様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １７ 日現在

機関番号：６２６１５
研究種目：若手研究
研究期間：2019～2020
課題番号：１９Ｋ２０２６２

研究課題名（和文）トラブルシューティング・予測のための大規模ネットワークシステムログからの知識抽出

研究課題名（英文）Knowledge mining of large-scale network operational data for troubleshooting and predictive analysis

研究代表者

小林　論（Kobayashi, Satoru）

国立情報学研究所・アーキテクチャ科学研究系・特任研究員

研究者番号：４０８２４１０７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では大規模ネットワークの運用における障害の原因対応支援を目的に、因果推論に基づく運用データの自動解析技術の研究を行った。運用ログデータのテンプレートをオンラインに自動生成して高速な分類を行うログ解析フレームワークamulog、およびログや計測データについての因果探索を高速に行う因果解析フレームワークlogdagを開発・公開し、運用データからより実用的に文脈的情報の自動抽出を行う一連の技術を実現した。

研究成果の学術的意義や社会的意義
ネットワーク運用データに見られるシステムの振る舞いが、データからより直感的な関係性として示されあるいは視覚化することが可能となり、ネットワーク障害の実態を素早く把握し対処する上で大きな助けになると期待される。また特にログ解析に関する知見はネットワーク以外のシステム運用分野にも応用可能であり、データドリヴンなシステム運用自動化を助ける重要な技術として幅広く活用されることが期待される。

研究成果の概要（英文）：In this research, we proposed a new analysis approach for network operational data on the basis of causal inference, to help troubleshooting of large-scale networks. We proposed two analysis frameworks: "amulog", a general log analysis framework to estimate log templates and classify messages with them in online processing, and "logdag", a causal analysis framework designed for network operational data including logs and measured SNMP data. The proposed technology will help operators to extract contextual information in operational data automatically.

研究分野：情報ネットワーク

キーワード：ネットワーク運用　データマイニング　因果推論　転移学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

現代の情報社会を支えるネットワークシステムは年々その重要性を増しており、システムの安定した継続的運用が必要不可欠となっている。システムの継続的運用には、オペレータがシステムから出力される運用データを分析することでシステムの振る舞いを把握することが必要である。しかしシステムの規模の拡大や複雑化に伴い運用データの規模と多彩さが増大しており、効率的な活用のため自動解析技術への需要が高まっている。

これまで運用データを用いた自動解析は、異常検知、異常箇所特定、原因究明といった観点から行われてきた。特に障害の原因究明を実現するためには、値の増減などのデータの表面的な変化の情報だけでなくそれらの関係や意味などシステムの振る舞いを説明することのできる文脈的情報を抽出する必要がある。既存技術ではこの文脈的情報の 1 つとして、主にイベント間の関係情報の解析を時系列的相関に基づいて行なっている。しかし相関解析では擬似相関に相当する冗長な関係情報が多数発生し、それらによりトラブルシューティング上重要な情報が埋もれてしまうということが発生する。オペレータにとってより実用性の高い情報提供を実現するためには、正しく情報を抽出するのみでなく、得られる情報がより高い情報価値、あるいは具体性をもつことが求められている。

２．研究の目的

本研究では、大規模ネットワークの運用データにおいてイベントベースの時系列因果解析による文脈的情報の取得を機械的に行うためのフレームワークの確立を目指す。これにより、ネットワークシステムのトラブルシューティングにおいてより直感的な情報の提供やその機械的な取捨選択が可能となり、システム運用の効率を大きく向上することができると期待される。

３．研究の方法

研究代表者は本研究以前より、運用データとして特にシステムログに着目し、ログの時系列イベント間の因果構造を有向グラフの形で明らかにする技術の提案を行なってきた。本研究では、(1)運用データの時系列イベントとしての性質を用いたデータ間連携、(2)長期間の運用データに対応する知識ベースの構築、の 2 つの視点からこの技術の拡張を行い、スケーラブルかつ実用性に長けた運用情報の抽出・解析技術を実現する。

(1) 運用データの時系列イベントとしての性質を用いたデータ間連携: 運用データにはシステムログ以外にも、SNMP により計測された数値データなどが存在する。これらはフォーマットや数値的特性が異なるため連携は容易ではない。現状ではデータの連携はアドホックな手法、もしくはパラメータのチューニングによって実現されることが多く、新たな種類・形式のデータを解析対象として追加する作業コストが大きいという点でスケーラビリティに欠ける問題がある。本研究では、多くの運用データがその時系列上の顕著な変化などの特徴に着目することで時系列イベントとして扱うことができる点に着目する。これらの時系列イベントをシステムログ上のイベントと同様に因果解析の対象とすることで、システムログだけでなく他の運用データをも対象とした横断的かつスケーラブルな関係分析を実現する。

(2) 長期間の運用データに対応する知識ベースの構築: 因果情報はあくまでイベント間の関係の有無を示すものであり、その関係のトラブルシューティング上の重要性を判別することはできない。これまでの因果解析では多数の自明な情報が検出され、それらが重要な因果情報を埋もれさせることで情報価値の低下を招いていた。本研究では因果情報の変化や前後関係に着目することで、因果情報の具体性を高めることを考える。すなわち、時系列上広範囲のデータを対象とする探索的な因果解析を実現し、得られる因果関係の変遷を知識ベースとして集約・表現することを目指す。

４．研究成果

本研究では運用データの時系列因果解析による文脈情報抽出を行うための一連の技術を備えた解析フレームワークを開発し、オープンソースで公開した。1 つはシステムログを自動生成されたテンプレートによって分類し時系列として扱うことを可能とするフレームワーク amulog である。このフレームワークは図 1 に示す構成により、既存研究で多数開発されるテンプレート生成手法を汎用的に動作させ、かつそれらに必要な前処理・高速化などを一元的に取り入れることができる。これにより、テンプレート手法の比較や連携などを用意かつ柔軟に行うことが可能となり今後のより高精度なテンプレート生成手法の開発という研究上の観点、および実際に用いる環境・データに適した手法の選択という実用上の観点、その双方についての貢献により今後のログ自動解析技術の導入・開発が促進されることが期待される。

もう 1 つは、amulog から読み込むシステムログ時系列に加え SNMP 計測データなどを入力として因果解析を行うことが可能なフレームワーク logdag である。このフレームワークはそれ

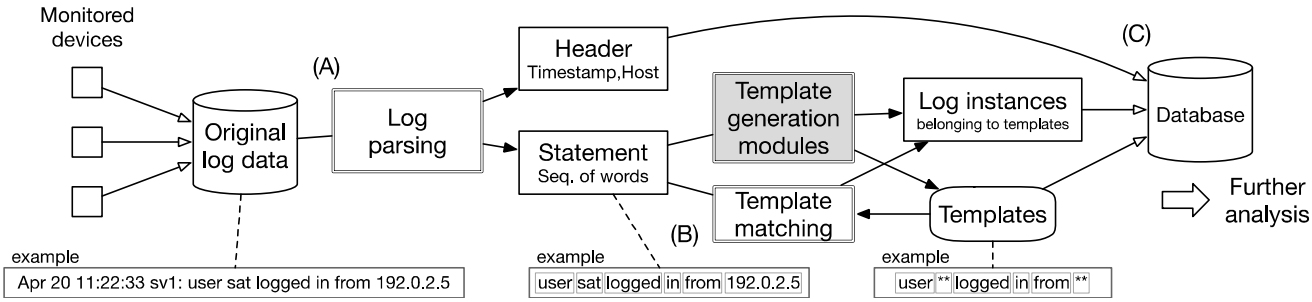

図 1: amulog の動作概要

ぞれの入力を時系列イベントデータに変換し、時系列の前処理や事前知識を用いた枝刈りによって高速な因果解析を実現する。得られた因果情報は入力の種類やその検出頻度を元に検索や照合、分類などが可能である。これにより、運用データ間の因果解析がより実用的に活用可能となり、障害発生時にそのシステムの振る舞いを素早く把握する、あるいは障害範囲やその原因を特定する上で直感的な情報をオペレータに提供することが可能となる。

これらの技術の開発過程において、以下の学術的知見・成果が得られた。

(1) **テンプレート生成手法に依存しない汎用ログ解析フレームワーク** amulog **の設計**

ログテンプレート生成手法はこれまで多数提案されているがそれらは多様な分野・前提の元で開発されており、手法間の比較や連携が難しいという問題があった。これに対し、ログテンプレート生成を共通の前提の元に扱うための汎用的フレームワーク amulog **の設計と実装を行った。**amulog **の特徴は、ログ中の自由記述文字列を記号により分離した単語列として一貫して扱うことで処理の効率化と解析の利便性を向上している点にある。**

図 1 **に示すように、**amulog **はまず収集されたログデータについてルールベースのログ前処理ツール** log2seq **を用いてヘッダ処理と単語分割を行い、その後既存技術をモジュールとして導入しテンプレート推定を行う。この時複数のテンプレート生成手法の連携を行うことが可能であり、特に本研究で提案した既知のテンプレートと一致するメッセージを高速に判別・分類するテンプレートマッチング手法と組み合わせることで既存のテンプレート生成手法を大きく高速化することが可能である。得られるヘッダ情報、テンプレートおよびメッセージ情報をデータベースに格納し、解析時には適切な検索** API **を提供する。**amulog **および** log2seq **は** Python **で実装を行い、**GitHub **で公開中である。また本成果は国際会議** CNSM2020 **において採録された。**

(2) **転移学習によるログテンプレート生成手法**

ログテンプレートの推定において、メッセージ中の単語間の位置関係などの情報を学習して利用する Conditional Random Fields (CRF)**などを用いた教師あり学習による構造学習アプローチ[A]は、特に精度面で既存のクラスタリング手法よりも優れた手法であることがわかっている。しかし教師あり学習には十分な数の教師データが必要であり、オペレータにとって手作業で教師データを作ることは負担の大きい作業である。本研究ではベンダやネットワーク環境などの異なるデータを教師として学習することを可能とするクロスドメインでの教師あり学習アプローチを考える。これにより、オープンソース機器のコードから自動抽出したテンプレートを教師データとして商用機器のログテンプレート推定に用いるなど教師データ作成の負担を大きく軽減し、かつ高い精度での推定が可能となる。**

本研究ではまず推定対象のドメインの教師データを用いないトランスダクティブ転移学習手法として、Bridged Refinement[B]**を用いたテンプレート生成手法を提案した。この手法により一定の精度改善が見られたが、機器間の共通のプロトコル以外の機能について大きく精度が低下する等の問題も見られた。本成果について電子情報通信学会** IA **研究会で口頭発表を行った。**

次に、推定対象のドメインの教師データを少数併用することでドメイン間の知識転移をより高精度に行う転移学習アプローチについて、ニューラルネットを用いた自然言語分野の手法[C]を応用するテンプレート推定手法 LogDTL **の提案を行った。この手法では図** 2 **に示すように、ドメイン間の知識転移を複数のニューラルネットの学習におけるパラメータ空間の共有・分離によって表現している。また個別のニューラルネットは** CRF **をベースとする構成であり、ログテンプレートの構造は単語と文字の** 2 **つの側面で特徴付けて扱われる。この手法について、**SINET4 **のログデータのテンプレート生成に** APAN-JP **で運用されるオープンソースルータ** Vyatta **のログデータおよび** Vyatta **のソースコードから自動生成したテンプレートを追加で用いる形での評価を行った。**LogDTL **を用いることで、**LogDTL **の部分実装である** CRF **と比較して特に同ドメインの教師データ数が少ない条件でテンプレート別単語精度が** 77.2%**から** 87.1% **へと改善されるなど、他ドメインの知識を精度改善に有効活用できていることが明らかになった。**LogDTL **は** Python **で実装を行い、**GitHub **で公開中である。また本成果は国際会議** AnNet2021 **において採録された。**

(3) **ネットワークトポロジ知識を用いた因果解析の効率化手法**

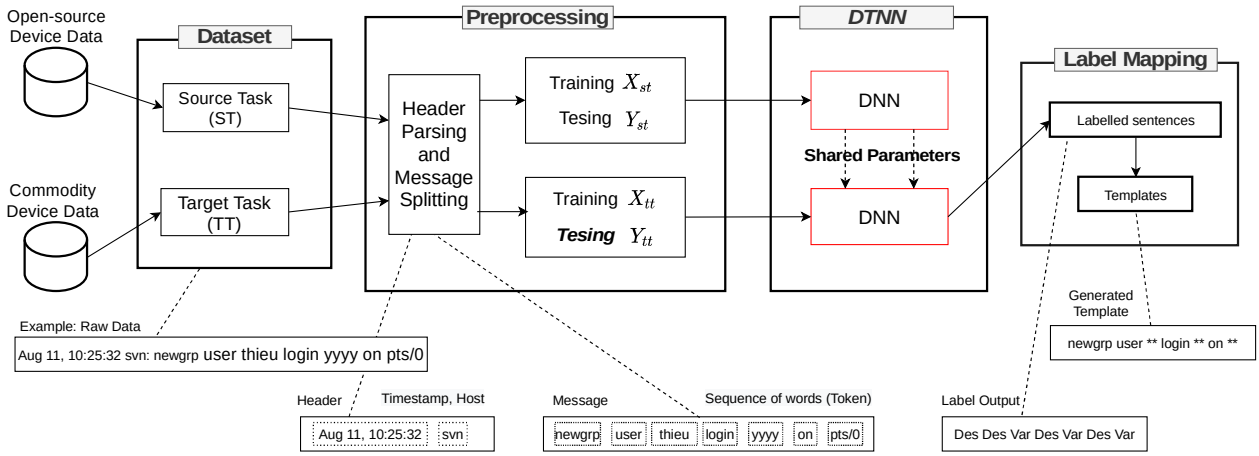

図 2: LogDTL の構成概要

因果解析には処理時間が入力ノード数に強く依存するという問題が存在する。本研究で目指すログと監視データなどを併用する因果解析では因果解析に与えるノード数が数倍に増加しうることから、この処理時間の問題が無視できなくなる。そこで本研究では、ネットワークのトポロジ知識を用いた因果解析の効率化技術の研究を行った。

オペレータが運用データの照合を行う際には、機器の振る舞いに関連が生まれる範囲は基本的に同種の機能や接続している機器であろう、という暗黙の知識を用いる。これと同種の考え方により因果解析の候補因果エッジをあらかじめ枝刈りすることで、因果解析の計算量を削減することができる。しかしこの枝刈りを不用意に行うと、因果フローが寸断され因果解析を用いた原因究明の実現上問題となる。特にシステムの振る舞いにはログに記録されないもの(未観測潜在要因)が多数存在するため、それらに中継されうる因果の枝刈りを防ぐ必要がある。

本研究ではネットワークのドメイン知識としてトポロジ上の機器接続情報、およびログイベントのプロトコルレイヤ情報の 2 つを用いた枝刈りルールに基づく因果解析の効率化手法を考案した。この枝刈りルールはドメイン知識に反するエッジ候補の枝刈りを行うが、その際未観測潜在要因による因果の仲介を一定範囲で許容しうる設計となっている。この技術を用いてログの因果解析を行った結果、処理時間を 74%削減することに成功した。これは従来のエリア分割に基づく手法と比較しても 16%の削減に相当し、エリア分割手法に見られたエリア境界の問題を解決しているなど精度面の改善も見られた。本成果は当該分野のトップ国際会議 CNSM2019 において採録された。

(4) DirectLiNGAM によるログ因果解析技術の検討

従来の因果解析で用いていた stable-PC アルゴリズムには同一時系列のイベントが複数存在する場合に関連する因果が得られなくなる問題点があることが因果解析の過程で判明した。この問題により、特に SNMP の計測データを入力として追加しても因果情報による障害チケットのカバー率が増加しないという問題に発展した。これに対し、同問題を回避可能な新たなログ因果解析技術の研究を行った。

本研究では DirectLiNGAM[D]を用いた因果解析手法を提案した。DirectLiNGAM は、因果効果を定量的に扱うことが可能であり、PC アルゴリズムに依存せず、かつ(3)で示した事前知識を用いた因果解析の効率化技術を応用可能な技術である。この手法は正規分布以外のデータについて成立する LiNGAM と呼ばれる因果モデルを用いている。ログ因果解析で用いるイベント時系列は一般に正規分布ではなくポアソン分布などに近く、このモデルを利用することができる。この DirectLiNGAM を用いることで、因果解析によるトラブルチケット関連因果検出率は、stable-PC アルゴリズムの 62%から大きく改善され 88%となり、stable-PC アルゴリズムの問題点は DirectLiNGAM を用いることで改善可能であるということが確認できた。

＜引用文献＞

[A] S. Kobayashi, et al. “Towards an NLP-based Log Template Generation Algorithm for System Log Analysis”, CFI, p.4, 2014

[B] D. Xing, et al. “Bridged refinement for transfer learning,” ECML PKDD, pp.324–335, 2007.

[C] Z. Yang, et al. “Transfer learning for sequence tagging with hierarchical recurrent networks,” arXiv, 2017.

[D] S. Shimizu, et al. “DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model”, Journal of Machine Learning Research, 12, 1225–1248, 2011.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Kazuki Otomo, Satoru Kobayashi, Kensuke Fukuda, Hiroshi Esaki	４．巻 E102-D, no.9
２．論文標題 Latent Variable based Anomaly Detection in Network System Logs	５．発行年 2019年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 1644-1652
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2018OFP0007	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Satoru Kobayashi, Kazuki Otomo, Kensuke Fukuda	４．巻 -
２．論文標題 Causal analysis of network logs with layered protocols and topology knowledge	５．発行年 2019年
３．雑誌名 Proceedings of the 15th International Conference on Network and Service Management (CNSM 2019)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.23919/CNSM46954.2019.9012718	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Satoru Kobayashi, Yuya Yamashiro, Kazuki Otomo, Kensuke Fukuda	４．巻 -
２．論文標題 amulog: A General Log Analysis Framework for Diverse Template Generation Methods	５．発行年 2020年
３．雑誌名 Proceedings of the 16th International Conference on Network and Service Management (CNSM 2020)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Thieu Nguyen, Satoru Kobayashi, Kensuke Fukuda	４．巻 -
２．論文標題 LogDTL: Network Log Template Generation with Deep Transfer Learning	５．発行年 2021年
３．雑誌名 Proceedings of the 6th IEEE/IFIP International Workshop on Analytics for Network and Service Management (AnNet 2021)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計3件（うち招待講演　1件／うち国際学会　0件）

1．発表者名
山城 裕陽，小林 諭，福田 健介，江崎 浩

2．発表標題
Bridged Refinementによるログテンプレート推定手法の検討

3．学会等名
電子情報通信学会IA研究会

4．発表年
2019年

1．発表者名
Satoru Kobayashi

2．発表標題
Causal Analysis of Network Log Events

3．学会等名
JFLI Workshop 2020 on Next Generation Networking（招待講演）

4．発表年
2020年

1．発表者名
徳備彩人，大友 一樹，小林 諭，福田 健介，江崎 浩

2．発表標題
ネットワークログデータへの自動文書ラベリングの提案

3．学会等名
電子情報通信学会IA研究会

4．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
amulog https://github.com/cpflat/amulog
logdag https://github.com/cpflat/logdag

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	山城　裕陽 (Yamashiro Yuya)		2019年度にRAとして雇用 ログ解析の業務に従事
研究協力者	大友　一樹 (Otomo Kazuki)		2020年度にRAとして雇用 ログ解析の業務に従事

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関