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解析支援システム

MieLog: Log Information Browser using Information Visualization and Text Mining

高田 哲司
Tetsuji TAKADA

電気通信大学 SVBL
(〒182-8585 東京都調布市調布ヶ丘1-5-1 E-mail: zetaka@csrs.is.uec.ac.jp)

ABSTRACT. It is necessary for system administrator to inspect computer log-files. As the Internet becomes essential in our life, system administration task get a more important than before.
An inspecting computer log-files is a one of the basic task in system administration. There are, however, mainly two problems on that task, namely the difficulty in recognizing the log information by human and extracting a valuable information from a massive amount of log information. Most of administrators, therefore, do not carry out it sufficiently. In this project, we developed a visual log browser named "MieLog". It assists system administrators to inspect computer log-files. MieLog makes use of two methods to help an administrator inspecting log information. These methods are an information visualization and text mining. Information visualization is useful to reducing the recognition load in textual information. It also makes MieLog a highly interactive system. Text Mining enables to extract the valuable information without a domain specific knowledge. We give three examples that illustrate the ability to extract the valuable log messages easily from the computer log.

1 背景

今日，インターネットを介した電子メールやWebの利用は"普及"という段階を過ぎ，もはや社会基盤となりつつある．その一方で，この社会基盤を快適かつ安全に使用するために必要不可欠なのだが、あまり注目されていない一面がある．それはこれらの基盤を提供する計算機の管理作業である．

この作業は，ネットワークに接続しているすべての計算機に必要となる作業であり、なんらかのサービスを提供しているのであれば，その重要性はなおさら高いものとなる．しかし，一方でこの作業には専門的な知識と高度な技能が必要とされるため，この作業を実施できる計算機管理者は決定的に不足している．このような状況から，多くの計算機運用環境では維持管理作業が確実に行われていないと推測される．

また，近年ではこれに"セキュリティ"すなわちコンピュータウィルスや計算機への不正侵入，計算機上での不正行為という問題が多数発生しており，事態の悪化に拍車をかけている．この問題の特殊な点は，自分が意図したものでなくても加害者になりうることである．つまり，"自分の計算機は不正侵入されてもかまわない"という安易な決断が，自分自身の計算機だけでなく，自身の計算機を通じて他の計算機にまで被害を及ぼす恐れがあるためである．

このような状況から，計算機の維持管理作業はますますその重要性を増しつつある．しかし，その作業は地道で手間のかかる作業が多く、その作業支援が望まれている．またこの作業は，セキュリティの観点から"望まれている"という段階から"当然すべき行為"となりつつあり，その事実は「不正アクセス行為の禁止等に関する法律」[1]の第五条にも明記されている．

このような背景をふまえ，本プロジェクトでは計算機の管理作業の中でも最も基本的な作業である"ログ情報の調査"に焦点をあて，その作業を支援するシステムとして"見えログ"の開発を行った．本論文では，ログ情報ブラウザ"見えログ"について，その開発目的，システム概要と使用例について述べる．

2 目的

本プロジェクトでは前述の背景をふまえ，システム管理者によるログ情報の調査を支援するシステム"見えログ"を構築した．見えログは，その作業を支援するために「情報視覚化」と「テキストマイニング」という2つの技術を利用している．

ここでいうログ情報の調査とは「Operating System(以降OSと略す)やサーバプログラム，アプリケーション等が生成するログ情報を対象とし，その情報の中から計算機管理上有用な情報を探索/抽出する行為」を指す．本章では，このような作業を支援する理由と，これらの技術を利用した理由の二点について述べる．

まず始めに，ログ情報の閲覧/調査作業を支援する理由について述べる．

計算機の管理作業とは，すなわち計算機を利用して提供している機能やサービスが正常に稼働していることを確認する作業である．

これを行うためには，サーバプログラムやシステムが出力する「ログ情報」を閲覧し，その情報の中になんらかの異常を示す記録(メッセージ)がないことを確認する必要がある．なぜならば，サーバプログラムやOSの多くは，自身の動作に異常が発生した場合にそれを計算機管理者に通知する機能を持たないからである．そのかわりに，そのほとんどが自身の状態や異常事象をログとして記録している．この機能は，現在，主として使用されているOSや電子メール，Web等のサービスを提供するサーバプログラムには，ほぼ間違いなく提供されている機能である．

また，セキュリティ強化のために導入される監視システムも，そのほとんどは計算機の状態を時々刻々とログと

して記録するだけである[1]．つまりセキュリティ監視を目的としたシステムですら，記録されたログ情報をのちに監査，すなわち調査することを前提に作られている．これらの事実から，ログ情報の調査がいかに大事な作業であるかということを認識することができる．

最後に，この作業が “終わりのない作業” であることについて述べる．ログ情報の調査作業には終わりがない．つまり，日々継続してこの作業を行うことが必要である．なぜならば、サービスを提供するサーバプログラムやOSは稼働し続けるため，ログ情報は生成され続けるとともに，これまでにまったく異常事象が発生しなかったとしても，この先，異常が発生しないという保証はないからである。

またセキュリティ上の観点から，この作業頻度を高める必要があると言われている．つまり頻繁にログを調査し，セキュリティ上の問題が発生していないことを確認する必要がある．なぜならば，セキュリティ違反に関しては，時間的な側面が重要であり，可能な限り迅速に不正行為を認識し、その対処を行うことが望まれるからである．

これらのことから，計算機の管理作業においてログ情報の調査は重要な作業であることは明らかである．したがって，この作業を支援することは大きな意味を持つと言える．

次に2つの主要技術，すなわち「情報視覚化」と「テキストマイニング」を本作業の支援に利用した理由について述べる．

まずはじめに，ログ情報を調査する際の問題点を以下にあげる．

- 文字としての記録
- 膨大な情報量
- 情報源の多様性と偏在性
- 異常事象抽出の困難さ

まず1つめは文字としての記録である．

ログ情報は，その多くが文字による平文で記録される．一方，これを調査するのは人間である．したがってログ情報を把握するためには，人間が記録されたログ情報を一つ一つ読んで，その内容を理解しなければならず，その作業負荷は高い．また，仮にそれらの情報を読んだとしても，一度に知ることのできるログ情報は膨大な量のログの一部分であり，概要を把握するのは困難である．

つまりログ情報は文字として記録されるがゆえに2つの問題がある．それは認識負荷が高いことと，概要把握が困難であるということである．

2つめの問題は，その膨大な情報量である．

ログ情報は，OSやサーバプログラムが稼働しているかぎり継続して生成されるため，増加の一途をたどる．したがってログ情報は，膨大な量になる傾向が高い．

一方，このログ情報を調査するのは “人間” である．したがって「膨大な量のログ情報」と前述の「読んで理解する」という2つの点から，調査者は “膨大な量のログ情報を読んで理解する” という単調な作業を長時間強いられることになる．これこそが，計算機管理者がログ情報の調査を敬遠する主たる原因であると考える．

3つめは情報源の多様性と偏在性である．

ログ情報は，その記録されている情報の内容がそれぞれ異なり，またその記録形式も統一されていない．したがって計算機管理者は，ログを記録する主体ごとにどのような情報がどのように記録されるのかを知っている必要がある．

また，ログ情報の存在場所も一意には決まっていない．つまり，ログが記録されるログファイルの存在場所(ディレクトリ)は，OSやサーバプログラム毎に異なる．計算機管理者は，ログ情報の存在場所に関する知識も必要とされる．

最後の問題は，異常事象を抽出することの困難さである．

ログ情報にはさまざまな情報が記録されるが，そのすべてが異常を示す情報ではない．むしろ異常事象を示すログメッセージは，膨大な量のログ情報の中にごく少数存在すると言われている[2]．

また，異常事象を示すログ情報が不明確であることも問題としてあげられる．現状では，異常事象発生時にどのようなログメッセージが記録されるかが明確になっていないことが多く、またそれらは個々の異常事象や計算機の運用環境にも依存する．したがってこれらの理由から，一般的に使用可能な異常事象抽出規則の構築も困難であるといえる．

さらに場合によっては，複数のログ情報から断片的な情報を得たうえで，それらを総合して判断を行わなければ異常事象を認識できない場合もある．

このように，計算機管理におけるログ情報の調査作業にはさまざまな問題が存在する．そこで本プロジェクトではこれらの問題を改善するため「情報視覚化」と「テキストマイニング」を利用する．

情報視覚化とは，情報を図として抽象化して表示することにより，人間の認識特性を生かし，情報への理解をより早くより深くする手法である[3]．我々はこの技術を利用することにより，ログ情報に対する認識負荷の問題，すなわち前述の “文字としての記録” と “膨大な情報量” という二つの問題の改善を目指す．

ログ情報が視覚化されることにより、調査者は文字記録としてのログ情報を読むという手段のほかに，抽象化された図的表示を見るという手段を得ることになる．つまり，文字情報を読む必要性は依然としてあるが、その前に図的表示を閲覧し，ログ情報に対する理解を深めることが可能になる．

また情報が図化されることにより，文字による情報提示法よりも，より多くの情報を一度に見ることが可能になる．さらに情報視覚化に際してログ情報が解析されるため、ログ情報が持つ “特徴” を知ることも可能になる．これらの利点を利用することにより，調査者のログ情報に対する認識負荷を軽減可能にする．

一方，テキストマイニングとは，様々な観点からデータを解析し，有用な知識や情報を取得しようとする技術であるデータマイニングの一手法である．データマイニングとテキストマイニングの違いは対象となるデータにある．データマイニングは，その処理対象がデータベース内のテーブルに格納されたデータであるのに対し、テキストマイニングは，形式化されていない通常文書を処理対象とする．

我々は，ログ情報にテキストマイニングを適用することにより，いくつかの特徴情報を抽出し，それを利用した異常事象の抽出を行う．その理由は，前述の通り異常事象を抽出するための規則を構築することが困難だから

[1] 一部のシステムにはmailや音にてシステム管理者に通知する仕組みがある

である．抽出すべき異常事象がすべて既知ならば，その知識を基になんらかの方法で異常事象を抽出することは可能であるが、現実はそうではない．また，この抽出規則が計算機の運用環境や管理者の知識にも依存することも知られている．さらに，これまでには知られていない異常事象がこの先発生する可能性もある．

したがって，既知の異常事象以外にも異常事象と推測されるログメッセージを抽出する仕組みが必要となる．そこで本プロジェクトでは，テキストマイニングを利用した出現頻度に基づく異常事象の抽出手法を利用する．この手法の大きな利点は，計算機管理者に事前知識を必要としない点である．

テキストマイニングにより抽出される特徴情報とは出現頻度である．現在，ログに記録されているログ記録時刻，ログの出力主体，ログメッセージ中の単語を基準としてその出現頻度を取得している．これらの情報を利用することで異常事象と推測されるログメッセージを抽出可能と考える理由は，異常事象を示すログメッセージが膨大な量のログ情報の中にごく少数存在する [2] からである．

また，これとは逆に異常事象が大量のログメッセージとして出現する場合も考えられる．この場合は「大量のログメッセージが突発的に出力される」とか「ある時点を境に大量のログメッセージが出力され始める」という平時とは異なるログ出力パターンとして出現すると考えられる．このような場合にも出現頻度に基づく抽出は有効であると考えられるとともに，この場合には視覚化による図的表現によってもその抽出が可能である．

次章では，これらの目的をもとに開発されたログ情報ブラウザ “見えログ” の詳細について述べる．

3　見えログ: システム概要

本章では，見えログのシステム概要と特徴情報の抽出そしてその視覚化手法について説明を行う．

図1は見えログの処理概要図である．この図からわかる通り，見えログは大きく3つの処理から構成されている．以下ではそれぞれの処理について説明を行う．

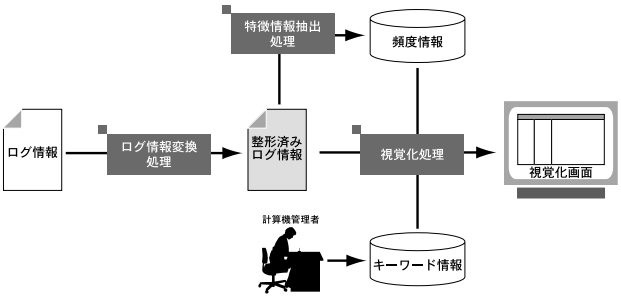

図 1: 見えログのシステム概要

a) ログ情報変換処理

ログ情報変換処理は，様々な記録形式と情報を持つ種々のログ情報を，中間フォーマットに変換する．この中間フォーマットは4つの情報から構成されており、UNIX系OSのsyslogが出力するログメッセージの記録フォーマットに似たものとなっている．我々はこのフォーマットを汎用ログフォーマットと定義する．汎用ログフォーマットの形式とその変換例を図2に示す．

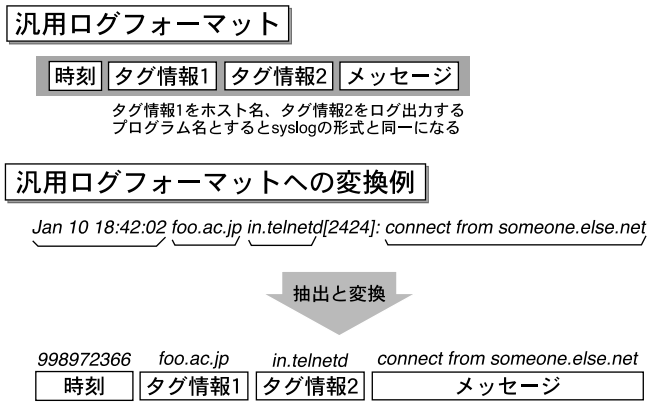

図 2: 汎用ログフォーマットとその変換例

4つの情報のうち，時刻は秒数として表現されるが，それ以外は任意の文字列として定義される．この処理を設けた理由は，ログ情報の多様性と偏在性の問題を改善するためである．この処理を用いることにより，ログ調査者はログファイルがどこに存在するかを気にする必要がなくなる．またログ情報がある一定の枠組みにおいて変換されるため，記録形式の多様性についても一定の改善を見ることができる．またこの変換により，記録形式の異なる複数のログ情報を時刻を基準として一つのログ情報に統合することが可能になる．この統合化されたログ情報を調査することにより，各時刻における複数のログからの情報を容易に知ることが可能になり，複数のログ情報間の関連が明確になる．結果として，異常事象を抽出する際の問題である “総合的な判断の必要性” を改善することが可能になる．

b) 特徴情報抽出処理

見えログでは，汎用ログフォーマットという中間フォーマットに変換されたログ情報から特徴情報を取得する．本処理部で取得する特徴情報は以下の通りである．

- タグ情報に基づく頻度情報
  タグ情報ごとにログ出力数を集計する (図3)．

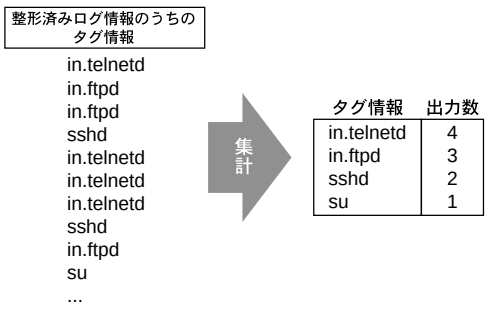

図 3: タグ情報に基づく特徴情報の抽出

- 時刻情報に基づく頻度情報
  時刻情報をもとに，ユーザが決定した時間間隔毎のログの出力数と毎時，毎曜日を単位としたログ出力数を集計する (図5)．すなわち，ログが出力された全時間帯におけるログメッセージの出力傾向と，1日ならびに1週間単位における周期的特徴を得ている．

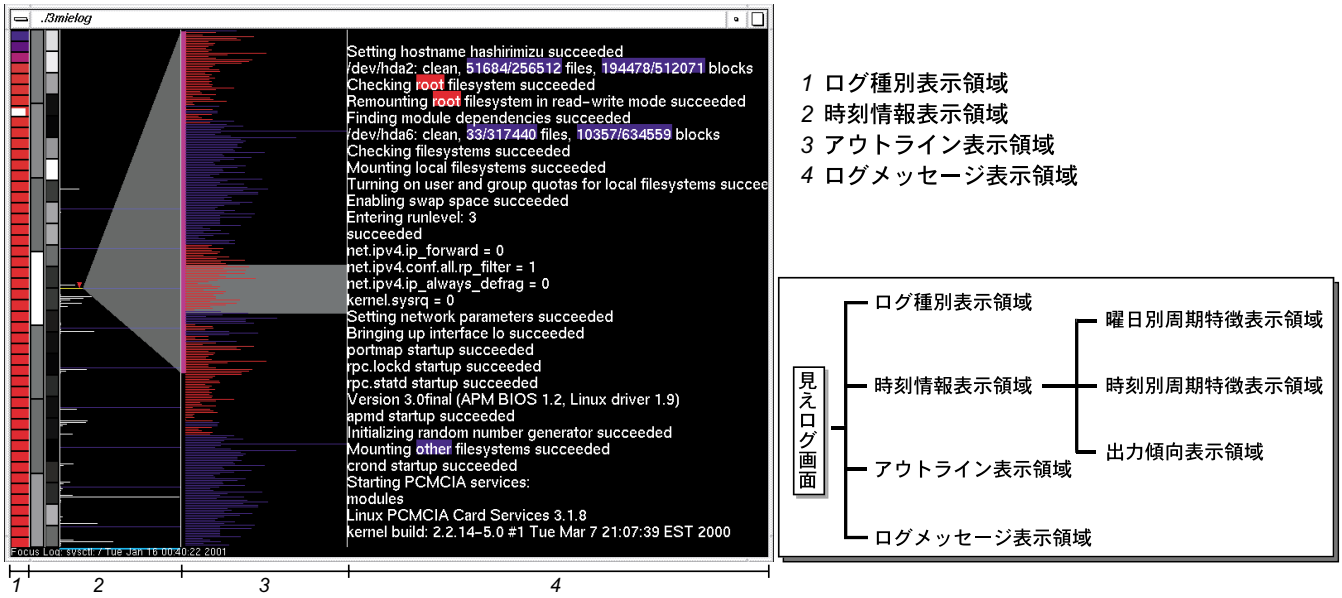

図 4: 見えログの視覚化画面

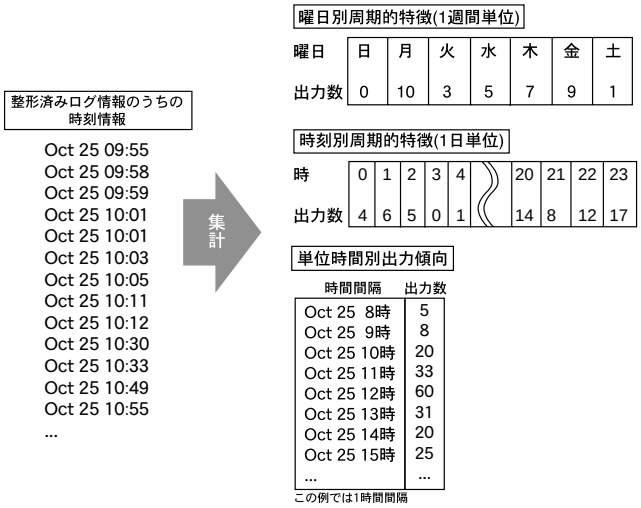

図 5: 時刻情報に基づく特徴情報の抽出

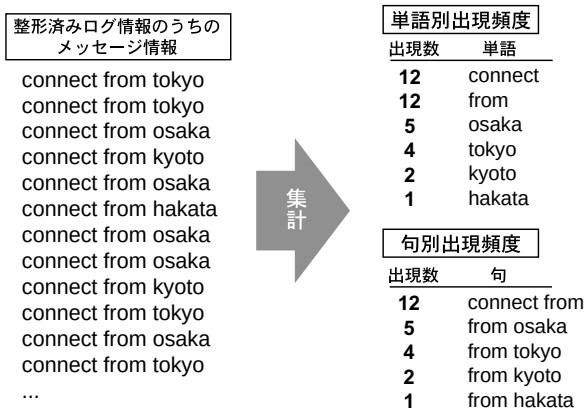

図 6: メッセージ情報に基づく特徴情報の抽出

- メッセージに基づく頻度情報

  メッセージ情報からは，メッセージを構成する各単語と句の出現数を集計している (図6)．ここでいう句とは連続した二単語からなる “句” と定義する．句による出現数を求めた理由は，出現頻度の高い単語同士が，出現頻度の低い句を構成する可能性があるからである．

c) 視覚化処理

視覚化処理部では，図1からもわかる通り中間フォーマット化されたログ情報と特徴情報およびキーワード情報を基にログ情報を図化する．見えログの視覚化画面を図4に示す．

画面は横方向に4つの部分領域に分割することができ、それぞれ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呼ぶ．また時刻情報表示領域内はさらに3つの領域に分割されており，左からそれぞれ曜日別周期特徴表示領域，時刻別周期特徴表示領域，出力傾向表示領域と呼ぶ．

ここでは各表示領域における視覚化手法について説明する．

1. **ログ種別表示領域**

ログ種別表示領域では，特徴情報抽出処理で得られたタグ情報に基づく頻度情報を図化してユーザに提示する．視覚化方法は，タグ情報を縦方向に格子として表現する．また，各格子の色によってタグ情報の出力数を表現しており，その出力数が多いほど青，少ないほど赤色となっている (図7)．

2. **時刻情報表示領域**

時刻情報表示領域は，前述のようにさらに3つの領域に分割されている．この領域の左側にある2つの格子状領

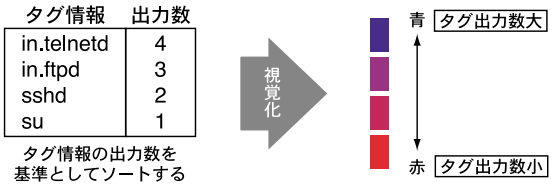

図 7: ログ種別表示領域の視覚化手法

域が，曜日別ならびに時刻別周期特徴表示領域となっており，それは格子の数が 7 個 (曜日別 日〜土) と 24 個 (時刻別 0〜23 時) であることからも容易に認識できる．

この格子状表示の視覚化手法は前述のログ種別表示領域と同様である．ただし，格子の色付けのみ時刻情報表示領域とは異なっており、この表示領域では赤から青の色遷移ではなく，白黒のグレー階調を使用し，白いほど出力数が多く，黒いほど出力数が少ないことを意味している．

出力傾向表示領域の視覚化手法は，時間軸が上から下へ，出力数が左から右への基準軸を持つヒストグラムとして図化している (図 8)．これによりログ情報の各時刻における出力状況を容易に把握可能にしている．

なお画面中央部に存在する半透明の三角形は，時刻情報表示領域と次に説明するアウトライン表示領域の関連性を示している．見えログでは，調査者が現在注目しているメッセージを “注目ログ” として定義し，そのログメッセージが画面中央部に表示されるという表示法を採用している．したがって，画面中央部に表示されている時間帯のログ情報が，アウトライン表示領域やログメッセージ表示領域に表示される．つまり半透明の三角形は，現在の注目ログの出力時刻が出力傾向表示領域のどこに該当するのかを表すとともに，その出力時刻と同じ時間帯に出力されたログ情報は，アウトライン表示領域のどの部位になるのかを示している．

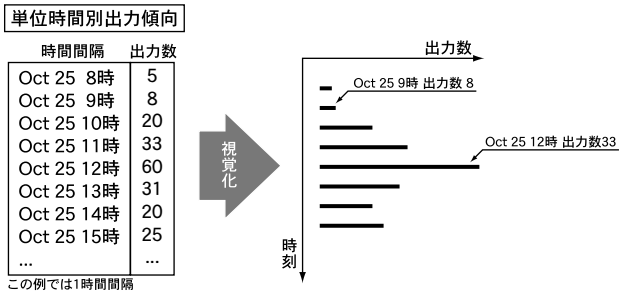

図 8: 出力傾向表示領域の視覚化手法

3. アウトライン表示領域

アウトライン表示領域は，ログメッセージの概略を提示する．

この領域における視覚化方法は極めて簡単である．各メッセージをそのメッセージ長に対応する長さの線として描画する．また線の色はログ種別表示領域で定義された色と同様であり，各ログメッセージのログ種別に対応する色が各線に着色される．つまり，アウトライン表示は文字として提示されたログ情報を “遠目” から見たようになる (図 9)．

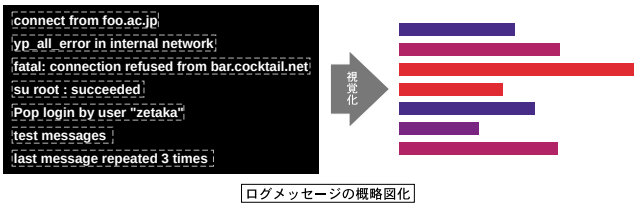

図 9: アウトライン表示領域の視覚化手法

これにより，ログメッセージ群を “図” や “パターン” として認識することが可能になる．したがって，文字による情報提示よりもより多くのログメッセージを一度に閲覧することが可能になるとともに，一連のログメッセージの概要把握を可能にする．また線の色がログ種別毎の出力数に依存しているため、色によって出力数が多いログ種別からのメッセージか否かを即座に判断可能である．これにより，ログ情報を直接読む前にいくつかの点において異常事象と推測されるメッセージか存在するか否かを判断することが可能になる．

また，この領域にも画面中央部に半透明の枠が存在する．この枠はアウトライン表示領域とログメッセージ表示領域との関連を示している．つまり次に説明するログメッセージ表示領域に表示されているログメッセージ群がアウトライン表示領域のどの部分に該当するのかを示しているのである．

4. ログメッセージ表示領域

ログメッセージ表示領域では，エディタやページャでログ情報を見るのと同様，文字によりログ情報を提示している．ただし単にログ情報を文字で表示するだけでなく，注目すべきログであることを示す単語をハイライト表示する (図 10)．

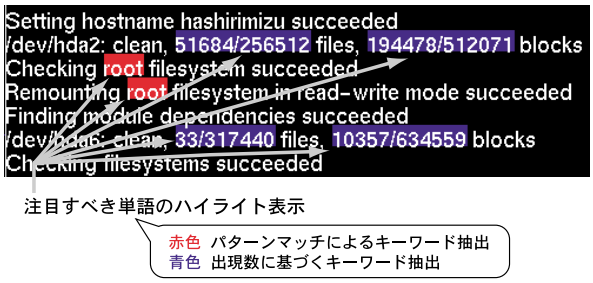

図 10: ログメッセージ表示領域におけるハイライト表示

このハイライト表示は 2 種類あり、色で分別されている．赤色のハイライト表示は，既定のキーワードに一致した単語を表している．この既定のキーワードは，計算機管理者の経験や知識をもとに事前に定義されるものである．一方，青色のハイライト表示は，特徴情報抽出処理にて得られた単語別の頻度情報を利用し，調査者が決めた基準値よりも出現頻度が少ない単語をハイライト表示する．これらの機能により，すでに異常事象を示していることが明らかなログや出現頻度の低い単語を含んでいるログ，すなわち異常事象を示していると推測されるログの抽出を支援する．

4　調査事例と対話的機能

(1) 調査事例

本章では見えログを使って異常事象と思われるログを発見する方法と，見えログが持つ対話的機能について述べる．以下では，3つの例を用いて異常事象と思われるログの発見方法を説明する

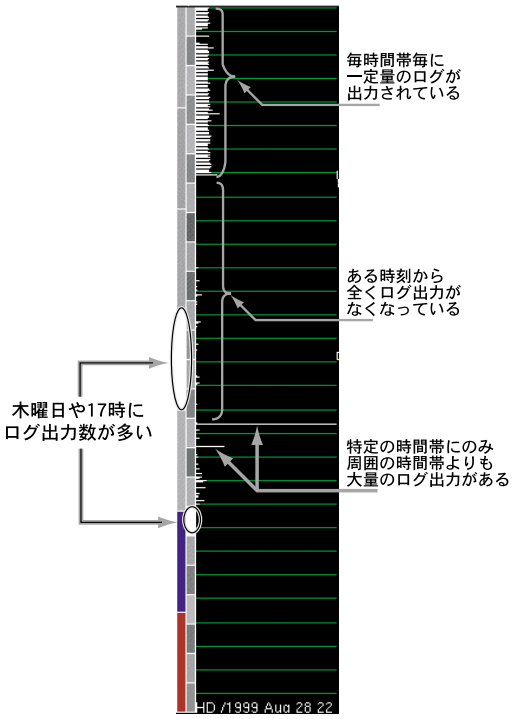

図 11: 時刻情報に注目した調査事例

例1)　図11は見えログの時刻情報表示領域の視覚化例である．この表示から，ログ情報が持つ時刻情報から異常事象と推測されるいくつかの現象を発見することが可能なことを示す．

時刻情報表示領域は3つの領域からなっているのは前述した通りだが，図11の左側の2つの格子状領域，つまり曜日，時刻別周期特徴表示領域を見ると，曜日，時間ともに一ヶ所だけ白色の濃い格子が存在する．これはその曜日ならびに時間帯に大量のログが出力されたことを表している．したがってこれを見た調査者は，その詳細，すなわちその時間帯に出力されているログを調査する必要があると言える．なぜならば，なんらかの異常により大量のログメッセージが出力された可能性があるからである．

一方，出力傾向表示領域を見るといくつかの出力傾向が容易に認識できる．それは以下の通りである．

- ログが記録され始めてからしばらくの間は一定量のログが出力され続けていた
- ある時刻を境にログが出力されなくなった
- 特定の時間帯にのみ大量のログ情報が出力されていた

これらの情報から，ログが出力されなくなった時刻周辺の詳細を調査し，なぜログ情報が出力されなくなったのかを調査する必要がある．また，ログが大量に出力された時間帯についてもその大量に出力されているログ情報の詳細について調査を行う必要があると言える．

このように，ログ情報が持つ時刻という面に注目し、そこから特徴情報を抽出して視覚化することにより，異常事象と推測される事象を抽出できることを示した．これにより調査者は，ログ情報を読まなくても調査すべき契機を得ることができる．

例2)　図12は3つのアウトライン表示領域の視覚化例である．左側は通常時と思われる視覚化例，中央と右側はあきらかに異常と推測される視覚化例である．

左側の視覚化例を見ると，画面中央と下部に同じような出力パターンをもつログの存在が認識できる．また同じく画面下部には，その前後のログメッセージとは明らかに異なるメッセージの存在が認識できる．これらから，発見した出力パターンのログ情報がどのような内容か，またその前後にも繰り返し出力されているかということを調査する必要があるかもしれない．また，前後のログメッセージとは明らかに異なるログメッセージは，異常事象を示していると推測されるため，必ずその詳細を調査する必要があるといえる．

また中央の視覚化例では画面中央より下の部分にほぼ同一色，同一長のメッセージが繰り返し出力されているのがわかる．これはなんらかの原因により，あるプログラムから繰り返しログメッセージが出力されていると推測できる．線の色が青いのはこのログメッセージが繰り返し出力されたために，そのメッセージに該当するログ種別の出現頻度が高くなってしまったためである．これは明らかに平時とは異なる事象なので，その詳細を調査する必要があるといえる．

最後に右側の視覚化例だが、これも中央の視覚化例と同様に平時とは明らかに異なるパターンのログ出力である．ただし中央の視覚化例と異なるのは，メッセージの出力パターンが中央の視覚化例と違ってバラバラであり、またその線の色も赤いことから，そのログ種別自体の出現頻度も低いということがわかる．明らかに平時とは異なる事象なので，その詳細を調べる必要があるといえる．

このようにアウトライン表示は，ログメッセージをパターンとして認識可能にすることで、実際に個々のログメッセージを読む前に何らかの異常事象を発見する契機を与えるものである．

また，この例1,2からわかるように，見えログは出現頻度の低いものに限らず，出現頻度は高いが、異常事象と推測されるものの抽出も可能である．それは，情報視覚化を用いて情報を図化して提示していることに起因する部分が多いといえる．

例3)　図13はログメッセージ表示領域の一部の視覚化例を3つ挙げている．これらはすべて同一ログ情報の同一メッセージ部分を視覚化した例である．これら3つの違いは，頻度情報に基づき低出現頻度の単語をハイライト処理しているか否かである．出現頻度に基づく低出現頻度の単語を知りたいユーザは，まずはじめに基準値となる出現頻度を決定する．するとその基準値よりも出現頻度の低い単語が青色でハイライト処理され，どのログメッセージが低出現頻度の単語を含んでいるかを容易に知ることが可能となる．

左の視覚化例は出現頻度による低出現頻度の単語抽出を行っていない例である．画面下部にパターンマッチングにより抽出された単語が赤色でハイライトされている．

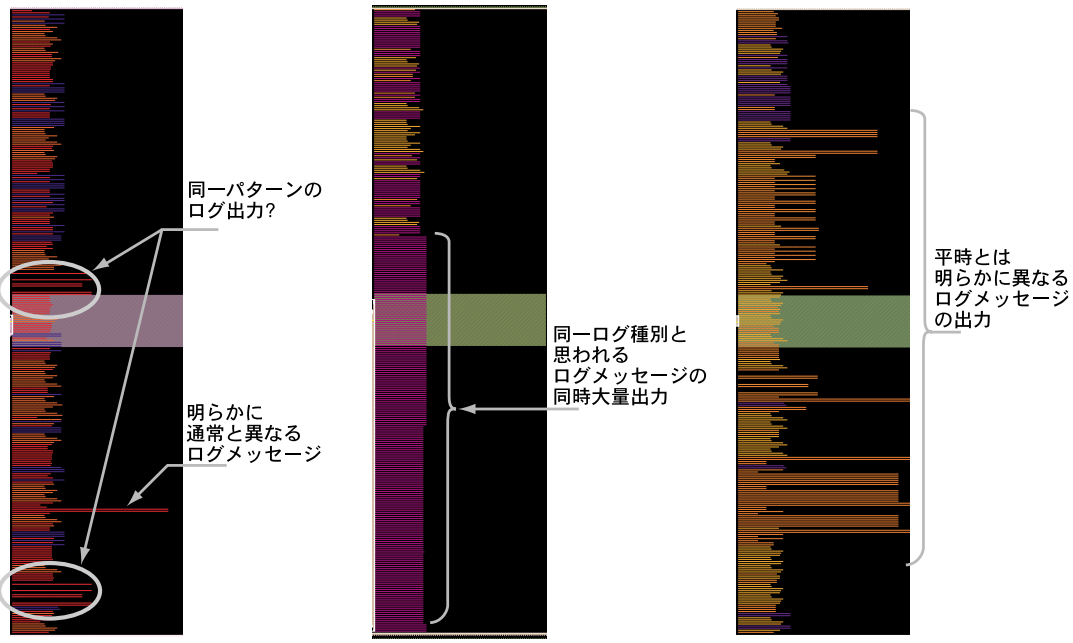

図 12: アウトライン表示に注目した調査事例

これは中央ならびに右の視覚化例でも同様である．中央ならびに右の視覚化例は出現頻度に基づく単語抽出を行っている例である．中央は基準値が小さい値の場合，右は基準値が中程度の場合の視覚化例である．この2つから，右の視覚化例の方が多くの単語をハイライト表示していることが容易に認識できる．

このように，ある一定の基準値をもとに低出現頻度の単語を抽出しハイライト表示することによって，調査者が明確な異常事象抽出規則を持っていなくても疑わしいと推測されるログメッセージを発見することが可能になる．

(2) 対話的機能

見えログでは様々な対話的機能を持っており、調査者が様々な観点からログを閲覧し，異常事象と推測されるログの抽出を支援する．それらの機能の多くは，図化されている情報を直接操作することで実行できる．以下にその機能例を挙げる．

- ページャ機能
- 低出現頻度の単語抽出のためのしきい値決定機能
- タグに基づくフィルタリング処理
- 時刻に基づくフィルタリング処理
- メッセージアウトラインに基づくフィルタリング処理
- 単語や句に基づくフィルタリング処理

5 見えログの開発状況

平成12年度未踏ソフトウェア創造事業では開発者一人で開発作業を遂行した．しかし，残念ながら昨年度の期間内では諸般の事情により目標としていた開発項目をすべて完了することができなかった．

しかし，本プロジェクトはProject Managerの許可を得て，平成13年度も継続して未踏ソフトウェア創造事業として開発が行われている．今年度は昨年度に完了しなかった開発項目に加えて，対話的処理の拡張と実時間モニタ機能の追加という二大機能の追加を目指し，鋭意開発進行中である．

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，その開発にいたった背景からその仕組み，そして調査事例について述べた．

計算機による種々のサービス提供が重要になるにしたがい，計算機の管理作業はその重要度を増している．その中でも基本的な作業であるログ情報の調査を支援するシステムは，今度ますます重要になると考えらる．しかしその一方で，ログの調査作業には膨大な量の文字情報の把握と，計算機管理上注目すべきログの抽出が困難であるという問題が存在しており，これらがその作業遂行を阻害していた．

そこで見えログでは，情報視覚化とテキストマイニングを用いてログ情報の調査作業を支援する．情報視覚化を用いて膨大な量の文字情報把握を支援し，テキストマイニングを用いることで，明示的な規則がなくても，異常事象と推測されるログメッセージの抽出を支援可能にした．また本論文では，3つの調査事例を示し，これらを通じてこれらの機能が異常事象と推測されるログメッセージの抽出に有効であることを示した．

今後は，システムの更なる機能拡張を継続するとともに，システムの有用性評価を行う予定している．

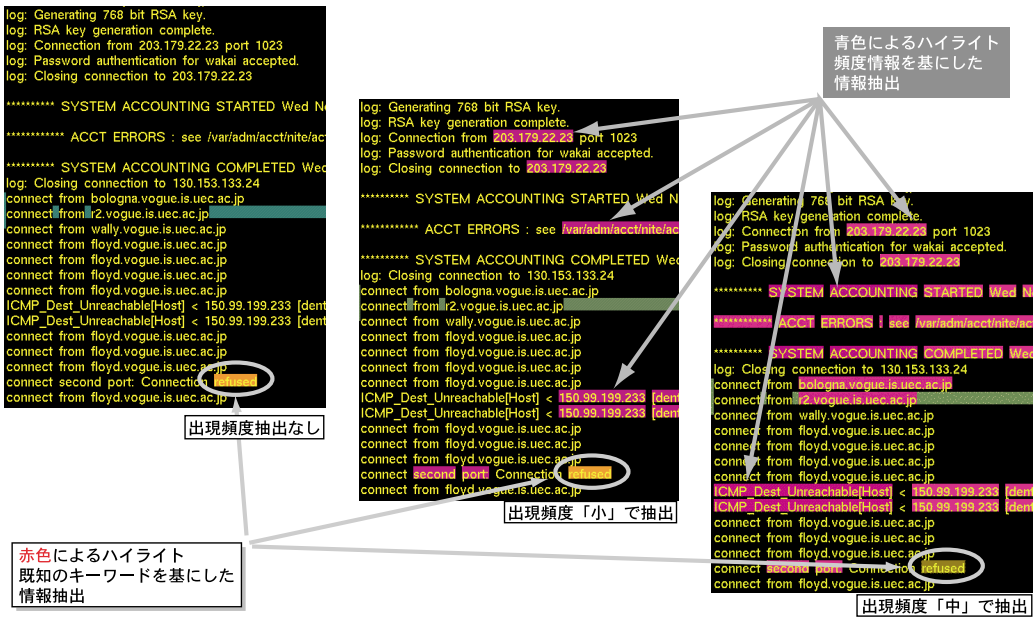

図 13: 単語の出力数に基づいた調査事例

7 参加企業及び機関

平成12年度 未踏ソフトウェア創造事業における開発は本人のみで行われた．平成13年度 未踏ソフトウェア創造事業では，以下の企業から開発支援を受けている．

- SEER INSIGHT SECURITY 株式会社

8 参考文献

[1] 警察庁: 不正アクセス行為の禁止等に関する法律, http://www.npa.go.jp/hightech/fusei_ac2/houann.htm, (Aug. 1999).

[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] 平川，安村(編): ビジュアライゼーション bit別冊 ビジュアルインターフェース – ポストGUIを目指して，pp.29-44，共立出版 (1996).

[4] Greve, G.C.F.: The Xlogmaster (1998). http://www.gnu.org/software/xlogmaster/

[5] Eick, S.G., Nelson, M.C. and Schmidt, J.D.: Graphical Analysis of Computer Log Files, *Comm. of ACM*, Vol.37, No.12, pp.50-56 (1994).

[6] Lee, W. and Stolfo, S.: Data Mining Approaches for Intrusion Detection, *Proc. 7th USENIX Security Symposium* (Jan. 1998).

[7] 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).