2005－DSM－36 (15)
2005／3／18

トラフィック特徴抽出と可視化による P2P ファイル共有通信検出支援システムの構築

戸川 聡 * 金西計英 ** 矢野米雄 ***

* 徳島大学大学院工学研究科
** 徳島大学高度情報化基盤センター
*** 徳島大学工学部

概要：Peer-to-Peer（P2P）型ファイル共有通信を利用した著作権侵害が問題となっている．インターネット上のファイル共有コミュニティで流通する大半のデータは，音楽 CD や DVD から抽出された著作物であり，ほとんどは著作権者の許諾を得ないまま共有されている．また，企業や大学のキャンパスネットワークでは，P2P ファイル共有通信そのものを禁じている場合がある．しかし既存のポートフィルタリングなどでは P2P ファイル共有通信を制限することは難しい．その結果，管理者はトラフィックを常時監視し，P2P ファイル共有通信を検出しなければならない．本稿では，ネットワーク管理者がおこなう P2P ファイル共有通信の検出作業を支援するためのシステムを構築した．そして，実際に P2P ファイル共有プログラムが発するトラフィックを可視化し，実験をおこない，有効性を検証した．

Peer-to-Peer File Sharing Detection Assistance System Using the Traffic Activity Extraction and Visualization

Satoshi Togawa*, Kazuhide Kanenishi** and Yoneo Yano***

*Graduate School of Engineering, University of Tokushima
**Center for Advanced Information Technology, University of Tokushima
***Faculty of Engineering, University of Tokushima

Abstract: In this research, we have proposed the assistance system for peer-to-peer traffic detection. Recently, an illegal file has been exchanged with peer-to-peer file exchange software. These files are extracted from music CD and DVD. Most files do not obtain the copyright person's approval and are open to the public. Neither enterprise nor the Campus Network user of the university must acquire these files from the problem in morality. However, the illegal file is actually acquired via Campus Network. The network administrator should observe the users' peer-to-peer communication. In this paper, first of all, We explain a problem of peer-to-peer file sharing system. Next, we explain the assistance system for peer-to-peer file sharing traffic detection. Finally, we conclude it.

1 はじめに

Peer-to-Peer（P2P）通信によるファイル共有が問題となっている．これを実現するソフトウエアとして，WinMX[1] や Winny[2]，BitTorrent[3]，Share などが存在する．一般的にこれらのソフトウエアは，公開対象としたディレクトリ中に存在するファイル群をインターネット上の不特定多数の利用者に公開する．公開可能なファイルの種別に技術的制約はない．コンピュータシステム上にファイルとして存在可能であれば，そのデータ内容に関わらず公開が可能となる．

P2P 共有ソフトウエアを用いて，ファイルをインターネット上に公開する行為そのものに違法性はない．しかし，音楽 CD から抽出した楽曲データや DVD から抽出した動画データなど，著作権法で保護される著作物をデータファイル化し，著作権者の許諾を得ず公開，共有することはできない．これらの行為は，著作権法で規定される公衆送信権，送信可能化権の侵害にあたる．

P2P ファイル共有の利用を制限する場合，データ転送元となるホストが不特定多数かつ可変であることから，IP アドレスベースの通信フィルタリングは効果的ではない．さらに Winny や Share は標準的な待ち受けポート番号を持たず，ランダムに設定された TCP ポートにて接続を待ち受ける．したがっ

てポート番号ベースのフィルタリングも利用制限に効果的ではない．

このため既存のフィルタ技術ではP2Pファイル共有の制限は困難である．したがって，ネットワーク管理者がP2Pファイル共有の制限を試みる場合，管理するネットワーク内のP2Pノードから受発信されるP2Pトラフィックを検出し，個別に対応しなければならない．

現在，P2Pファイル共有通信検出のためにSnort[4]などの侵入検知システム（Intrusion Detection System:IDS）を使用できる．本来Snortは不正侵入検知のために構築されたシステムだが，ルールを記述することでP2Pトラフィック検出に応用できる．しかし，ルール記述文法が複雑であることから管理者への負担が大きい．また，パターンに適合しないとP2Pトラフィックとして管理者に通知しないため，取りこぼしなどの誤検出は避けられない．

そこで本研究では，トラフィック特徴抽出と可視化によるP2Pファイル共有通信検出支援システムを提案する．監視対象ネットワークから送受信されるトラフィックをもとに特徴量を抽出しモデル化する．生成されたモデルを可視化し特徴マップを生成する．管理者は特徴マップを参照することで，定常のトラフィック傾向の俯瞰が可能となる．定常状態の把握により，低頻度で発生する特異状態に気づきやすくなる．この結果，ログ情報調査時に「あたり」をつけやすくなり，調査負担を軽減できる．

以下本稿では，2章でP2Pファイル共有通信の現状について述べ，3章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章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章で試作システムの概要を述べ，その後実証実験の概要と考察を述べる．

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形態

現在主流のP2Pファイル共有通信は，以下の2つに分類できる．

Hybrid型： 図1にHybrid型P2P通信の例を示す．リソース探索機能およびノード探索機能は中央サーバに依存し，リソース交換はノード間にて行う方式である．中央サーバに蓄積した索引にて検索を行うため，高速なリソース検索が可能となる．ノードとなるクライアントは，中央サーバから示された相手ノードとファイル交換を行う．

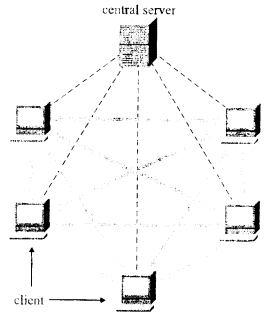

図1: Hybrid型P2P通信モデル

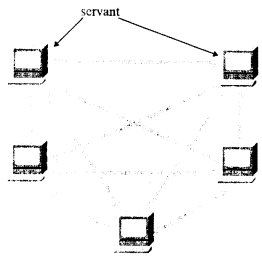

図2: Pure型P2P通信モデル

Pure型： 図2にPure型P2P通信の例を示す．Pure型はリソース検索機能やノード探索機能を管理する中央サーバを持たない．これらの機能は各ノードに実装される．あるノードがリソースを探索する場合，対等に存在する他ノードの連携により探索が実現される．各ノードはサーバ機能，クライアント機能の両方を実装するため，サーバントと呼ばれる．

Hybrid型P2Pファイル共有を実装している代表的ソフトウエアとしてWinMXが存在する．また，Pure型P2Pファイル共有を実装する代表的ソフトウエアとしてWinny，Shareが存在する．

2.2 Pure型P2Pファイル共有通信の特徴

現在代表的になりつつあるPure型P2Pファイル共有ソフトウエア「Share」を例に，そのトラフィック特徴を述べる．Pure型P2Pファイル共有通信の発生時からファイル取得までを4フェーズに分け，それぞれの特徴を述べる．

2.2.1 コミュニティ参加フェーズ

インターネット上に存在するP2Pコミュニティに自分ノードを参加させるフェーズである．P2Pコ

ミュニティに参加するためには，まず初期ノードリストを入手する．初期ノードリストには既にコミュニティに参加しているホストの IP アドレスとポート番号が暗号化され記載される．宛先ポート番号には標準値が存在しない．Share の待ち受けポートはソフトウエア導入時にランダム選択される．この結果，初期ノードリストに記載される宛先 IP アドレスと宛先ポート番号に一貫性は存在しない．

コミュニティに参加しようとする自ノードは，事前取得した初期ノードリストのうち数ノードを対象にコネクションを生成する．前述の理由より，宛先 IP アドレスおよび宛先ポート番号には一貫性がない．このため表層的には多数のランダムな宛先に対しコネクションが生成される．初期値設定にも左右されるが，初期コネクションから 2〜4 ノードを選択し上流ノードに設定する．

2.2.2 待機フェーズ

上流ノードが選択された後，待機フェーズとなる．待機フェーズではおおむね 2〜4 程度の上流ノード，および同数程度の下流ノードとリンクを確立し，定常状態に入る．この間，自ノードを中心とした宛先ノードの増減は少ない．しかしリンクを確立した上流ノード，下流ノードとの間で継続的にリンク情報が交換される．しかし下流ノードからのリソース検索要求が不定期に依頼されるため，自ノードへのコネクションが増減する．

2.2.3 検索フェーズ

取得するリソースを探索するため，検索語を入力し検索する．この段階では既に上流，下流の検索リンクが確立しているため，接続コネクションの大きな変動はない．

2.2.4 取得フェーズ

検索結果として示されたリソース一覧から，ファイル取得を行う．Winny は 1 ヶ所のノードからファイル転送を行うのではなく，コミュニティ全体から複数選択された転送元から分散的にファイル転送を行う．したがって取得フェーズにおいては接続コネクションの増加が観測できる．同時に転送元に選択された相手ノードから，自ノードに対し継続的なデータ転送が行われる．

2.3 フィルタリングによる利用制限の検討

ネットワーク利用者への通信制限実現のために，ポートベースや IP アドレスベースのフィルタリング技術によるアクセス制限手法が存在する．本節では既存技術であるフィルタリングを用いた P2P ファイル共有通信の制限を検討する．

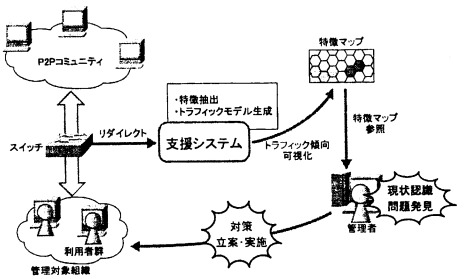

図 3: 検出支援モデル

Hybrid 型 P2P ファイル共有通信は中央サーバが単一障害点となるため，中央サーバへの経路を遮断すれば理論上容易にリソース検索機能を遮断できる．このため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は一定の効果が期待できる．しかし現実には複数の中央サーバが運用されているため，これらをすべて網羅することは難しい．

Pure 型 P2P ファイル共有通信の制限は，さらに困難をともなう．Pure 型 P2P ファイル共有利用者は，インターネット上に無数に存在する P2P サーバントのうち，いずれか 1 つに接続できれば P2P ファイル共有コミュニティに参加できる．さらに Freenet や Winny では，待ち受けポート番号がランダムに生成されるため，既存のポートフィルタ技術は適用できない．

3 検出支援モデル

2 章では P2P ファイル共有通信の現状について述べた．P2P ファイル共有通信形態を Hybrid 型，Pure 型にそれぞれ分類し，特に今後主流になりつつある Pure 型 P2P ファイル共有通信につき，トラフィックを表層現象としてとらえた場合の特徴を述べた．本章では，P2P トラフィック検出支援モデルについて述べる．

3.1 検出支援の枠組み

2.3 で述べたように，P2P ファイル共有通信の利用制限に既存技術であるポートフィルタリングを用いることは難しい．特に Pure 型 P2P トラフィックの特性であるランダムな宛先ポート選択が，フィル

タリング技術の適用を困難にしている．

しかし見方を変えれば，P2P ファイル共有トラフィックは広範囲な宛先 IP アドレスと宛先ポートを対象に通信を行っていることがわかる．従来のクライアント- サーバ型通信のように，コネクション確立後に同一宛先ポート間にて通信を行うのではない．頻繁に宛先 IP アドレス，宛先ポートを変更し，自ノードに接続を試みる他ノードも頻繁に発生することがわかる．

監視対象トラフィックの送信元，宛先の各 IP アドレス，ポート番号および発生頻度とそれぞれの間の通信量を抽出し可視化すれば，管理者は監視対象のトラフィック傾向を俯瞰可能となる．通常は点ないし線として表出する特徴マップ上に，一定の平面を占める部分が表出すれば，通常とは異なるトラフィックが発生したことを示唆できる．

図 3 に本研究で提案する検出支援モデルを示す．利用者群が行う通常の通信は，キャンパスネットワーク内各所に設置される Layer2 スイッチを経由する．スイッチに実装されるポートミラーリング機能により，監視対象トラフィックを本研究で提案する支援システムにリダイレクトする．支援システムは，取得したトラフィックから特徴量を抽出しモデル化する．本研究では，これをトラフィックモデルと呼ぶ．トラフィックモデルには，送信元，宛先の IP アドレス，ポート番号をインデックスとして，コネクション生成数と単位時間当りのデータ転送量を集積する．

生成されたトラフィックモデルを特徴マップとして可視化する．管理者は特徴マップを参照することで，管理対象組織における定常のトラフィック傾向が俯瞰でき，定常状態を把握することで低頻度で発生する異常事象の発見が可能となる．

4 モデル構成と可視化

4.1 トラフィックモデルの構成

本研究で用いるトラフィックモデルの構成について述べる．

特徴ベクトルは送信元 IP アドレスごとに生成される．宛先 IP アドレス・ポート番号をインデックスとして，単位時間あたりのデータ転送量，コネクション生成数を特徴量として保持する．

n を送信元 IP アドレス，m を宛先 IP アドレスとすると，トラフィックモデルは次式で表現できる．

$$\boldsymbol{A'} = \begin{pmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} & \cdots & \boldsymbol{A_{1n}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{22}} & \cdots & \boldsymbol{A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A_{m1}} & \boldsymbol{A_{m2}} & \cdots & \boldsymbol{A_{mn}} \end{pmatrix} \quad (1)$$

トラフィックモデルの各要素には，送信元ポート番号，宛先ポート番号をインデックスとして，単位時間あたりのデータ転送量，コネクション生成数が保持される．

x, y をそれぞれ送信元ポート番号，宛先ポート番号とすると，トラフィックモデルの一要素は次式で表現できる．

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B_{11}} & \boldsymbol{B_{12}} & \cdots & \boldsymbol{B_{1y}} \\ \boldsymbol{B_{21}} & \boldsymbol{B_{22}} & \cdots & \boldsymbol{B_{2y}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{B_{x1}} & \boldsymbol{B_{x2}} & \cdots & \boldsymbol{B_{xy}} \end{pmatrix} \quad (2)$$

また，格納される特徴量として，a を単位時間あたりのデータ転送量，b をコネクション生成数すると，各要素は次式で表現できる．

$$\boldsymbol{B} = \{a, b\} \quad (3)$$

この結果トラフィックモデルは，送信元 IP アドレスごとに割当てられる特徴ベクトルを集合させた多次元ベクトル集合であると言える．

4.2 自己組織化マップによる可視化

トラフィックモデルは 4.1 で述べたとおり多次元のベクトル集合である．一般に人間が直感的に認識できる次元空間は三次元までである．このため，管理者にトラフィックモデルをそのまま提示しても，人間の空間認識能力をはるかに超えるため，直感的な認知が難しい．

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．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 試作システムの概要

本章では，実証実験のために構築した試作システムについて述べる．図 4 に試作システムの構成を

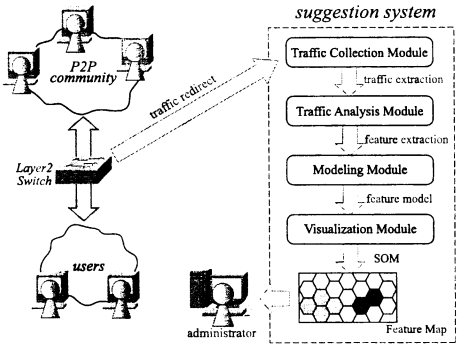

図 4: システム構成

表 1: 実験環境

CPU	Intel Pentium4 2.4GHz
Memory	640 Mbytes
HD	40 Gbytes
OS	Linux (kernel 2.4.18)

表 2: 実験データ件数

種別	件数
実験データ件数	5,091,953 件
特徴ベクトル生成数	149,694 件

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」「モデル化部」「可視化部」から構成される．以下，各部の概要を述べる．

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．L2 スイッチのポートミラーリング機能により，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

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 モデル化部

ベクトル空間モデルで定義されるトラフィックモデルを生成する．送信元 IP1 つに対し，宛先 IP・ポート番号数が次元となる多次元ベクトルを生成する．本稿ではこれを特徴ベクトルとよぶ．モデル全体では特徴ベクトルが全送信元 IP 数分集積されたベクトル集合となる．ベクトルの各要素には宛先 IP アドレス・宛先ポート番号別にパケット出現回数とパケットサイズを集積する．

また，モデル化部では重み付けをおこなう．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 フラグを設定する．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が高いと判断し，トラフィックモデル中での存在を強化しておく．

5.4 可視化部

得られたトラフィックモデルを SOM アルゴリズムを用いて可視化する．SOM アルゴリズム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た特徴マップが生成される．PUSH フラグが設定されたパケットなど，特に特徴を持つ特徴ベクトルはクラスタとして表出する．このため管理者に対し，管理対象ネットワークに発生した特異トラフィックへの気づきを支援できる．

6 実験と考察

6.1 実験環境

試作システムに実験データを入力し特徴マップ生成をおこなった．表 1 に実験環境を示す．

ある組織に許可を得て，2005 年 2 月 1 日にその組織内の端末が受発信したすべての IP パケットを収集し，実験データとした．なお，実験期間中 1 台の端末にて意図的に「Share」を動作させ，適当なデータファイルをダウンロードした．表 2 に実験データ件数および処理過程で生成された特徴ベクトル数を示す．

6.2 考察

図 5 に実験で生成した特徴マップを示す．

1 つの特徴マップは 20 × 16 の 320 要素を持つ．それぞれの要素には比較的多く出現した特徴ベクトルが表出する．今回の実験で生成された特徴ベクトル総数は 149,694 件であるため，約 0.2%の大規模通信が表出することになる．

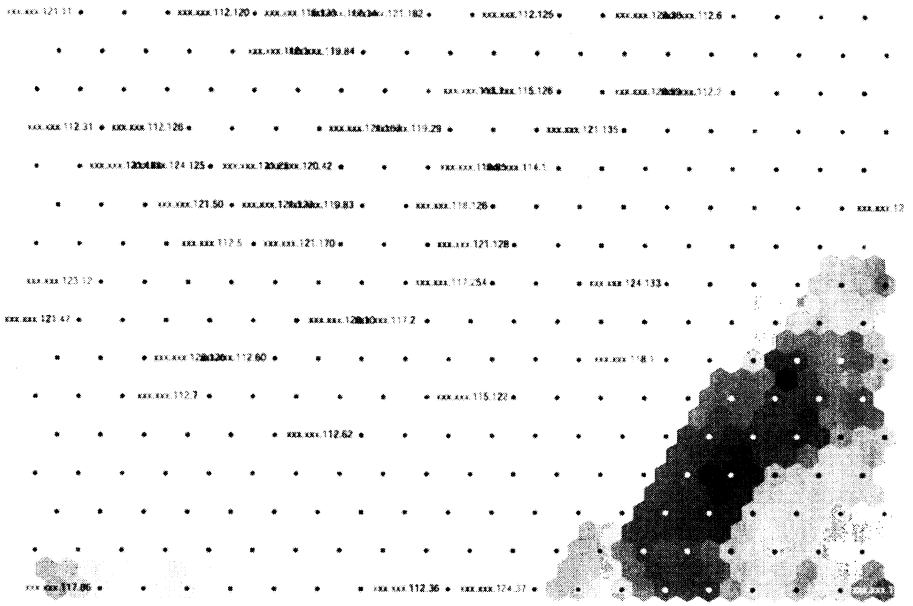

図 5: 特徴マップ

特に広範囲の宛先 IP および宛先ポートに対して通信をおこなっている送信元 IP のベクトルは自己組織化されクラスタとして表出されている．実験的に P2P ツール「Share」を動作させたホストのトラフィックに関して，マップ右下に大規模クラスタとして表出していることが確認できる．P2P など，PUSH フラグを設定されているトラフィックに関して，重み付けによりベクトル自体が有する特徴量を強調しているため，より明確なクラスタとして表出されている．実際当該クラスタを構成する宛先 IP への TCP フラグを確認したところ，過半数以上の TCP セグメントにおいて PUSH フラグが設定されていた．

7 まとめ

本稿では，企業や大学のキャンパスネットワークで行われる P2P ファイル共有通信の問題について述べ，これらの 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の困難性について述べた．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 トラフィック検出のための支援モデルを提案した．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，多次元モデルの認識限界を下げトラフィック傾向の俯瞰を可能にするために行う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 トラフィックのより明確な提示を試みる．

参考文献

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

[2] Winny Web Site, http://www.geocities.co.jp/SiliconValley/2949/

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

[4] Snort Web Site, http://www.snort.org/

[5] 藤井聖, 中尾嘉宏, 中村豊, 藤川和利, 砂原秀樹, “フローを用いた特定トラフィック検出システムの運用”, 第 31 回分散システム/インターネット運用技術研究会, 2003.