ネットワークの利用動向からの異常検知手法について

Aberrant Detection from Behavior of Campus Network Traffic

金西 計英†，戸川 聡‡，松浦 健二†，光原 弘幸††，矢野 米雄††
Kazuhide KANENISHI †, Satoshi TOGAWA ‡, Kenji MATSUURA †,
Hiroyuki MITSUHARA †† and Yoneo YANO ††

marukin@cue.tokushima-u.ac.jp, doors@shikoku-u.ac.jp, matsuura@ait.tokushima-u.ac.jp,
mituhara@is.tokushima-u.ac.jp, yano@is.tokushima-u.ac.jp

† 徳島大学 高度情報化基盤センター
†† 徳島大学大学院 ソシオテクノサイエンス研究部
‡ 四国大学 経営情報学部

† Center for Advanced Information Technology, The University of Tokushima
†† Institute of Technology and Science, The University of Tokushima
‡ Faculty of Management and Information Science, Shikoku University

概要

Peer-to-Peer（P2P）型ファイル共有ソフトウェアの利用によってもたらされる問題は，座視できない状況となっている．ネットワークの帯域の圧迫という問題もさることながら，Winny などのファイル共有ソフトウェアに感染する暴露ウイルスによる情報漏洩のリスクが深刻な問題となっている．多くの大学や企業では，P2P ファイル共有ソフトウェアの利用を禁止している．現実には，P2P ファイル共有ソフトの利用を制限することは困難である．結果，管理者はトラフィックを常時監視し，P2P ファイル共有ネットワークの存在を認識しなければならない．本稿では，P2P 通信の検出作業に対する，トラフィックのマイニングと可視化による支援を提案する．管理者への支援という観点に立ち，トラフィックの可視化をおこなうトラフィックマイニングツールを開発した．また，試作環境での実験を通し，提案手法の有効性を検証した．

キーワード

管理者支援，トラフィックマイニング，P2P，可視化，自己組織化マップ(SOM)

1. はじめに

キャンパスネットワークとそこを流れる情報は，大学の教育・研究活動にとって必要不可欠なものとなっている．情報インフラの重要性は高まり，その利用も増大している．キャンパスネットワークの利用の増大は，それを管理する管理者の負担も大きくする．さらに，ネットワークに関連する技術は，常に新しくなっており，管理

者は自らの担当するキャンパスネットワークを適切に運用するため，そうした新しい技術への対応が求められる．様々な面から，管理者のネットワーク管理に対するスキル向上の要求は高まっている．管理者に対する要求は増加する傾向にあり，管理者の負担が減少する兆候はみられない．管理者の負担を軽減するような試みが無いではないが，そうした試みが充分な効果を上げているとはいえない．そのために，本稿では，キャンパスネットワークの管理者の，ネットワーク運用の負担を軽減するための手法について提案する．本研究は，ネットワークの管理者の作業を補完するツールの提供を目指している．具体的には，キャンパスネットワークの通常とは異なるトラフィック上の振る舞いを管理者へ通知することで，ネットワークのトラフィック監視作業を支援するようなツールを開発している．本ツールを，トラフィックマイニングツールと呼ぶ．

ネットワークのトラフィック上の異常の検知という点で，我々は，Peer-to-Peer（P2P）通信によるトラフィックに着目した．多くのキャンパスネットワークにおいてP2P 通信が問題になっているからである．中でも P2P 通信を用いたファイル共有が問題なっている．従来，WinMX[1]，Winny[2]，BitTorrent[3]，Share 等が存在し，最近では Perfect Dark[4]などのより匿名性を高めたファイル共有ソフトウェアも開発されている．

P2P ソフトウェアの利用が問題になるのは，一つは，帯域の圧迫という点にある．P2P の利用によって，対外接続のネットワークの帯域が占有されてしまう．そのため，DNS，電子メール，HTTP 等の通信が阻害され，学内の様々な活動に影響が出ることが予測される．

次に，ファイル共有ソフトを用いて，音楽，映画，画像等，あらゆる種類のコンテンツが共有されることが挙げられる．著作権の保護対象である著作物を，違法に流通させているものが多い．著作権者の許諾を得ず，コンテンツを複製・配信することは問題がある．

しかし，より深刻な問題は，ファイル共有ソフトウェアの使用による情報漏洩のリスクである．Antinny などの暴露型コンピュータウイルスは，感染したコンピュータ内のリソースをアーカイブし，ファイル共有ネットワークへ公開する．この結果，使用している PC のローカルストレージに保存されている様々な情報が，ファイル共有ネットワークへ流出する．個人のプライベートな動画像，自衛隊や警察の機密資料等がファイル共有ネットワークに流出した情報漏洩事件が社会問題となったことは記憶に新しい[5,6]．

無論，ファイル共有ネットワークへの情報漏洩は，ファイル共有ソフトが直接の原因ではない．しかし，ファイル共有ソフトの利用が，暴露ウイルスの罹患を促し，情報漏洩のリスクを高めていることは間違いない．

キャンパスネットワークにおいて，ファイル共有ソフトの利用を規制している事例が多い．キャンパスネットワークの管理者は，利用者のモラルにのみ頼るのではなく，さまざまな対策を講じている．しかし，キャンパスネットワークからファイル共有行為を完全に停止させることは容易ではない．

管理者が P2P ファイル共有ソフトを制限する場合，まず，パケットフィルタの利用が考えられる．しかし，P2P ノードは，不特定多数かつ可変であるため，IP アドレスによるフィルタリングは困難である．さらに，Winny や Share 等の自律分散ノード集合で共有ネットワークを構成するタイプのソフトは，標準的な待機ポートを持たない．P2P ノードはランダムな TCP ポート番号で接続を待ち受ける．そのため，ポート番号によるフィルタリングも困難である．ファイル共有ソフトは，常に，管理ツール等の裏を掻く方法を求めており，管理者とファイル共有ソフト開発者とのいたちごっこの様相を呈している．最終的に，管理者は，フィルタリング等を併用しつつ，キャンパスネットワーク内のトラフィックを自ら監視し，P2P 通信を見つけ，対処しなければならない．

以上のような点を踏まえ，我々のトラフィックマイニングツールは，キャンパスネットワーク内からおこなわれている P2P ファイル共有通信の検出を支援する．トラフィックマイニングツールは，キャンパスネットワークのある時点のトラフィック全体の振る舞いを示したモデル情報を生成する．その中で特徴的な振る舞いを持つトラフィックに対し，特性を強調する．最終的に，トラフィックの全体傾向を俯瞰可能な特徴マップとして可視化し，管理者に提示する．特徴マップにより異変に気づいた管理者は，対象を絞った調査をおこなうなどの作業に取り掛かることができる．

以下本稿では，2 で P2P ファイル共有通信の現状について述べ，3 でこれらファイル共有通信の検出支援モデルについて述べる．4 で本研究で使用するトラフィックモデルの構成と可視化手法を述べ，5 で試作システムの概要を述べ，6 で後実証実験の概要と考察を述べる．最後に 7 で全体のまとめをおこなう．

2. P2P ファイル共有通信の現状

2.1. P2P ファイル共有の通信モデル

P2P は，そのインデクスの管理形態の違いによって以下のように分類される．

Hybrid 型：図 1 に Hybrid 型の通信モデルを示す．ファイル所在情報である索引を保持するインデックスサーバと，実体ファイルを保持するノード群から構成される．あるノードがファイルの入手を試みる場合，目的ファイルの所在をイ

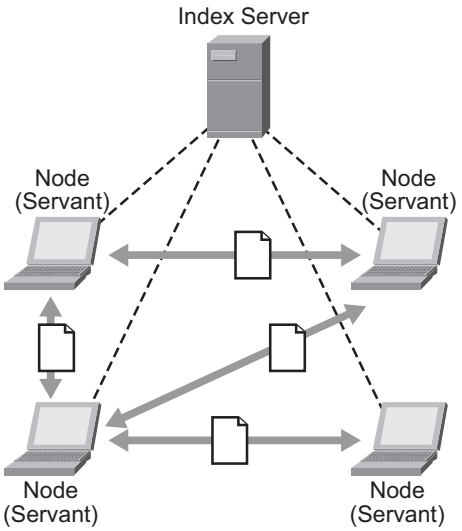

図1 P2P ファイル共有モデル

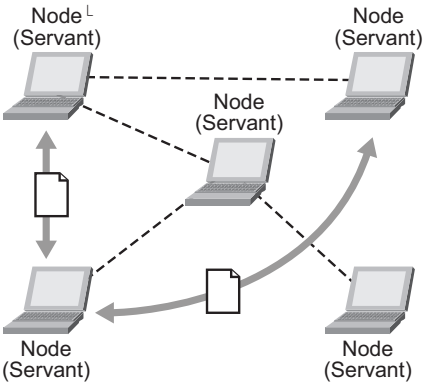

図2 Pure 型 P2P ファイル共有モデル

ンデックスサーバに問い合わせる．インデックスサーバは当該ファイルの所在情報を探索元ノードに返す．探索元ノードはファイルを所有するノードとコネクションを確立し，目的ファイルを入手する．

Pure 型：図 2 に Pure 型の通信モデルを示す．Pure 型は索引情報を保持するインデックスサーバを持たない．ファイルやノードの探索機能はノード自体に実装される．ノードがファイルを探索する場合，近接するノードに探索要求を発行する. これを繰り返すことで探索要求は P2P コミュニティ内に伝搬する．あるノードが要求に合致するファイルを所有していた場合，そのノードはファイル所在情報を返す．所在情報を受信した探索元ノードは，ファイルを所有するノードとコネクションを確立し，実体ファイルを入手する．

SuperNode 型：SuperNode 型は，Hybrid 型を進化させたものと考えることができる．インデックス情報を，複数のスーパノードに保持させる．このことによって，インデックスが膨大になっても共有ネットワークが破綻しない．SuperNode 型のスケーラビリティは高い. また. 複数のスーパノードによる構成から，耐障害性も高くなっている．

Hybrid 型アプリケーションとして WinMX，BitTorrent 等があり, Pure 型アプリケーションとして Winny, Share, Perfect Dark 等がある．また，Super Node 型としては KaZaA，Skype 等が挙げられる．

2.2. Pure 型 P2P ファイル共有の通信特性

Pure 型 P2P ファイル共有の通信特性を明らかにするため，以下のような予備実験をおこなった．本実験の目的は，P2P ファイル共有プロトコルの内容を明らかにすることではない. 送信元から見た Pure 型ファイル共有ソフトウェアの利用によってネットワークに現れる表層的な振る舞いを明らかにすることである．

実験用 PC に Share をインストールし, P2P のトラフィックを発生させた．今回，比較対象として Web 閲覧のトラフィック情報も収集した．これは，一般的なキャンパスネットワークにおいて, 利用の大部分は Web アクセスが占めると想定されるためである．

実験時間は, それぞれ 15 分間程度実験機から送信される IP パケットを収集し, 解析をおこなった. 表 1 に Share の計測結果を, 表 2 に Web 閲覧における計測結果を示す．

表 1 予備実験結果（Share）

IP パケット送信数	16,009
TCP PUSH フラグ付与件数	5,056
宛先 IP アドレス数	299
宛先 TCP ポート数	254
宛先ホスト名名前解決率	0.1%

表 2 予備実験結果（Web 閲覧）

IP パケット送信数	5,322
TCP PUSH フラグ付与件数	469
宛先 IP アドレス数	34
宛先 TCP ポート数	2
宛先ホスト名名前解決率	92.6%

まず，Share の IP パケット送信数は 16,009 件であり，Web 閲覧の約 3 倍である．IP パケット送信数に占める TCP PUSH フラグ付与率は，Share が約 31.6%，Web 閲覧が約 8.8%である．Share から見た宛先の TCP ポート番号は 1 番から 65535 番まで一様に広がっていた. なお, Web 閲覧の宛先 TCP ポート番号は，80 番と 443 番のみであった．この実験から，Share は通常の Web 閲覧に比べ大量の IP パケットを送信し，TCP PUSH フラグの付与率が高いことが分かる．これは，広範囲な宛先 IP アドレス，

および均等に散らばった宛先 TCP ポート番号に対しコネクションを確立した結果だといえる.

また，宛先ホストに関する名前解決についても解析をおこなった．その結果，Web 閲覧における宛先ホストの名前解決率が 92.6%だったのに対し, Share の場合は 0.1%であることが分かった．Web 閲覧時において宛先ホストを指定するとき，ほとんどの場合 FQDN を含む URL が指定される．そのため，Web 閲覧では，宛先 IP アドレスを入手するために，DNS の名前解決が発生する．一方，P2P ファイル共有の場合, 宛先ホストを直接 IP アドレスで指定する場合がほとんどであるため, DNS 名前解決が発生していないことが分かる.

2.3. フィルタリングによる利用制限

P2P ファイル共有ソフトを制限する方法として, まず，パケットフィルタリングが想定される．フィルタリングによる制限の有効性について検討する．予めルールを記述しておく静的なフィルタリングについて考える.

まず，Hybrid 型 P2P ファイル共有通信では，ファイル共有ネットワークの構成から見たとき，インデックスサーバが単一障害点となる．インデックスサーバへの経路を遮断すれば，理論上リソース検索機能を遮断できる．そこで，既存のフィルタリング技術は，Hybrid 型 P2P ファイル共有通信の制限に関して，一定の効果が期待できる．また，SuperNode 型の P2P ファイル共有通信においては，ファイル共有ネットワークの規模によるが（スーパノードの数等），Hybrid 型と同様に，一定の効果があると考えられる.

一方，Pure 型 P2P ファイル共有の制限を，フィルタリングでおこなうことは困難である．前節で述べた通り，Pure 型 P2P ファイル共有の通信では，広範囲な宛先 IP アドレスと, ランダムな宛先 TCP ポートに対しコネクションが張られる．このため予めルールを記述しておくフィルタリングの活用は困難である.

2.4. フロー情報解析による異常トラフィック自動検出

NetFlow[7]，sFlow[8]等を用いてフロー情報を収集，解析し，P2P トラフィック等をトラフィック上の異常事象として検出する手法が存在する.

フロー情報の収集粒度は，フローレコードのサンプリングレートによって決定される．フローを収集しようとするルータ，もしくは L3 スイッチの性能にもよるが，一般的に，1:100 から 1:1000 程度のサンプリングレートが用いられる．フロー情報をもれなく収集するには，サンプリングレートを 1:1 にすればよい．しかし，当該機器への負荷を考慮すれば，これは非現実的である．このためフローによるトラフィック収集は，あくまでもサンプルの抽出に止まる. 解析しようとするトラフィックを，完全に収集することは難しい．つまり，フロー情報解析による異常トラフィック検出には，一定の効果が期待できるものの，精度上，推定や予測の誤差による誤検出を避けることはできない.

また，NetFlow，sFlow 等のフロー収集機能は，キャリアクラス，大規模エンタープライズクラスの L3 スイッチに実装されているが，廉価な L2 スイッチには実装されていない．そのため，キャンパスネットワーク内の任意の計測点に対し, フロー収集を実現することは難しい.

2.5. その他の P2P ファイル共有ソフト検知手法について

P2P のトラフィックに対応するために，フィルタリングやフロー解析以外も，いろいろな手法が提案されている[10,11]. IDS(Instruction Detection System)と呼ばれる検知システムは増えている．例えば，Winny の通信を検出するため，FPGA を用いた手法が提案されている．FPGA を用いることで高速な処理が可能となり，Winny のトラフィックの発見に有効である．しかし，高速で正確な検知が可能となっている一方で，FPGA を用いているため，システムの動作の変更を施すことは容易ではない. また，システムの設定設定も，必ずしも容易とはいえない．さらに, 現状では P2P のトラフィックの発見というよりも，Winny に特化しており，他の P2P ファイル共有ソフトを検知するためには，ファイル共有ソフト毎に FPGA を用意する必要がある.

3. トラフィック情報からの異常発見の支援

3.1. トラフィックマイニングと可視化による異常発見の支援

我々は，ネットワーク管理者のトラフィックの異常発見を支援する枠組みの構築を目指している．管理者は，自らが管理するネットワークの状態を，常に，監視している．その上で，何らかの異常を見つけた場合，対処を進めている.

管理者がおこなっている監視作業は，ネットワーク機器からログを収集し, ログを解析することに因っている．多くのログは，あるフォーマットに従ったテキスト形式である．管理者は，大量のテキスト情報の中から，何らかの意味を読み取る．そこで，我々は，このログ情報をクラスタリングし，その上で可視化し，提示するツールを提供する. 各種のログからの情報を可視化することで，管理者にとって情報を解釈する負担が軽減される，と考

えるからである．つまり，管理者への情報の提供について，機能的な処理を加えることで，管理者の支援が可能になる[13]．

一方で，動的なフィルタリングの開発のように，システム自身が異常に対処する自動化ツールを提供することで管理者の支援をおこなう立場も存在する．しかし，我々は，自動化ではなく，あくまでも，管理者の支援を考えている．つまり，管理者にとっての作業の遂行上存在する様々な負荷の軽減を目指しているのであり，管理者に代わる機能の提供を目指しているのではない．これは，支援の立場の違いであり，どちらの立場が優れているといった問題ではなく，それぞれの技術を高める必要があると考える．

また，我々の提案する手法は，トラフィックの全体の傾向を対象としている．TCP パケットのヘッダを用いるだけである．これは，通信における個人のプライバシーに配慮してのためである．

図 3 に，我々の提案する異常検出の支援モデルを示す．本稿では，これら一連の流れから実現されるトラフィック特徴強化と可視化の流れ基づく異常発見支援の手法を，トラフィックマイニングと呼ぶ．その上で，トラフィックマイニングの支援ツールをトラフィックマイニングツールと呼ぶ．トラフィックマイニングツールは，「トラフィックモデル生成」「特徴強化」「可視化」の処理をおこない，管理者の異常検知を支援する．

3.2. トラフィックモデル生成

トラフィックマイニングをおこなうために，トラフィックマイニングツールでは，キャンパスネットワークの送信トラフィックを収集し，以降の処理の基盤となるデータセットを作成する．このデータセットは，トラフィックの全体傾向を内部表現したものである．

P2P ファイル共有を検出するため把握すべきことは，広範囲なコネクションを持ち，TCP PUSH フラグ付与率が高いトラフィックが存在するか否かを見つけることだといえる．さらに，宛先 IP アドレスが名前解決の結果得られたものではなく，直接得られたものが多ければ，そのトラフィックを発生しているクライアントは P2P ファイル共有を行っている可能性が高い．

そのため，ネットワーク内部から外部に対するコネクション生成状況が把握可能な情報抽出をおこなう．また，コネクションごとの TCP PUSH フラグ付与状況，および宛先 IP アドレスにおける名前解決の試行状況を抽出する．抽出した特徴量を，送信元 IP アドレスを要素とし，単位時間ごとにモデル化したデータセットを作成する．本作業で得られたデータセットを，トラフィックモデルと呼ぶ．

3.3. 特徴強化

クラスタリング処理において，データの前処理は重要である．一般に，得られたモデルのデータに対し，正規化と，特徴の強調をおこなう必要がある[9]．我々も，トラフィックモデルに対し，特徴量の強化を試みる．分散したコネクション状態を持ち，TCP PUSH フラグが付与された要素に重み付けをおこなう．さらに宛先 IP アドレスに関する名前解決が試みられていない要素についても

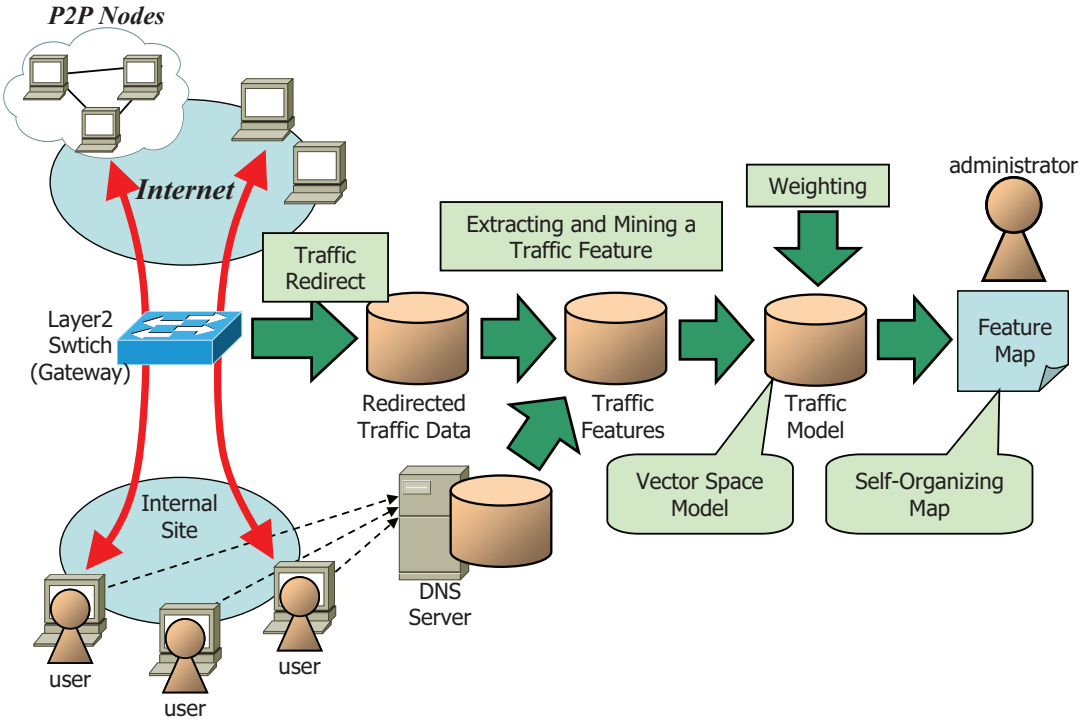

図 3　検出支援モデル

重み付けをおこなう．これにより，P2P ファイル共有をおこなっている可能性を持つ要素を強調することができる．そして，他のトラフィックに埋没する P2P トラフィックを浮上させ，P2P トラフィックの存在を，管理者に気づかせることができる．

3.4. 可視化

可視化の過程では，単位時間で集積されたトラフィックモデルを可視化し，管理者に提示することをおこなう．本研究で扱う監視は，全体傾向把握とその変化による異変発見である．このため管理者への情報提示は一目で全体状況が把握できることが望ましい．人間にとって，テキストとして情報を読むよりも，構造化された情報を一定の形式に従って図示した方が，情報を把握しやすくなることは良く知れている．そのため，我々は情報の提示に，グラフ状の矢線図等を使用する．つまり，情報の可視化は，情報提供の手段として，一定の効果があると考える．

その上で，管理者にとって，単位時間毎の状況が把握できればトラフィック全体の俯瞰が可能となり，変化の追跡も容易となる．

さらに，トラフィック情報を提供する場合，個人のプライバシーに配慮する必要がある．詳細な情報を収集すれば，解析の精度が上がる可能性があるものの，通信の秘密の保護を侵害する恐れがある．そのために，我々は情報の可視化を用いる．個々のデータを扱うのではなく，あくまでもトラフィック全体の俯瞰図を提供する．個々の情報が表出するこのないように配慮した．

4. トラフィックモデルの生成と可視化

4.1. トラフィックモデルの構成

トラフィックモデルは単位時間におけるトラフィック特性を定量的に集積しなければならない．我々は，トラフィックモデルをベクトル空間モデル（Vector Space Model:VSM）を用いて表現する．トラフィックモデルは，一定時間のトラフィックの状態を表した特徴ベクトルということになる．モデルを構成する要素ベクトルには送信元 IP アドレスが対応し，特徴量として宛先 IP アドレスとその出現量を集積する．なお，各要素の特徴量は，パケットサイズ，TCP PUSH フラグの出現量，および DNS 名前解決の有無により重みを付け，その特徴を強化する．

要素ベクトルを x，宛先 IP アドレスごとの出現量を a_1〜a_n とすると，要素ベクトルは次式で表わされる．

$$x = \{a_1, a_2, \cdots, a_n\} \tag{1}$$

トラフィックモデルは，生成されたすべての要素ベクトルを集めたものである．トラフィックモデルを D，要素ベクトルを x_1〜x_m とするとトラフィックモデルは次式で表わされる．

$$D = \{x_1, x_2, \cdots, x_m\}^T \tag{2}$$

これにより，ネットワーク内ノードから送信されるトラフィック特性を，要素ベクトル x のベクトル集合で表現できる．そして，組織内のノード間のトラフィックの状態の類似性を，要素ベクトル間の余弦尺度として算出することができるようになる．その上で，要素ベクトルを，ある通信の特徴を表現したものと考えるとき，組織内のトラフィックの類似性を計算することで，結果的に，計算結果から特徴的なトラフィックを抽出することができると考える．ベクトルモデルは，組織のトラフィックから特徴と呼ぶものを，計算によって導き出すための基盤を提供する．

4.2. 自己組織化マップによる可視化

生成されたトラフィックモデルは多次元ベクトル集合として構成されている．これは送信元 IP アドレスと宛先 IP アドレスの関係が，多次元空間上の分布として表現できることを意味する．人間は基本的に三次元までの空間は直感的に把握可能だが，それ以上の多次元空間の把握には困難を伴う．

自己組織化マップ(Self-Organizing Map:SOM)は，2 層のニューラルネットワークで構成される教師なし競合学習モデルである．入力層に入力される要素ベクトルを x，出力層の各ユニットに連結される参照ベクトルを m_i，処理時間を t とするとき，SOM は以下の処理をおこなう[12]．

$$m_i = \{\mu_{i1}, \mu_{i2}, \cdots, \mu_{in}\}^T \tag{3}$$

Ⅰ　参照ベクトル $m_i(i = 1, \cdots, n)$ を初期化する．

Ⅱ　次式に従い，全ノードから入力 x への最近傍ノード c を探す．これはユークリッド距離が $\|m_i - x\|$ が最小となるノードである．

$$c = \arg\min_i \|m_i - x\| \tag{4}$$

Ⅲ　次式に従い，探索したノード c の参照ベクトル m_c とその近傍ノードを更新する．

$$m_i(t+1) = m_i(t) + h_{ci}(t)[m_i(t) - x(t)] \tag{5}$$

ここでは α は学習率係数である一般に 0〜1 の範囲をとる．h_{ci} は近傍関数であり，次式でしめされるガウス関数である．ここでは r_c，r_i はぞれぞれノード c，ノード i の位置ベクトルを示す．

$$h_{ci}(t) = \alpha(t) \cdot \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)} \right) \quad (6)$$

Ⅳ　ⅡからⅢの過程を繰り返す.

SOM はデータ間の幾何学的構造を可能な限り保った状態で二次元平面に写像する．同時にクラスタリングをおこなう．この結果，管理者は平易な二次元平面にて管理対象組織のトラフィック傾向の俯瞰が可能となる.

5. トラフィックマイニングツールの概要

本章では，トラフィックマイニングツールについて述べる．図 4 にシステムの構成を示す．本システムは「トラフィック収集部」「トラフィック解析部」「DNS クエリ解析部」「モデル化部」「可視化部」の各モジュールから構成される．以下，各モジュールの概要を述べる.

5.1. トラフィック収集部

トラフィック収集部では，監視対象ネットワークが受発信するすべての IP パケットを収集･蓄積する. 今回は，キャンパスネットワーク内の L2 スイッチとの連動を想定している．状態のモニタリングをしたい箇所の L2 スイッチのポートミラーリング機能により，スイッチレベルのトラフィックに対し, 獲得する IP パケットを本システムにリダイレクトする．トラフィック収集部は，導入システムの Ethernet カードを promiscuous mode に設定し，リダイレクトされた IP パケットを収集する.

5.2. トラフィック解析部

トラフィック解析部では，収集したトラフィック情報に対し，特徴ベクトルを作成するための前処理をおこなう．収集された IP パケット群を解析し，送信元 IP アドレス，送信元ポート番号，宛先 IP アドレス，宛先ポート番号，パケットサイズ，フラグを抽出する.

5.3. DNS クエリ解析部

DNS クエリ解析部では，トラフィック解析部と同様に，収集したトラフィックデータに対する，前処理をおこなう．DNS 関連の重み付けのための前処理をおこなう．キャンパスネットワーク内のクライアントが利用する DNS サーバが，インターネット上に散在する DNS サーバへ問い合わせる DNS クエリの結果を収集する.

この結果，キャンパスネットワーク内からインターネット上の DNS に対して，名前解決要求が発行された FQDN とその IP アドレスが対となったクエリ情報が得られる.

5.4. モデル化部

モデル化部では，トラフィックモデルである特徴ベクトルを生成する．具体的には，トラフィックに現れた各送信元 IP に対し，トラフィックに現れた全ての宛先 IP が対応した次元となる多次元ベクトルを，要素ベクトルとして生成する．特徴ベクトルは，収集した単位時間における, 送信元 IP のトラフィックの状態が集積されたベクトル集合となる. 各要素ベクトルには宛先 IP アドレス

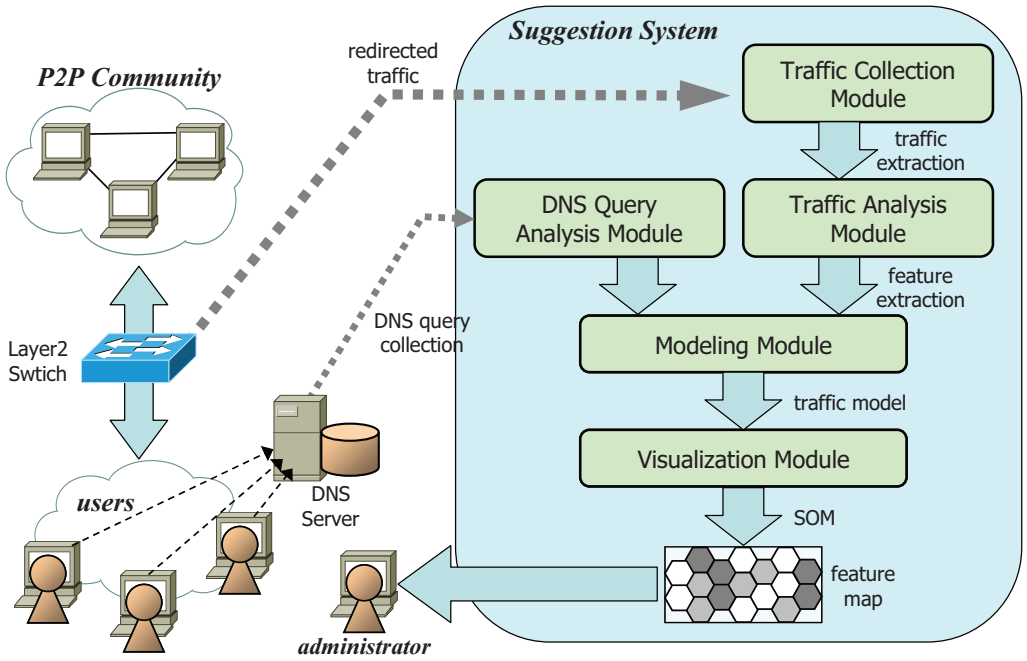

図 4　システム構成

ごとに, パケット出現回数とパケットサイズを集積する.

また，モデル化部では，この特徴ベクトルに対する重み付けを行う．Share などの P2P プログラムやストリーミングプログラムではパケット送信時にPUSH フラグが設定される．このため，PUSH フラグが設定されたパケットはP2P やストリーミングによるトラフィックである可能性が高い．さらに，P2P アプリケーションは，接続する相手ノードの IP アドレスを直接指定しコネクションを生成する．このため DNS による名前解決が発生しない．そこで，上の 2 種類の特徴に合致するデータに対して，その値を一定割合で増やすことで重み付けをおこなう.

なお, 我々は, 要素ベクトルを, 宛先 IP だけではなく, 宛先 IP アドレスとポート番号の組み合わせで, 出現量を収集することも検討した．より詳細なモデルを作成することで，結果の精度も上がると予測したためである．また，システム的には，IP アドレス，IP アドレス＋ポート番号のどちらでも，要素ベクトルを作成することが可能である．しかし，予備的な実験をおこなった結果，IP アドレス＋ポート番号による出現量で作成した要素ベクトルも用いても，SOM の精度は，期待したほど上がらないことが分かった．さらに，IP アドレス＋ポート番号を用いた場合，計算量も増加する．そのために，システム全体の効率を考えると, 宛先 IP アドレスのみで要素ベクトルと構成した方が良いと判断した．以下の実験においても，要素ベクトルは，宛先 IP アドレスを基に作成した.

5.5. 可視化部

得られた特徴ベクトルに対し, SOM アルゴリズムを適用する. SOM により抽出されたパケット群が自己組織化され，似た特性を有する特徴ベクトルが集約される．SOM は，結果を 2 次元のベクトルを生成することでき，これを 2 次元平面上に写像したものが特徴マップと呼ばれ，データを可視化したものとして利用される．特徴マップにはトラフィックの全体状態が表現される．その中で，ある同一の特徴を持ったベクトルの集合が，マップ上にクラスタとして表出する．管理者は特徴マップを眺め，マップ上のクラスタに注目することで，管理対象のネットワーク上の何らかの特異トラフィックの存在に気付くことができる.

6. 実験と考察試作システムの概要

6.1. 実験環境

我々はトラフィックマイニングツールの試作システムを作成した．試作システムを用いて，管理者の支援が可能かどうか評価を試みた．具体的には，試作システムに対し, 実験データを入力し特徴マップ生成をおこなった．生成された特徴マップに対する評価や，特徴マップ生成に掛かる時間等を調べた．表 3 に実験環境を示す.

表 3　実験環境

CPU	Intel Pentium4 3.2GHz
Memory	1 GBytes
HD	300 GBytes
OS	Linux (Kernel 2.4.18)

表 4　実験データ件数

種　別	件　数
実験データ件数	1,423,592 件
要素ベクトル生成数	16,356 件

ある組織の協力を得て，2006 年 11 月 20 日に，1 時間分の組織のトラフィック，全 IP パケットを収集し，これを実験データとした．なお，実験中 1 台の端末において意図的に Share を動作させ，任意のデータファイルをダウンロードした．表 4 に実験データ件数および処理過程で生成された要素ベクトル数を示す.

今回の実験で用意したデータには，P2P のトラフィックが潜んでいる．そのために，生成された特徴マップにおいて P2P のトラフィックが，管理者が注目できるような形で表出されているかどうかが問題である. この点が，特徴マップの性能を図る一つの目安となると考える.

6.2. 考察

図 5 に今回の実験で生成した特徴マップを示す．前述の通り IP アドレスごとの通信量および PUSH フラグ出現率，および，DNS による名前解決情報を重みとして付加した後，可視化した結果を示している．また，処理結果において特徴を持つノードにラベルを表記している.

この特徴マップは 20×16 の 320 ノードを持つ．それぞれのノードは，要素ベクトルに対応している．今回の実験で生成された要素ベクトルの総数は16,356 件であるため，約 2%の要素ベクトルが現れていることになる．大規模な通信をおこなっている IP アドレスが, 現れているといえる．特に広範囲な宛先 IP，および宛先ポートに対して通信をおこなっているような送信元 IP を持つ要素ベクトルは，自己組織化されクラスタとして表出している.

実験にあたり, 見かけ上単一 IP アドレスを持ち対外的に複数の宛先 IP アドレスと通信をおこなう NATBOX のような機器のトラフィックも，特徴マップでクラスタとして表出されることが予測された．そうした場合，管理者は特徴マップ上のクラスタが P2P なのか，NATBOX

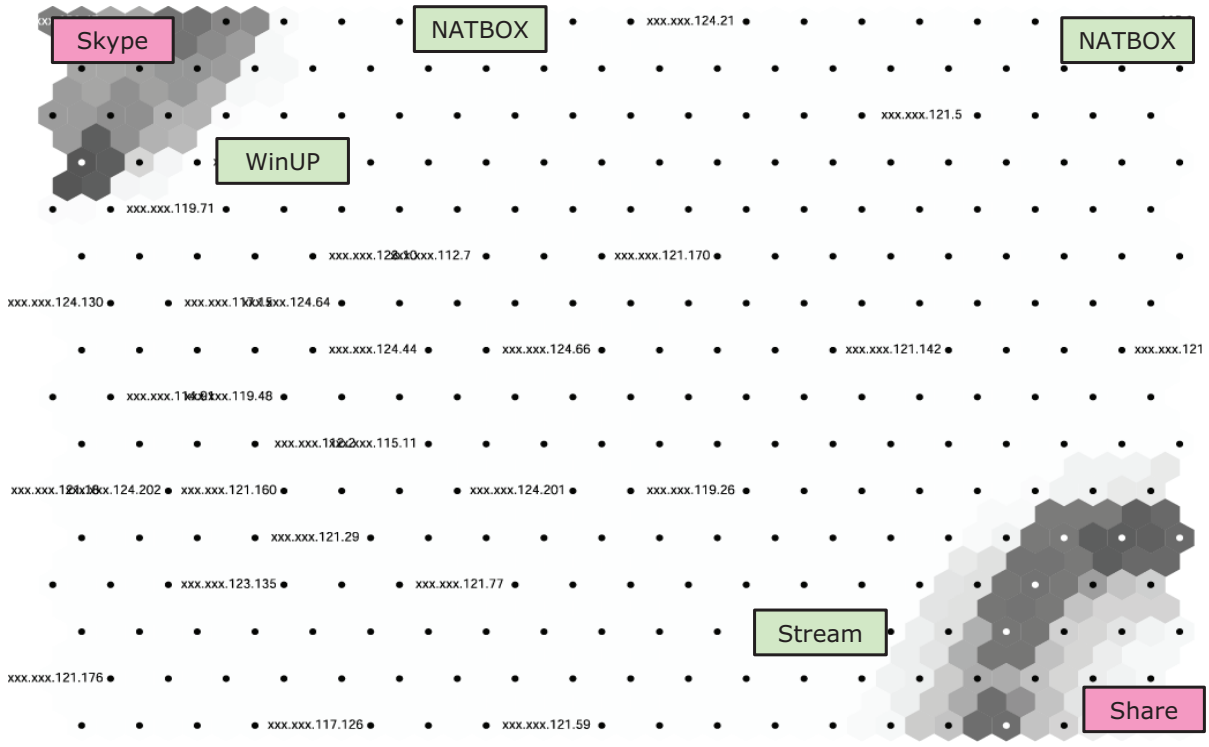

図 5　実験結果からの特徴マップ

のものなのか，一々確認しなければならなくなる．図 5 において，NATBOX はクラスタとなっておらず，他の一般の IP アドレスと同様に認識されている．

一方，意図的に発生させた Share のトラフィックは，クラスタとして表出していることが分かる．さらに，実験中，偶然，ある利用者が使用していた Skype のトラフィックもクラスタとして表出していることが分かる．今回生成した特徴マップは，DNS による名前解決情報を重みとして加えたものである．P2P クライアントは相手ノードに対し直接 IP アドレスを指定してコネクションを生成するため，DNS による名前解決が発生しない．この特徴を収集し，重み付けに利用することで Share および Skype に関係するトラフィックを，明確なクラスタとして表出させることができることが確かめられた．

Skype も，現段階において，特徴マップ上にクラスタとして表出する可能性が高いことが分かった．このことから，我々の提案する手法において，Winny 以外の P2P の通信一般も，表出可能であることが分かった．

なお，我々が提案しているのは，このような形で，トラフィック状態を可視化した俯瞰図の提供である．先に述べた通り，本図を提供することで，管理者によるトラフィック上の異常検知に，役だつと考えるからである．無論，図の解釈は主観的な問題であり，個人差が存在する．また，特徴マップを見て意味を読み取る，といった訓練が必要である．しかし，管理者の主観的な評価の問題は存在するものの，今回の結果からは，特徴マップの表現力には，一定の成果があったといえる．筆者らは，それぞれキャンパスネットワークの管理に係わっており，今回得られた特徴マップを見たとき，P2P トラフィックの存在について理解することができた．このことは，特徴マップに一定の有効性があることを示していると考える．今回の結果から，我々の提案するトラフィックマイニングツールについて，一定の可能性を示したと考える．

7. まとめ

本稿では，ネットワークの管理者が，ネットワークの異常を検知するトラフィックマイニングと，トラフィックマイニングツールについて述べた．我々は，管理者の異常検知のタスクを支援することを目指している．

我々は，ネットワークの異常検知において，P2P トラフィックの発見に着目した．まず，キャンパスネットワークでの P2P 通信の問題について述べた．これら P2P トラフィックを，既存のフィルタリング技術で制限することは困難であることについて述べた．その上でキャンパスネットワーク内から受発信される P2P トラフィックを検出する手法を検討し，管理者がおこなう P2P トラフィック検出支援のために，トラフィックマイニングと可視化による監視支援モデルについて述べた．さらに支援モデルを実現するために必要なトラフィックのモデル化手法について述べ，トラフィック傾向の俯瞰を可能にするための可視化手法について述べた．

また，本提案の有効性を検証するために実装した試作

システムについて述べ，実証実験の結果である特徴マップを示し考察をおこなった.

今回の実験において，P2P ファイル共有のトラフィックのみを分離し可視化することはできなかったものの，P2P トラフィックと非 P2P トラフィックとを比較的明確に分離表示することができた. むしろ, P2P については，クラスタリングが可能であることが分かった. その上で，管理者に提示する情報として，一定の有効性が認められると考える．今後は重み付け手法の改良などにより，特徴マップ上での P2P ファイル共有トラフィックの明確な提示を試みる.

謝　辞

本研究の一部は, 平成 21 年度科学研究費補助金基盤研究（B）（課題番号 19300283）の支援を受け行われたものである.

参考文献

[1] WinMX Web Site, http://www.winmx.com/

[2] 金子勇，“Winny の技術”，アスキー，2005.

[3] BitTorrent Web Site, http://bittorrent.com/

[4] Perfect Dark@ウィキ, http://www21.atwiki.jp/botubotubotubotu/

[5] Internet Watch，“海上自衛隊の「秘」情報が Winny で流出”，http://internet.watch.impress.co.jp/cda/news/2006/02/23/10993.html

[6] 毎日新聞，“愛媛県警：4400 人分の情報流出かウィニー介して”，http://www.mainichi-msn.co.jp/shakai/jiken/news/20060320k0000m040096000c.html

[7] NetFlow, http://www.cisco.com/warp/public/732/Tech/nmp/netflow/index.shtml

[8] sFlow.org, http://www.sflow.org/

[9] 石川博，“次世代データベースとデータマイニング”，CQ 出版社, 2005.

[10] 藤井聖，中村豊，藤川和利，砂原秀樹，“通信先ホスト数の変化に注目した異常トラヒック自動検出手法の提案と評価”，信学論，Vol.J88‐B，No.10，pp.1922‐1933，2005.

[11] 佐藤友暁，伊丸岡修哉，深瀬政秋，“キャンパスネットワークにおける Winny 検知手法の FPGA 実装”，学術情報処理研究，No.12，pp.68-76，2008.

[12] Kohonen, T. “Self-Organizing Maps, 3rd ed.”, Springer, Heidelberg，2001.

[13] Yoshida, k., Katuno, S., Ano, S., Yamazaki, K., Tsuru, M. “Stream Mining for Network Management”，IEICE Trans. Communication E89-B(6)，pp.1774-1780，2006.