CEPを用いたストリーム・データ分析

桑田　修平，中川慶一郎

ストリーム・データ分析をベースとするサービスがビジネスの世界で注目を集めている．このサービスは，時々刻々と変化する株価情報，センサやGPS衛星を通して得られる位置情報等々，流れるように発生するデータ（ストリーム・データ）をリアルタイムに活用するサービスである．1件当たりミリ秒レベルの低遅延処理を可能とするCEP（Complex Event Processing）と呼ばれるデータ処理基盤によって高度な分析が容易に実施できるようになった結果，さまざまな分野への適用が現在検討され始めている．本稿では，ストリーム・データ分析に関して従来のデータ分析との違いや分析事例を示すとともに，分析基盤であるCEPについて解説する．

キーワード：ストリーム・データ，データ分析，Complex Event Processing

1. はじめに

情報技術の発展により，ツイッターやGPS（Global Positioning System）機能付きの携帯電話などを通して，時々刻々と変化していく情報をリアルタイムに取得し活用できる時代になった．そして，一昔前には想像もつかなかったことができるようになった．例えば，メーカの品質管理担当者は，ツイッター上の誰かのつぶやきから新製品に不具合があったことを知り，即座に対応策を打てるようになった．また，災害が起きたとき，GPSを通して得られる車の移動ルートから通行可能な道路を素早く把握できるようにもなった[1]．

このような例は，情報を取得した時点ですぐにそれを具体的な行動として反映できた結果，新たな価値や知恵が生まれた事例と見ることができる．先の例でいうと，車の移動情報を把握し，車の移動ルートと地図とを即座に重ね合わせることができた結果，瞬時に現在通行可能な道路が分かるようになったと捉えることができる．

ここでのポイントは，車の位置情報など，連続して随時発生する“ストリーム・データ”と呼ばれるデータを分析して知恵に結び付けている点にある．なお，ストリーム・データとは，次のような特徴を持つデータのことを表す[2]：

- データ・サイズや発生タイミングの予測が難しい，
- 蓄積しようとするとすぐに大容量になる，
- さまざまなデータ構造をもつ（構造化/非構造化），
- 誤りや欠損を含む．

このような特徴を持つストリーム・データは，人や物の位置情報や金融における板情報（株価やその株の注文状況が分かるデータ），他には各地の気温情報など世の中に溢れている．ストリーム・データ分析とは，そのようなデータを対象にリアルタイムな分析を行うことであり，新たな価値を生み出す源泉としてビジネスの世界で注目を集めている[3]～[6]．

ただし，データ処理という観点からいえば，ストリーム・データ処理自体は決して新しいものではなく，例えば異常検知のように従来から“リアルタイム処理”や“オンライン処理”などといった枠組みですでに考えられてきている．

では今なぜ注目を集め，ストリーム・データ分析をベースとしたサービスが展開され始めているのかというと，それはリアルタイムなデータ処理基盤の発展に伴って，単なるデータ“処理”ではなく新たな価値をもたらす高度なデータ“分析”が実施可能となり，ストリーム・データ分析がもたらす新たな市場に企業が大きな期待感を持ち始めたからである．ここでリアルタイムなデータ処理基盤は，データをリアルタイムに取得可能とするセンサ等の入力装置や取得したストリーム・データを高速に処理する基盤技術（CEP）[7]などを組み合わせて開発することが考えられている．現在，国内外の企業がCEPを用いたストリーム・データ分析を基とするサービスの市場開拓に積極的に乗り出し始めている．

本稿では，ストリーム・データ分析について，どのような適用がビジネスレベルで考えられているのかを具体的な分析事例を通して紹介するとともに，ストリ

くわた　しゅうへい，なかがわ　けいいちろう
(株)NTTデータ
〒135-8671 江東区豊洲3-3-9

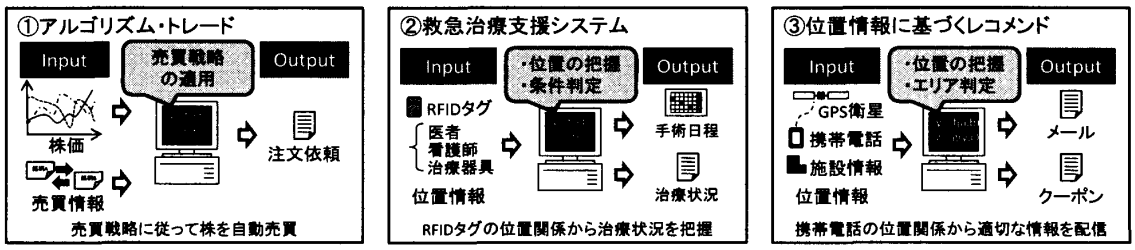

図1 ストリーム・データ分析事例のイメージ

ーム・データ分析に共通する枠組みを述べる．また，ストリーム・データ分析を支える基盤技術であるCEPの概要やCEPを用いて実際に我々が開発したサービスについて紹介し，最後に今後の展望について述べる．

2. ストリーム・データ分析

まず，ビジネスの世界においてこれまで行われてきた従来型のデータ分析との違いを明確にしておく．大きな違いは，対象とするデータの種類が異なる点である．すなわち，従来型のデータ分析が静的データ，つまり，いったんデータベースに蓄積したデータを主な分析対象としているのに対し，ストリーム・データ分析は先に述べたとおり時々刻々と傾向が変化していく動的なデータを主な分析対象とする．したがって，顧客の購買履歴を分析することでマーケティングに活用するといった分析は従来型のデータ分析として位置づけられる．

本節では，ストリーム・データ分析の3つの具体例を紹介したのち，それらに共通する枠組みを示すことで，ストリーム・データ分析とはどのようなものか解説する．

2.1 事例①：アルゴリズム・トレード

最初に紹介する事例は，欧米ではすでに普及し始めている「アルゴリズム・トレード」である．

株の売買において重要なポイントの1つは，タイミングを見逃さないことである．そこで，株の売り買いを決定する基準のうちですでに定式化されているものについては事前にプログラムを作成しておき，システムに自動で売買の判断をさせることとなる．このように，ある決められた売買戦略（アルゴリズム）に従って，株の売買（トレード）を行うサービスがアルゴリズム・トレード[8]である．アルゴリズム・トレードでは，逐次発生する板情報がストリーム・データであり，板情報に対してあらかじめ用意しておいた売買戦略をリアルタイムに適用することがストリーム・データ分析に対応する．

アルゴリズム・トレードが金融業界に与えたインパクトは，単にシステムが自動で売買できるようになったということだけでなく，株の売買が人手ではもはや対応不可能なスピードで行われるようになったことにある．1件当たり数ミリ秒での売買が成立可能となり，これに対応できないと売買のタイミングを逃がしてしまう．結果として，処理性能が劣る売買システムや売買のタイミングを掴めないアルゴリズムを抱える証券会社は当該市場から退場を余儀なくされることとなり，システムの処理性能や機能（売買戦略）自体が企業の存亡に直結するといっても過言ではない状況になった．

アルゴリズム・トレードは株の売買業務を大きく変え，それがCEPによって実現されていたことから，ストリーム・データ分析やCEPが一躍注目を集め始めたのである．

国内においては，東京証券取引所が2010年1月にシステムを刷新し（システム名：arrowhead），これまでは数秒かかっていた注文受付処理が数ミリ秒レベルにまで高速化（低遅延化）した．これを受けて，各証券会社は自社の売買システムの更改を急ピッチで進めている状況である．

2.2 事例②：救急治療支援システム

2つ目は医療分野での適用事例である．具体的には，

- 救急治療室のあらゆる場所にセンサを設置する，
- 医者・看護師・検査技師等の医療スタッフにRFID（Radio Frequency IDentification）タグを持たせる，
- 医療器具にRFIDタグを取り付ける，

ことで，医療スタッフや医療器具の位置をリアルタイムに把握し，互いの位置関係や医療スタッフと医療器具との接触状況をもとに，事前に定めておいた判定ルールに従って治療の進行状況を自動生成するサービスである．

緊急を要する救急の現場において，医療スタッフや医療器具の稼動状況を自動で把握可能とするだけでなく，治療内容の記録や手術スケジュールの計画作成までをシステムが支援することで，限られた人財/資財/設備を効果的に活用できるようになる．その結果，救急患者を“たらい回し”にするようなことが回避される．

救急治療支援システムにおいては，各センサから取得される医療スタッフや医療器具の位置情報がストリーム・データに相当する．また，ストリーム・データに対して，位置情報や医療スタッフと医療器具との接触状況から治療内容を判断するルールを逐次適用することがストリーム・データ分析に対応する．ここで，判断ルールは事前に定義されていることが前提である．

2.3 事例③：位置情報に基づくレコメンド

3つ目の事例は，GPS衛星を通して得られる携帯電話の位置情報をリアルタイムに分析する事例であり，例えば，NTTドコモが実施している「ドコモ・ワンタイム保険」の保険案内メール・サービス[9]が挙げられる．これは，携帯電話の位置をリアルタイムに把握することで，空港に到着した瞬間に保険の案内メールを携帯電話へ配信するサービスである．

具体的には，携帯電話と空港の位置関係をリアルタイムに監視し，両者がある一定の間隔内になったその瞬間に，定められた内容を該当する携帯電話に配信する．逐次変化していく位置情報がストリーム・データに相当し，配信条件を満たすかどうかを常に監視することがストリーム・データ分析に相当する．

このようなサービスは，位置情報に基づくレコメンド・サービスと捉えることができ，次のような事例も考えられている．すなわち，ある小売店舗の位置を中心として，事前に設定した大きさの同心円状のエリアに入った携帯電話の利用者のみに対して，該当する小売店舗の情報をリアルタイムに配信する．“今ここにいる貴方だけに”お得な情報を配信するサービスである．携帯電話の利用者の普段の移動ルートや嗜好を踏まえて，配信する情報を利用者ごとに変えることも考えられる．

2.4 共通する枠組

以上に紹介した3つの事例にはある共通点を見いだすことができる．それは，背景知識やあらかじめ収集しておいた過去の履歴データからルールやパターンといったモデルを事前に抽出・作成しておき，次々と発生するストリーム・データに対して先に抽出したモデルをリアルタイムに適用するという枠組みである．

アルゴリズム・トレードの事例では，過去の売買履歴やトレーダの経験等から売買戦略を設定し，逐次得られる板情報などに対してその売買戦略を適用している．また，救急治療支援システムの事例では，医療スタッフと医療器具との接触を判定する条件や医療器具と治療内容との紐付けを行うルールを事前に定めておき，リアルタイムに把握される医療スタッフなどの位置関係にそれらのルールを適用している．位置情報に基づくレコメンドの事例についても同様の枠組みで表現可能であり，空港に着いたかどうかを判定するルールを，時々刻々と変化する携帯電話の位置情報に適用している．

図2に共通する枠組みのイメージを示す．図2の上側の部分がモデルを抽出する部分，図2の下側が抽出したモデルをストリーム・データに対してリアルタイムに適用する部分である．なお，適用するモデル内にそのモデル自身を逐次更新する仕組みを取り入れておくことも可能である．

ここで，モデルの抽出にあたっては，データの背後にある文脈（コンテクスト）をいかにして読み取るかがポイントとなる．つまり，先に挙げた保険案内メール・サービスのような場合には，情報の配信目的が場所に強く依存するため，比較的容易に利用客の文脈が把握可能である（国際線の空港に来た人は海外に行く可能性が高い等々）が，場所に依存しない場合には位置情報以外の情報（例えば，個人の嗜好情報）や高度な分析モデルが必要となる．

図2で示されるような枠組みは，情報活用の観点からいうとプロアクティブな情報活用として位置付けられる[10]．文献[10]では4つの情報活用のタイプが示されており，プロアクティブ型のポイントとして，“プロアクティブ”の名のとおり，“一歩先を行く”情

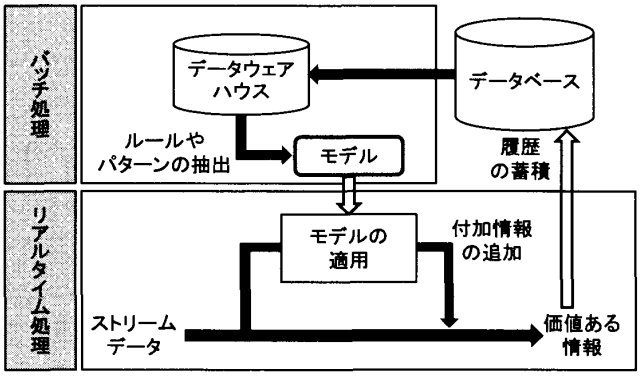

図2 プロアクティブ型の情報活用イメージ

報活用/データ分析が挙げられている．一歩先を行くためにはストリーム・データ分析の適用シーンにおける文脈を的確に読み取る必要があり，上述したとおり高度なデータ分析技術が求められることとなる．

アルゴリズム・トレードが金融業界に大きなインパクトを与えたように，ストリーム・データ分析は，“プロアクティブ”という観点で，他の分野や他のサービスにおいても大きなインパクトを与える．

3. CEP : Complex Event Processing

2節の最後で，ストリーム・データ分析は，逐次発生・取得されるストリーム・データに対して，事前に作成したモデルをリアルタイムに適用するという枠組みで捉えることができることを示した．CEPとは，まさにストリーム・データを受け取り，リアルタイムにモデルを適用するデータ処理基盤である．これまでは分析を実施するたびにプログラムを自前で組む必要があった．これに対して，CEPを利用すれば，極端にいえば適用するモデルのみを記述するだけでリアルタイムなデータ分析が可能となる（あるプロジェクトにおいては，手組みによるシステム開発と比較して開発期間が1/2以下になったという話を筆者らは聞いたことがある）．

本節では，ストリーム・データ分析が注目を集める原因となったCEPについて，その考え方や構成について説明する．

3.1 複合イベント処理

CEPの概念を最初に説明する．CEPとは「複合イベント処理」と訳され，ある条件を満たすような複数のイベントが同時に発生した際に，あらかじめ定めておいた処理を行うことを意味する．アルゴリズム・トレードの事例を用いて説明すると，複合イベント処理とは，例えば，

イベント1：企業Aの株価がX円以上になる，

イベント2：企業Bの株価がY円以下になる，

と定義したとき，イベント1とイベント2が同時に起こったとき，企業Cの株を売るという処理を行うことを指す．複数個の“イベント”が“複合”して発生したとき，事前に定めておいた“処理”を行うことが複合イベント処理である．イベント駆動アーキテクチャ（Event-driven architecture）と類似した概念であるが，さまざまな事象を対象とした複雑/高度な条件設定を想定しているのがCEPの特徴である．

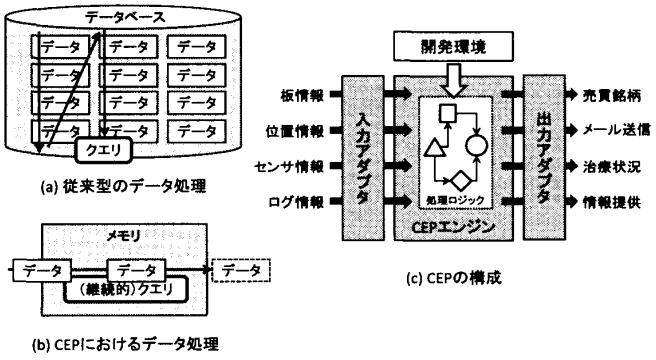

図3 CEPの処理イメージと構成

3.2 従来のデータ処理との違い

次に，従来のデータ処理とCEPにおけるデータ処理との違いについて説明する．従来のデータ処理では，データベースに蓄積したデータ1つ1つに対してクエリ（問い合わせ）を適用する（図3(a)）．これに対して，CEPでは，逐次発生するストリーム・データ1つ1つに対してクエリ（モデル）をメモリ上で逐次適用する（図3(b)）．ストリーム・データが発生するたびに継続的にクエリの適用が行われることから「継続的クエリ」と呼ばれる[1]．CEPによるデータ処理においてはデータベースを介する必要はなく全ての処理がメモリ上で行われるため，低遅延な処理が可能となる．したがって，ストリーム・データに適用するクエリは，メモリ上に格納できるサイズである必要がある．

データ処理の観点からはクエリと表現されることが多いが，データ分析の観点からいうとクエリとはストリーム・データに適用するモデルに他ならない．これは，CEPに関する研究開発が，主にデータベース技術の研究者によって行われてきたことによる．

3.3 CEPの構成

現在，さまざまなCEP製品が提供されているが，Stanford大学など大学での研究成果[7]をベースに開発されたものが多い．したがって，源流が限られているためにCEPの構成はどの製品でも大体同じようなものとなり，大きく次の3つのコンポーネントからなる（図3(c)参照）：

①入力アダプタ：ストリーム・データを取得し，CEPエンジンが処理可能なデータ形式に変換する処理を行う．金融における板情報や株価データなど，CEPへの取り込みが事前に想定されるデータについては入力アダプタを事前に用意している製品もある．入力アダプタが無い場合には，製品に付属のAPIを

用いて独自に開発する必要がある．

② CEP エンジン：入力アダプタによって形式変換されたストリーム・データに対して，事前に設定した処理を実行する．各製品には製品独自の開発環境が用意されており，入力データの時刻同期などを気にする必要なく，具体的な処理内容のみを記述するだけで良い．大雑把にいうと，入力ストリームのデータ形式，処理の内容，出力ストリームのデータ形式を記述するだけで分析内容をコーディングすることができる．

③出力アダプタ：CEP エンジンで処理した結果を，出力先が受取り可能な形式に変換する．入力アダプタと同様，想定される出力先についてはアダプタを事前に用意している製品もある．アダプタが無い場合には独自開発する．CEP はルールの適用結果を出力するのみで，適用結果に基づいて具体的なアクションをとる場合には，別の仕組みが必要であることに注意する．

ここで，②の分析内容のコーディングに関して，多くの製品については CQL（Continuous Query Language）や CCL（Continuous Computation Language）などと呼ばれる各製品特有の専用言語を用いて処理内容を記述する必要がある（一部で標準化の動きが見られるが活発ではない[11]）．ただし，SQL と非常に似通った言語である場合が多く，SQL の経験者であればそれほど苦労することなく処理コードが書けるようになる．

その他，製品によっては各種の統計処理を行う関数があらかじめ用意されているものもあり，高度な分析内容が簡単に記述できるようになっている．また，過去に作成した C/C++ や Java のコードを UDF（User Defined Function）として CEP エンジンから呼び出して利用できる仕組みを用意している製品もある．各製品ともバージョンアップが継続して行われており，可用性の向上や統計関数の組み込みなど充実化が図られている．

4. CEP の取り組み事例

2 節で紹介した事例は，海外での事例や検討段階のものであり，国内での実施例はまだ少ないというのが現状である．ここで最後に，CEP を用いて実際に構築した異常検知サービスを紹介する．金融分野以外での CEP によるストリーム・データ分析の実施例としては，国内での先駆け的な取り組みであるといえる．最初に，異常検知サービスが動作する具体的なシステムについて述べた後，我々が独自に開発した異常検知ロジックと試験的に導入した際の結果を簡単に紹介する．

4.1 橋梁モニタリングシステム（BRIMOS®）

CEP を用いて構築した異常検知サービスは，NTT データが提供している橋梁モニタリングシステム（BRIMOS®：BRIdge MOnitoring System）[12] 上の一機能として動作する．ここで BRIMOS® とは，高速道路などの橋梁の各所に取り付けセンサから，センサを取り付けた箇所の変位や加速度などの物理量を逐次取得/監視することで，その橋梁に異常が生じたかどうかをリアルタイムに判断するシステムである（概要を図 4 上部に示す）．

具体的には，橋梁に取り付けた光ファイバセンサから 125 Hz/1 Hz の周波数で波長データを取得し，物理量への変換式を逐次適用することで，4 つの物理量（変位，加速度，傾斜，ひずみ）を収集する．そして，収集した物理量に対して，以下の分析を行う．

1. 桁姿勢監視，
2. 振動/応力長期解析，
3. 車重/車種推定．

リアルタイムに橋梁を監視することで，災害時の異常検知や平常時の早期異常把握，さらには車両通行状況の解析による点検・補修の優先度検討などを支援する．

4.2 CEP による異常検知システムの構成

上に示した 3 つの分析機能に追加する形で，CEP を用いて異常検知機能を開発した．具体的には，図 4 下部にあるとおり，波長データをストリーム・データとして CEP エンジンに取り込み，CEP エンジン内において，物理量への変換式，および，異常検知ロジックを適用し，検知結果をディスプレイ画面に出力する構成とした．

4.3 遅延相関に基づく異常検知アルゴリズム

現在数多くの異常検知手法が提案されているが[13]

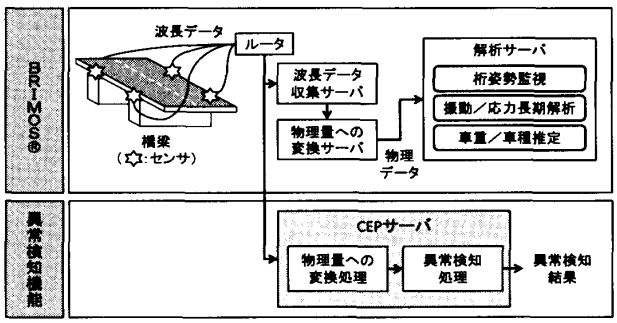

図 4 BRIMOS® と CEP による異常検知機能の構成

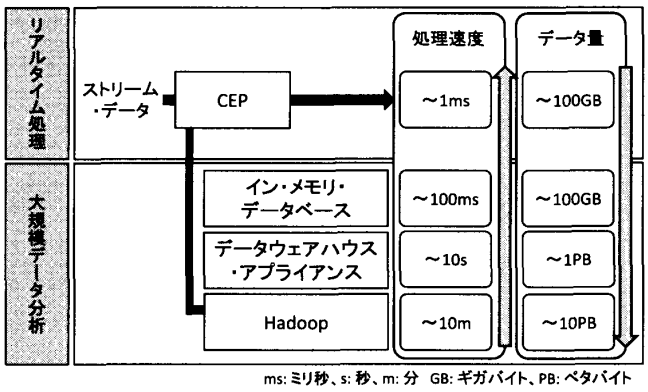

図5 大規模リアルタイム・データ分析基盤

～[15]，橋梁の各所に取り付けたセンサから物理量が得られることを考慮し，遅延相関[16]に基づく異常検知アルゴリズムを独自に開発した．具体的には，変位量を対象に，異常値が発生していないかどうかをセンサごとに判断するだけでなく，センサ間の遅延相関の関係を見ることで，橋梁の動きを総合的に見て異常が発生したかどうかを判断する．今回の実施例では，事前に定めた仮説のもと，センサの故障と思われる異常や地震や風などの外力が原因と思われる異常といった複数の異常タイプをあらかじめ設定しておき，それらの中から1つの異常タイプを随時選択するモデルを過去の履歴データから構築した．

開発した異常検知機能を実際の橋梁に適用した結果，処理速度の面では，異常検知ルールの適用に約0.1秒かかった．これは，変位量の発生間隔である1秒と比較して十分な処理速度である．また，異常検知精度の面では，実際に地震の発生した時刻における異常検知結果を見てみると，外力が原因と思われる異常タイプとして異常検知されていることが確認できている．

5. 今後の展望とまとめ

本稿では近年注目を集め始めたストリーム・データ分析と，それを実現するための基盤技術であるCEPを紹介した．今後，CEPの適用事例が増え始めることで認知度がさらに高まっていくであろう．そして近い将来，ストリーム・データ分析をベースとしたサービスは，情報活用サービスの1つの大きな柱となる可能性を大いに秘めている．

最後に今後の展望に関して述べると，分析対象となるストリーム・データの規模や分析目的に合わせて，CEPを含む統合的なデータ分析基盤を利用したサービスが生まれてくるものと思われる（図5参照）．つまり，リアルタイムに分析する必要がある場合にはCEPを利用し，長期的な傾向を分析する場合にはイン・メモリ・データベースを利用するなど，データの規模や求められる処理速度に合わせてストリーム・データを分析することで，より深い文脈理解に基づくサービスが提供されていくことが想定される．

他には，数値データ（構造化データ）とテキストデータ（非構造化データ）を合わせたストリーム・データ分析が今後行われていくことも考えられる．ただし，数値データとテキストデータを合わせたデータ分析技術は，現段階では新サービスの創出につながるレベルにはまだ至っていない．世間にインパクトを与えるようなストリーム・データ分析を実現するためには分析手法の発展を待つ必要があるであろう．

参考文献

[1] HONDA広報発表：
http://www.honda.co.jp/news/2011/4110428.html

[2] S. Chakravarthy and Q. Jiang, Stream Data Processing: A Quality of Service Perspective, Springer-Verlag, 2009.

[3] ITアーキテクト，“大量のデータをリアルタイムで監視して処理する複合イベント処理「CEP」,” Vol. 23, pp. 128-141, IDG, 2009.

[4] 月刊DBマガジン2009年8月号記事，“Products Focus「Sybase CEP/Sybase RAP-The Trading Edition」,” pp. 186-189，翔泳社，2009.

[5] 日経コンピュータ2010年9月15日号記事，“徹底取材CEP（複合イベント処理），” pp. 78-81，日経BP社，2010.

[6] 日経コンピュータ2011年4月14日号記事，“大量の情報を即時分析するストリームコンピューティング，” pp. 78-82，日経BP社，2011.

[7] D. Luckham, The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems, Addison-Wesley Professional, 2002.

[8] 田中隆博，“データ分析に基づく売買執行を実現するアルゴリズム取引，” 人工知能学会，Vol. 24, No. 3, pp. 376-384, 2009.

[9] ドコモ・ワンタイム保険保険案内メール・サービス：
http://www.nttdocomo.co.jp/service/convenience/onetime_insurance/about/index.html

[10] NTTデータ技術開発本部ビジネスインテリジェンス推進センタ編著：BI（ビジネスインテリジェンス）革命，NTT出版，2009.

[11] N. Jain, S. Mishraet, et al., “Towards a streaming SQL

standard," In Proceedings of The Vldb Endowment, Vol. 1, No. 2, pp. 1379-1390, 2008.
[12] BRIMOS®: http://www.nttdata.co.jp/services/s090421.html
[13] V. Chandola, A. Banerjee and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, Vol. 41, Issue 3, Article 15, July 2009.
[14] 山西健司, データマイニングによる異常検知, 共立出版, 2009.
[15] T. Ide, A.C. Lozano, N. Abe and Y. Liu, "Proximity-Based Anomaly Detection using Sparse Structure Learning," Proceedings of 2009 SIAM International Conference on Data Mining (SDM 09), pp. 97-108, 2009.
[16] Y. Sakurai, S. Papadimitriou and C. Faloutsos, "BRAID: Stream Mining through Group Lag Correlations," ACM SIGMOD Conference, pp. 599-610, Baltimore, Maryland, June 13-16, 2005.