データストリーム処理基盤を用いた高速プローブ情報収集・分析

中村 暢達・喜田 弘司
藤山 健一郎・今井 照之

要　旨

自動車（移動体）からセンターにアップロードされるデータを大量に収集し、その分析結果を活用した高度なテレマティクスサービスへの期待が高まりつつあります。その実現のためには、多数のデータソースから、高速にデータを収集、分析する情報通信技術が重要です。本稿では、このような大規模データ収集・分析を実現するデータストリーム技術について紹介します。さらに、本技術を適用し、5万台の車両からの位置・速度データを処理し、渋滞情報をほぼリアルタイムに提供するシステムを試作し、技術検証しましたので、その結果を報告します。

キーワード

●データストリーム ●大規模データ収集・分析 ●テレマティクス ●渋滞

1. はじめに

情報通信技術の発展により、いつでもどこでも高度な情報サービスを享受できるユビキタス情報社会が到来しつつあります。自動車においても、各種車両センサーデータ、ドライバーデータを利用したテレマティクスサービスが拡大してきています。車両の位置、速度などのデータをセンターにアップロードし、センターで集計することで、各道路の渋滞状況を把握することが可能となります。渋滞状況に基づいて経路案内するなどのサービスは、ドライブが快適になるだけでなく、省エネルギー、環境負荷軽減など社会的にも大きなメリットがあります。

このようなサービスを実現する上で、各車両データを収集・分析するような情報通信処理が不可欠ですが、車両数が多くなると、その処理負荷は大きく、システム構築に多大なコストを要します。NECにおいては、多数のデータソースから、効率的にデータを収集、分析する基盤の研究開発を進めており、この基盤を使うことで、大規模なテレマティクスサービスを低コストに構築することが可能となります。

本稿では、このような大規模データ収集・分析基盤であるデータストリーム処理基盤の技術を紹介し、また本基盤を用いて試作した渋滞情報提供システムについて述べます。

2. データストリーム処理基盤

ログデータ、プレゼンス情報など、たとえ1つのデータは小サイズ（数百バイト）であっても、数百万箇所からデータを収集すると、秒あたり数十万件、数十万メガバイトのデータをセンターで処理することになります。通常のデータ処理の場合、このデータをデータベースに格納し、バッチ処理を行います。大量データを処理するには、高性能計算機が必要となり、センターシステムの構築には多大なコストがかかります。さらに処理すべきデータが増大すれば、想定時間以内にバッチ処理が終了しなかったり、データ溢れが発生し、データを消失するなどの問題が発生します。

そこでNECでは、**図1** に示すように、データベースを用いずに、データを貯めることなく、順次データをパイプライン的に分散処理するデータストリーム処理技術の研究開発を進めています。データを収集する過程で、データをサンプリング、フィルタリング、クレンジング、統計計算などの処理をあらかじめ行い、センターサーバの負荷を軽減します。

図2 にデータストリーム処理基盤のアーキテクチャを示します。各処理単位をノードと呼び、そのノードを組み合わせることで、1つのアプリケーションを構築します。各ノードは、スレッドであり、非同期、独立に動作しますが、ノードマ

ネージャと呼ばれるモジュールが、処理の開始・停止、処理速度の制御など全ノードの管理を行います。

ノードは、同一マシン内にあっても、異なるマシンにあってもよいです。異なるマシンにあるノード間で通信する場合には、送受信用のノードを使って、処理データの受け渡しを行います。同一マシン内にあるノードにおいては、共有メモリを使った本基盤独自のメモリ管理を使って、データの受け渡しをします。このメモリ管理は、小さなサイズのデータを逐次処理することに最適化しており、高速なデータの受け渡しおよびキュー管理を実現します。

現時点では、本基盤はC++のクラスライブラリの形で提供しています。開発者は、このクラスライブラリを利用し、ノードごとにプログラムを実装します。各ノードでは、処理のトリガーとなる条件の指定と、データ処理のプログラムを記述します。そして、ノードの接続を指定しますが、複数の接続先を指定することも可能です。たとえば分岐することで、1つのデータソースで異なるサービス、データのアーカイブを並列的に行う、あるいは集約することで、統合分析するなどを実現できます。

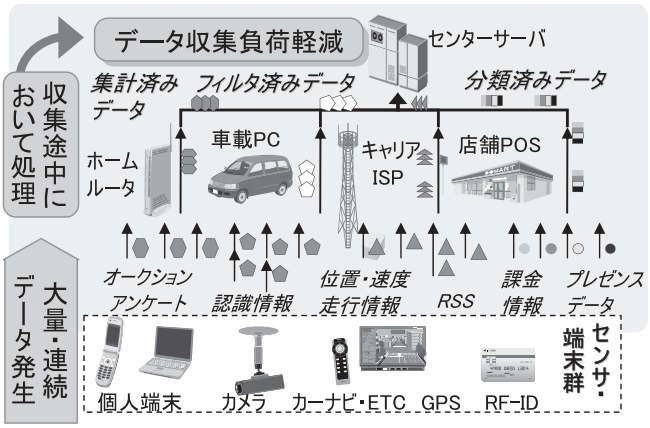

図1　データストリーム処理

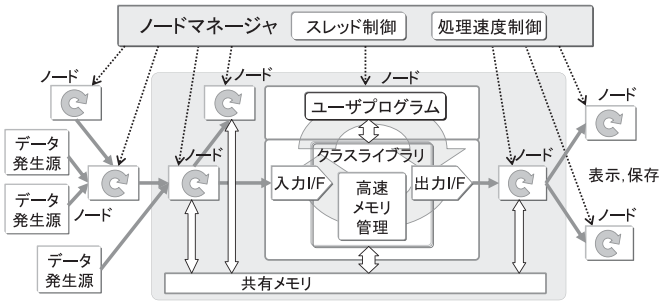

図2　データストリーム処理基盤アーキテクチャ

3. 渋滞状況提供システムの試作

上述のデータストリーム処理基盤を用いて、大量の車両・移動体から、位置、速度情報を収集することで、きめ細かく、かつ鮮度の高い渋滞状況を提供するシステムを試作しました。

従来、渋滞状況を把握するには、道路にセンサーを設置して情報を収集、加工する方法が一般的でした。一度設置すれば、対象道路の情報は確実に収集できる利点がある一方、費用の点からすべての道路に設置することは難しく、提供範囲のきめ細かさという点では課題があります。しかし、現在GPS(Global Positioning System)および無線通信（携帯電話）は広く利用されており、各車両の位置・速度のデータを容易に収集できるインフラは整いつつあります。本システムでは、各車両・移動体から、位置・速度などのデータを収集することとします。

渋滞情報を生成するために、各車両の位置・速度のデータを次のように処理します。

1) 車両情報の通信
2) モザイクマッチング方式による各車両の地図データ上の道路への対応付け
3) 計算コスト可変近似方式によるデータのサンプリング
4) 各道路区間に対する渋滞度の算出

モザイクマッチング方式とは、道路ネットワークをあらかじめ緯度経度のグリッド分けした格子に対応付けをしておき、車両の位置(GPS)データと道路データとを高速に照合する方式です。走行道路区間、交差点、車線（上り、下り）の判定を行います。

計算コスト可変近似方式とは、収集データの解析速度・精度の調整のために、解析する収集データのサンプリングレートを変更する手法です。可能な限り情報量を減らさない、つまりサンプリングを行わないことが望ましいですが、通常、同じ道路を同時に走っている車両のデータは類似しており、サンプリングによりデータを捨てたとしても解析に影響は少ないです。そこで、道路を地理的な接続関係でグルーピングし、解析に十分なデータがあるグループはサンプリングレートを低くします。

図3　渋滞状況提供システム画面例

上記の方式を含む大規模渋滞状況提供システム試作しました。奈良県奈良市および生駒市の約2,000本の道路からなる道路ネットワーク上に、市販の交通シミュレーションソフトウェアNETSIMを使って仮想的に車両を走らせて、その位置、速度、方向を解析サーバに送信します。 解析サーバではマップマッチングを行い、道路区間ごとの渋滞度を算出します。試作システムの画面例（拡大図）を**図3**に示します。地図上の数字および道路の描画の色で平均速度を表します。太線が渋滞、太点線がやや低速、細線が順調であることを示します。

4. 性能評価

本システムの性能を評価するために、1台のPC(CPU: Pentium4 3.0GHz、RAM: 512Mバイト、OS: Windows 2003 Server)において、1) 道路数を2千本に固定し、車両台数を変化させた場合と、2) 車両数を3万台に固定し、道路数を変化させた場合の実験を行いました。その結果を**図4**および**図5**に示します。

従来方式では、モザイクマッチング方式を用いずに、単に照合処理する方式であり、車両台数、道路本数の増加に対し、両方とも線形に処理時間が増加します。一般に、処理対象とする地域が広がれば、車両台数と道路本数ともに線形に増えますので、対象地域の広さに対する従来方式に基づくシステムの処理時間オーダーは$O(n^2)$となります。一方、本方式は、車両台数の増加に対しては処理時間が線形に増加しますが、道路本数に関しては処理時間にほぼ変化はありません。また、車両台数に対する変化も、従来方式に基づくシステムに比べて、約15％程度になります。よって、対象領域の広さに対する提案方式に基づくシステムの処理時間のオーダーは道路本数を無視できるほどグリッドサイズを小さくすることで$O(n)$と考えることができ、従来方式に基づくシステムと比べ、高い

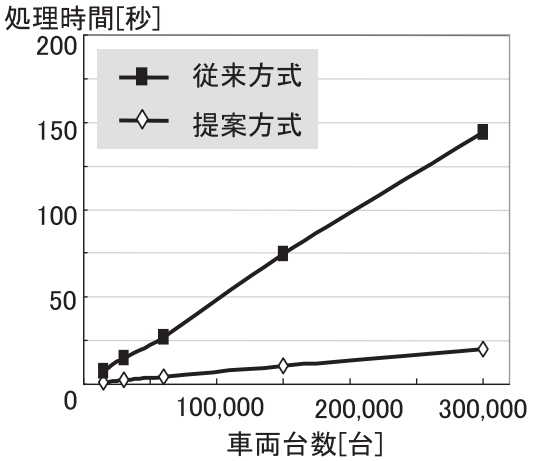

図4　車両台数を変化させた場合の処理時間の変化

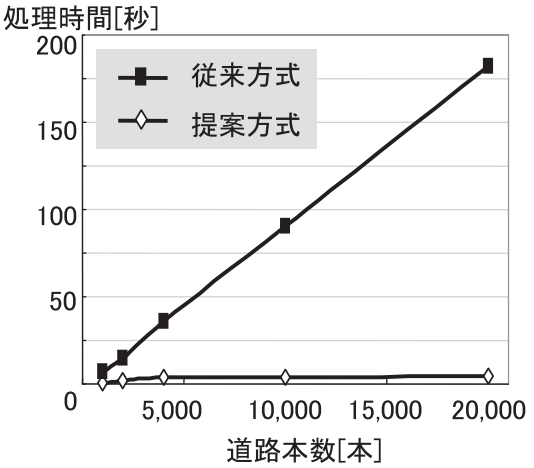

図5　道路区間数を変化させた場合の処理時間の変化

スケーラビリティを持つことが分かります。

5. おわりに

本稿で説明したデータストリーム処理技術に関連する研究としては、MIT、ブラウン大のAuroraプロジェクト、スタンフォード大学のSTREAMプロジェクトが著名です。これらの研究においては、クエリ言語、処理アーキテクチャ、分析アルゴリズムの研究が盛んですが、理論の研究が中心です。一方、製品としては、Apamaが金融分野への適用でよく知られていますが、大規模適用での処理性能に関する実績は知られていません。また、内部統制に関連して、ログ収集関連製品の動きが活発ですが、これらは、収集に特化しており、収集データを分析し、リアルタイムに活用することは考えられていません。

データストリーム処理基盤は、大量かつ連続発生するデータを収集・分析する技術に関し、上記関連技術に対し、計算機資源の効率性、高スループット性で優位であることをめざし、研究開発中のものです。これまでに、渋滞状況提供システムにおいて、秒5万件の車両データをマップマッチング処理し、基盤の性能をシミュレーション実証してきました。今後は、基盤そのものの機能・性能の向上を進めるとともに、実証実験を実施し、実用性を確認していく予定です。大量データの収集・分析という特徴を活かし、高度なITSサービスの実現に貢献したいと考えています。

執筆者プロフィール

中村 暢達
サービスプラットフォーム研究所
主任研究員

喜田 弘司
サービスプラットフォーム研究所
主任

藤山 健一郎
サービスプラットフォーム研究所

今井 照之
サービスプラットフォーム研究所