様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １８ 日現在

機関番号：１２１０２
研究種目：挑戦的萌芽研究
研究期間：2013～2015
課題番号：２５５４００９３

研究課題名（和文）Big Data向メモリ管理技術とData Mining性能の関係

研究課題名（英文）Memory management for Big Data Mining

研究代表者
吉田　健一（YOSHIDA, Kenichi）

筑波大学・ビジネスサイエンス系・教授

研究者番号：４０３４４８５８

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：SPAM filter, P2Pソフト検出技術など、大規模なオンラインデータの解析技術を研究してきた研究実績をベースに、本提案では、1)メモリ管理技法とメモリ管理性能の関係、2)メモリ管理性能と各種データ分析精度の関係、の２つを明らかにし、3) 効率の良いデータ分析手法を開発する事を目的とした研究を実施した。
Ｈ２５年度には上記研究背景・目的のもと、メモリ管理性能の解析方法を検討し、オンラインでリアルタイムで動作中のマイニングプログラムのメモリ効率を計測する手法を考案し、国際会議で発表し、Ｈ２７年度は、上記メモリ管理技術を使ったマイニングシステムの高速化に成功し、国際会議で発表した。

研究成果の概要（英文）：Based on the past research experience of on-line network threat analysis, such as SPAM filtering and DDoS finding, we have studied 1) a new memory management strategy which is suitable to be used in on-line network analyzers, 2) relationship between the performance of data mining method and memory management method.
The results of this study are reported in 1) a reviewed conference paper which propose a new memory management strategy whose characteristics is the functionality to report its memory management performance, and 2) another reviewed conference paper which reports an analyzer of browser based DDoS attacks as an application of developed memory management strategy.

研究分野：データマイニング

キーワード：データマイニング　メモリ管理　Big　Data

１．研究開始当初の背景

提案者はこれまで SPAM filter, P2P ソフト検出技術、違法コピービデオ検出技術と、大規模なオンラインデータの各種解析技術を研究してきた。また、これらの研究の中で高速で大規模なデータを効率良く処理するためのメモリ管理技術を開発してきた。これらの研究の中で開発してきたメモリ管理技術 LessFU は、その他の代表的なメモリ管理技術(LRU, FIFO, 2Q, Random など)に比較して少ないメモリ使用量で効率良くデータを記憶できる。

このような研究は、例えば SPAM filter は携帯電話会社で SPAM 除去に実際に使われており、P2P ソフト検出技術もネットワーク機器メーカとプロトタイプを共同研究する等、LessFU を使ったデータ解析が実務上十分な結果を提示できる事を示してきたが、応用面を志向した研究であったため、理論的側面の検討や、更なる応用の検討など研究の余地を残していた。

２．研究の目的

上記を背景に、本研究は下記２項目を研究目的とした。

a. メモリ管理技法とメモリ管理性能の関係の解析
b. メモリ管理性能とデータ分析精度の関係の解析

昨今 Big Data の応用が注目されているが、メモリ管理とデータの分析精度の関係の研究は充分行われているとは言えない。例えば電子商取引やネットオークションにみられるような社会インフラとしてのインターネットの重要性の増加に伴い、ネット取引に関する様々なデータが蓄積され、マーケティングへの利用を目的とした分析方法の研究が進んでいる。そのような 応用を考えた場合、効率良いシステムを開発するにはキャッシュメモリに記憶された部分データを使った効率良いデータ分析手法の開発は極めて重要である。

３．研究の方法

(1) 研究の出発点として、提案者がこれまで SPAM filter, P2P ソフト検出技術、違法コピービデオ検出技術の研究の過程で開発してきた Hash2 メモリ管理方式のメモリ管理性能の分析を行った。この目的のためにメモリ管理性能をリアルタイムで計測する手法を提案し、国際会議で発表した。

(2) 上記計測手法をベースに各種アプリケーション（主として DDoS 検出などを行うデータマイニングアプロケーション）を用いてメモリ管理性能とデータ分析精度の関係を解析し、その結果、非常に効率の良い DDoS 発見手法を考案し、やはり国際会議で発表した。

４．研究成果

(1) LessFU のメモリ管理性能解析

LessFU はリアルタイムで流れてくるネットワークのストリームデータをキャッシュする仕組みとして用いた場合に、過去に出現したデータをキャッシュから捨てて初めてのデータと誤認する弱点があった。この弱点を補うために反対の（初めてのデータを既出のデータと誤認する）弱点を持つ Bloom Filter と組み合わせて用いることで、実行中のメモリ管理の性能をリアルタイムで把握する仕組みを考案した（図 1)。

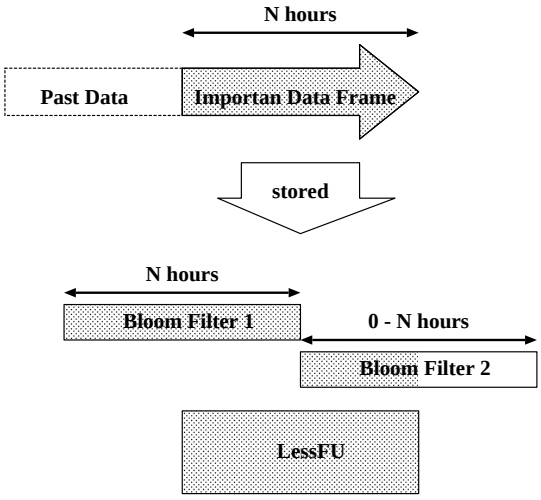

図 1: LessFU と Bloom Filter によるメモリ管理

図 2 は提案方法により LessFU により管理されたメモリのキャッシュミス率とリコール率を調べた結果である。横軸はメモリの容量であり、容量が増えるにつれ、キャッシュミス率が減少し、リコール率が増加することが見て取れる。

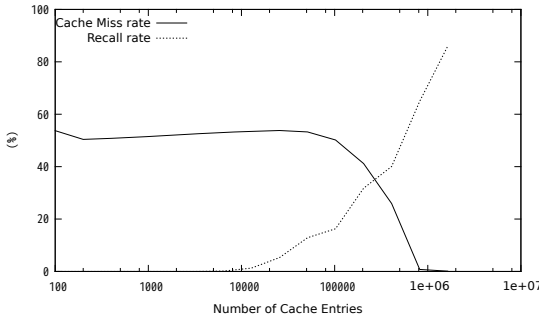

図 2: キャッシュミス率とリコール率

メモリ容量が増加するにつれてキャッシュミス率が減少し、リコール率が増加することは当然のことであり、図２に示された傾向自体は目新しいものではないが、動作中のメモリ管理性能がリアルタイムで計測できる事の利点は大きい。すなわちネットワークのデータは所謂冪則に従うだけでなくバース

ト特性を持つなど、その挙動を理論的に分析することが困難である。ミス率やリコール率を把握せずにデータ解析結果を信用する事はできないため、図 2 のような結果がリアルタイムに把握できる事は応用上の観点から大きな利点である。

(2) 効率の良い DDoS 発見手法への適用

近年ネットワークに流れる大量のデータをリアルタイムで解析するデータマイニング手法への期待が高まっている。中でも繰り返し出現する新規の network threat に対する自動防御、すなわち ZERO day attack への自動防御への期待は大きい。

本研究では前述のように提案したメモリ管理手法を用いる対象として、この ZERO day attack の自動検出を取り上げ検討した。中でも昨年大きな問題となった WWW ブラウザーを利用した DDoS 攻撃の検出を具体例に取り上げ、本研究で提案した上記メモリ管理手法の評価を行った。

図 3 はデータマイニングの手法を用いて ZERO day attack を検出するアイデアを示している。WWW ブラウザーを用いた新規の DDoS 攻撃であっても、動作として多数の攻撃者が少数の攻撃対象にパケットを送付する・少数の攻撃者が仲間を増やすために多数のクラック対象計算機にパケットを送り付ける、といった動作は共通であり、通信相手の数（「異なり数」と呼ぶ）をリアルタイムで計測する事で検出可能である。

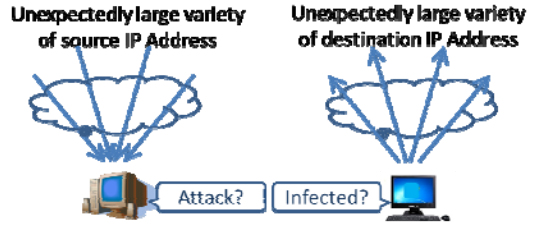

図 3: 異なり数を用いた DDoS 検出

図 4 はこのアイデアを使って実際にネットワークを流れるパケットを解析し、異なり数を調べたものである。

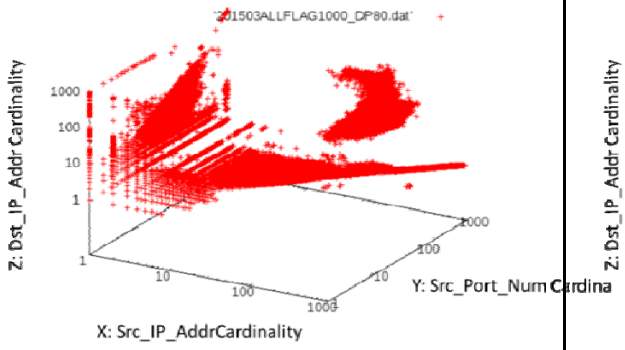

図 4:異なり数解析結果

多数の通信相手と IP パケットを交換する計算機として DDoS 攻撃に関係する計算機が検出できる。これは本研究で検討した、メモリ管理技術が実用上極めて有効な特徴を持っている事を示している。

一方この手法の欠点は、異なり数の計算負荷が高くメモリ使用量も大きく、40Gbps などの高速回線の分析には適していなかった事にある。本研究では、この処理速度の問題にも取り組んだ。

高速回線を処理するためにフィルタ処理を行ってデータ量を削減する事は素直なアイデアに思えるが単純なランダムフィルタリングなどではトラフィックデータの特徴が変わってしまい、解析結果正しくなくなる現象は良く知られている。そこで本研究では各種フィルタリング技法を用いた場合のメモリ負荷や DDoS 検出性能の計測を行い、適切なフィルタリング方法の提案を試みた。

図 5 は SYN flag を使ってフィルタリグした結果である。データ量は 1/20 まで減り高速処理は達成されるが図示されたデータ形状を見てもわかる通り解析結果が変わってしまい、DDoS 検出には適さない事がわかった。

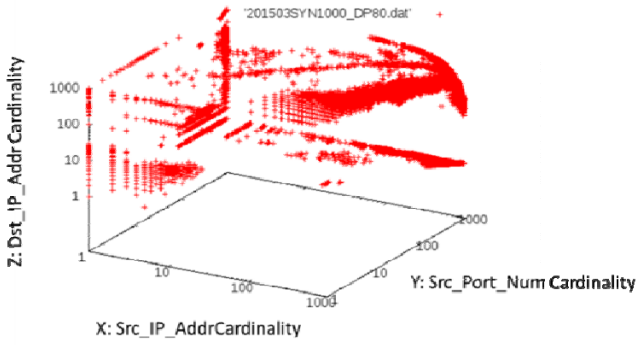

図 5:異なり数解析結果(SYN フィルタ実施)

図 6 は ACK flag を使ってフィルタしたものである。図の形状が変わらない事が示す通り DDoS 検出性能は保たれたが、データ量は削減されず、高速処理は達成されなかった。

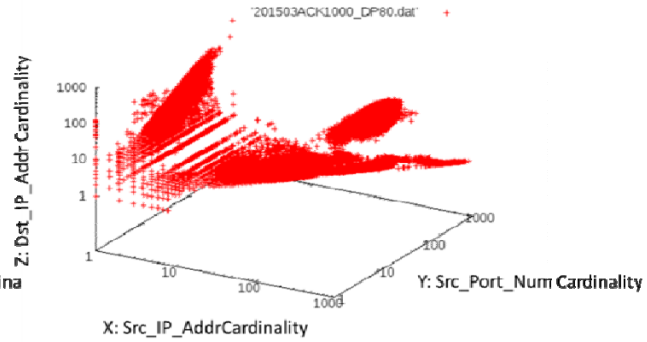

図 6:異なり数解析結果(ACK フィルタ実施)

図 7 は FIN/ACK でフィルタした結果である。データ量が大幅に削減(1/250)された結果、図は違った形状に見えるが、特徴はフィルタ処理を実施しなかった図 4 と類似しており、DDoS 検出性能の劣化も少ない、高速回線の分析に適したフィルタ手法である事がわかった。

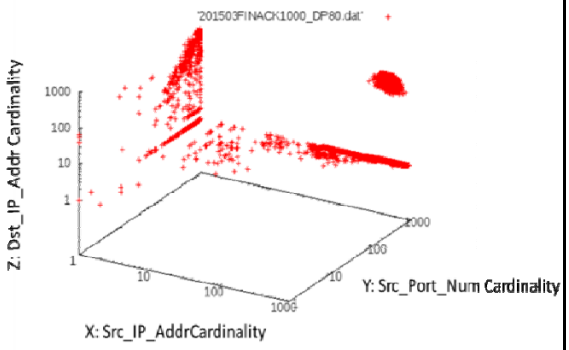

図 7:異なり数解析結果(FIN/ACK フィルタ実施)

以上の結果は WWW ブラウザーを使った DDoS 攻撃と言う近年急速に注目を集めている新種の DDoS 攻撃を例にした ZERO say 攻撃の自動発見手段(FIN/ACK でフィルタ処理した異なり数解析法)という点で実用的にも大きな結果であるだけでなく、メモリ性能や計算負荷を計測しながらデータマイニング手法のデータ分析性能を計測する事（本萌芽研究のテーマそのもの）の重要性を示しており、具体的な結果を得たことより、本萌芽研究は提案の目的にそった成果を得たと考える。

５．主な発表論文等

〔雑誌論文〕（計　4 件）

① 藤原和典、佐藤聡、吉田健一、ルート DNS サーバへのクエリ数の削減、信学論(B)，査読有， Vol.J98-B， No.6， pp.497-508 (2015)
http://www.ieice.org/cs/jpn/JB/search/summary.php?id=j98-b_6_497&category=B&year=2015&lang=J

② Yoshida Kenichi, MEMORY MANAGEMENT FOR BIG DATA MINING -- CACHE HIT RATE ESTIMATION OF LESSFU, Elsevier Procedia. Elsevier, 査読有, Vol.17, pp 114-12, (2014)
doi:10.1016/j.protcy.2014.10.214

〔学会発表〕（計　8 件）

① Mori Shinichi, Sato Akira, Yoshida Kenichi, Enhancing performance of cardinality analysis by packet filtering, Proc. of International Conference on Information Networking, pp.23-28 (2016.1.13) Kota Kinabalu (Malaysia)
DOI: 10.1109/ICOIN.2016.7427068

② 吉田健一，ネットワークデータの異なり数計測とその応用，サイエンティフィック・システム研究会，(2015.10.29)，ホテルオークラ神戸(兵庫県神戸市)，http://www.ssken.gr.jp/MAINSITE/event/2015/20151029-joint/index.html

③ 佐藤聡、小川智也、新城靖、吉田健一、筑波大学におけるハニーポットを用いた不適切な SSH アクセスの収集とその解析、インターネットと運用技術研究会 2014-IOT-25(17)， 1-6， (2014-05-15)，ホルトホール大分(大分県大分市)

④ Yoshida Kenichi, MEMORY MANAGEMENT FOR BIG DATA MINING -- CACHE HIT RATE ESTIMATION OF LESSFU, Conference on Electronics Telecommunications and Computers (2013-12-05) Lisbon(Portugal)

⑤ Takeshi Mitamura, Kenichi Yoshida, Cardinality in Big Data -- Examples in L3&L7 Network --, IA workshop 2013, (2013.10.10), Seoul(Korea), http://www.ieice.org/ken/program/index.php?mode=program&tgs_regid=db58dff9c900fa7b3d8fa250ac4fea8472ee63d676c8261483a7ee5a24534cf2&tgid=&layout=&lang=eng

６．研究組織

(1)研究代表者

筑波大学・ビジネスサイエンス系・教授
吉田 健一 (YOSHIDA, Kenichi)
研究者番号：40344858