2B1-R-3-7

相関ルールに基づく近傍サーバログ分析

Association Rule-based Analysis of Neighborhood Server Logs

北川　哲平[*1]　福井　健一[*2]　鈴木　聖人[*3]　冨士井　裕之[*3]　山口　慶大[*3]
Teppei Kitagawa　Ken-ichi Fukui　Masato Suzuki　Hiroyuki Fujii　Keita Yamaguchi
沼尾　正行[*2]
Masayuki Numao

[*1]大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
[*2]大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University
[*3]NTT 西日本 技術革新部 研究開発センタ
Research and Development Center Network Service Group, Nippon Telegraph and Telephone West Corporation

In our information society, it is a critical issue to assure the safety of the Internet by monitoring network health and troubleshooting network faults. When troubleshooting a network fault, operators manually analyze server logs that describe events observed by it. To support troubleshooting network faults, we proposed an assosiation analysis method considering neighborhood servers, importance of logs and non-periodicity of logs. We applied our proposed technique to server logs collected from a network verification experiment. In the result, we succeeded in detecting logs that cause network faults.

1. はじめに

現在の情報化社会において, ネットワーク監視によりインターネットの安全性・安定性を確保することは重要な課題である. 現在, 多くの通信事業者では SNMP (Simple Network Management Protocol) を用いたネットワーク自動監視が行われており, 異常が起きているサーバを検出することがある程度可能になっている. また, 異常検出の課題である精度の向上を目指し, 統計的な手法を用いた研究 [Yamanishi 05, Hirose 09] が盛んに行われている. 一方, サーバ異常の原因診断はサーバが出力する自らの状態の記録であるサーバログをオペレータが手作業により解析することで行われている. しかしながら, サーバログは膨大であるため人手による解析は困難である. さらにインターネットの普及に伴いネットワークの複雑化・大規模化が進み, オペレータによる原因診断はますます困難になっている. そのため, 効率的にサーバ異常の原因診断を行える手法の開発が求められている.

本研究ではネットワーク監視により異常が検出されたサーバ (以降, 異常サーバと呼ぶ) とその近傍サーバのログを統合して相関分析を行う. これまで, Qui ら [Qui 10] は大規模ネットワークの監視システムの開発を目的として, サーバログの相関分析を行った. この研究では各サーバごとに相関分析を行っていたため, 異常原因が異常サーバの近傍サーバであった場合には相関ルールを抽出できない. また, Turner ら [Turner 07] はネットワークの全サーバログを統合して, サーバログの相関分析を行った. しかしながら, 時間周期など利用して枝刈り行っても尚, 異常と関係がない相関ルールが多数生成され, 適切なルールが埋もれてしまう問題があった. これは離れたサーバで出力される原因とは関係ないサーバログの擬似相関によるものである. そこで本研究では, 相関分析を行う際に対象の異常サーバとの距離を考慮し, 異常サーバと離れたサーバのログほど重みを下げることで対処する. さらに, サーバログの非周期性や意味的な重大性も考慮して重み付けすることで異常原因の候補サーバログを抽出する手法を提案する.

連絡先: 福井 健一, 大阪大学産業科学研究所,
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1,
Tel:06-6879-8426, Fax:06-6879-8428,
E-mail:fukui@ai.sanken.osaka-u.ac.jp

2. サーバログ

本研究で扱ったサーバログの例を表 1 に示す. サーバログには日時, 出力サーバ名, 警報レベル, アラーム内容の 4 つの情報が含まれている.

今後, 時刻 t にサーバ $s_i(i = 1, 2, \ldots, m)$ でアラーム内容 $a_j(j = 1, 2, \ldots, n)$ のサーバログが出力されたとき, このサーバログを $L(s_i, a_j, t)$ と表記する.

表 1: サーバログ例

日時	サーバ	警報レベル	アラーム内容
1/1 0:00:10	s_1	通常	config が変更されました.
1/1 0:00:11	s_2	注意域	認証サーバが応答しません.
1/1 0:00:11	s_2	警戒	待機系システム：Down
1/1 0:00:12	s_3	重要警戒	インタフェース状態：Down
1/1 0:00:12	s_3	危険	インタフェース情報取得失敗
1/1 0:10:11	s_1	クリア	リンクアップ

警報レベルはアラーム内容の重大性を表す指標であり, 重大性が高い順に “危険”, “重要警戒”, “警戒”, “注意域”, “通常 ” の 5 段階で表される. その他にはサーバが異常から復旧した際には警報レベルには “クリア ” と表される. 次にアラーム内容はサーバの状態に関する詳細情報を記述するフリーフォーマットのテキストである. アラーム内容は数千種類あり, ネッ

トワーク運用に与える影響の大きさによって各アラーム内容にあらかじめ警報レベルが割り当てられている.

3. 提案法：サーバ異常の相関分析法

サーバ異常の原因診断を行う際に, オペレータが最も関心を払うサーバログを異常ログ $L_E(s_e, a_e)$ と定義する. 本研究では, $L_E(s_e, a_e)$ と相関の高いサーバログを異常原因の候補サーバログとして抽出する. オペレータは異常が検出された時刻の近傍で発生している $L_E(s_e, a_e, t)$ の前後のサーバログを重点的に解析する [Qui 10]. 特に時刻 t の近傍かつ, サーバ s_e と近いサーバのログに着目している.

1.[サーバログの共起]
サーバログ系列を一定時間間隔 K ごとに区切り, $L_E(s_e, a_e)$ と $L(s_i, a_j)$ が同じ区間内に含まれていれば時間的に近いと判断する (図 1). $L_E(s_e, a_e)$ と $L(s_i, a_j)$ が共起している区間の割合 (確信度) を次式で表す.

$$Co(L_E(s_e, a_e) \Rightarrow L(s_i, a_j)) = \frac{count(L_E(s_e, a_e) \cap L(s_i, a_j))}{count(L_E(s_e, a_e))} \quad (1)$$

ここで, $count(L_E(s_e, a_e))$ は $L_E(s_e, a_e)$ を含む区間の数である.

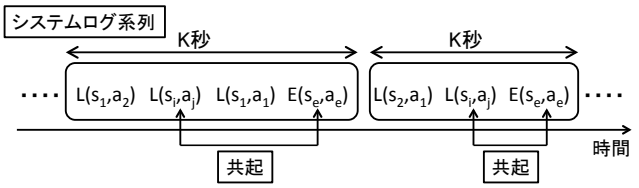


図 1: サーバログの共起概略図

2.[近傍サーバの考慮]
サーバ間の距離をホップ数 (サーバ間通信で経由する中継サーバ数) で定義する. s_e と s_i のホップ数 $hop(s_e, s_i)$ は最短経路探索アルゴリズムのダイクストラ法で算出する. s_e の s_i に対する影響は距離が離れるほど指数関数的に減衰すると仮定し, 次式で重み付けを行う.

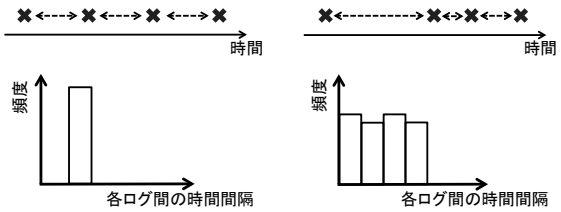
$$T_{s_e}(s_i) = \begin{cases} \exp\{-\alpha[hop(s_e, s_i) + 1]\} & (s_i \neq s_e \text{のとき}) \\ 1 & (s_i = s_e \text{のとき}) \end{cases} \quad (2)$$

3.[サーバログの非周期性]
周期的に出力されるサーバログは実際には関係のないサーバログとも共起しやすい. そこで周期的なサーバログの重みを下げる必要がある.

まず $L(s_i, a_j)$ の時間間隔を算出し, 時間間隔 C ごとのヒストグラムを作成する. 本研究では情報エントロピーを用いて周期的なサーバログの重みを下げる. サーバログに周期性があれば図 2(a) のようなヒストグラムになり, エントロピーは小さくなる. 一方, サーバログに周期性がなければ図 2(b) のようなヒストグラムになり, エントロピーは大きくなる. $L(s_i, s_j)$ の非周期性を次式の $NP(L(s_i, a_j))$ で定義し重み付けを行う.

$$NP(L(s_i, a_j)) = -\sum_t \frac{m_t}{n_{L(s_i, a_j)} - 1} \log \frac{m_t}{n_{L(s_i, a_j)} - 1} \quad (3)$$

ここで, m_t はサーバログ間の時間間隔が t のときの頻度である. $n_L(s_i, a_j)$ はサーバログ $L(s_i, a_j)$ の総出力回数である. 時間間隔の数は総出力回数より 1 少ない.



(a) 周期的なサーバログ (b) 非周期的なサーバログ

図 2: サーバログの時間間隔に関するヒストグラム例

4.[サーバログの重大性]
サーバログの重大性 $A(a_j) \in [0, 1]$ は予め指定された警報レベルにより, 5 段階で評価する. 重大性が高いほど注目すべきサーバログであるので, 重大性が高いサーバログは重みを上げる. ただし, 警報レベルがクリアの場合は $A(a_j) = 0.2$ とした.

以上から, サーバ異常の相関分析法として, 異常ログ $L_E(s_e, a_e)$ に対する各サーバログ $L(s_i, a_j)$ の相関の高さを次式の重要度 $Priority_E(L(s_i, a_j))$ で定義する.

[重要度の定義]

$$Priority_E(L(s_i, a_j)) = Co(L_E(s_e, a_e) \Rightarrow L(s_i, a_j)) \times T_{s_e}(s_i) \times NP(L(s_i, a_j)) \times A(a_j) \quad (4)$$

4. サーバ異常の相関分析結果および評価

4.1 ネットワーク検証実験の概要

本研究でサーバ異常の相関分析を行ったサーバログデータは, 総実験時間数 727 時間のネットワーク検証実験において得られたデータである. この検証実験は人為的にネットワーク異常を起こしサーバログの挙動を確認するために行われた. 検証実験に用いられたネットワークは通信気業者の商用ネットワークを模擬して構築されている (図 3). 検証実験ではサーバに対して, 再起動, コンフィグの修正, LAN ケーブル抜き差し, サーバ周辺機器の電源オフなどの操作を行った.

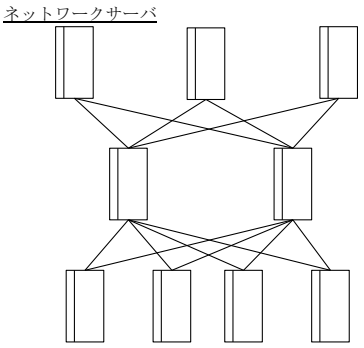


図 3: ネットワーク検証実験の模式図

検証実験の主要サーバ 17 台から得られたサーバログに対して, 前処理として, ネットワーク運用と明らかに関係がないサーバログを削除し, 242,895 個のサーバログを得た. このサーバログ系列に含まれるアラーム内容は 71 種類, アラーム内容とサーバの組み合わせは 324 種類であった.

4.2 評価実験の概要

警報レベルの高い 3 種類のサーバログ (表 2) を異常ログとし, 検証実験から得られたサーバログ系列に対してサーバ異常

の相関分析法を適用した．評価実験はオペレータ経験年数が 2 年以上, かつ同程度のネットワークに関する知識をもつオペレータ 3 名で実施した．

表 2: 評価実験で対象とする異常ログ

ID	出力時間	異常ログ L_E	警報レベル
E1	613:34:40	$L_{E1}(s_5$, RP(1) 隣接関係の状態変化)	重要警戒
E2	134:54:20	$L_{E2}(s_6$, RP(1) インタフェースダウン)	重要警戒
E3	661:54:56	$L_{E3}(s_7$, インタフェース状態：Down)	重要警戒

[**評価実験 1**] 通常の原因診断の状況を模擬的に再現し, オペレータに異常ログと相関のあるサーバログを抽出してもらう. このオペレータ判断による相関があるサーバログを, 提案法による候補サーバログがどの程度網羅できているのかを評価する．

各オペレータに異常ログ $L_E(s_e, a_e, t)$ と相関があるサーバログを全サーバログデータから手作業で全て抽出してもらい, 相関があると判断した理由を述べてもらった．このとき, オペレータにはあらかじめネットワーク構造の情報を与えた．

[**評価実験 2**] 提案法により抽出された候補サーバログが異常ログと相関があるかをオペレータに判断してもらう．提案法による候補サーバログの妥当性と共に, 実験 1 による通常の原因診断では発見できなかったが, 妥当だと考えられるログの評価を行う．

異常ログ $L_E(s_e, a_e)$ に対する各ログ $L(s_i, a_j)$ の重要度を算出し, 候補サーバログを抽出した．このとき, 各パラメータを式 (1) では $K = 10$(秒), 式 (2) では $\alpha = 2$, 式 (3) では $C = 60$(秒), とした．

次に，オペレータには式 (1) の $count(L_E(s_e, a_e) \cap L(s_i, a_j)) \geq 3$ を満たす全サーバログが異常ログと相関があるかを判断してもらい, 相関があると判断した場合には, その理由を述べてもらった．相関があるかの判断を行う際に参考にしたサーバログを回答してもらう．また, オペレータにはあらかじめネットワーク構造の情報を与えた．

なお, 評価実験は実験 1, 実験 2 の順でを行い, 実験 2 の候補サーバログが実験 1 の手作業のログ抽出に影響を与えないようにした．

4.3 実験結果

4.3.1 提案法の精度・被覆率・発見率による評価

提案法の有用性を定量的に評価するために, 以下の精度・被覆率・発見率の算出を行った．

1. 精度

精度は提案法により検出されたデータの正確性を表す指標である．本研究では提案法による候補サーバログ S 個に対する精度 (*precision*) を次式で定義する．

$$precision = \frac{y_2}{S} \tag{5}$$

ここで, y_2 は評価実験 2 でオペレータ x 名以上が異常ログと相関があると判断したサーバログの個数である．

2. 被覆率

次に, 被覆率は提案法により検出されたデータがどれだけ正解データをカバーしているかという網羅性の指標である．本研究では提案法による候補サーバログ S 個に対する被覆率 (*coverage*) を次式で定義する．

$$coverage = \frac{R}{y_1} \tag{6}$$

ここで, y_1 は評価実験 1 でオペレータ x 名以上が異常ログと相関があると判断したサーバログの個数, R は y_1 に含まれ, かつ提案法による候補サーバログに含まれるサーバログの個数である．

3. 発見率

発見率は提案法による候補サーバログから相関があるサーバログを新たに発見できる能力を表す指標である．発見率 (*discovery*) を次式で定義する．

$$discovery = \frac{y_{discovery}}{S - R} \tag{7}$$

ここで，$y_{discovery}$ は評価実験 1 でオペレータ x 名以上から抽出されず，かつ評価実験 2 でオペレータ x 名以上が相関があると判断した候補サーバログの個数である．

各異常ログ L_{E1}, L_{E2}, L_{E3} に対して, 精度・被覆率・発見率を算出した．また, オペレータが相関があると判断した人数を 1 名以上と 2 名以上の場合でそれぞれ算出を行った．算出結果を表 3, 表 4 に示す．

表 3: 提案法の精度・被覆率・発見率の評価 ($x = 2$ 人の場合). 括弧内はサーバログの個数を表す．

異常ログ	精度	被覆率	発見率
L_{E1}	40%(4/10)	0%(0/2)	40%(4/10)
L_{E2}	50%(5/10)	80%(4/5)	17%(1/6)
L_{E3}	60%(6/10)	50%(1/2)	56%(5/9)

表 4: 提案法の精度・被覆率・発見率の評価 ($x = 1$ 人の場合). 括弧内はサーバログの個数を表す．

異常ログ	精度	被覆率	発見率
L_{E1}	80%(8/10)	0%(0/12)	80%(8/10)
L_{E2}	90%(9/10)	28%(9/32)	0%(0/1)
L_{E3}	70%(7/10)	17%(1/6)	67%(6/9)

提案法ではオペレータ 2 名以上が相関があると判断した場合の精度は全ての異常ログに対して 40%を超えている．また, オペレータ 1 名以上が相関があると判断した場合の精度は全ての異常ログに対して 70%以上である．つまり, 提案法による候補サーバログは大半が L_E と相関がある可能性が高いサーバログを取り出すことに成功していると言える．

次に, 表 3 の L_{E2}, L_{E3} の被覆率に着目する．L_{E_2}, L_{E3} の被覆率はともに 50%以上であり, 提案法による候補サーバログにより実験 1 でオペレータ 2 名以上が相関があると判断したサーバログを半数以上を網羅できた．また, L_{E1} の被覆率については, 提案法による候補サーバログには実験 1 でオペレータが相関があると判断したサーバログと類似の意味をもつサーバログが全て含まれていたが、今回はログの種類の一致で判断しているため、被覆率が 0%になってしまっている。

発見率に関しても L_{E1}, L_{E3} で全て 40%を超えており, 提案法によりオペレータに新たな知見を与えることに成功していると言える．一方, L_{E2} の発見率は 0%である．これは提案法による候補サーバログ 10 個のうち 9 個が通常の方法で発見できているためである。

4.3.2 提案法により抽出された異常原因

異常ログ $L_{E2}(s_6,$ PR(1) インターフェースダウン) に対する評価実験 2 の結果を表 5 に示す[*1].

表 5: 提案法により抽出された L_{E2} と相関が高いログ. 人数欄は L_{E2} と相関があると判断したオペレータ数 (3 人中) を表す.

順位	提案法による候補サーバログ	人数
1	$L(s_6,$ RP(1) 隣接関係の状態変化)	**3**
2	$L(s_6,$ IP(2) インタフェースがダウンしました.)	1
3	$L(s_6,$ インタフェースダウン (リンクダウン) の発生)	**2**
4	$L(s_6,$ IP(2) インタフェースがアップしました.)	0
5	$L(s_6,$ RP(1) インタフェースが pointtopoint になりました)	**3**
6	$L(s_6,$ インタフェースダウン (リンクアップ) の発生)	**2**
7	$L(s_6,$ NW(1) アラームセット)	1
8	$L(s_6,$ NW(1) アラームセット解除)	1
9	$L(s_6,$ 周辺機器電源オフ)	**2**
10	$L(s_7,$ RP(2) ピア状態が変化しました. 現在の状態= idle)	0

オペレータ 2 名以上が相関があると判断したサーバログの $L(s_6,$周辺機器電源オフ)(9 位) に着目する. このサーバログはサーバ s_6 の周辺機器の電源がオフになったことを表す. オペレータは s_6 の周辺機器の電源がオフになったことを原因として, s_6 が正常に運用できなくなったために L_{E2} が出力された可能性が高いと推察している. また, 検証実験では人為的な異常としてサーバ周辺機器電源オフを行っており, L_{E2} の異常原因として $L(s_6,$周辺機器電源オフ) が抽出されているのは妥当であると考えられる. $L(s_6,$周辺機器電源オフ) は実験 1 でオペレータ 1 名に抽出されていたが, 他のオペレータ 2 名は抽出できていなかった. さらに, 式 (1) の確信度のみで L_{E2} と相関が高いサーバログを抽出した場合, $L(s_6,$周辺機器電源オフ) は上位 10 個のサーバログには含まれていなかった. これは $L(s_6,$周辺機器電源オフ) の頻度が低く, 警報レベルも低かったためであると推察される.

この結果から, 提案法によりオペレータ複数名が抽出できなかった異常原因の可能性が高いサーバログの抽出に成功したと言える.

4.3.3 提案法と確信度のみの場合の比較

異常ログ $L_{E3}(s_7,$インターフェース状態：$Down)$ に対する評価実験 2 の結果を表 6 に示す. また, 式 (1) の確信度のみを用いて L_{E3} と相関が高いサーバログを抽出し, 評価実験 2 の結果を参考にして抽出されたサーバログが異常ログと相関があるかを評価した (表 7).

表 6: 提案法により抽出された L_{E3} と相関が高いログ

順位	提案法による候補サーバログ	人数
1	$L(s_7, a_{37}$(インタフェース状態：Administratively down))	1
2	$L(s_8, a_{31}$(インタフェース状態：Down))	**2**
3	$L(s_7, a_6$(インタフェース状態：Up))	**3**
4	$L(s_2, a_{50}$(インタフェース情報取得失敗))	**2**
5	$L(s_2, a_{31}$(インタフェース状態：Down))	**2**

表 6 と表 7 を比較すると, 上位 5 個に含まれるサーバログが大きく異なることがわかる. 確信度のみの場合では, オペレータ 2 名以上が相関があると判断したサーバログは上位 5 個中 1 個である. 一方, 提案法による候補サーバログでは, オペレータ 2 名以上が相関があると判断したサーバログは上位 5 個中 4 個であり, 提案法は確信度のみの場合に対して明らかな優位性がある. これは, 提案法では近傍サーバを考慮して重み付けを行っているので, 異常と関係のない離れたサーバのログの重要度を下げているためである.

表 7: 確信度のみにより抽出された L_{E3} と相関が高いログ

順位	確信度のみにより抽出されたサーバログ	人数
1	$L(s_7, a_{37}$(インタフェース状態：Administratively down))	1
2	$L(s_5, a_{50}$(インタフェース情報取得失敗))	0
3	$L(s_2, a_{50}$(インタフェース情報取得失敗))	**2**
4	$L(s_4, a_{31}$(インタフェース状態：Down))	0
5	$L(s_6, a_{50}$(インタフェース情報取得失敗))	0

この結果から, 提案法では近傍サーバと異常サーバの距離を考慮して重み付けを行ったことで, 確信度のみの場合より有益なサーバログを抽出できることが確認できた.

5. まとめ

本研究では, ネットワーク検証実験から得られたサーバログデータ系列について, 相関分析を行う際に近傍サーバと異常サーバの距離・サーバログの非周期性・サーバログの重大性を考慮して重み付けを行う手法を提案し, 異常ログと相関が高いログを抽出した.

評価実験の結果, 提案法による候補サーバログの大半が異常ログと相関があることが確認できた. さらに, 提案法によりオペレータが抽出できなかった異常原因の可能性が高いサーバログの抽出に成功した. また, 確信度のみの場合と比較して, 提案法の優位性が見られた.

今後の課題として, 忘却型のアルゴリズムを導入することで, ネットワークの非定常性に対応できると考えられる. また, 予め同じ単語を含むアラーム内容を類似サーバログとしてにクラスタリングし, 提案法により候補サーバログを抽出すれば, ユーザが異常ログと相関のあるログをより容易に把握できると考えられる. 提案法により抽出された相関ルールをデータベース化し, オペレータがサーバ異常の原因診断を行うための補助システムのへの発展が期待できる.

参考文献

[Hirose 09] Hirose, S., Yamanishi, K., Nakata, T., and Fujimaki, R.: Network Anomaly Detection based on Eigen Equation Compression, in *Proc. the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD09)*, pp. 1185–1193 (2009)

[Qui 10] Qui, T., Ge, Z., Pei, D., Wang, J., and Xu, J. J.: What Happened in my Network? Mining Network Events from Router Syslogs, in *IMC '10 Proc. the 10th Annual Conference on Internet Measurement*, pp. 472–484 (2010)

[Turner 07] Turner, A., Kim, H. S., and Wong, T.: Automatic Discovery of Relationships Across Multiple Network Layers, in *INM '07 Proc. the 2007 SIGCOMM Workshop on Internet Network Management*, pp. 230–235 (2007)

[Yamanishi 05] Yamanishi, K. and Maruyama, Y.: Dynamic syslog mining for network failure monitoring, in *Proc. the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD05)*, pp. 230–235 (2005)

*1 企業秘密保持のため, プロトコル名は伏せている. RP は Routing Protocol, IP は Internet Protocol, NW は Network を表しており, 括弧内の番号は小区分の識別子である.