様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月３１日現在

機関番号：１２６０８
研究種目：挑戦的萌芽研究
研究期間：2010～2012
課題番号：22650017
研究課題名（和文）　ウェブ上の大規模ストリーミングデータを用いた実世界リアルタイム分析基盤
研究課題名（英文）　Real-time physical world analysis with large-scale web streaming dat
研究代表者
鈴村　豊太郎（SUZUMURA　TOYOTARO）
東京工業大学・大学院情報理工学研究科・客員准教授
研究者番号：70552438

研究成果の概要（和文）：

本研究では、複合的な Web サービスから、公開 API を通して得られる情報をリアルタイムに分析することによって、実世界の動向のリアルタイム把握を可能にするシステム StreamWeb を構築した。また、効率的かつ高性能なデータ処理基盤を構築すると共に、様々なストリーム処理系の形態や GPU の活用による高速化、クラウドとの協調処理に関する研究を行った。

研究成果の概要（英文）：

We propose a real-time Web monitoring system called “StreamWeb”. The StreamWeb system allows developers to easily describe their analytical algorithms for a variety of kinds of Web streaming data without worrying about the performance and scalability, and provides real-time and scalable Web monitoring for massive amounts of data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,700,000	0	1,700,000
2011 年度	700,000	210,000	910,000
2012 年度	500,000	150,000	650,000
年度			0
年度			0
総　計	2,900,000	360,000	3,260,000

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：Ｗｅｂサービス　　ストリームコンピューティング

１．研究開始当初の背景

近年、Web においては、マイクロブログの Twitter や通常のブログサービス、Mixi や Facebook などのソーシャルネットワーキングサービス（SNS）、写真共有サイト Flickr など、多くのデータが公開 API を通して、ほぼリアルタイムに近い形で取得できるように時代になってきた。　これらのサービスを利用するユーザーは膨大であり、例えば、

140 文字の制限のある“つぶやき”サービスを提供する Twitter では現在 4600 万人(2009 年 8 月）が全世界で利用されていると言われている。 また、データ量も膨大であり、Twitter サービスが提供している Streaming API と呼ばれる API を用いてデータを取得した場合、最も多くデータが間引かれている場合においても 1 日 1GB は越えるが、実際にはその 10 倍～100 倍のデータが流れている予測される。

２．研究の目的

本研究では、これらの複合的な Web サービスから、公開 API を通して得られる情報をリアルタイムに分析することによって、実世界の動向のリアルタイム把握を可能にするシステムを構築する。例えば、Twitter の例であれば、Twitter から出てくるリアルタイムな情報を収集することによって、各ユーザーにとってその発言があまり重要でないと思われる発言（例えば、風邪っぽい、熱がある、病院にいく、などの発言）でも、集合知を利用することによって、例えば、インフルエンザの兆候や、経済状況（例： 失業した。。。などの発言）などの世界の状況をリアルタイムに把握することが可能になる。また、Twitter の発言と Twitter ユーザーのプロフィールの現在地情報を組み合わせることによって、Google Map の地図上に、指定したキーワード（例えば、インフルエンザ）を表示するシステムを構築することによって、どの地域でどのようなことが起きているかをリアルタイムに表示することができる。本研究においては、スケーラブルな実行処理系の構築手法を研究し､かつ Twitter だけなく多様な情報ソースを活用したより高精度な実世界のリアルタイム分析手法に関する研究を行う。

３．研究の方法

膨大なウェブデータをリアルタイムに実行するソフトウェア処理系を構築する。活用する Web サービスからのストリーミングデータを増やせば詳細な分析が可能になるが、ソフトウェア処理系として、リアルタイム分析を可能にするには、分散計算環境を利用した高いスケーラビリティと低いレイテンシが要求される。我々は既存のストリーム処理系(例: IBM System S）を拡張する形でこれを実現する処理系を構築する。

また、Twitter などの膨大なストリームデータから、インフルエンザや流行などの現実世界のトレンドをリアルタイムに分析する仕組みを実現する。応用としては、インフルエンザなどの予兆の発見、リアルタイムトレンドの発見、地震やゲリラ豪雨のリアルタイム検知、犯罪キーワード検知、リアルタイムの経済指標（失業率）など様々な応用が考えられる。 また、一つの情報ソース（例えば、Twitter） などでは精度の低い情報がある場合があるので、それらを複数の情報ソースを活用することによって、情報を高精度化する。

本プロジェクトでは、ストリーム処理系の処理形態や最適化技術などこれまでに行われた来なかった挑戦的な研究テーマに取り組んだ。

まず、GPGPU （汎用的なグラフィックプロセッシングユニット）をデータストリーム処理系に適用する研究は新規性が高く、Web などのリアルタイムにおいても、例えば、重い行列演算などに適用することによって、低レイテンシを実現することができる。これによって、Twitter の発言から、リアルタイムに類似するユーザーを発見するために必要な重い内積計算などを高速化することができる。

また、Web における膨大なデータにおいては、時間帯やイベント発生の有無によって到着するデータレートが大きくばらつく可能性があり、Bursty な状況にどう対処するか、そして、低データレート時にどうマシンを有効活用するかがソフトウェア処理系としては課題がある。

次に、マシンの有効活用のための技術として、ストリーム処理とバッチ処理の動的な資源割り当て、および 高データレート時には処理を間引く Load Shedding 技術に関する研究を行う。Load Shedding 技術とは、ストリーム処理において、リソース（計算資源、ネットワーク資源）に対して処理が間に合わない場合に、処理を間引く技術である。確率的に間引く手法が最も単純であるが、我々は処理が間に合わない場合には、SSD などの高速ストレージに一旦退避し、比較的データレートが低い時に SSD 上のデータの計算を行う手法を提案する。また、データレートによって、ストリーム処理（オンライン）とバッチ処理（オフライン）に割くリソースを動的に変化させることによって、システム全体の可用性を向上させる技術の研究開発を行う。

４．研究成果

「Web 上の大規模ストリーミングデータの実世界リアルタイム分析基盤」を実現する効率的かつ高性能なデータ処理基盤を構築するため主に以下の４点に関する研究を行った。各研究テーアの概要を成果を述べる。

(1) データストリーム処理とバッチ処理の

統合実行環境に関する研究
データストリーム処理ではメモリ上データのみを参照し逐次処理を行うため、データ全てを参照するような詳細な処理は行えず、必然的に従来型のデータを全て蓄積してから処理を行うという、MapReduce モデルを具現化した Hadoop のようなバッチ処理との共存が必要となる.また、データストリーム処理が対象とするアプリケーションでは、負荷が予測不可能に変動するものが多いが、負荷に関わらず応答速度をなるべく低減しなければならない.本研究では、データストリーム処理と Hadoop によるバッチ処理を統合実行し、動的負荷分散機構と負荷予測機構によって負荷に応じてそれぞれの処理に適切な計算資源を動的に与えることで、計算機の実行効率の向上とデータストリーム処理の応答速度の向上が可能となるシステムの提案と構築、評価を行った.この統合実行環境により、台数を固定して実行した場合に比べ、データストリーム処理のレイテンシの発散を防ぎつつ、計算環境の CPU 使用率を 47.77%から 72.14%にまで向上させた。

(2) GPGPU による変化点検知の高速化： ストリーミングデータに対してリアルタイムに計算処理などの操作を行う事が出来るデータストリーム処理においてレイテンシが特に求められるアプリケーションである異常・変化点検知アルゴリズム SST (Singular Spectrum Transformation)の実装と，高性能計算分野において比較的安価かつ低消費電力で高いパフォーマンスを発揮する事で注目を集めている GPGPU による性能最適化を行った．SST で最も計算のボトルネックとなっている SVD 演算 (Singular Value Decomposition)を GPU にオフロードする事で処理全体の速度向上を図り，SST のウィンドウサイズ 1000 において 12.44 倍の高速化を達成した.

(3) クラウドを用いた Elastic なストリーム処理系の実現
データストリーム処理を用いたリアルタイム性の高いアプリケーションを運用する場合、データレートの増大に対して、レイテンシの発散を抑えることが重要である。しかし現実的には、計算資源は有限であるため、データレートが一定以上になった場合に、レイテンシが発散することは避けられない。本研究ではそのような状況下で、クラウド環境上に処理を委譲することで、レイテンシの発散を抑える手法及びアーキテクチャを提案した。一方で、クラウド環境を利用するには経済的なコストを要するため、アプリケーションのレイテンシと経済的コストのトレードオフが発生する。本研究ではこれを最適化問題として扱い、コスト最適なスケジューリングを実行することで、クラウド環境の利用コストを最小化する手法も同時に提案する。また、そのシステムを実クラウド環境である Amazon EC2 を用いて構築し評価を行い、常にクラウド環境を利用する構成と比べて、経済的コストを 80%削減することに成功した。

(4) Load Shedding による近似計算を補完するデータストリーム処理システム
データストリーム処理においては,Twitter ストリームに対するリアルタイム自然言語処理システムのように，大容量のストリームデータを低レイテンシで処理することが求められるようなアプリケーションがある．しかし，例えば Twitter においてツイートの量が一時的に増大するような時，入力データレートがシステムの処理能力を超えてしまい処理レイテンシが増加すると，アプリケーションのサービスレベル・アグリーメントを満たせない場合がある．そのような過負荷時に入力データの一部を削除する Load Shedding という手法があるが，レイテンシを確保する代わりに計算精度は落ちてしまい，後に別のアプリケーションで同じ計算結果を利用したい時に問題となってしまう．我々は Load Shedding によって削除されるデータと不完全な計算結果の両方をストレージに保持し，システムの処理能力に余裕があるときに前者のデータを再度読み込んで処理を施し，後者の値と集約することによって計算結果を補完する処理機構を提案する．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１１件）

1. 雁瀬優、上野晃司、鈴村豊太郎 「グラフ分割を用いた大規模 2 部グラフのデータストリーム処理」 情報処理学会論文誌（ACS論文誌第 35 号)，2011 年（査読付き）
2. 石井惇志、鈴村豊太郎「クラウドを利用したElasticなデータストリーム処理」情報処理学会 SACSIS 2011 先進的計算基盤システムシンポジウム(査読付き)，2011 年 5 月（査読付き）
3. Shunsuke Nishii and Toyotaro Suzumura, ″Design and Implementation of Distributed Speech Recognition

System″, CUTE 2012, 2011/12, Korea（査読付き）

4. Tomoaki Oiki and Toyotaro Suzumura, ″Tonegawa: Highly Scalable Distributed Web Server with Data Stream Processing″, IEEE ICC 2011(International Conference on Communications) Workshop T2 FutureNet IV, 2011/6 （査読付き）
5. Toyotaro Suzumura and Tomoaki Oiki, ″StreamWeb: Real-Time Web Monitoring with Stream Computing″, IEEE ICWS (The 9th International Conference on Web Services), 2011/7, Washington DC, US（査読付き）
6. Atsushi Ishii and Toyotaro Suzumura, ″Elastic Stream Computing with Cloud″, IEEE CLOUD 2011 (The 4th International Conference on Cloud Computing), Research Track, 2011/7, Washington DC, US（査読付き）
7. 森田康介、鈴村豊太郎 「データストリーム処理を用いた変化点検知アルゴリズムSSTのGPUによる性能最適化」電子情報通信学会技術研究報告. DE, データ工学 110(107), 19-24, 2010-06-21（査読付き）
8. 松浦紘也、鈴村豊太郎 「データストリーム処理とバッチ処理における動的負荷分散」電子情報通信学会技術研究報告. DE, データ工学 110(107), 69-74, 2010-06-21（査読付き）
9. 西井俊介、鈴村豊太郎 「データストリーム処理系を用いたスケーラブルな音声認識」 インターネットコンファレンス 2010 , 2010 年 10 月 （査読付き）
10. 石井惇志、鈴村豊太郎 「クラウドを利用したElasticなデータストリーム処理の検討」 インターネットコンファレンス 2010 WIP セッション （査読付き）、2010 年 10 月
11. 松浦紘也、雁瀬優、鈴村豊太郎 「データストリーム処理系System SとHadoopの統合実行環境」情報処理学会第 22 回コンピュータシンポジウム COMSYS 2010 （査読付き） , 2010 年 12 月

〔学会発表〕(計５件)

1. 西井俊介、鈴村豊太郎「データストリーム処理によるインクリメンタルグラフ処理に向けて」電子情報通信学会データ工学研究会（査読なし）, 2011 年 5 月 30 日慶応大学
2. 松浦紘也、上野晃司、鈴村豊太郎「データストリーム処理におけるGPU統合型CPUの予備的評価」, 情報処理学会 HPC(High Performance Computing) 研究会, 2011 年 10 月 6 日、京都大学
3. 上野晃司、鈴村豊太郎 「データストリーム処理におけるGPUタスク並列を用いたスケーラブルな異常検知」 GTC Workshop Japan 2011, 2011 年 7 月 22 日、東京ミッドタウンホール（六本木）
4. 竹野創平、上野晃司、雁瀬優、鈴村豊太郎 「Wikipedia の編集履歴を用いた大規模 2 部グラフのデータストリーム処理」, 情報処理学会 HPC(High Performance Computing) 研究会, 2011 年 11 月 29 日、北海道大学

６．研究組織

(1)研究代表者

鈴村 豊太郎（SUZUMURA TOYOTARO）
東京工業大学・大学院情報理工学研究科・客員准教授
研究者番号：70552438

(2)研究分担者

なし

(3)連携研究者

なし