様式 C-19

科学研究費補助金研究成果報告書

平成２３年　５月１４日現在

機関番号：１５４０１
研究種目：若手研究（Ｂ）
研究期間：２００９～２０１０
課題番号：２１７００００６０
研究課題名（和文）　統計的トラッフィク解析モデルの開発と情報セキュリティへの応用

研究課題名（英文）　Development of statistical traffic models and their application to information security

研究代表者
岡村 寛之（HIROYUKI OKAMURA）
広島大学・大学院工学研究院・准教授
研究者番号：１０３１１８１２

研究成果の概要（和文）：本研究課題では機械学習モデルとして利用されていた HMM (Hidden Markov Model) とネットワークトラッフィクの到着過程モデルとして利用されていた MAP (Markovian Arrival Process) の関連を明らかにし，それらを融合したトラッフィク解析指向の統計的学習モデルの開発を行った．また，大量のトラッフィクデータを解析する目的でデータ系列に対して並列化可能な HMM および MAP の学習（推定）アルゴリズムの開発を行った．さらに，構築したアルゴリズムの情報セキュリティに対する応用を提案した．

研究成果の概要（英文）：The project developed statistical models for network traffic analysis with bot hidden Markov models (HMMs) and Markovian arrival processes (MAPs) by revealing the mathematical relationship between HMMs and MAPs. We also considered parallelization of the parameter estimation algorithm for the proposed HMM-MAP-based models to handle a long-term observation for network packet arrivals. Moreover, we proposed the applicable example of the developed parameter estimation algorithm towards the information security.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,500,000	450,000	1,950,000
２０１０年度	1,300,000	390,000	1,690,000
年度			
年度			
年度			
総　計	2,800,000	840,000	3,640,000

研究分野：性能評価，ディペンダブルコンピューティング
科研費の分科・細目：情報学・計算機システム・ネットワーク
キーワード：隠れマルコフモデル，マルコフ型到着過程，トラッフィクモデル，EM アルゴリズム，一般化 EM アルゴリズム，変分近似，並列計算，マルチコア CPU

１．研究開始当初の背景

インターネットをとりまく環境は年々変化している．インターネットの通信速度は増加の一途をたどり，現在，多くの企業の基幹ネットワークは 10Gbps 帯域のネットワークに置き換わりつつある．一般家庭においても 100Mbps 帯域の光ファイバーを利用したブロードバンドインターネットが広く普及しており，ネットワーク上のトラッフィクは劇的に増加している．また，新しい技術やサービスの導入による変化も起こっている．IPv6 への移行や，NGN (Next Generation Network) による新しい IP 電話網の構築などが構想されており，これらの技術を導入することで IP ネットワークを流れるパケッ

トがこれまで以上に多様化すること考えられる．

ネットワークの高速化およびサービスの多様化に伴うトラッフィクの大規模化と多様化は，現在行われているトラッフィクの監視およびその解析に大きな影響を与える．その影響の一端として，ネットワークセキュリティではこれまで以上の厳格さが要求されることが予想される．例えば SIP (Session Initiation Protocol) の利用など，新たな技術導入による新たな脆弱性悪用に関する危険性，高速ネットワークで莫大な量の情報が漏洩する危険性など，高速化・多様化に伴って多くのリスクを抱えることになる．将来の IP ネットワークに対して安全・安心を確保するためには，このようなリスクを回避する技術開発が必要であり，高速化・多様化が目前に迫っている現在，そのような技術を確立することは急務である．

既存のトラフィック監視・解析技術は大きく 2 つに分類できる．1 つは，トラフィックからパケットヘッダー情報などを抜き出し，既存のデータベース（例えばウィルスデータベース）に登録してある情報とマッチング処理を行うような知識ベースの解析であり，もう 1 つは，同様な情報を統計的に処理して異常を発見する統計を利用した解析である．特に，統計による処理は未知の不正アクセスやコンピュータウィルスに対抗する手段として有効であり，実際に隠れマルコフモデル (HMM: Hidden Markov Model) やベイジアンネットワーク (BN: Bayesian Network) などの統計解析手法が学際的および実用的な観点から数多く研究されている．本研究課題では統計ベースの解析技術に注目する．

一方，従来からの待ち行列理論におけるトラッフィクモデリングではインターネットトラッフィクにおける長期依存性を表現するため，1980 年代から MAP (Markovian Arrival Process) の利用が行われている．MAP は伝統的なポアソン過程によるトラッフィク（到着過程）モデリングの自然な拡張であり，MAP を入力とした様々な待ち行列モデルに関する解析が行われている．MAP は任意の点過程を任意の精度で近似できるが，内部パラメータを多く含むため，観測されたトラッフィクデータから MAP パラメータを特定する手法として EM (Expectation-Maximization) 法が提案されている．MAP に対する EM 法は 1980 年代に議論され，現在も MAP やその拡張クラスである BMAP (Batch MAP), MMCPP (Markov-Modulated Compound Poisson Process) のパラメータ推定に対して有効な手段として研究されている．

２．研究の目的

本研究課題では，統計的処理によるトラッフィク解析に注目するが，先に記述した待ち行列解析を利用したトラッフィク理論を統合した理論を展開することにより，トラッフィク解析に最適化された監視・解析技術の確立を行うことを目的とする．

研究の全体像は，(i) 学習モデルとしての HMM および BN と，確率点過程のモデルである MAP, BMAP, MMCPP を統合した，トラッフィク解析指向の統計的学習モデル群の構築，(ii) トラッフィクデータを用いた効率的な並列学習アルゴリズムの確立，(iii) トラッフィク異常検知アルゴリズムの開発から構成される．

(i) では HMM および BN の学習モデルと MAP 系の点過程モデルの対応を行う．MAP と HMM は確率的に多くの類似点を持ちながら，歴史的に利用されるフィールドが大きく異なっていたため，HMM と MAP の関係を明らかにして，HMM を点過程のモデルとして利用することや，MAP を学習モデルと利用することは行われていない．本研究課題では，申請者自身による MAP パラメータ推定に関する先行研究を通じて，HMM, MAP, BMAP, MMCPP 間の関係について言及し，各モデルの表現能力に関する分類を行う．さらに，HMM と MAP の対応関係から，従来から議論されている HMM と BN の関係を経由して，点過程モデリングに利用可能な連続時間上の BN 表現を与える．

次に分類した学習モデル群に対して，(ii) では効率的な学習アルゴリズムの構築を行う．ここでは，最尤法に基づいた手法とベイズ法に基づいたものをそれぞれ考える．従来の最尤法に基づく学習（推定）理論では，これらのモデルに対して EM 法が適用されている．しかしながら，大量のトラッフィクデータに対する学習のスケーラビリティを確保するため，本研究課題では一般化 EM 法 (GEM: Generalized EM) の適用を考える．GEM の適用は学習アルゴリズムの並列化を実現し，大量のデータからの学習を可能とする．一方，ベイズ的アプローチでは変分ベイズを用いた高速な推定について検討する．

(i) と(ii)の成果に基づいて，トラッフィクを監視・解析し，統計的な異常を検知するアルゴリズムの開発を行う．従来と大きく異なる点は，トラッフィクの点過程としての性質とパケットのペイロードの情報を統合的に利用できる点であり，既存の MAP に対する待ち行列解析と融合することで，パケット溢れ等と統計的異常の相関など従来では行えなかった分析を行うことができる．これは，今後多様化する IP ネットワークに対して有効に機能する可能性を有していることを意味している．

３．研究の方法

(i) 学習モデルと点過程モデルの統合および分類

既存の学習モデルである HMM と MAP は多くの部分で共通点を持っている．先行研究において，連続型分布を出力とする HMM と MAP がほぼ同様な表現能力を持っていることが分かっている．一方で，HMM を連続時間上に展開したものは MMCPP と同じとなることも示されている．これらの既存の関係に基づいて，HMM と MAP 系の関連をより詳細に調査する．この分類結果から得られる利点の一つとして，HMM を使った点過程表現，あるいは MAP を使った学習モデルの構築ができることである．例えば，HMM の出力をアーラン分布にしたものは MAP あるサブクラスと等価になるが，この HMM に基づいた点過程モデルを用いることで，従来の MAP では解析不能であった，大量のトラッフィクに対する解析（フィッティング）が可能となる．また left-to-right 構造の HMM は BN で表現することが可能であるが，この関係を MAP にまで広げる．

(ii) 学習アルゴリズムの開発（最尤法）

(i) で導出した関係および各 HMM 系，MAP 系の学習モデルに対する学習アルゴリズムを再考する．最尤法に基づいた学習アルゴリズムは HMM に関しては Baum-Welch 法が広く使われている．ここでは，それをさらに発展させて，大量のトラッフィクデータを利用した学習が可能となるような高速化を行う．具体的には，一般化 EM 法（GEM）の適用を行う．GEM は未観測データを変分近似によって保管した上で EM アルゴリズムを適用する手法である．ここでの作業では，GEM を HMM 系および MAP 系の学習（推定）アルゴリズムに適用することで，従来の forward-backward アルゴリズムと呼ばれる体系から，並列化可能なアルゴリズム体系に大きく構造を変更する．学習アルゴリズムの並列性は大量のデータを扱う際に非常に強力なツールとなる．

(iii) 学習アルゴリズムの開発（ベイズ法）

(ii) の作業に引き続いて，ベイズ法の適用を行う．最尤法と異なりベイズ法はパラメータ推定値に対する事前情報を取り入れることができる．そのため，統計的異常のような希少な確率で発生する事象を取り扱う際には有効な手段である．ここでは，(ii) で発展させた GEM による学習アルゴリズムを変分ベイズによって再記述する．GEM は未観測データの一部を変分近似する手法であるのに対して，変分ベイズはパラメータの事後分布を含めて変分近似を適用する手法である．

(iv) 情報セキュリティに関する応用

(i)，(ii)，(iii) の結果を統合して，情報セキュリティに関する応用を行う．具体的には特定のコンピュータウィルスのパターンに関する調査を行い，タイプ別の分類を行う．またその分類に対して学習モデルの適用を行う．つまり，実際のトラッフィクデータを学習データとして，安全性の定量評価に対する汎化能力を数値実験により検証する．

４．研究成果

平成２１年度は主として，HMM（Hidden Markov Model，隠れマルコフモデル），MAP（Markovian Arrival Process，マルコフ型到着過程）を中心としたトラッフィクモデルの構築およびその推定手法の再検証を行った．

HMM のシンボル出力を位相型分布としたトラッフィクモデルを構築し，従来の MAP によるトラッフィクモデリングとの対応および表現能力の検証を行った．HMM の出力分布に最も単純な位相型分布であるアーラン分布を適用することで，一般的な MAP とほぼ同程度の表現能力を持ち，且つ，大量のトラッフィクデータから高速にパラメータを推定するアルゴリズムを構成できるモデルの構築に成功した．

次に，従来では，HMM および MAP のパラメータ推定には EM アルゴリズムが用いられてきたが，大量のトラッフィクデータを扱うことを考慮して，これを一般化 EM アルゴリズムの枠組みで再構成した．ここでの一般化 EM アルゴリズムは変分近似で用いられている変分原理を利用して中間的な状態を表現する手法となっている．総合的な計算量は従来の EM アルゴリズムよりも増加するが，共有メモリ型の並列計算の適用が容易な構造をもつ．実際に，マルチコアの CPU を実装したワークステーションを用いて，一般化 EM アルゴリズムを並列実装したところ，従来の逐次型の EM アルゴリズムと比べて計算時間（実測値）を大幅に減少させることができた．

平成 22 年度は HMM（Hidden Markov Model）の学習アルゴリズムの精緻化と統計的異常検知アルゴリズムへの応用を行った．平成 21 年度の成果の一つとして，一般化 EM（Expectation-Maximization）アルゴリズムによる HMM の学習アルゴリズムの再記述と，効率的な学習アルゴリズムの並列化がある．この成果をさらに応用面に関して発展させ，CHMM（Continuous HMM）の枠組みにおける一般化 EM アルゴリズムの構成，および通常の

HMM よりも多状態を表現できる階層化 HMM (Hierarchical HMM) に対する一般化 EM アルゴリズムを構成した．さらに，学習の精度向上のための工夫として，観測系列をそれぞれ独立に扱うアルゴリズム以外にも，ある程度の系列長をまとめて扱うアルゴリズムを構築した．また，CHMM の応用として，トラッフィク監視データから統計的な異常検知を行うための基礎を成す学習アルゴリズムの構築を行った．具体的には，トラッフィクモデルとしてよく利用されているマルコフ型到着過程と呼ばれる確率過程を CHMM で構成し，構築した並列化手法による実データからのパラメータ推定（学習）を行った．結果として，従来の最尤法による学習と比較して，ほぼ同じ学習効果を保つこと，効率的な並列化による計算時間の短縮（プロセッサ数８に対して８分の１の時間短縮）が実現できることを確認した．また，変分近似によるベイズ学習のアルゴリズム構築にも着手し，理論的なアルゴリズムの構成に成功した．

また推定アルゴリズムの成果の一部をホームページに掲載し，フリーツールとして配布している．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計７件）

1. H. Okamura, T. Dohi and K.S. Trivedi, A refined EM algorithm for PH distributions, Performance Evaluation, 2011（掲載決定）.
2. 山口祐介，岡村寛之，土肥正，変分ベイズによるパラメータ相関を考慮した事後分布の近似計算，日本応用数理学会論文, vol. 21, no. 1, pp. 73-88, 2011.
3. H. Okamura and T. Dohi, Estimating computer virus propagation based on Markovian arrival processes, Proceedings of 16th IEEE Pacific Rim International Symposium on Dependable Computing (PRDC-2010), pp. 199-206, 2010.
4. H. Okamura, T. Dohi and K.S. Trivedi, An improvement of EM algorithm for PH distributions with group data, Proceedings of 4th Asia-Pacific International Symposium on Advanced Reliability and Maintenance Modeling, pp. 532-539, 2010.
5. H. Okamura, H. Kishikawa and T. Dohi, Application of deterministic annealing EM algorithm to Markovian arrival parameter estimation, Proceedings of 2010 Symposia and Workshop on Ubiquitous, Autonomic and Trusted Computing, pp. 352-357, 2010.
6. Y. Yamaguchi, H. Okamura and T. Dohi, Variational Bayesian approach for estimating parameters of a mixture of Erlang distributions, Communications in Statistics -Theory and Methods, vol. 39, no. 3, pp. 2333-2350, 2010.
7. H. Okamura and T. Dohi, Faster maximum likelihood estimation algorithms for Markovian arrival processes, Proceedings of 6th International Conference on Quantitative Evaluation of Systems (QEST 2009), pp. 73-82, 2009.

〔学会発表〕（計４件）

1. 河合理恵，岡村寛之，土肥正，一般化 EM 法による連続型隠れマルコフモデルの学習アルゴリズム，平成 22 年度電気・情報関連学会中国支部第 60 回連合大会，2010 年 10 月 23 日，総社市.
2. 河合理恵，岡村寛之，土肥正，一般化 EM を用いた MAP パラメータ推定の並列化に関する一考察，2010 年度待ち行列シンポジウム「確率モデルとその応用」, 2010 年 1 月 17 日-19 日, 京都市.
3. 河合理恵，岡村寛之，土肥正，HMM の学習アルゴリズムの並列化に関する一考察, 京都大学数理解析研究所研究集会不確実・不確定性下での意思決定過程, 2009 年 11 月 12 日，京都市.
4. 河合理恵，岡村寛之，土肥正，一般化 EM 法を用いた並列処理による HMM 学習の高速化，平成 21 年度電気・情報関連学会中国支部第 60 回連合大会，2009 年 10 月 17 日，広島市.

〔その他〕
ホームページ等
mapfit:
http://www.rel.hiroshima-u.ac.jp/okamu/index.php?option=com_content&view=article&id=14&Itemid=31

６．研究組織

(1)研究代表者

岡村 寛之（HIROYUKI OKAMURA）
広島大学・大学院工学研究院・准教授
研究者番号：１０３１１８１２

(2)研究分担者
(3)連携研究者