

システムの挙動分析のためのリアルログ変換手法の実装及びその評価

秋山 馨^{1,a)} 阿部 洋丈^{1,b)} 加藤 和彦^{1,c)}

概要: 昨今のシステムにおいては高いスケーラビリティが重要な指標として求められ、そのためのシステムの実行時ログの分析はもはやシステム運用者にとっては必須であるといえる。しかし、実際に大規模システムのログをシステムが巨大になる前に入手するのは困難であるため、大規模ログデータでの検証は難しく、スケーラビリティを担保するような新手法の評価に課題がある。本研究では、実際にシステムが稼働している中で得られたシステム稼働ログを元にして、様々なスケール条件でのログを自動生成する手法を提案する、また、それに基づきシステム実行を行うことにより、様々なシステムの挙動分析を汎用的に行う手法を確立することを目標とする。

1. はじめに

近年の情報サービスは従前のモノリシックな構成ではなく、内部的に独立した細かい単位のサービスに分割されたマイクロサービスなアーキテクチャが主流となりつつある [1]。

このようなアーキテクチャでは Kubernetes などに代表されるようなオーケストレーションツールを用いてシステムが動作している最中に計算資源を動的に変化させることによりスケーラビリティを担保するシステムが増加している。このようなシステムを運用する過程で得られる実際のログを用いたシステムの評価は様々な場面で盛んに行われているが、このような評価は実際のログを有している場合にのみ実行が可能であり、この点で制約があるといえる。

一方、システムに与えられるイベントを妥当に生成させることによって実際のログに近いようなログを得て、それを用いてシステムの評価を行うような研究領域はほとんど未開拓であるといえる。

本研究ではこの部分に焦点を当て、システムの実行時ログはその裏に存在するユーザーなどのエージェントから発生したものと考えることができるため、ある程度の精度でログ集合からその裏にあるエージェントのグラフ構造を数理的なモデルにより推定できるという仮説に基づき、そのようなモデルの実装及び評価を行う。この提案は実際のシ

ステムの実行時ログから、そのシステムがスケールした際の挙動を分析するような手法を構築する際にも有用であると考えられる。

2. ログの生成手法

現実存在する様々なシステムの分布は、Barabási-Albert モデルによく従うことが知られている [2]。本研究ではこれを根拠とし、リアルログを元にして、その裏に存在する Barabási-Albert モデルの構造の推定を行う。

具体的には、初期値としてある大きさの Barabási-Albert モデルに基づいたグラフを有し、そのグラフに対してリアルログを元にしてエッジを追加することにより、リアルログに近いようなログを生成する。この際にエッジを追加する頂点を、実際のログにおいて各ノードが参照された回数に基づき決定することで Barabási-Albert モデルの構造を保ったままグラフを拡張することが可能である。

また、図 1 に示すように、今回のログ生成手法は、実際のログを必要とせずいくつかのパラメーターだけで良いモデル化が可能なデータに対するマルチエージェントアプローチと、実際のログを、その裏にあるモデル構造を考えずに生成するリアルログ変換アプローチとを組み合わせたハイブリッドアプローチとなっている。これ以外の 2 アプローチについても、今後の研究において実験を進めて、どのようなデータに対して各アプローチが上手く適合するかについても調査する予定である。マルチエージェントアプローチはいくつかのパラメーターを指定するだけでログ生成が可能であり、一方リアルログ変換アプローチは実際のログのみを情報源とするので追加の指定が必要ないという

¹ 筑波大学
University of Tsukuba

a) akiyama@osss.cs.tsukuba.ac.jp

b) habe@cs.tsukuba.ac.jp

c) kato@cs.tsukuba.ac.jp

利点がある。今回のハイブリッドアプローチはこれらを組み合わせた手法であり、実際のログをマルチエージェント的なモデルに変換することが可能であるような場合に特に有用であると考えられる。

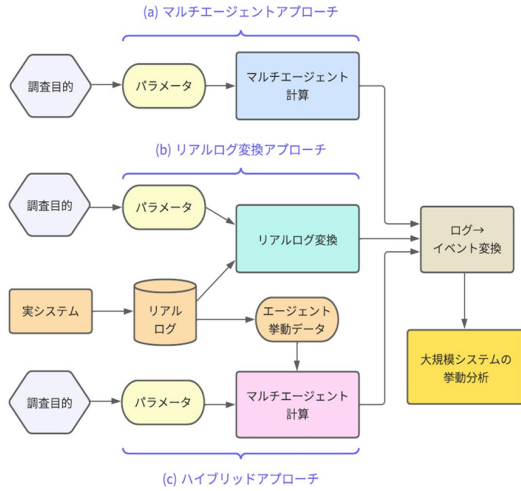


図 1 本研究における 3 アプローチ

3. 評価実験の概要

前述のログを追加生成するような数理モデルに基づいた実装を行ったうえで、図 2 に示すような実験を計画した。ここで実際に用いるログとしては 1 年超にわたり収集された実際の Flickr での行動ログデータを用いた。これにはいつどのユーザー ID のユーザーが、どのファイルに対してアクセスを行ったのかという情報が記録されている。

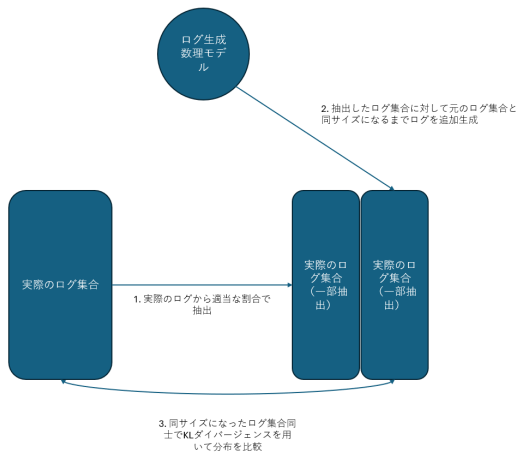


図 2 評価実験の概要

ここで、KL ダイバージェンスは、確率分布 P と Q の類似度を表すことのできる指標であり [3]、離散的な確率分布 P と Q に対しては以下のように定義される。

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \quad (1)$$

4. 評価結果

元のログから n 割のログを残し、残りの $10 - n$ 割のログを生成した場合の KL ダイバージェンスを記録する実験を、 n を 1~10 の整数値で動かし、各 n で 1000 回の実験を行った結果を図 3 に示す。

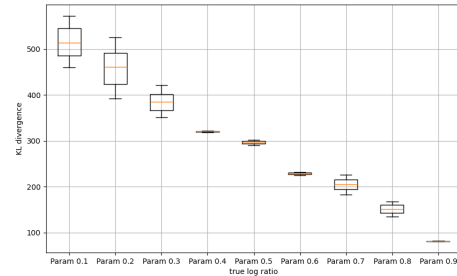


図 3 評価実験の結果

概ね実際のログの割合が高いほど KL ダイバージェンスが小さくなっていることが分かるが、一部のケースでは実際のログの割合が少なくても、それより多い割合のケースよりも良い生成を行うことができていたことが示唆される。

5. 展望

今後は実験をさらにに行い、どのようなケースで良い結果が得られる傾向にあるのかを調査する予定である。また、今回示した結果は提案モデルの理論的な評価であるが、これをより実際のシステムに適用したときの効果などを調査するために文献 [4] で提示されているような、ディスクや計算機の消費電力といったシステム的な観点での指標の変化が、実際のログ集合に基づくイベントと、生成したログを合わせたログ集合に基づくイベントとの間でどの程度生じるのかを調査する予定である。

参考文献

- [1] Newman, S.: *Building microservices*, " O'Reilly Media, Inc." (2021).
- [2] Albert, R. and Barabási, A.-L.: Dynamics of complex systems: Scaling laws for the period of Boolean networks, *Physical Review Letters*, Vol. 84, No. 24, p. 5660 (2000).
- [3] Kullback, S. and Leibler, R. A.: On information and sufficiency, *The annals of mathematical statistics*, Vol. 22, No. 1, pp. 79–86 (1951).
- [4] Machida, F., Hasebe, K., Abe, H. and Kato, K.: Analysis of Optimal File Placement for Energy-Efficient File-Sharing Cloud Storage System, *IEEE Transactions on Sustainable Computing*, Vol. 7, No. 1, pp. 75–86 (online), DOI: 10.1109/TSUSC.2020.3037260 (2022).