Syslog+SNS分析による ネットワーク故障検知・原因分析技術

木村 達明（きむら たつあき）/竹下 恵（たけした けい）
豊野 剛（とよの つよし）/横田 将裕（よこた まさひろ）
西松 研（にしまつ けん）/森 達哉（もり たつや）※
NTTネットワーク基盤技術研究所

本稿では，ネットワーク機器から生成される億単位のSyslogデータに対し，事前知識なしに自動学習で分析し，短時間で原因特定を可能とするログ分析技術，およびSNS全体で１％に満たない故障関連情報を，リアルタイム・高精度に抽出するSNS分析技術について紹介します．

ビッグデータ分析技術のネットワークオペレーションへの応用

近年のIP系サービス提供を支えるネットワークは，複数のベンダによるさまざまな装置で構成されています．また，トラフィックの急速な増大などに対応するために，日々，管理対象の機器が増加し，ネットワーク構成も変化するなど，非常に複雑化しています．ネットワーク上で提供されるサービスは，複数のネットワーク事業者・端末等の組合せで成り立っているため，故障・品質劣化時の原因を特定することが難しくなっています．

これらの背景を受け，ネットワーク運用者の負荷は増大しており，故障検出や故障対応を効率的に行う技術やプロアクティブに運用を行うための技術が求められています．

本稿では，サイレント故障の早期検知や故障の予兆検知を目的に，膨大かつ非定型な情報である，ネットワーク装置のログ（Syslog）の情報やSNS（Social Networking Service）の情報を分析することで，従来のネットワーク監視システムではとらえきれなかった故障事象や，その原因を早期に検出する技術について紹介します．

ログ分析技術

ネットワーク運用者が保全業務で確認する範囲は，装置からのTrap情報に加え，トラフィック情報や機器のCPU・メモリ使用率，Syslogに代表される装置ログなどさまざまですが，中でも装置ログは各機器の詳細な情報を含むため，故障対応時や設定変更時などに機器の状態を把握するうえで有用です．しかし，ログを有効に活用するためには，以下の課題があります．

・ログメッセージには重要度が高いものから低いものまでさまざまなタイプが存在しており，加えて，機器増大等により大規模化したログ情報の中から，故障対応や予防保全に有用な情報を効率的に抽出する仕組みが必要
・ログ形式はベンダやサービスに依存しており，ログの意味を理解するためには，個々の形式ごとに専門の知識やノウハウが必要

これらの課題に対し，機械学習を適用することで，ログのフォーマットやベンダ情報などの事前知識を全く使わずに，ログ間の関係性や異常性を自動的に抽出可能とするログ分析技術について述べます（**図１**）．ログ分析技術は，ログテンプレート抽出，ログ生起特徴量の抽出，ロググルーピング，異常イベントの可視化の４つのステップから構成されます．

⑴　ログテンプレート抽出

ログデータにはIPアドレス，PID（Process ID），インタフェース名などのパラメータが多く含まれており，そのままのかたちでは，メッセージ間の関

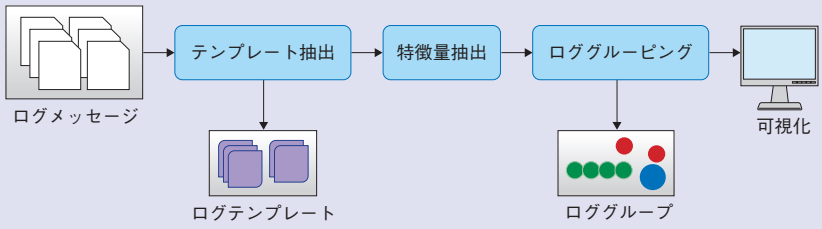

図１　ログ分析技術

※　現，早稲田大学

係や個々のログの異常性を分析できません．そこで，パラメータ部に含まれる単語は，ほかの単語と比較して相対的に出現頻度が低い性質を利用し，ログメッセージからパラメータ部を除いたログテンプレートを自動生成します．これにより，ログをテンプレート単位でまとめて扱うことができ，互いの関係を把握しやすくなります（**図 2**）．

⑵　ログ生起特徴量の抽出

ログの中には，ユーザのセッション切断のたびに発生するログなど，定常的に高頻度に発生するものや，cron ジョブ*1 や定点監視のログのように，頻度は低くても毎日同頻度で発生するものがあります．これらはその生起パターンから正常なログと考えられます．そこで，これら頻度と周期性の 2 つの生起パターンを用いてログごとに特徴量化を行います．

⑶　ロググルーピング

ログによっては，同時生起性の高いものが存在し，その組合せによりネットワーク機器の詳細な状態把握が可能となります．例えば，ルータの再起動時には各種プロセスの初期化に伴うメッセージが同時に発生しますが，これはまとめてルータ再起動というイベントととらえることができます．こうした関係をとらえることで，大規模なログをイベントという意味あるログの集まりに情報圧縮します．ログテンプレートの生起を行列で表現し，NMF（Non-negative Matrix Factorization：非負値行列因子分解）[1] を適用することで，イベントに変換した情報やロググループを得ます（**図 3**）．

⑷　異常イベントの可視化

ログイベントの可視化方法を**図 4** に示します．横軸は時間，縦軸は前述のグルーピングの種類を表しています．グラフ内の丸はロググループまたはログテンプレートの時間における生起を表しています．**図 5** の実施例では，色やシンボルの形状によってホストと一意に紐付いており，異なるホストでのログの生起の区別が可能となります．縦軸のソートの方法については，最下部には頻度が一定値以上高いログテンプレートをプロットし，その上部には周期性の高いログテンプレートをプロットします．また，頻度によりソートされたロググループを，そのログ内に含まれるログ数に応じてプロットします．あらかじめ頻度や周期性の高いログを別で表すことで，時間に関係なく発生するロググループを区別して閲覧することができます．従来のテキストベースのログ閲覧ではできなかった，数千，

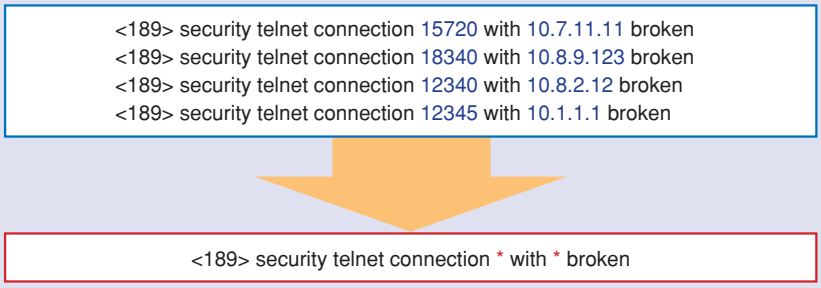

図 2　ログテンプレート抽出イメージ

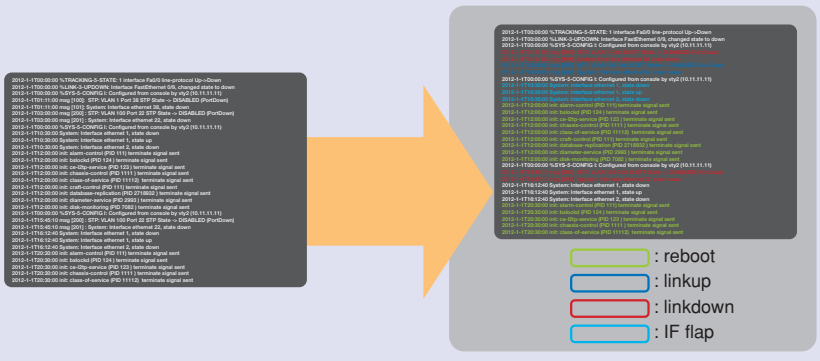

図 3　ロググルーピングのイメージ

*1　cronジョブ：スクリプトを自動実行するためのデーモンプロセス．

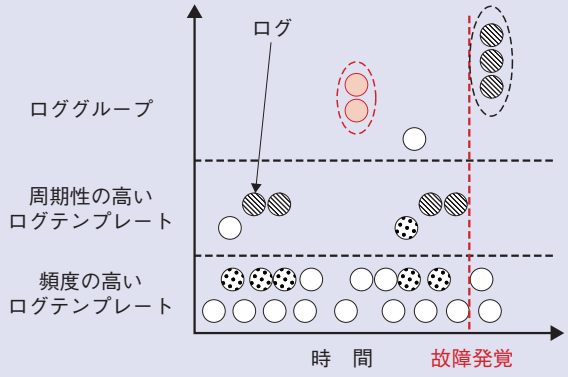

図 4　ログイベントの可視化方法

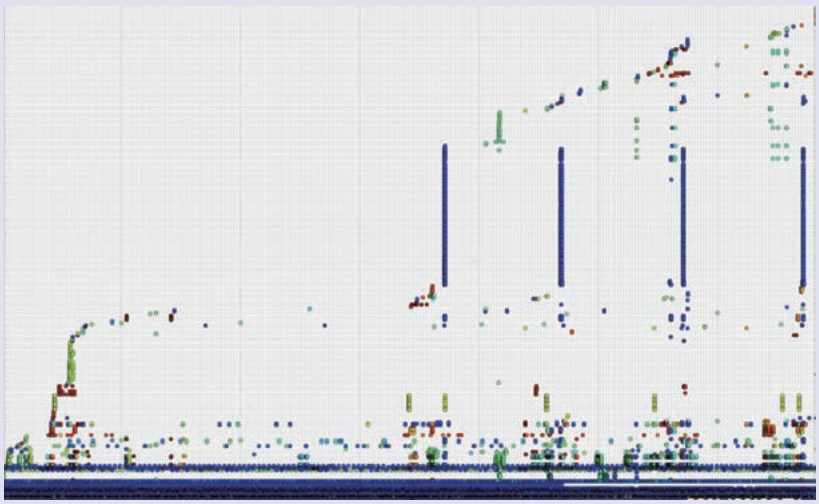

図 5　実施例

数万行のログを一画面へ可視化することが可能となり，どの時間にログが発生したのかを視覚的に分かりやすくとらえることができます．さらに，頻度や周期性を用いたソートにより，普段発生しないタイプのログを確認でき，これらがロググループとしてまとめられることで，実際にネットワーク上で発生したイベントとの紐付けが容易となります．

Twitter分析技術

自社のネットワークの問題を監視する場合には，ネットワーク内の装置によって対応していますが，サイレント故障のように対応できない故障も存在します．また，品質劣化などは，お客さまにどのような影響が発生しているか把握が難しいケースもあります．これらのネットワーク側で把握が難しい情報に関し，SNS上でのお客さまの声を分析することで，ネットワークに問題が起こっていることを直接検出する方法を検討しています．これまでの検討から，リアルタイム性の高いTwitter[(2)]では，特にモバイル系のサービスにおいて，お客さまがネットワークへの不満をリアルタイムに記述してくれていることが分かっています．

ある故障発生時のTwitter上での故障に関するつぶやき数を**図 6** に示します．故障が起こる前はほぼなかったツイートが，故障が起こった瞬間に盛り上がり，復旧とともに落ち着いていく様子が分かります．

このように，Twitter上に故障情報ツイートが増えていないかをリアルタイムに監視することで，故障が起こった際の状況把握や，これまで見つからなかった問題の検知を目指しています．

取り組むべき課題

Twitter上では，日々さまざまな話題がつぶやかれており，その数は1日4億ツイートを超えています[(3)]．そのうち，15～20%のツイートが日本語であるため，1億近くの日本語のツイートが日々投稿されていることになります．しかし，その中に含まれる故障情報に関するツイートはごくわずかなので，①膨大なツイートから適切に故障

情報のみを取り出すこと，②故障エリアを特定するために，投稿者の位置を推定すること，この2つが必要となります．

高精度に故障に関連するツイートを抽出する技術

故障に関連するツイートを抽出するための既存技術として，キーワード検索が挙げられます．しかし，故障を表現するキーワードは一般的な表現が多いという問題があります．例えば，「遅い」「重い」などの表現は，故障のときに使われます（ネットワークが遅い，サーバが重いなど）が，故障でないときにも使われてしまう（販売が遅い，端末が物理的に重い）ため，これらを見分ける必要があります．

そこで，**図7**のように，キーワード検索の結果に対して，教師あり機械学習アルゴリズムであるSVM（Support Vector Machine）による分類器にかけて判定を行うことを検討しています．教師あり機械学習では，教師データとして，目視によって選んだ故障に関する情報のツイート，および故障とは関係ないツイートを与え，各教師データツイートに対して，それぞれの単語を

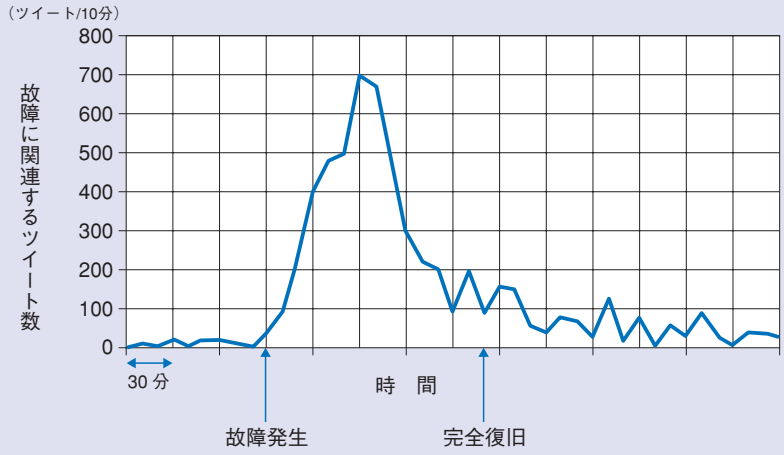

図6　ある故障発生時のツイート数の変化

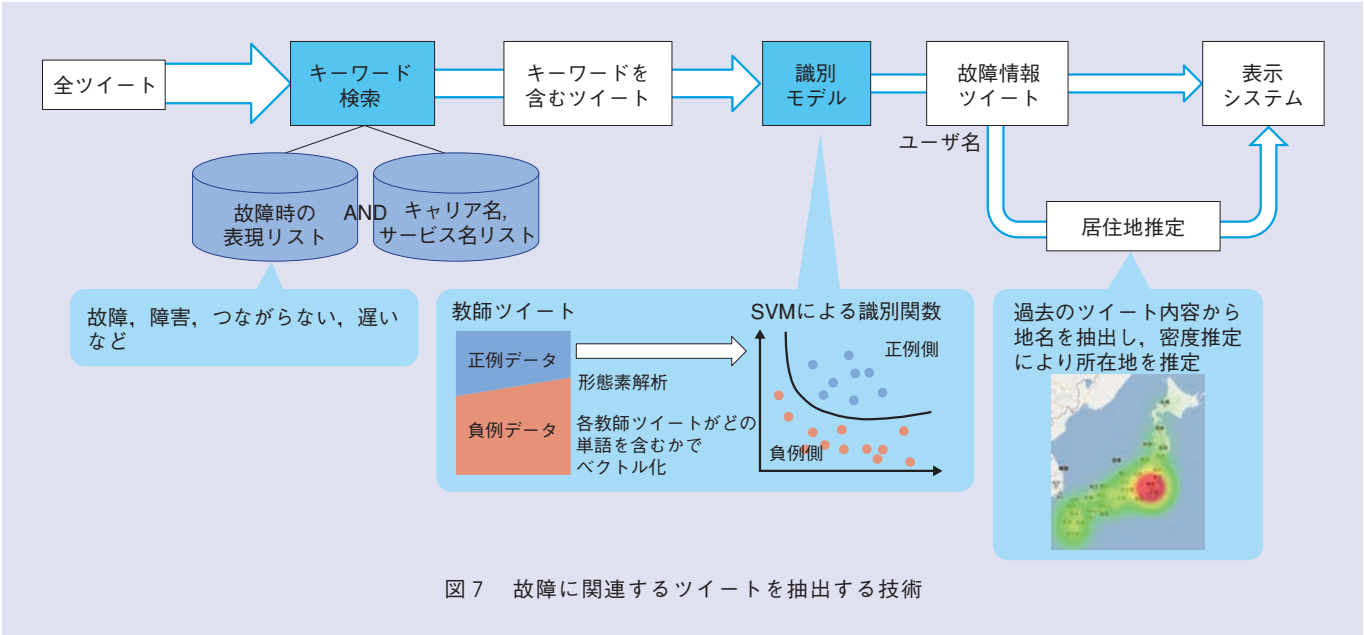

図7　故障に関連するツイートを抽出する技術

含むかどうかの変数のベクトルとして与えることで，正解データと不正解データをもっとも良く分けるような関数を作成します．この方法では，各ツイートの単語の共起関係を考慮して故障情報ツイートかどうかを判定することができるため，単純なキーワードで検索するよりも精度が向上します．

この方法の性能評価のため，公開されている故障情報に対して2011年11月から1年分のツイートを対象として，これらの情報を取得できるかを実験しました．この期間には，公開されている故障は6回起こっていました．10分間に100ツイートが発生し続けている間を故障状態だと判定するようにしたところ，いずれの手法でも実際の6回の故障は検出できており，故障の見逃しはありませんでした．しかし，キーワードのみで抽出した場合には誤検出（故障と判定したのに，実際には故障でなかったケース）が94件もあったのに対して，機械学習でフィルタリングした場合には5件となり，機械学習によるフィルタリングが有効であることが分かりました．

故障発生場所を推定する技術

Twitterには，GPS等で位置情報を付ける仕組みはあるものの，ほとんどの人は位置情報を付けていません．そのため，故障が起こっている場所を知るには，投稿内容などからその人が発言した位置を推定する必要があります．既存の手法では，方言等の言語の偏りを用いているため，関東・関西といったエリアでの推定を実施していますが，故障の検出には，最低でも県レベルの精度での推定が必要となります．

そのため，私たちは座標情報を用いた高精度な居住地の推定法を検討しています．位置情報を付けていないユーザでも，駅名や地名はつぶやきに含まれています．1つの地名だけで判断すると，自分がいない地域の話題を出した際などに推定を誤ってしまうといった問題がありますが，Twitterは自分が今行っていることや興味がある物事を記述するサービスですので，ユーザは自分の近くの話題を出す頻度が自然と多くなります．この特徴を利用して，ユーザの過去のツイートに含まれる地名情報の座標をカーネル密度推定*2という技術により重ね合わせることで，その人の生活圏を高精度に推定できると考えています．

実際にGPS等で位置情報を付けている人に対して，位置情報を抜いて上記の技術を使用した結果，半数のユーザは位置情報と25 km以下の誤差で推定できることが分かりました．

今後の展開

本稿では，非定型なビッグデータを対象に，サイレント故障の早期検知や故障の予兆検知を目的とした，ログ分析技術とSNS分析技術について述べました．ログ分析技術は，現在，事業会社と共同で，予兆検知の観点からサンプルデータを利用して，有効性の評価を進めています．SNS分析技術は，サイレント故障検知や故障発生時のユーザ影響の早期把握ツールとして，デモ等により，ユースケース確立に向けた提案活動を行っています．

*2 カーネル密度推定：統計学において，確率変数の確率密度関数を推定する手法の1つ．

■参考文献

(1) D. D. Lee and H. S. Seung：“Learning the Parts of Objects by Non-Negative Matrix Factorization，”Nature，Vol.401，No.6755，pp.788-791，1999.
(2) https://twitter.com/
(3) http://blog.twitter.com/2013/03/celebrating-twitter7.html

（後列左から）西松　研/ 木村　達明/ 豊野　剛/ 森　達哉（右上）
（前列左から）竹下　恵/ 横田　将裕

端末，サービスの多様化，ネットワークの大規模化により，ネットワークの運用は非常に複雑になっています．私たちは，複雑さを解消し，安定したネットワーク運用を支えるためのデータ分析技術の研究開発に取り組んでいます．

◆問い合わせ先
NTTネットワーク基盤技術研究所
通信トラヒック品質プロジェクト
TEL　0422-59-3061
FAX　0422-59-6364
E-mail　nishimatsu.ken@lab.ntt.co.jp