L-019

ネットワークの長期観測のための高効率トラフィック情報格納法

Efficient Traffic-data Storing Method for Long-term Observation of Network

有馬 亮[1)]　佐藤 彰洋[1)]　笹井 一人[2)]　北形 元[2)]　木下 哲男[3)]

Ryo Arima　Akihiro Satoh　Kazuto Sasai　Gen Kitagata　Tetsuo Kinoshita

1. はじめに

複雑化，大規模化したネットワークを効率良く管理するためには，ネットワークの出入り口を流れるトラフィックを保存し，活用する方法が有効である．しかしながら，単位時間にネットワークの出入り口を流れるトラフィックは膨大であり，また過去のデータを参照することもあるため，長期間に及ぶ大量のトラフィック情報を管理・運用する技術が必要がある．更に，これは大量のリソースを消費し，管理者にとって新たな問題となっている．

そこで本稿では，大規模なトラフィック情報の管理・運用を支援することを目的とし，効率良くトラフィックを保存する蓄積過程と，必要な情報を高速に抽出する手法を提案する．2 章では関連技術について述べ，3 章で提案手法の詳細を説明する．そして最後に，まとめと今後について述べる．

2. 関連技術

2.1 トラフィック収集の蓄積過程

図 1 に，トラフィック収集システムの例を示す．この図では，対外接続ルータにトラフィック収集システムが接続しており，トラフィック収集システムはネットワークの出入口，すなわち内部（建物側）とインターネット間を流れるトラフィックをコピーし，保存している．同様の構成である東北大学電気通信研究所では，トラフィック・ログファイルを 3 ヶ月分保存すると，そのデータサイズは 1TB に達し[1]，大量のリソースを消費している．これを解決するためには，トラフィックの保存により適した圧縮法を用いて，効率的に保存することが必要とされる．

2.2 解析に必要な情報の抽出過程

管理者が解析に必要な情報を抽出するときには，トラフィック・ログファイル全体を走査して，特定

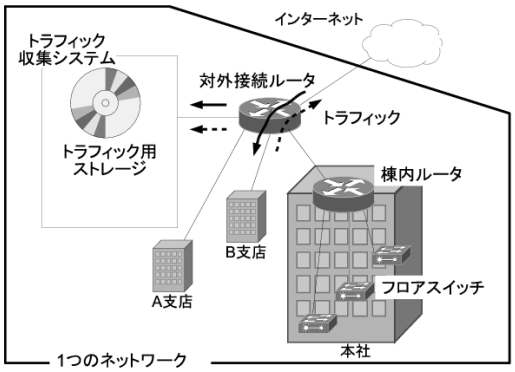

図 1　トラフィック収集システムの例

のパケットのような必要なデータを収集した後，収集したパケットから実際に解析に用いる情報を抽出するという手順が必要である．更に，効率的なトラフィックの保存について考えたとき，トラフィック・ログファイルは圧縮して保存されていることが多い．そのため，上記の手順にログファイルの復号という手順が加わると考えられる．

ところで，トラフィック・ログファイルを圧縮するときには，圧縮するログファイル全体を走査する必要がある．そのため，圧縮するときにログファイル全体を走査し，情報抽出のときに再び全体を走査することになり，無駄な処理がある．

3. フローに着目した圧縮法と効果的な抽出法の提案

3.1 蓄積過程

効率的な圧縮を実現できる上，解析に用いる情報の提供を高速に行える，フローに着目した圧縮法を試作した．

本研究では，フローを次式で表されるように” 宛先と送信元の MAC アドレス，IP アドレス，ポート

1) 東北大学大学院情報科学研究科
2) 東北大学電気通信研究所
3) 東北大学サイバーサイエンスセンター

[1] GZIP を用いて圧縮した後のサイズ．元のデータはもっと大きい．

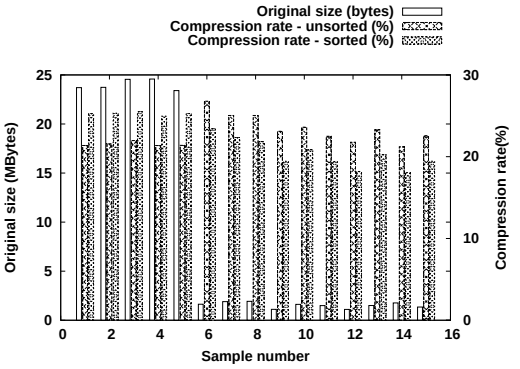

図 2　GZIP を用いた場合の圧縮率の比較

番号の全てが一致するパケットの集合” と定義する．

$$F = \{p|(p_{srcmac} = p^F_{srcmac}) \cap (p_{dstmac} = p^F_{dstmac}) \cap (p_{srcip} = p^F_{srcip}) \cap (p_{dstip} = p^F_{dstip}) \cap (p_{srcprt} = p^F_{srcprt}) \cap (p_{dstprt} = p^F_{dstprt})\}$$

式において，F はあるフローの集合，p はこれから評価するパケット，p^F はあるフローに属するパケットを表している．パケットによっては IP アドレスやポート番号の情報を持たないものがあるが，これらのデータは 0 とする．

パケットにおけるヘッダ部分は構造的であり、同一フロー内のパケット間では高い類似度が得られる。IP ヘッダを例に挙げると，送信元/宛先アドレスは IP ヘッダ全体の 1/3 を占めており，これらが一致すると，同一フロー内におけるパケット間で類似度は高くなる．また，バージョンやプロトコル番号なども一致するため，更に高い類似度が得られる．同じことがフレームヘッダ，セグメントヘッダにも当てはまるため，前述の定義に従って並べ替えを行うと，効率的な圧縮が可能なデータを作成できる。更に、構造的ではないペイロードを分離して別々に圧縮することで、更なる圧縮率の向上が可能になる。

3.2 蓄積過程の評価

蓄積過程における提案手法の有効性を確認するために，並べ替えたログファイルとそうでないものを，それぞれ GZIP（1.3.12）を使って圧縮し，圧縮率を比較した．図 2 に結果のグラフを示す．

グラフより，元のサイズが 24MB 前後の場合の圧縮率は，提案手法を用いると，用いない場合に比べて 4（%）程度悪いことがわかる．一方，元のサイズが 2MB 前後の場合は，提案手法を用いたほうが用いない場合より 3（%）程度良い事がわかる．

3.3 抽出過程

管理者がどのような動機で提案システムに情報取得を要求する動機は様々考えられる．そうすると要求も多様になり, 要求それぞれに応じた情報を高速に提供するためには，効率の良い抽出法が必要である．

そこで本研究においては, まず管理者のトラフィック情報に対する要求を，動機を含めたプロセスとしてモデル化する．このモデルに基づいたデータ構造によって，蓄積過程における圧縮法を拡張し，様々な要求に対応可能なトラフィック情報格納法を確立する．

具体的には，特定のプロトコルのパケットを参照したいという要求であれば，圧縮時にプロトコルの索引を作成しておくことにより，管理者に対して高速な情報提供を可能にする．

様々な操作を効率よく行えるようなデータ構造では，そのサイズがデータ自身のサイズを大きく上回ることがあるという問題があるが，この問題は，簡潔データ構造を用いることで解決できる可能性がある [1].

4. おわりに

本稿では，トラフィック情報の管理・運用において管理者の負担軽減を目的とし，フローに着目した圧縮法及びトラフィック情報の効果的な抽出法を提案した．更に，実際にトラフィック・ログファイルを圧縮することにより，蓄積過程における提案手法の有効性を示した．今後は，より効率的に圧縮できる手法の検討が必要である．また，抽出過程における要求のモデル化について検討する．その上で，抽出過程を含めたトラフィック情報格納システムを実装し，一般の圧縮技術との圧縮率の比較や，情報抽出にかかる時間の測定等の評価実験を行う．

参考文献

[1] 定兼邦彦，” 超簡潔データ構造”，電子情報通信学会誌，Vol.92，No.2，pp.97-104，2009