした．本システムによって得られる利点を以下にあげる．

- ログ情報の統合
  見えログでは汎用ログフォーマットとして定義した中間形式を使用することで複数のログ情報を 1 つに統合可能にしている．この特徴により，複数のログ情報を一度に調査することが可能になり，ログ情報を個々に調査する手間を省くことが可能になる．また，複数のログ情報間の時間関係が明確になるという利点もある．
- 異常事象を表すと推測されるログ情報の抽出支援
  見えログでは異常事象を表すと推測されるログ情報の抽出のため，テキストマイニングを用いてログ情報から特徴情報を抽出した．この情報を利用することで，文字情報を読んで調査するだけでは把握が困難な時刻情報や大域的な見地からのログ調査を可能にする．また本手法では，事前に情報抽出規則やキーワードを定義する必要がないため，調査作業に必要となる知識的な要件を低くおさえることが可能になり，初級システム管理者でもログ情報の調査を行うことが可能になる．
- 情報視覚化を用いた情報提示
  特徴情報抽出処理で有用な情報が得られたとしても，それらの情報をユーザが把握し，疑わしいと推測されるログ情報の抽出に利用できなければ意味がない．見えログでは抽出した特徴情報を情報視覚化を用いて図化して提示しているため，特徴情報の把握が容易であり，ログ情報に対する認識負荷を軽減することが可能である．また 4 章でも述べたが，情報視覚化によってログ情報が図としてユーザに提示されることで，頻度情報だけでは抽出することの困難な異常事象の抽出が可能になる．つまり情報視覚化を用いることにより，頻度情報をもとにした手法以外にも異常事象の抽出手段を与えることが可能になるとともに，その判断を人間が対話的に行うことが可能になる．
  また特徴情報と文字による表示の関連性が視覚的に明示されているため，図的表現内から異常事象を抽出した際に，その事象に関する詳細情報が即座に得られるという利点もある．

5.2 今後の課題

本節では，見えログにおける今後の課題について述べる．

1 つ目の問題は，ログ情報を注意深く “見る” 必要があることである．この問題に対し，出現頻度の低いログ情報の存在を音で通知する方法や，ログ情報を要約し，見るべきログ情報量を減少させることが必要であると考えている．

2 つ目は，定型処理の自動化である．未調査のログ情報抽出や，ログ情報を自動的にスクロールすることによるログ情報の閲覧，異常事象と推測されるログ情報抽出の基準値設定など，調査時に必要とされる定型作業を自動化することが望ましいと考えている．

3 つ目は，抽出された情報と異常事象を表すログ情報間の妥当性である．見えログでは出現頻度をもとに異常事象であると推測されるログ情報を抽出しているが，それらが必ずしも一致しているとは限らない．抽出されたログ情報と異常事象との妥当性について評価し，その結果をもとに異常事象の抽出処理を改善する必要がある．現状においてもログメッセージに含まれる数値などは重複して出力される可能性が低く，それゆえに数値情報のほとんどが異常事象を表すログ情報として出力される問題が発生しており，頻度解析の対象を限定するなどの対策が必要であると考えている．

表 1 見えログの性能指標
Table 1 An index of MieLog's process time.

処理名称	処理時間（sec）
汎用ログフォーマットへの変換処理	2.63
特徴情報抽出処理	49.54
特定キーワードを持つログ情報の抽出処理	38.50

4つ目としては，実行速度の改善と大規模ログ情報への対応があげられる．

表 1 は，5 MB（45005 行）の syslog ファイルを調査対象とした場合における見えログの各処理時間である．測定に使用した計算機は RedHat Linux 6.1J（CPU: Intel PentiumIII 650 MHz Memory: 64 MB）である．なお特定キーワードを持つログ情報の抽出処理とは，対話的処理の一例として取り上げた処理である．この処理を取り上げた理由は，全ログ情報を走査するため対話的処理の中でも最も時間のかかる処理であり，ログ情報の規模が直接処理時間に影響するからである．

この結果から，大量のログ情報調査を行う際には特徴情報抽出処理および全ログ情報を走査する処理において相当量の処理時間が必要となることが分かる．ただし，汎用ログフォーマットへの変換処理およびテキストマイニングによる特徴情報抽出処理は起動時にのみ行う処理であり，ログ情報の調査作業中に必要となるわけではない．また本論文で述べた他の対話処理の多くは実用的な時間で動作する．今後の課題として，より大規模のログ情報を対話的に処理可能にするよう処理内容や実装方法を見直す必要がある．

最後の問題としてログ情報監視変換処理の強化があげられる．現在，本処理は個々のログ情報に対するコマンドとして実装されており，見えログを使用する前に，コマンドを実行してログ情報を汎用ログフォーマットに変換し，統合する必要がある．この一連の作業を自動的に行うシステムを構築する必要がある．

5.3 関連研究

関連研究について述べる．

SeeLog[4]は本システム同様にログ情報の調査支援を目的としたシステムである．しかし，その表示はアウトライン表示だけであり，文字表示によるブラウジングは不可能である．また視覚的表現による情報抽出は多くの既定規則に基づいて行われているため，その規則を定義する必要もある．さらに時刻情報は，各ログ情報が出力された時間帯としてしか認識できないという問題がある．

Swatch[8]や Logsurfer[9]，syslog-ng[10]はいずれもキーワードを事前に定義したうえでパターンマッチングを用いて既知の異常事象と推測されるログ情報を抽出する手法である．これらのシステムの問題点は3つある．1つ目は，情報抽出のためにキーワードを定義しなければならず，事前知識が必要となることであり，2つ目は，結果として得られる情報が文字として提示されるため認識負荷の問題が残ることである．3つ目は抽出されたログ情報が個別に提示されるため，それらの中に異常事象を表すログ情報が存在したとしてもその付近で発生している事象やその異常事象に関連すると推測される他の事象の調査はそれぞれ別に行わなければならないという問題がある．

Xlogmaster[11]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法は文字による表示であるため，ログ情報に対するユーザの認識負荷は軽減されず，概要把握を行うことも困難であるという問題がある．

なおログ情報の調査にデータマイニングを応用する研究はすでにいくつか存在する[5),6)]．これらの研究結果からも不正侵入検知やログ情報の解析に対するデータマイニングの有効性は指摘されている．

6. おわりに

本論文では，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“見えログ”について述べた．

ログ情報調査における問題点は，ログ情報の膨大な量とシステム管理者が注目すべきログ情報の不明確さにある．そこで本研究では，注目すべきログ情報は大量のログ情報の中に少数埋もれているという特徴に注目した．そこで我々は，テキストマイニングを用いてログ情報から種々の特徴情報を抽出し，異常事象と推測されるログ情報の抽出を支援する手法を提案した．また得られた特徴情報を情報視覚化を用いて図化して提示するとともに，対話的に調査可能なブラウザシステムとして構築した．

見えログでは種々の形式で記録されているログ情報を汎用ログフォーマットに変換して利用する．これにより複数のログ情報の調査を一度に行うことが可能になると同時に，ログ情報間に存在する時間関係も明確になるという利点がある．これによりログ情報収集時の問題点と複数のログ情報を調査をする際の作業負担を軽減することが可能になる．

本システムを用いることにより，人間が文字情報を読むだけでは認識することの困難な種々の情報を加味しながらログ情報の調査を行うことが可能になり，従来よりも確実に異常事象を表すログ情報の抽出が可能になる．

今後の課題としては，本手法で抽出されるログ情報と調査者が注目すべき情報間の整合性を評価し，その結果を抽出手法の改善に反映させる必要がある．またログ情報量が増加するに従い人間が閲覧して調査することが困難になるため，いかにして注目すべきログ情報を限定するかを考慮する必要がある．さらに調査作業の自動化についても考慮する必要がある．

参 考 文 献

1) 川又英紀，志度昌宏：セキュリティは万全か，日経コンピュータ，pp.124–148 (Dec. 1997).
2) 佐々木元也：ファイアウォールを見直す：ログのチェックは不可欠確実にアタックを発見し対策を，日経 Internet Technology，pp.90–94 (Aug. 1999).
3) 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/hightech/enquete/
4) Eick, S.G., Nelson, M.C. and Schmidt, J.D.: Graphical Analysis of Computer Log Files, *Comm. ACM*, Vol.37, No.12, pp.50–56 (1994).
5) Lee, W. and Stolfo, S.: Data Mining Approaches for Intrusion Detection, *Proc. 7th USENIX Security Symposium* (Jan. 1998).
6) Cox, K.C., Eick, S.G., Wills, G.J. and Brachman, R.J.: Visual Data Mining: Recognizing Telephone Calling Fraud, *Journal of Data Mining and Knowledge Discovery*, Vol.1, No.2, pp.225–231 (1997).
7) James, A.H.: Audit Log Analysis Using the Visual Audit Browser Toolkit, Computer Science Department U.C. Davis Technical Report (CSE-95-11) (1995).
8) Hansen, S.E. and Atkins, E.T.: Automated System Monitoring and Notification With Swatch, *USENIX 7th System Administration Conference* (Nov. 1993).
9) Ley, W. and Ellerman, U.: Logsurfer (1995). http://www.cert.dfn.de/eng/logsurf/
10) BalaBit: The free software company, syslog-ng. http://www.balabit.hu/products/syslog-ng/
11) Greve, G.C.F.: The Xlogmaster (1998). http://www.gnu.org/software/xlogmaster/
12) 那須川哲哉，諸橋正幸，長野 徹：テキストマイニング—膨大な文書データの自動分析による知識発見，情報処理学会学会誌，Vol.40, No.4, pp.358–364 (1999).
13) 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).
14) Adriaans, P. and Zantinge, D.（著），山本英子，梅村恭司（訳）：データマイニング，共立出版 (1998).

(平成 12 年 5 月 1 日受付)
(平成 12 年 10 月 6 日採録)

高田 哲司（学生会員）

2000 年電気通信大学大学院情報システム学研究科情報システム運用学専攻博士課程修了．工学博士．情報視覚化の研究に従事．特に情報視覚化，不正侵入検知に関心がある．IEEE/CS 会員．

小池 英樹（正会員）

1991 年東京大学大学院工学系研究科情報工学専攻博士課程修了．工学博士．同年電気通信大学電子情報学科助手．1994 年同大学大学院情報システム学研究科助教授．現在に至る．1994〜1996，1997 年 U.C. Berkeley 客員研究員．情報視覚化の研究に従事．特に視覚化へのフラクタルの応用，情報検索システム，パーセプチュアルユーザインタフェース，情報セキュリティに興味を持つ．1991 年日本ソフトウェア科学会高橋奨励賞受賞．IEEE/CS，ACM，日本ソフトウェア科学会各会員．