解析優先度を考慮した情報追跡ログ解析作業の迅速化

小崎真寛[†]　西岡千文[†]　岡田謙一[††,†††]

情報漏洩対策として企業等に導入される一般的な市販ツールでは，企業内のクライアント PC を監視した際の結果はログとして取得・保存され，このログを解析することで漏洩源の特定や事後対応が可能となる．万が一情報漏洩インシデントが発生した際，このログ解析作業を早急に行なうことは極めて重要である．素早く漏洩原因を特定することで被害の拡大を防ぎ，関連組織へ即座に報告・対応することは会社の信用失墜をできるだけ防ぐことが可能であるためである．従来，ログ解析を迅速化するためのログ解析支援としては解析者に対してテキスト形式でログを提示する手法が主流であった．それに対して，本研究ではログの各行に対してアルゴリズム化された手法で解析優先度を付与し，この解析優先度を解析者に対して視覚的な提示を行うことで解析を支援する．本手法によって機密情報を含んでいる確率が大きいログデータは視覚的に目立つように明るく大きく提示され，逆に機密情報を含んでいる確率の小さいログデータは目立たないように暗く小さく表示される．本論文では，評価として，組織内での具体的な業務をシミュレーションし作成したログを用いた評価実験を行うことで，解析優先度を付与することによって解析を迅速化できることを示す．

A Visualization Method for Log Analysis Considering the Analysis Priority

MASAHIRO KOZAKI[†]　CHIFUMI NISHIOKA[†]
KEN-ICHI OKADA[††, †††]

In this paper, we describe a visualization technique for rapid log analysis. Recently, information leakage is a growing public concern. There are many monitoring tools that monitor operations in each client PC and get monitoring logs. Our proposed technique enables administrators to determine which line of log data should be analyzed and which should not by referencing the analysis priority. In consequence, our method contributes to the prevention of information leakage and rapid report of leak causes to related organization. Lastly, we verify through an experiment that our method is useful for providing rapid analysis.

[†]慶應義塾大学院理工学研究科
Graduate School of Science and Technology, Keio University

[††]慶應義塾大学理工学部
Faculty of Science and Technology, Keio University
[†††]科学技術振興機構
Japan Science and Technology Agency

1. はじめに

情報漏洩インシデントは依然として深刻な問題であり，企業側の対策も多くの手法が実施されている中，その対策は大きく事前対策と事後対策に分類できる．前者は情報漏洩を未然に防ぐことを目的とした対策であり，例えば市販されている一般的な情報漏洩対策ツール 1) は，社内外にある機密データを検出・監視・保護することが可能である．また後者は日常的にクライアントを監視しそのログを取得しておくことで，漏洩インシデントの発生後も迅速な漏洩原因の特定・関連組織への報告に役立てることを目的としている．ここで事後対策に着目すると，万が一情報漏洩が発生した際に早急にクライアント監視ログの解析作業を進めることは極めて重要である．素早く漏洩原因を特定することで被害の拡大を防ぎ，関連組織へ即座に報告・対応することは会社の信用失墜をできるだけ防ぐことが可能であるためである．そこで本論文では，クライアント監視ログの解析を迅速化する可視化手法を提案する．具体的には，まず管理者がヒトの手で解析を行うログデータの各行に，予めシステムの判断で解析優先度を付与しておく．従来のクライアント監視ツール 1), 2) は取得したログを解析者に対してテキスト形式で提示するため，解析作業に手間と時間がかかるという問題があった．それに対して本提案では，システムによる解析優先度をもとにログの各行を視覚的に提示する．機密情報を含んでいる確率が高い文書に関する行は解析者に対して視覚的に目立つように提示し，逆に機密情報を含んでいる確率の低い文書に関する行は目立たないように提示することで，解析作業にかかる手間を簡略化し，解析時間の短縮を図る．本手法の有効性は，文書操作などのユーザ操作をシミュレートして作成したログを用いた評価実験を行うことにより検証した．以降，本論文の構成は以下のとおりである．まず 2 章で本研究と関連する従来研究を紹介し，3 章で提案の概要，4 章で解析優先度付与のアルゴリズムに関して詳述する．5 章では評価実験について記載し，最後に 6 章を本論文の結びとする．

2. 関連研究

漏洩対策としてユーザの操作を監視し機密情報の追跡を実現する製品が多くのベンダーから発表されている．一般的には，これらのツールの多くは導入されたクライアントの挙動を監視し，ログデータとして記録する．ログデータは定期的に管理者のもとへ集約され，その後解析作業に用いられる．クライアントを監視した際に取得する情

報はツールによって様々であるが，例えば MaLion 3 3) では以下のような情報が取得される．

操作されたマシン，ログインユーザ，アプリケーションの起動，デバイス操作 (USB メモリ，CD/DVD/FD ドライブ)，印刷操作，送受信メール，ファイルアクセス （読込み，書込み，移動，コピー，名称変更，削除)，Web アクセス，アクティブウィンドウ，モバイル PC 操作，共有フォルダー，クリップボード

また，SKYSEA Client View 4), 5) においては MaLion 3 と同様の監視データに加え，Web ダウンロード/アップロード，FTP ダウンロード/アップロード等のデータも取得している．さらに，InfoCage 6) では PC を監視した際のログをテキスト形式での表示方法によって定量的に把握が可能である．ゆえに本論文では上記のような情報を含むログデータは取得することが可能であることを前提とし，以下ではログ解析にフォーカスし詳述していくこととする．
従来の監視ツールにおけるログ解析の特徴は以下のようである．まず，ログデータを分析・統計し解析者に提示するものがある 7), 8)．これにより，解析者は組織全体をマクロ視点でとらえたユーザの行動やログの記録内容を視覚的に判断することが可能となる．次に，ログを表形式で解析者に提示し，解析者が検索・絞り込みを行っていくことで解析を行うものがある 6), 7), 9) ．両者とも解析に工夫を施しているが，それぞれ分析・統計データでログ全体の概要を把握した後と検索・絞り込みを行った後に，詳細な情報を得るためテキストベースのログデータを解析する必要がある．そして，既存のツールの殆どがこのときのログデータを表形式で提示している．

3. 解析優先度を考慮したログ解析

万が一企業内の機密情報が漏洩するという事態が発生した際には，ログの迅速な解析作業が重要となってくる．しかし既存ツールによる PC を監視した際に生成されるログは，ほとんどの場合テキスト形式であり解析に多くの手間と時間を要する問題があった．そこでこのような膨大なテキストログを「可視化」することで解析速度や解析精度を高め，セキュリティに貢献するといった研究が注目されつつある 10), 11)．本論文では,解析者が解析を行なうテキストログの各行に予めシステムの判断による「解析優先度」を付与し，この解析優先度を可視化に反映させることで解析の迅速化を図る．ここで解析優先度を「文章に機密情報が含まれる確率」と定義し，機密情報を含んでいる確率が大きいログデータは解析者に対して視覚的に目立つように明るく大きく提示し，逆に機密情報を含んでいる確率の小さいログデータは目立たないように暗く小さく提示する．このように可視化することで解析者は本来解析する必要のないログデータと解析すべきログデータを直観的に把握できるようになり，解析作業の手間を省き解析時間を短縮できる．さらに，多くの既存ツールでは機密文書に対する編集や複製が行われた新規文書も同様に機密文書として扱っている．つまり編集・複製された情報がどの程度の機密性を持つかどうかとは無関係に新たな機密文書とみなされおり,結果として監視対象となる文書数は徐々に増大してしまう.対して本手法では，ある文書がどの程度の機密性を持つかという情報を数値化し，可視化に反映させることで膨大なログデータの中から機密情報を含んでいる確率の高いものを優先的に解析することができる．

4. 解析優先度付与アルゴリズムおよびログ可視化手法

本章ではログの解析優先度の付与に関するアルゴリズムについて詳述する．まず，以降の文中で用いる用語を以下の通り定義する．

- **機密箇所最小ユニット CMU**（Confidential Minimum Unit）：文書内の機密情報で，それ以上細分化できない最小範囲の情報
- **機密箇所最小ユニット含有数期待値 ECMU**（Expected number including CMU）：文書内に含まれる CMU 数の期待値
- **機密箇所含有確率 PIC**（Probability Including Confidential unit）：文書内に CMU が含まれている確率

4.1 アルゴリズム概要

本提案では，事前準備として管理者が予め各機密文書に ECMU を付与しておくことを想定している．各クライアント内で文書に対する編集や複製が行われるたびに，対象となった文書の ECMU を新たに更新していく．機密情報の編集・複製が行われる可能性のあるクライアント内のファイル操作としてはコピー・アンド・ペーストとカット･アンド･ペーストを想定しており,各場合における ECMU の算出方式は 4.2 節に詳述する．各文書の PIC は当該文書に付与されている ECMU の値から算出する．CMU を含む数が 0 個のものは一般文書，1 個以上のものは機密情報と考えることができるので，ECMU の値が 1 以上の場合は PIC=1 とし，ECMU の値が 1 未満の場合は PIC = ECMU とすることができる（図 1)．最終的に PIC をシステムが判断した解析優先度と考え，可視化に反映させる．

4.2 ECMU の算出

ECMU の値が $ECMU_{source}$ の文書（複製元文書）から $ECMU_{destination}$ の文書（複製先文書）に複製を行った場合を考える（図 2)．ここで複製元文書における，複製の対象となった箇所の情報量と，複製元文書全体の情報量の比が $m:n$ となった時，複製対象

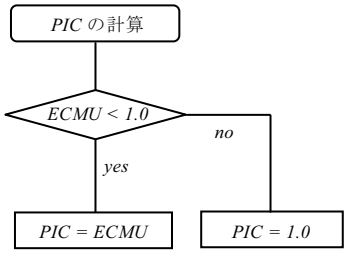

図 1　*PIC* の計算

となった情報に含有される CMU 数の期待値 ΔECMU は，

$$\Delta ECMU = \frac{m}{n} ECMU_{source} \quad (1)$$

となる．ゆえに，複製操作後の複製先文書の ECMU の値 $newECMU_{destination}$ は，

$$newECMU_{destination} = ECMU_{destination} + \Delta ECMU \quad (2)$$

である．複製元文書の ECMU に関しては複製操作としてコピー・アンド・ペーストが行われた場合には変更はないが，カット・アンド・ペーストが行われた場合にはカットされた箇所が複製元文書からは削除されるため，

$$newECMU_{source} = ECMU_{source} - \Delta ECMU \quad (3)$$

が与えられる．

ところで上記のような ECMU 算出方式では，同じ情報の複製を行ったとしても ΔECMU の値はその都度異なってしまい，同一内容の文書であるにも関わらず異なる ECMU の値が付与されてしまう可能性がある．例えば，図 3 のように ECMU が 2.4 の文書（文書 A）から ECMU が 1.6 の編集途中の文書（文書 B）にコピー・アンド・ペーストで複製を行ったとする．このとき複製箇所（以下，複製箇所 A）が文書 A 全体の半分を占めていたとすると，ECMU が 1.6 であった文書 B の ECMU は新たに 2.8（1.6 + 2.4 / 2）に更新される．そして文書 B 内における複製箇所 A をさらに新規に作成した文書 C（文書 C の初期 ECMU = 0.0）に複製したとすると，新規の文書 C には新たに ECMU = 1.4（0.0 + 2.8 / 2）が付与されることとなる．一方で，文書 A から複製箇所 A を別の新規文書 D（文書 D の初期 ECMU = 0.0）に複製したとすると，この新規文書 D の ECMU は新たに 1.2（0.0 + 2.4 / 2）という値に更新される．

ここで文書 C と文書 D は完全に同内容の文書であるにも関わらず，付与されている ECMU の値が異なっている．このように同内容の文書の ECMU の値に整合性がとれていない問題は，複製箇所 A が複数回複製の対象となっているが，機密箇所 A に含まれる CMU 数の期待値である ΔECMU の値が一致していないことに起因している．

そこで過去の複製データとそのΔECMU を保持しておき，複製操作実行のたびに照合を行ことでΔECMU の整合性を確保する．図 4 は整合性確保アルゴリズムの詳細である。まず，ある文書中における複製の対象となった領域のデータ（複製データと名付ける）をデータベースに保存しておく（図 5）．また，このデータベース中にある個々の複製データを登録データと呼ぶことにする．例えば，過去にある文書中の ”The idea is confidential.” という一文を複製し，そのΔECMU が 0.4 だった場合，登録データは表 1 の 1 行目のようになる．そして複製が行われるたびに，その複製データを完全包含する登録データがデータベースの中にあるかをチェックする．存在した場合には，その登録データの ΔECMU と (複製データのサイズ) / (登録データのサイズ) の積を求めることで複製範囲に含まれる CMU 数の期待値を求めることができる．そして，この値と式 (1)で算出したΔECMU の値のうち小さい方の値をΔECMU の値として採用する．小さい方の値を採用することによって，機密文書に付与される PIC の誤差が大きくなってしまう反面，一般文書に対しては誤差を少なく抑えることができる．一方で大きい方の値を採用した場合には，先とは逆に機密文書に対する誤差は少なくなるが，一般文書に対する誤差が大きくなってしまう．一般に，全文書に占める機密文書の割合に比べて非機密文書の占める割合が大きいと推察される．そこでログ全体の ECMU，PIC 誤差を少なくできるという理由から小さい値を採用することとした．なお，複数の登録データと一致すると判定された場合には，より過去に登録されたデータの方が ΔECMU に含まれる誤差が小さいと考えられることから，最も過去の登録データを一致する登録データとみなす．そして，完全一致のデータを完全包含する登録データがないと判定された場合には，一致率がスレッショルドを超える登録データが存在するかのチェックを行う．例えば複製データを意味する文字列 “This is confidential idea.” と，登録データを意味する文字列 “The idea is confidential.” との一致率は，複製データの 4 単語のうち登録データにも登場する単語が 3 つあることから 3 / 4 = 75%と算出する．今回はこの一致率の数値が 90%を超えるものを一致とみなす．複製データと 90%以上一致する登録データが複数あると判定された場合には，一致率が

最も高い登録データの ΔECMU と (一致する部分のデータサイズ) / (登録データサイズ) の積と, 式 (1) から算出される ΔECMU のうち小さい方の値を ΔECMU とする.

最後に，一致率がスレッショルドを超える登録データも存在しないと判定された場合には，式 (1) で算出される ΔECMU を採用する.

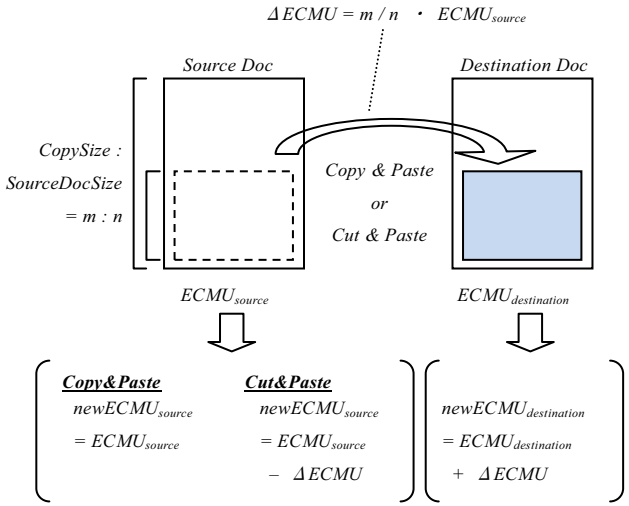

図 2　*ECMU* の計算

4.1　ログの提示手法

ログは従来からログ解析で用いられている表形式で提示した．その表形式のログに PIC を反映させることで，解析すべきデータとそうでないものを視覚的に表現する．PIC はセルの色と縦幅，フォントサイズで解析アプリケーションの UI 上に提示することとした．PIC と UI 上での提示方式の対応の詳細は図 6 の通りである．PIC の値が小さいものほど暗く小さな提示を行い，大きいものほど明るく大きな提示を行う．図 7 は従来型のテキスト形式でのログ提示手法の一例である．対して，提案システムを適用した図 8 では，PIC の値を視覚的に反映させることで解析者はどのログから優先的に解析していけば良いのかについて直感的に把握できる．

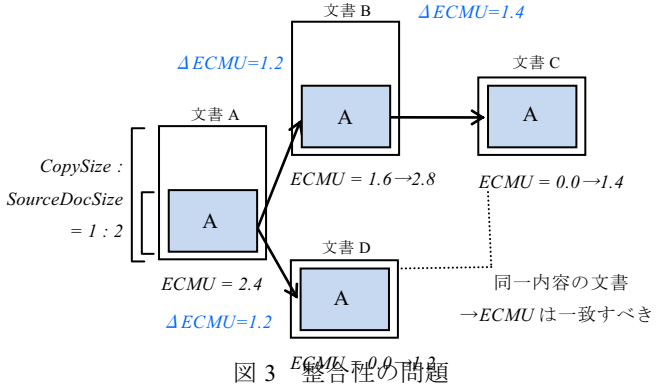

図 3　整合性の問題

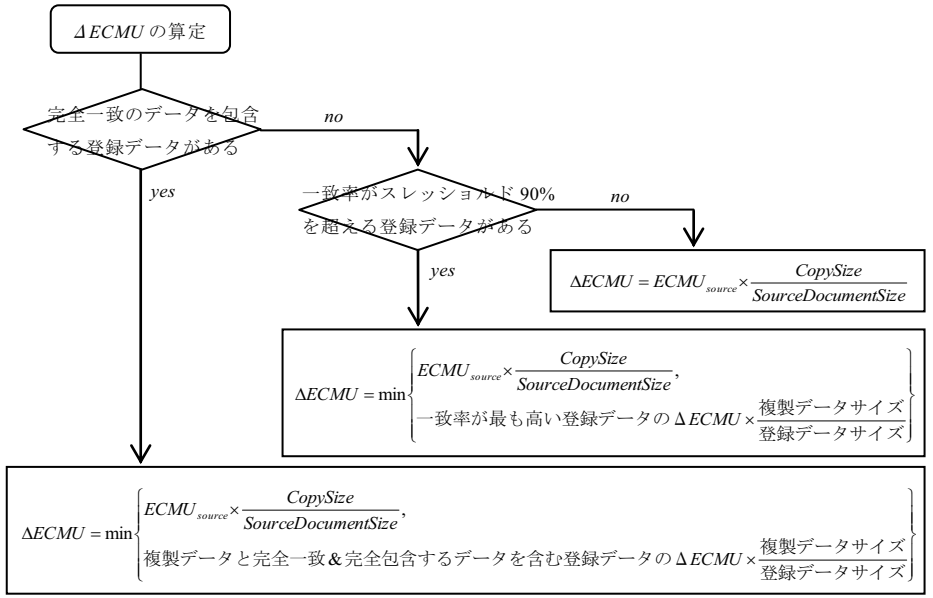

図 4　整合性確保アルゴリズム

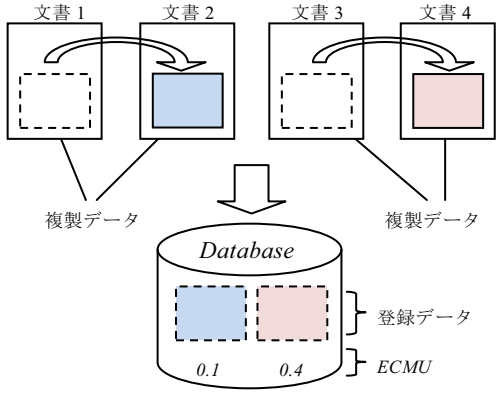

図 5　複製データと登録データ

表 1　複製データとその ΔECMU の値

複製データ	ΔECMU
The idea is confidential	0.4
Our propose is…	0.1
:	:

PIC	セルの色	セル幅	フォントサイズ
0.0≦PIC<0.2	RGB(100,100,100)	12	8
0.2≦PIC<0.4	RGB(150,150,150)	14	8
0.4≦PIC<0.6	RGB(175,175,175)	16	9
0.6≦PIC<0.8	RGB(200,200,200)	18	10
0.8≦PIC<1.0	RGB(225,225,225)	20	11
PIC=1.0	RGB(255,255,255)	20	12

図 6　PIC の値とログへの視覚的反映方法

4.2　CMU，ECMU，PIC についての補足

上記で CMU とは文書内の機密情報であり，それ以上細分化できない最小範囲の情報だと定義しているが，そもそもそういったものは本当に定義できるのだろうかという疑問がある．しかし CMU は ECMU を厳密に定義できると仮定した場合に，その導出のために「便宜的に」仮定した値である．従って CMU の目的は、ECMU および PIC の算出方式をアルゴリズム化することであり，文書の機密度という曖昧な概念を厳密化するためのものである．では，ECMU や PIC は明確に定義できるのだろうか．本提案では，事前準備として管理者が各機密文書に ECMU を付与しておくことを想定している．例えば 1) はサンプルのドキュメントで学習・定義した特徴を認識し，機密データと非機密データの微妙な相違を識別できる．この他にも機密データを判別する技術は存在し，これらの技術を応用することで将来的には各文書に ECMU，またはそれに近い値を設定できるもの考える．機密情報を最小範囲に分割した場合の機密箇所の数（ECMU）を厳密にカウントし，初期設定を行うということは現実問題として難しい可能性があるが，例えば上記のような技術を用いて各機密文書の機密度を 0〜5 の 6 段階程度（0 : 一般文書，1〜5 : 機密文書，値が大きいものほど機密度が高い）で定義することは可能である．このように，機密情報に対して 6 段階程度に分けられたパラメータのいずれかを設定するのであれば十分に現実的であると考える. 本提案は，このように何らかの形で各文書に ECMU・PIC を付与できたと仮定した場合に，ユーザによるファイル操作によって PIC が伝搬し，それによって解析すべきデータとそうでないデータを視覚的に提示することを主眼に置いている．

5.　評価

ログにシステム判断 PIC を付与することによる解析支援効果の測定を行うためのユーザ実験を行った．本章ではその評価実験の内容と結果について記す.

5.1　実験方法

実験の手順としては，まず被験者に対して PIC 提示なしのログ，開発アルゴリズムを用いて算出された PIC を提示したログのいずれかを提示する．そして被験者は提示されたログの解析を行い，タスク 8 題に回答する．タスクは，あるソース文書内の特定の情報が指定された時点で指定された文書内に存在しているかを問う正誤問題である. 例えば,「2010/10/2 15:34 サーバ A からダウンロードした C:¥hoge.doc 内の文字列情報“This apple is blue”は，2010/10/11 9:00 時点で C:¥foo.txt 内に複製されて存在している.」という問いの正誤を回答する．被験者は 2 通りのログ提示に対して同様のタスクを行い，その際の正答率と解析所要時間を測定する．なお，被験者がログ提示に十分に慣れた状態で実験を行えるよう，それぞれのログ提示に対して実験前に予め

Time	Operation	Source Document
4:14:00	Open	C:¥Users¥client¥Documents¥Projec
4:19:00	Close	C:¥Users¥client¥Documents¥Projec
4:56:00	Open	C:¥Users¥client¥Documents¥Projec
2010/10/29 5:12	send	client8
5:18:00	Close	C:¥Users¥client¥Documents¥Projec
6:01:00	Close	C:¥Users¥client¥Documents¥Projec
6:16:00	Open	C:¥Users¥client¥Documents¥Projec
6:52:00	Open	C:¥Users¥client¥Documents¥Projec
6:55:00	Copy	C:¥Users¥client¥Documents¥Projec

図 7　PIC を反映しない場合のログデータ

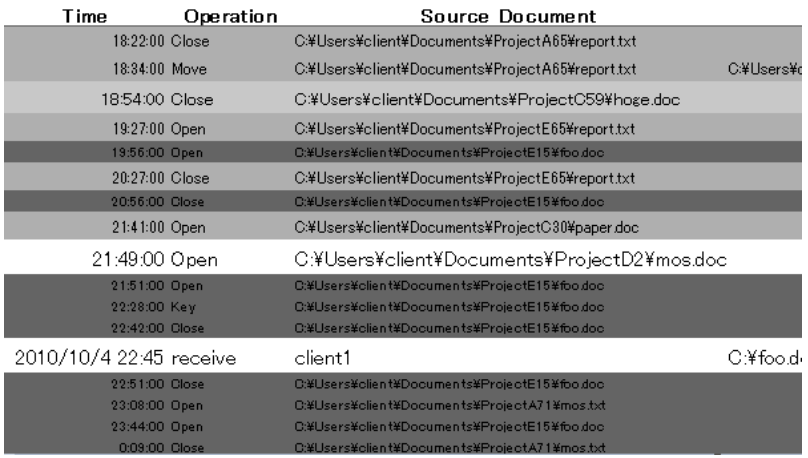

図 8　PIC を反映した場合のログデータ

タスク 4 題を練習問題として回答してもらった．また，本実験で用いたログは独自に構築したシミュレータによって生成されたクライアント監視ログであり，被験者は情報工学を専攻する大学生/大学院生 12 名である．

5.2 実験用ログの生成

本実験では Java によるシミュレータを用いて生成したクライアント監視ログを使用した．想定したシステムの構成は図 9 の通りである．Confidential Server には管理者が機密と定める文書が保存されているものとし， Share Server は文書を共有する目的で設置されたサーバである．各クライアント内での文書の編集・複製，交換などの通常業務（ユーザ操作）はパラメータ化し発生させた．シミュレーション期間は 3 カ月（1 カ月=4 週間，1 週間= 5 日，1 日= 5 時間で定義），クライアント数は 15 台，取り扱い文書は英文テキスト文書に限るものとした．また，各クライアントおよび各サーバの初期保有文書数は，各クライアントが一般文書 10，機密文書 0，Share Server が一般文書 0，機密文書 0，Confidential Server が一般文書 0，機密文書 10 である．なお，各クライアント内の初期保有文書である一般文書は空の文書であり，Confidential Server 内の機密文書は常用単語 5,000 語の中からランダムに選択し，長さ 3,000 文字になるよう羅列したもので，機密箇所 CMU はランタム箇所・ランタム長で 1 箇所定義した（ECMU の初期値=1）

先の条件下でシミュレーションを行った結果,全体での機密文書数は 357，一般文書数は 2768，全 15 台のクライアントの操作履歴である監視ログのサイズは平均 1022 行であり，実験ではこの中から 1004 行のログを使用した．

5.3 結果と考察

ログ解析実験の結果を表 2 に示す．PIC 提示なしの場合と比べ，開発アルゴリズムで算出した PIC を提示した場合には正答率には有意差がみられなかったものの，解析所要時間において有意差が確認できた（有意水準 1%）．従って，PIC を解析 UI 上に視覚的に提示することでログ解析を迅速化できるといえる．

ここで PIC 付与によって解析を迅速化できた要因を考察する．PIC を提示していない従来のログ解析であれば，被験者は時系列順に 1 行ずつログを解析し，文書操作や移動などの事象を把握して行く解析の流れとなる．しかし，システムの判断によってログの各行に解析優先度を付加されることによって，必ずしもすべての行を解析する必要がなく，優先度の高い行のログデータをまず解析し，それでも判断しかねる事象の場合のみ優先度の低いログデータの解析を行えばよいということになる．実際に実験の中でも，PIC 提示なしのログの場合には 1 行ずつ解析を進める被験者が多かった

のに対し，PIC 提示ありの場合には優先度の高いログのみ解析し，その後，判断しかねた箇所の解析や解析後の確認のために解析箇所を前の行へ戻すという作業を行う被験者がほとんどであった．

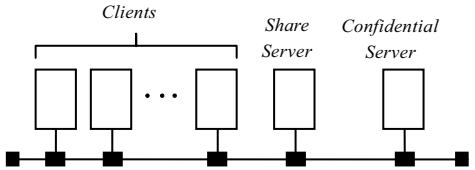

図 9　システム構成

表 2　実験結果

	正答率(%)	解析時間 (sec)
PIC 提示なしの場合のログ解析	91(11)	256(36)
PIC 提示ありの場合のログ解析	92(10)	177(47)

6. まとめ

近年の情報漏洩の深刻化にともない，ユーザ操作を監視することで漏洩対策を行うケースが多く存在する．実際にユーザを監視し，情報の所在をログデータ（クライアント監視ログ）として取得するツールがさまざまなベンダーからリリースされており，多くの組織に導入されている．これらのツールによって得られたログデータを解析するスピードは，漏洩の事前防止や事後対応の両面において非常に重要である．そこで，クライアント監視ログを解析する速度を向上することにより漏洩対策に貢献することが可能となる点に着眼した．本論文ではログの各行にシステムが判断した解析優先度 PIC を付与し，それを解析アプリケーションの UI 上に提示することでログの解析作業を支援する手法を提案した．具体的には解析優先度を各文書に機密情報が含有される確率 PIC と定義し，PIC の算出アルゴリズムの開発を行った．最後に独自に開発したシミュレータによって生成されたクライアント監視ログを用いた評価実験を行い，従来の表形式でのログに PIC を視覚的に提示することで解析をより迅速化することが可能であることを示した．

7. 参考文献

1)　シマンテック株式会社，Symantec Data Loss Preventation，
http://www.symantec.com/ja/jp/business/products/family.jsp?familyid=data-loss-prevention
2)　マカフィー株式会社，McAfee Host Data Loss Prevention，
http://www.mcafee.com/japan/products/data_loss_prevention.asp
3)　株式会社インターコム，情報漏洩＋資産管理ツール MaLion 3，
http://www.intercom.co.jp/malion/
4) Sky 株式会社，SKYSEA Client View，http://www.skyseaclientview.net/
5) Sky 株式会社，ファイル監視装置およびファイル監視プログラム，公開特許公報（A），特開 2009-15659 号（2009）．
6) NEC 株式会社，InfoCage，http://www.nec.co.jp/cced/infocage/index.html
7) FineArt Technology Co., Ltd，情報漏洩防止システム TotalSecurityFort，
http://www.fineart-tech.com/jp/
8) e-System, corporation，e-Tracker5，http://www.e-system.co.jp/etr5/index.html
9) MOTEX Inc，http://www.motex.co.jp/index.html
10)　Hideki Koike, Kazuhiro Ohno. 2004. SnortView: Visualization system of snort logs, workshop on visualization and data mining for computer security (VizSEC/DMSEC-04), 11th ACM Conf. on Computer and Communications Security (CCS 2005), ACM, pp.143-147, 2004.
11)　Deborah Estrin, Mark Handley, John Heidemann, Steven McCanne, Ya Xu, Haobo Yu. 2000. "Network Visualization with Nam, the VINT Network Animator," Computer, vol. 33, no. 11, pp. 63-68, Nov. 2000, doi:10.1109/2.881696