LL-017

統計的手法を用いた計算機ログからの異常検出
Anomaly Detection from System Log Messages by Using Statistical Analysis

橋爪 拓[†] 宮本 貴朗[‡] 泉 正夫[†] 福永 邦雄[†]
Hiroshi Hashidume Takao Miyamoto Masao Izumi Kunio Fukunaga

1. まえがき

現在，計算機やネットワークは社会にとって必要不可欠なものであり，Web や電子メールなどの主要なサービスの停止や障害は社会的な活動に支障をきたしかねない．そこで，システム管理者は計算機や主要なサービスが正常に稼働しているかを把握するために，計算機ログの定常的な分析作業を行う必要がある．

しかし，計算機ログの分析作業は重要であるにも関わらず，実際に実施されている割合は低いものとなっている．その理由として，利用者・機器の増加に伴いログが膨大な量となっており，管理者のログ分析負荷の増大が挙げられる．計算機ログの分析における従来手法としては見えログ [1] などがある．見えログはテキストマイニングを適用してログ情報の特徴抽出を行い，正常時の特徴と比較するものであり，異常と推測されるログ情報を情報視覚化を用い図化する．しかし，異常事象そのものは管理者が図化されたログ情報から判断しなければならないという問題点がある．

本稿では，ログ分析作業における問題点を解決するため，大量のログから異常なログを自動検出することを目的とする．検出のための特徴として，正常時には出現する頻度が低く，特定の単位時間に大量に出現するログは重要であることや，異常が発生すれば時刻ごとの出力数と連続して出現する行数が変化することに着目し，行の重要度・ログの時刻情報・ログの時間情報を用いる異常検出手法を提案する．

2. 異常検出

あらかじめ登録しておいた異常事象の特徴とログからの入力情報とを比較し，検出を行う不正検出とは異なり，異常検出はシステムやユーザの振舞いを監視し，通常と異なる振舞いを検出する．しかし，異常検出では正規ユーザの利用傾向が大きく変化した場合，不正行為でなくとも異常と判断されるため，正常なものを異常とみなす数（False Positive）が多くなる．また，False Positive を単純に削減すると異常なものを正常と判断する数（False Negative）が増加してしまう．どちらを優先するかは計算機環境や管理ポリシーなどに依存するため，本稿ではパラメータ設定により，異常検出対象を調節可能な検出手法を提案する．

3. 提案手法

異常を含むログは様々な特徴を持つが，本稿では三つの特徴に着目した検出手法を提案する．

[†]大阪府立大学大学院 工学研究科, Graduate School of Engineering, Osaka Prefecture University
[‡]大阪府立大学 学術情報センター, Library and Science Information Center, Osaka Prefecture University

3.1 行の重要度を用いる手法

システムにおけるログ分析においては，特定の単位時間にだけ大量に出現し，他の単位時間にほとんど出現しないログは重要であり，異常事象をあらわす一つの特徴であると考えられる．本手法ではテキストマイニング手法の一つである TF.IDF 法 [2] により，ログに出現する単語の重要度を求める．TF.IDF 法での単語 T_j の重要度 ω_j^i は，単語 T_j の対象日のログ D_i における出現頻度を tf_j^i，N 日分の全ログ中に単語 T_j が含まれる日数を df_j とすると

$$\omega_j^i = tf_j^i \cdot \log(N/df_j) \quad (1)$$

で表される．TF.IDF 法により，出現する単語の重要度を求めることができるが，ログ情報の異常検出においては，対象としているのは単語そのものではなく，異常を示している行である．そこで，本手法では行に含まれる単語の重要度の平均を行の重要度とする．

検出対象のログに含まれる行に対して，行の重要度の高い順にソートし，横軸に行の数，縦軸に重要度をとり，行の重要度の分布として表現する．また，行の重要度を行の重要度の分布曲線・横軸・縦軸に囲まれた面積が 1 となるように正規化する．本手法では，行の重要度の分布により抽出される行数を変化させるために，上位からの面積比に閾値を設け，閾値内に含まれている行は重要度が高く異常を示している行と判断する．図 1 に，実際のログにおける行の重要度の分布の例を示す．図中の縦線は面積比が 0.8 のときの閾値である．

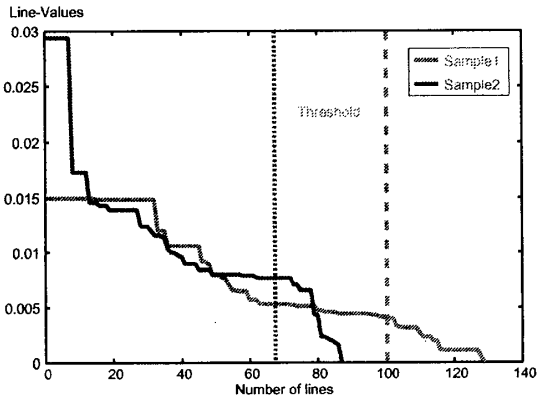

図 1: 行の重要度の分布の例

3.2 ログの時刻情報を用いる手法

ログの出現時刻に着目すると，ユーザのイベントにより記録される種類のログ出力数は平日の昼間多く，休日や夜間には少ないなどの時刻に依存する傾向を持つ．そこで，単位時間あたりのログの出力数を計測し，横軸に単位時間あたりのログの出力数，縦軸に頻度をとり分布として表現する．

時刻ごとの出力頻度分布は，正規分布に近似できると仮定し，予測区間を用いて異常を判別する．予測区間とは母集団のデータが存在することが期待される範囲のことを表し，与えられた分布が平均値 m，標準偏差 σ の正規分布であると仮定すると，$m \pm \beta\sigma$ で表される．ここでは，正規分布と仮定していることから，予測区間に含まれていない出力数が発生する確率は低い．そこで，あらかじめ正常なデータで時刻ごとの予測区間を求め，検出対象日の特定の時刻のログ出力数がその予測区間に含まれていなければ，正常時における発生確率の低い出力数であり，その時刻で異常が発生したと判断する．

3.3 ログの時間情報を用いる手法

ログの出現時間に着目すると，ひとつの異常が発生することにより，複数のログが出現する傾向があり，ログ情報は同一時間のタイムスタンプで出現するか，数秒離れた時間で出現する確率が高い．そこで，ログ出現間隔ごとに特有のログの連続行数に着目し，以下の手順で異常検出を行う手法を提案する．まず，ログの連続性をみるために，正常時のログ出現間隔ごとのログ出現数を計測し，連続行数の生起確率を求める．次に，検出対象となるログに対して，同様の手法を用い出現数を計測し，異常性を判断する．つまり，検出対象のとなるログの連続行数に対応する生起確率が低い場合，正常時に出現する確率の低い異常な連続行と判断する．

4. 実験

当研究室で実際に稼働している計算機から正常状態のサンプルとして 115 日間，検出対象の実験データとして 4 日間の計算機ログ (検証用ログ：data1 ～ data4) を収集した．検証用ログには平均すると約 2 割の異常が含まれていた．本手法を適用し，検証用ログから異常と判断した行を検出し，管理者が異常事象と判断した行と比較し，評価実験を行った．行の重要度を用いた異常検出結果を図 2 に示す．評価手法は横軸に False Positive，縦軸に異常を異常と正しく判断した割合である Hit Rate をとった ROC カーブを用いた．また，False Positive:30% のときの各手法の Hit Rate を表 1 に示す．

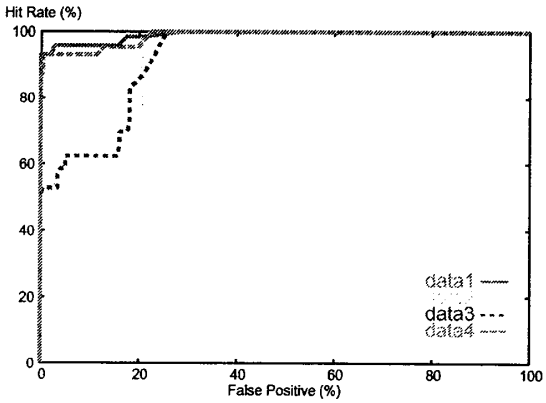

図 2: 行の重要度を用いる手法を適用した異常検出結果

提案した 3 つの手法を適用した異常検出結果では，時間情報を用いる手法は他の手法と比較すると，False Positive に対する Hit Rate の割合が低いものとなった．そこで，他の手法と組み合わせて適用することにより，検出率の向上を試みる．ここでは，手法の組み合わせにより，どの程度検出数が増加するかを検証する．行の重要度を用いる手法及び時刻情報を用いる手法での False Negative を対象として，時間情報を用いる手法の閾値を 5% 及び 10% としたときの検証用ログに対する検出増加数を表 2 に示す．表 2 での数値は，(複合的手法による検出増加数)/(行の重要度を用いる手法における False Negative)，及び (複合的手法による検出増加数)/(時刻情報を用いる手法における False Negative) を表している．

表 1: False Positive:30% のときの各手法の Hit Rate

	行の重要度	時刻情報	時間情報
data1	100%	90%	60%
data2	100%	89%	61%
data3	100%	85%	75%
data4	100%	100%	49%

表 2: 複合的異常検出結果

	行の重要度との複合		時刻情報との複合	
	5%	10%	5%	10%
data1	0/5	1/5	0/7	4/7
data2	14/20	15/20	0/6	0/6
data3	8/18	14/18	0/8	4/8
data4	1/3	1/3	0/0	0/0

5. 考察

行の重要度を用いる手法と時刻情報を用いる手法では，False Positive が 30% 前後でほとんどの異常を検出でき，すべてのログを一行ずつ検査することと比較すると管理者のログ分析作業負荷を約半分に軽減できる．また，False Positive が 10% 前後で Hit Rate:60% となり，重大な異常のみを通知する場合に有効である．したがって，多様な計算機環境や管理ポリシーに適応できる．

また，時間情報を用いる手法では，すべての異常の検出を行うと False Positive が 80% 程度まで増加するため，管理者のログ分析作業の負担として現実的ではない．しかし，表 2 を見ると，行の重要度を用いる手法で検出できなかった異常行が時間情報を用いる手法で検出されている．このことから，時間情報を用いる手法と行の重要度を用いる手法との組み合わせは有効であるといえる．

6. まとめ

本稿では，大量のログから異常なログを自動検出することにより，計算機ログの分析作業における負荷の軽減を目的として，計算機ログに出現する異常ログの特徴に着目し，行の重要度・ログの時刻情報・ログの時間情報を用いて，異常検出を行う手法を提案した．今後は他の計算機や異なる種類のログなどの多様なログへの対応，異常の度合に関してランク分けすることが考えられる．

参考文献

[1] 高田, 小池: “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”, 情報処理学会論文誌, Vol.41, No.12, pp.3265-3275(2000)

[2] 長尾 真: “自然言語処理” , 岩波書店（1996-04）