# ネットワークトラフィックデータ解析による異常発見

梅村恭司 (第 4 工学系, 情報環境コア)

## 1　はじめに

近年，P2P ソフトを介した個人情報や違法にコピーされた映画や音楽などのコンテンツの流出が問題となっている．また，ウイルスに感染した PC が不特定多数の相手に大量のメールを送りつけるなど，ネットワーク上で発生するこのような異常トラフィックを発見，対処することが安定的な運用を行う上で欠かせない．ここでは，本研究室が取り組んでいるネットワークトラフィックデータからの異常発見に関する研究について報告する．

## 2　多重観測による正常トラフィックパターン推定

異常トラフィックを検出するためには，まずは正常，あるいは異常なトラフィックパターンを定義する必要がある．本研究では，実際のトラフィックデータを用いてこれらのパターンを定義することを試みた．

一般に，ネットワークトラフィックには 1 日周期．1 週間周期の周期性が存在するとされ，定常状態に周期性を見出すことができれば，正常なトラフィックパターンを定義することは簡単である．しかしながら，トラフィックパターンに周期性を持たないネットワークも存在し，このようなネットワークの場合，そのトラフィックパターンが正常であるのか，あるいは異常が発生した結果として周期性を失ったのか判断することができない．よって，ネットワーク管理者が正常値を設定することや，異常トラフィック検出システムが定常状態から正常なトラフィックパターンを学習することも難しい．

そこで本研究では，トラフィックパターンに周期性を持たないネットワークにおいて，多重観測によって正常なトラフィックパターを推定する手法を提案し，提案手法の評価のために，豊橋技術科学大学 (以下，本学) のネットワークトラフィックを調査・分析し，提案手法に基づいて正常なトラフィックパターンを推定するとともに，正常なネットワーク利用状況下における通信ホスト数の最大値を検討した．そして，その結果得られた最大通信ホスト数を利用して，異常トラフィック検出システムを実装した．検出目的とした異常トラフィックは，通信ホスト数の異常な変化である．通信ホスト数を監視することで，従来の通信量に基づいた異常トラフィック検出手法では検出できない，不正アクセスを目的とすると思われるトラフィックや

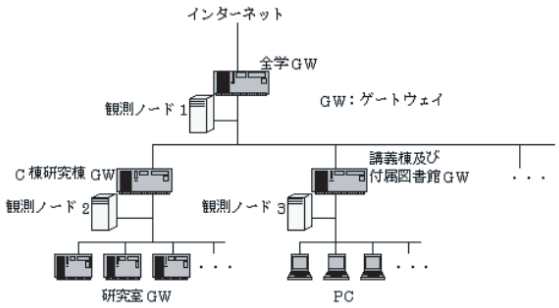


Fig. 1　観測システム概念図

P2P ファイル共有アプリケーションのトラフィックを検出することができる．トラフィックを取得するために，トラフィック観測用ノードを実装し，ゲートウェイ付近に設置してトラフィックを取得した．取得したトラフィックから，本学のネットワークトラフィックには一部を除いて一般的な周期性が存在しないことが分かった．そこで，本学において周期性の無いネットワークトラフィックが正常であるか異常であるかを議論するために，より局所的な 2 つのネットワークのトラフィックを分析した．その結果，局所的なネットワークトラフィックには明確な周期性が存在することが確認された．またトラフィックログを解析することで，局所的なネットワークトラフィックが全学のトラフィックに伝搬することも確認された．以上のことから，本学のトラフィックにおいて周期性が存在しないことは異常であり，一部の周期的なトラフィックが正常なパターンであると結論付けることができた．本研究ではさらに，上記の議論で得られた結論をもとにして，異常トラフィック検出システムを実装した．

## 3　文字列解析に基づくネットワークトラフィックデータからの異常発見

本研究では，時系列データの文字列表現手法を用いてネットワーク上の異常トラフィックを自動的に発見する方法として，マルコフモデルによる異常候補の列挙とクラスタリングによるはずれ値検出を組み合わせた方法を提案した．

異常発見方法には大別して，シグニチャ型とアノマリ型がある．シグニチャ型は，予め異常なパケットの特徴，例えばペイロード中に必ずある文字列が出現するといった特徴をルールとして記述しておき，パター

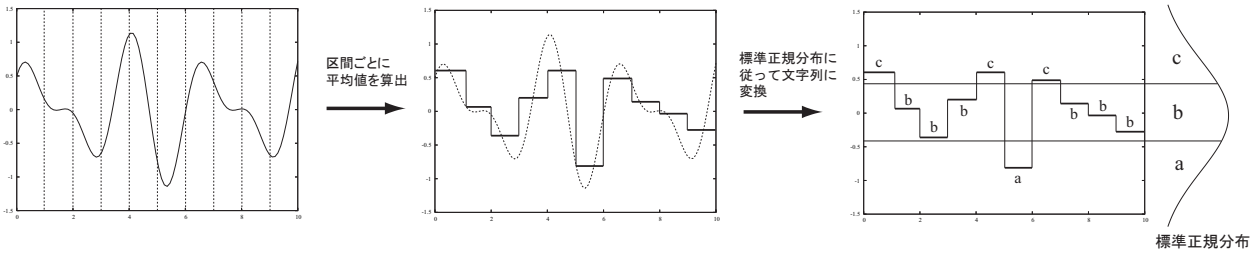


Fig. 2　SAX による時系列データの文字列への変換

ンマッチを行って検出する．一方，アノマリ型では，予め正常な通信パターンをモデル化しておき，これに外れる振る舞いを示す場合に異常と判断する．判断基準のルール化が容易ではないという欠点はあるものの，未知の異常へ迅速に対応できるという利点がある．

アノマリ型の発見手法としては，信号分析や時系列解析などによる方法が研究されているが，近年，データマイニング分野において，時系列データを文字列によって表現する SAX (Symbolic Aggregation approXimation) が提案され，いくつかの実験においてその有効性が示された．時系列データを文字列で表現することにより，自然言語処理分野で培われた文字列処理，言語解析などの有用な技術が適用可能となる．本研究では，この文字列表現による解析手法の一つとして，マルコフモデルとクラスタリング技術を用いた方法を提案する．マルコフモデルを用いた異常発見は，SAX を提案した Keogh ら [1] によって提案された．マルコフモデルとは，ある時点で出現するアルファベットの確率が，過去のアルファベットの出現に依存して決まるような確率モデルである．正常な通信データ (訓練データ) を用いてモデルを生成することで，任意の部分文字列の出現回数の期待値が計算できる．検証したい通信データ (テストデータ) における出現回数と比較することで，異常かどうかを判断する．以上が，マルコフモデルを用いた異常発見方法の基本的なアイデアである．

しかし，実際に我々がこの方法を用いて予備実験を行ったところ，誤検知，つまり正常であるのに異常であると判定されるデータが多く含まれてしまうことが分かった．このため，我々はマルコフモデルによって列挙されたデータを異常候補集合として扱い，この集合にクラスタリングを適用することで，より異常データである可能性の高いものを絞り込むこととした．正常データには類似性があるためクラスタを形成しやすい，よってクラスタに取り込まれないはずれ値を見つけ出すことが目的となる．本研究では，はずれ値検出に適した階層型クラスタリングを用いた．マルコフモデルを用いなくてもクラスタリングを適用することは可能であるが，一般に階層型のクラスタリングの計算コストは高く，データ長が大きくなると部分文字列数も多くなり，莫大な計算時間を要してしまうことになる．よって，マルコフモデルによる方法でクラスタリング対象を選別することで現実的な計算時間での処理を可能とすることができる．

実際のトラフィックデータを用いて，提案手法の定性的な評価を行ったところ，マルコフモデルを用いた方法のみでは誤検知が多いことが分かり，クラスタリングを適用することで誤検知を取り除く効果が期待できることが示された．いくつかの異常発見箇所では，異常原因が P2P ソフトやウィルスによるものであることも分かった．

## 4　まとめ

ネットワークトラフィックデータからの異常発見への取り組みは，現在のところ初期段階にある．今後も文字列処理による解析を継続するほか，視覚化ツールの構築などを行っていく予定である．

## 参考文献

[1] Keogh, E., Lonardi, S and Chiu, W. “Finding Surprising Patterns in a Time Series Database In Linear Time and Space.” In *Proc. ACM KDD’02,* pp.550–556 (2002)

## 発表論文

[1] 小塚雅洋，岡部正幸，梅村恭司，”分散強調システムによるトラフィッ ク測定システムの開発”，第 99 回システムソフトウェアとオペレーティング・システム研究発表会，情報処理学会研究報告 2005-OS-99，pp.99-104 2005 年 5 月

[2] 岡部正幸，三輪多恵子，梅村恭二，”文字列解析に基づくネットワークトラフィックデータからの異常発見”，インターネットカンファレンス，pp.67-74 2006 年 10 月