# クラウド運用のためのストリームマイニング

松　本　真　佑[†1]

本ポジションペーパーでは，計測データを用いたクラウド運用に対するストリームマイニングの適用について考察する．クラウドの保守・運用ではクラウドシステム自体から計測された結果が活用される．この計測結果は時系列に変化し終わりなく得られるストリームデータであり，高速かつリアルタイムに処理できることが求められる．クラウドシステムから得られる計測情報をストリームマイニングに適用することで，よりきめ細やかで高速なシステムの最適化を実現できると考える．

# Stream Mining for Cloud Operations

Shinsuke Matsumoto[†1]

## 1. はじめに

クラウドはその名に示すとおり「雲」のような存在である．本質を捉えにくく，またその全容を掴むことも容易ではない．Web2.0 という言葉と同様，一時はバズワードであるという批判もあったものの，現在では多岐に渡るクラウドサービスが登場するようになった．クラウドは仮想化や Web，分散技術などの様々な要素技術の集合で成り立つだけのコンピューティングパラダイムであるが，その組み合わせの強力さから IT 社会に新たな価値や可能性をもたらしつつある．このクラウドの台頭に伴い，ソフトウェア工学分野も要件・設計・開発・保守・運用いずれの開発フェーズに対しても，新たな課題に取り組んでいく必要がある．本ポジションペーパーでは，このうちクラウドの保守・運用について考える．

NIST による 5 つのクラウドの本質的性質の一つとして，“Measured Service” が定義されている．これはクラウドサービスそのものの利用状況を計測可能とする，という性質のことを意味する．計測対象は多岐に渡り，IaaS システムであれば仮想マシンの CPU 利用率やストレージの利用量が，PaaS や SaaS システムであればアカウントごとのサービス利用量などが含まれる．

Measured Service 特性による計測結果はクラウド利用者への課金モデルに反映されるほか，クラウドシステムの保守や運用のためにも用いられる．例えば，計測結果の可視化によるシステム利用状況の把握，監視対象メトリクスの上限下限等の閾値を超えた際の異常検出や通知のほか，仮想マシンの負荷状況に応じたシステムの自動最適化など様々な保守への活用方法が存在する．

このような計測情報を用いたクラウドの保守方法は，Measured Service 特性を持つあらゆるクラウドシステムに共通の手段であると同時に，課題であるともいえる．計測情報は大量のデータであることから高速で処理を，さらに時間的に変化するデータであることからリアルタイムに処理できる必要がある..

本ポジションペーパーでは，計測データを用いたクラウド運用に対するストリームマイニングの適用について考察する．ストリームマイニングとは時間的に変化する逐次的なストリームデータを，高速に処理するマイニング手法のことである[1)]．クラウドシステムから得られる計測情報をストリームマイニングに適用することで，よりきめ細やかで高速なシステムの最適化を実現できると考える．

## 2. ストリームマイニング

データストリームとは時間的に変化し終わりなく逐次的に得られるデータのことを指す．Web のアクセスログや POS による購買履歴などのログ情報が代表的な例である．また現在の社会でビッグデータと呼ばれるデータは，その名に表される「大量の・膨大な」という意味に限らず，暗に「逐次的な・リアルタイムな」という意味を含んだものも多く，ストリームデー

†1 神戸大学大学院システム情報学研究科
Graduate School of System Infomatics, Kobe University

タの一種であるとも見なすことができる．

ストリームマイニングはこの時系列に得られるストリームデータを，少ないメモリで高速に処理する手法のことを指す．なお，ストリームマイニングは特定のマイニング手法を指す言葉ではなく，ストリームデータを様々な工夫によって効率的に処理するマイニング手段の総称である．マイニング対象となる問題には，平均や総和，偏り，頻度計算などの単純な統計量に加え，クラスタリングや機械学習などのより高度なデータマイニング手法も研究されている[2)]．

ストリームマイニングの基本的なアイデアは，厳密な解を求めない点にある．高速なマイニング処理を実現するために，ストリームデータはメモリ上で計算が行われる．しかし，ストリームデータは時間的に終わりなく到着する無限のデータであるため，有限のメモリを節約する工夫が重要となる．頻度計算アルゴリズムの場合，代表値を用いて大まかに結果を集計することでメモリを節約する[3)]．また，マイニング結果を利用するアプリケーションの興味に応じて処理を簡略化する方法もある．例えば，対象がログデータでアプリケーションの興味がごく最近のストリームデータのみである場合，最新付近のデータを厳密に処理し，古い結果は要約しておくことで少ないメモリで効率的に計算を行う．

## 3. 議　　論

クラウドシステムから得られる様々な計測結果をストリームデータとみなし，ストリームマイニングを適用することでよりきめ細やかな保守が実現できる可能性がある．スケールアウトやスケールインと呼ばれる仮想マシンの自動最適化では，計算機リソースの利用量を逐次計測・監視し，あらかじめ設定した閾値を超えた際に自動的に仮想マシンを増加／現象させることでシステム全体の処理能力を最適化する．このような閾値を用いたクラウドの保守手段はストリームマイニングを用いなくても実現は可能である．これに加えストリームマイニングを用いれば，計算機リソース利用量の直近の変化傾向や，曜日や時間帯ごとの利用量のパターンをマイニングできれば，負荷の増大を予測しあらかじめ最適化しておくといった方法が実現できる．

一般的なデータマイニング手法と同様に，ストリームマイニングを適用する際には，アプリケーションの興味がどこにあるか，すなわちどのように計測結果を用いるかを事前に決定することが重要である．特にストリームマイニングでは，無限のストリームデータを有限メモリで処理するためにデータの要約化や確率的な計算が施されるが，その近似的な解が目的に合うかを判断して適用する必要がある．例えば頻度計算の場合，頻度が多い方に興味があれば少ない方の要約化を，少ない方に興味があれば多い方を要約化といった工夫が必要である．

最近では巨大なメモリを安価で確保できること，また SSD などの NAND 型メモリを用いた高速なストレージも登場していることから，計算結果を途中から再開可能で，かつ 1 パスで処理できるマイニング手法（総和や頻度計算，逐次学習アルゴリズムなど）であれば，比較的容易にストリームデータを処理できるといえる．

また現在では，MapReduce などの複数ノードによる並列処理技術が登場しており，様々な現場で利用されている．この並列処理手法とストリームマイニングは，大量かつ大規模なデータを高速に処理するという点で目的は同じであるが，対象データがバッチ指向かストリームであるかという違いを持つ．さらに現在，この並列処理技術とストリームマイニングを組み合わせたオープンソースプラットフォーム S4[⋆1] が登場している．S4 では並列処理実現の際の煩雑さや，マイニング手法適用の際の難しさをユーザから隠蔽することを目的として開発されており，単純な API の組み合わせだけでストリームマイニングを実現可能であると期待できる．

## 4. おわりに

本ポジションペーパーでは，計測データを用いたクラウド運用に対するストリームマイニングの適用について考察した．前述の通り，ストリームマイニングは「どの目的で利用するか」が重要である．ワークショップでは，クラウドシステムの保守現場の実態について，また計測情報をどのような形で活用されるかについて議論できれば幸いである．

## 参　考　文　献


1) 有村 博紀, 喜田 拓, “データストリームのためのマイニング技術”, 情報処理, 2005.
2) Charu C. Aggarwal, “Data Streams: Models and Algorithms”, Springer, 2007.
3) Toon Calders, Nele Dexters and Bart Goethals, “Mining Frequent Items in a Stream Using Flexible Windows”, Journal of Intelligent Data Analysis - Knowledge Discovery from Data Streams, Vol. 12, No. 3, 2008.


⋆1 http://oss.infoscience.co.jp/s4/s4.io/