連載

5分で分かる!? 有名論文ナナメ読み

鈴木　順（NEC）

Peter Bailis et al. : MacroBase : Prioritizing Attention in Fast Data

SIGMOD'17（2017）

SNS（Social Networking Service）の普及や産業システムのIoT（Internet of Things）化により，データセンタに大量のデータが集まるようになった．SNSでは1日に1兆件を超えるメッセージやイベントが処理されている[☆1]．またクラウドで提供されるIoTサービスでは，あらゆるログがデータセンタのサーバに蓄積されている．

今日ではこれらのデータを解析し，社会システムにおける障害の予兆検出や，異常時の原因解析を行うことが重要性を増している．そのためには大量のデータを効率的に処理できるITシステムが必要である．Apache Hadoop[☆2]やApache Spark[☆3]等のミドルウェアフレームワークはこれらの用途に用いられてきた．

しかし，これらのツールを活用して大量のデータから有用な事象の発見や原因の解析を行うには，データ解析者に高い技術スキルが要求される．なぜならデータに対し，どのようなクエリを用いて解析を行うか，選択演算や集約演算の設計をどうするかはデータ解析者に任されているからである．クエリセットを記述するためにはアプリケーションに対する理解とデータ解析の知識が必要である．また，大量のデータを人間である解析者が見て判断できる水準までデータ量を絞り込む調整や，システムが受信するデータの特性の変動に合わせたクエリの調整も必要である．

2017年のSIGMODにおいてスタンフォード大学のPeter Bailisらが発表したMacroBaseは，この課題の解決を目的としたミドルウェアフレームワークである．大量のストリームデータから着目すべきレコードを自動的に検出し，なぜそのデータに着目する必要があるか理由を提示する．これによりデータ解析者が選択演算や集約演算を駆使して試行錯誤でデータを絞り込む作業が不要となり，高い技術スキルがなくともデータ解析を行うことができるようになる．

MacroBaseを用いると，入力データストリームから着目すべき異常データと，そのデータを異常と判定した理由を提示することができる．データの各レコードは複数のmetricsおよびattributesから構成されている．たとえばモバイルアプリケーションのログデータではmetricsは消費電力でありattributesはデバイスやアプリケーションのバージョンである．MacroBaseはログデータの中から消費電力が異常なデータを検出し，異常データに特異なattributesの組を提示することによってそのデータが異常と判断される理由を説明する．それにより，本例では特定のデバイスとアプリケーションのバージョンの組合せが異常な電力消費と因果関係があることが示される．このとき提示されるattributesの組は，正常なデータでは出現頻度が低く，逆に異常なデータで出現頻度が高い組である．従来は大量のデータから解析者がクエリを組み合わせてこのよう

☆1 https://www.datanami.com/2015/09/02/kafka-tops-1-trillion-messages-per-day-at-linkedin/
☆2 https://hadoop.apache.org/
☆3 https://spark.apache.org/

な解析を行うことが困難だった．

MacroBaseのアーキテクチャを図-1に示す．Macrobaseは Macrobase's Default Analytics Pipeline（MDP）で構成されている．Ingestは Java Database Connectivity（JDBC）等により外部データソースからデータをMDPに投入する．Transformは必要に応じてデータに依存した変換を行う．たとえばデータに対する自己相関の取得や画像におけるオプティカルフロー処理が行われる．Classifyはデータの metrics を参照して異常データを検出する．Explainは異常データに特徴的な attributes の組合せを提示する．Presentはデータ解析者に対する表示機能である．

MacroBaseの論文では図-1に示したデータ解析のためのパイプラインアーキテクチャを提案したことに加え，大量のデータストリームから異常データを検出しその原因と考えられる attributes を提示するための3つの技術を提案した．1つ目の技術は Adaptable Damped Reservoir（ADR）であり，Classifyで用いられるデータ密度に基づく異常データの分類器をトレーニングするためのサンプルを作成する．ストリームデータに対するサンプリングでは，データ入力帯域に応じてサンプリングレートを調整する必要がある．ADRでは入力帯域の一時的な変化に過度に応答しない手法を提供する．2つ目の技術は検出異常データの説明となる attributes の探索手法である．あらゆる attributes の組合せを探索する場合，組合せ解析で探索空間が膨大となる．これに対しMacroBaseでは異常と検出されたデータに含まれる attributes の組合せのみを探索対象とするため，異常値として検出されるデータの数自体が小さいことを利用して組合せ探索の空間を抑制する．また3つ目の技術はExplainで提示する attributes の出現回数を数える Amortized Maintenance Counter（AMC）であり，ストリームデータにおけるトップN個の attributes のカウントを高速かつ省メモリで行うことを可能にする．

ここでMDPを構成する各機能モジュールは個々にユーザ定義モジュールと交換可能である．モジュールを入れ替えてもパイプラインを動作させるため，モジュール間のインタフェースが定義されている．一方，データ解析者が機能モジュールを作成することなくデータ解析が行えるよう，MacroBaseでは前述した3つの技術を備えたデフォルトのモジュールも提供している．

このようにMacroBaseは従来のミドルウェアフレームワークが提供していた解析機能とデータ解析者の要求のギャップを埋める自動解析技術に着目し，その領域に必要となるデータ処理技術を新たな視点から提案している点で興味深い．MacroBaseは実際，IoTシステムやクラウドシステムにおいて人が予期していなかった不具合とその原因を発見している．今後IoTが普及し，大量のデータがネットワークを通して収集可能となることで，このような大規模なデータの自動監視・解析ツールのミドルウェアが増え，より使いやすくなっていくだろう．

（2019年1月30日受付）

鈴木　順（正会員） j-suzuki@ax.jp.nec.com

2005年東京大学大学院工学系研究科電子工学専攻修士課程修了．2018年同情報理工学系研究科電子情報学専攻博士課程修了．博士（情報理工学）．2005年NEC入社．2012～2013年カリフォルニア大学バークレー校客員研究員．コンピュータプラットフォームのHW・SWの研究開発に従事．

図-1　MacroBase's Default Analytics Pipeline