NECOMA

Hashdoop

Hash-based Anomaly detection framework on Hadoop

ハッシュによるトラフィック異常検出基盤

ビッグデータ解析基盤を目指して

Hashdoop は、国立情報学研究所の研究者が中心となって研究・開発を行っている MapReduce 技術を利用したインターネットバックボーントラフィック異常検出基盤です．ビッグデータであるバックボーン中のパケットトラフィックには，スキャン，攻撃や機器の故障によるもの等の異常なトラフィックが存在しますが，多くの異常はバックボーンでの大多数の正常な通信により埋もれています．Hashdoop では IP アドレスをキーとするハッシュ処理を用いて，トラフィックデータを複数のサブフローへ分割し Hadoop 上で処理を行うことで，従来手法に比べて最大 15 倍の高速化および 20%の精度向上が可能であることを示しています．

トラフィック異常検出

インターネットバックボーントラフィック中の異常を検出するには，パケットやフローのシグニチャを用いた決定的な手法と，確率的なモデルに基づく手法の 2 つが知られています．前者は異常パターンをシグニチャとして最初に与えることで，精度の高い検出が可能ですが，未知の異常への対応や高速なバックボーンネットワークでの処理が困難という問題があります．確率的なモデルでは，正常な状態に対応するトラフィックの特徴を学習し，その正常な状態からのずれを異常として検出します．そのため，未知の異常の検出や大規模なデータ解析に適しています．国立情報学研究所では，今までに理論的背景が異なる複数の確率モデルに基づく異常検出アルゴリズムを研究開発してきました．Hashdoop はそれらの異常検出アルゴリズムを用いて異常検出を行う異常検出フレームワークとして開発を進めています．

Hashdoop = Hash + Hadoop

ビッグデータの処理基盤である Hadoop は MapReduce モデルによって実現されています(図 1)．既存の Hadoop を用いた解析では，データの分割を時刻に基づいて行いますが Hashdoop では，ネットワークトラフィックをアドレス情報により空間的に分割する点が異なります．Hashdoop では 2 段階の MapReduce 処理(トラフィックハッシュ，異常検出)を行うことで異常検出を行います．1 段目のトラフィックハッシュでは，入力となるトラフィック(パケットデータ)中の送信もしくは受信 IP アドレスをキーとするハッシュを計算します．各々のパケットはハッシュ値によって複数のサブフローに分割されます．

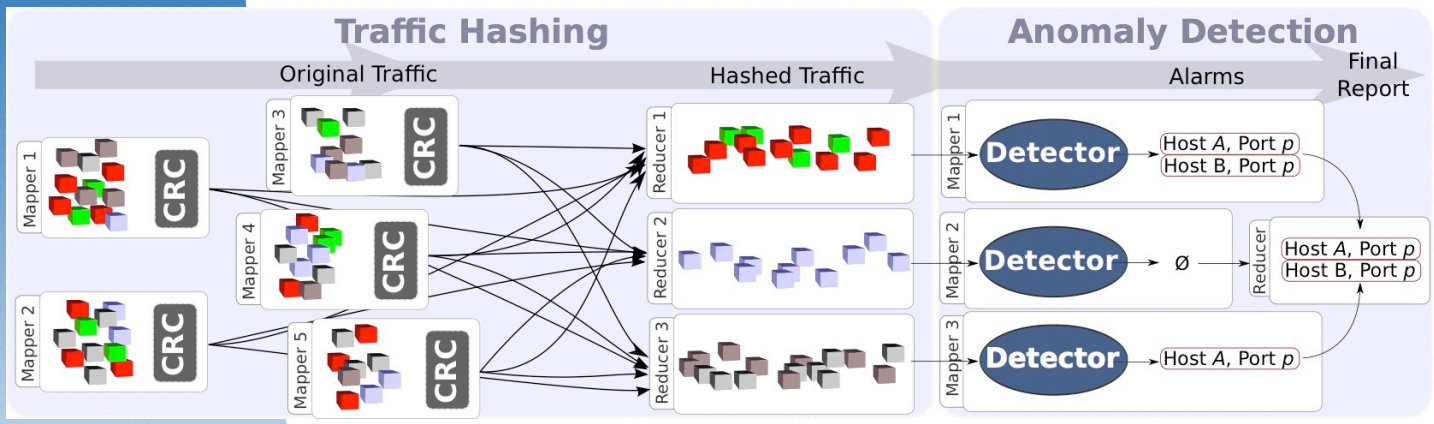


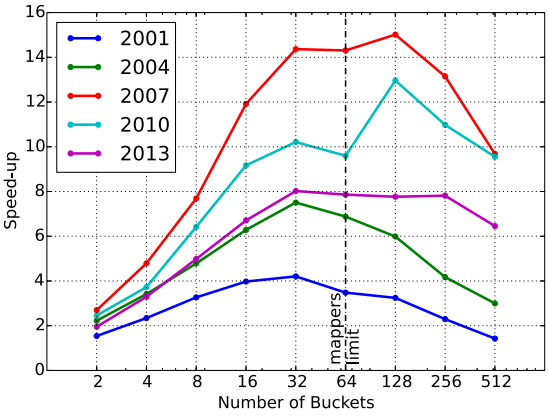
図 1 Hashdoop の構成

NECOMA

Hashdoop

Hash-based Anomaly detection framework on Hadoop

ハッシュによるトラフィック異常検出基盤



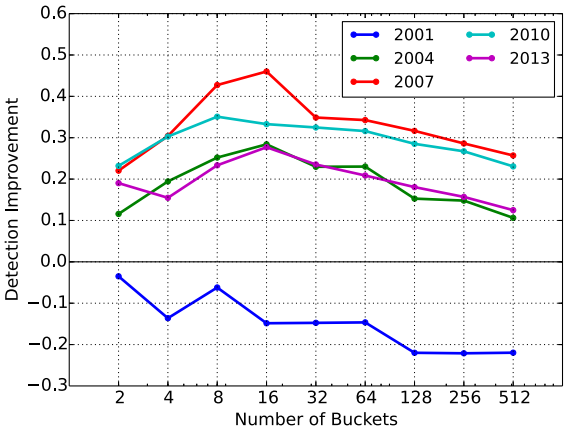


図 2 Hashdoop による性能改善：高速化(左)および精度改善(右)

MAWI データセット

WIDE プロジェクトが公開しているバックボーントラフィックデータコレクション．毎日 14:00-14:15(JST)のリンクトラフィックを pcap 形式で保存．
http://mawi.wide.ad.jp

NECOMA の研究活動ならびに成果についての詳細は、Web サイト www.necoma-project.jp をご覧ください。

サブフローは 2 段目の異常検出への入力となり，任意の異常検出アルゴリズムを適用することが可能です．ハッシュによるさらなる利点は，異常トラフィックをサブフローの一部に孤立化させ，それ以外のサブフローを正常なトラフィックとして取り扱いできる点にあります．これにより正常データからのずれを異常データとして分離することが可能となります．

Hashdoop の性能評価

図 2 は、6 台の計算機から構成される Hadoop クラスタを用いた，Hashdoop の性能評価(高速化および精度向上)を示したものです．評価用データとして 2001 年～2013 年のバックボーントラフィックデータ(MAWI データセット)を，異常検出アルゴリズムとして，パリ第 6 大学で開発された Astute アルゴリズムを使用しています．グラフの横軸はハッシュによって分割されたサブフロー数を，縦軸は 1 台の計算機を用いた場合と比較した際の性能改善を表しています．トラフィック量は年によって異なりますが，トラフィックのサブフローへの分割により，サブフローでのトラフィック量が少ない場合には Hashdoop の恩恵は少なく，トラフィック量の増加とともに性能改善の効果が大きくなります．最大では 15 倍の速度改善および 20%の精度向上という結果が得られています．速度だけでなく検出精度も向上しているのは，大規模バックボーンに潜む小さな異常をアドレス空間に基づくトラフィック分割によって効率良く発見できるハッシュの利点です．

今後の展望

NECOMA プロジェクトでは、バックボーンだけでなくエンドポイントやユーザにかかわるマルチレイヤのデータ収集を行っています．Hashdoop では現在バックボーントラフィックのみを対象としていますが，これらの多種多様なデータ分析を行う基盤へと拡張していく予定です．

GitHub にて Hashdoop のソースコードを公開しています．

https://github.com/necoma/hashdoop

【お問い合わせ先】
奈良先端科学技術大学院大学　情報科学研究科
インターネット工学研究室内
NECOMA プロジェクト事務局
Mail: fp7-necoma-pr@is.naist.jp
Web: http://www.necoma-project.jp

NECOMA
Nippon-European Cyberdefense-Oriented Multilayer Threat Analysis