# 研究レポート　No.6　～大規模データの解析手法～

2014 年 2 月 4 日　株式会社アイズファクトリー　http://bodais.jp/company/

## 概要

インターネット技術、センシング技術の発達などにより、日々膨大なデータが生成・蓄積され、ビッグデータと呼ばれている。このビッグデータを処理・分析する手法としては、どのようなものが存在するのであろうか。本稿では、大規模データの処理手法について、基本的な手法であるサンプリングについて解説したのち、近年主流となっている分散処理について概説する。

## 1.　はじめに

「ビッグデータ」とは、膨大なデータ量の増加とデータ形態の複雑化により、「既存の技術では管理するのが困難な大量のデータ」と考えられている。量的には、典型的なデータベースソフトウェアが把握、蓄積、運用、分析できる能力を超えたサイズのデータになる。質的には、ビッグデータを構成するデータの出所が多様であることに特徴がある。GPS での位置情報、絶え間なく変化し続ける温度等のセンサーデータ、カスタマーデータといった、様々なデータがある。そういった各データを連携させることで、更なる付加価値の創出も考えられる。異変の察知や近未来の予測等から、利用者個々のニーズに即したサービスの提供、業務運営の効率化や新産業の創出等が可能となる点に、ビッグデータの活用の意義があるものと考えられている[1]。こういった背景から、ビッグデータの解析ニーズが高まっている。本稿では、大規模データの解析手法について解説する。

## 2.　サンプリング

データ分析を行う場合、データ規模が大きいと、「労力」、「時間」、「コスト」、「処理スペック（マシン、解析ツール）」などの壁に突き当たる。例えば、視聴率を調べるために、全世帯の TV 視聴状況（全数調査）を調べることはできない。データの規模が大きすぎると、データ処理時間が膨大になる（分析できない）。国勢調査は、正確さが求められるため、全件調査であるが、結果発表まで数か月かかる。そのため、現実的な範囲内で、データ処理を行う必要がある。現実的なデータ処理手法として、サンプリングがある。ここでは、代表的なサンプリング手法であるランダムサンプリングについて解説する。

ランダムサンプリングとは、母集団から、一部対象者を抽出し、標本だけでなく母集団について一定の誤差範囲で推定する方法である。特徴として、統計学に基づき、誤差の範囲が把握でき、精度の保証を与えることができる点が挙げられる。母集団の全てが同様にサンプルに選ばれるチャンスを持っていて、完全にランダムに抽出される。これにより、偏りのないサンプリングが可能となり、母集団の推定を行うことができる。

・単純ランダムサンプリング：母集団からくじ引きの様に無作為に抽出する。 概ね乱数表を用いる。精度が高い抽出法であるが、母集団が大きくなると一連番号を振るだけでも大変な作業であり、実行は極めて煩雑かつ困難になる。

・系統抽出（等間隔抽出法）：母集団の最初のサンプルだけをランダムに抽出して、そこから等間隔にサンプリングを行う方法。乱数を発生するのは、最初のサンプル抽出のときだけであるので、単純ランダムサンプリングよりも手間が少ない。しかし、母集団の配列のなかに何らかの周期性があり、その周期が抽出間隔と同じ（整数倍）である場合、特殊な個体だけが選ばれる偏った標本になってしまう［2］。

このような問題点はあるものの、現実的な問題を解くときには、サンプリングしたデータを用いてモデルを構築することは、実用的にも十分意味のあることである。構築したモデルは全データに適用して予測などに活用される。ビッグデータに対しても、データをサンプリングすることで全データを分析したものとほぼ同じ結果をリーズナブルに導出することが可能という意見もある［3］。

## 3.　分散処理

Amazon は、購入される頻度の少ない商品に対する注文にも応えることで、成功を上げている。この例にもみられるように、ロングテールも加味した分析に対するニーズが高まってきている。ロングテールも考慮することになると取り扱うデータは大規模なものとなる。大規模データ処理の速度面、容量面の限界を克服すべく活用されているのが、分散処理の技術である。分散処理とは、データ処理手法の一種であり、プログラムの個々の部分が同時並行的に複数のコンピュータ上で実行され、それらが互いに通信しあう形態である。ペタバイトクラスのデータになると、ディスク装置の応答速度の制約から単体のマシンでの取り扱いは現実的ではなく、例えばマシンクラスタによる分散処理が必要となってくる［4］。

CAP 定理は、分散コンピュータシステムのマシン間の情報複製に関する定理である。ノード間のデータ複製において、同時に、一貫性（Consistency）、可用性（Availability）、分断耐性（Partition-tolerance）の保証を提供することはできない［5］。この定理によると、分散システムはこの 3 つの保証のうち、同時に 2 つの保証を満たすことはできるが、同時に全てを満たすことはできない。しかし、現実には、この定理の範囲内で問題に対応している。

ビッグデータの分散処理には、「分散したデータ」を「常に素早く」「深く分析」することが求められる。「分散」「速い」「深い」という 3 つの要件は基本的にトレードオフの関係にある。不確実で曖昧さが残る中で迅速な判断を下すニーズに応えることと同時に、常に判断材料を集め続けて、遅滞なく判断するように設計しなくてはならない。この問題点を克服する技術が昨今注目されるようになってきている。

## 4.　分散処理の代表的技術

Google は 2003 年に大規模分散ファイルシステム“Google File System (GFS)”についての論文［6］、2004 年に大規模分散プログラミングモデル“MapReduce”についての論文［7］をそれぞれ発表した。これらの内容をもとに、2005 年に、当時 Yahoo! Inc.のエンジニアであった Doug Cutting 氏によって開発が進められ、現在は、Apache Software Foundation (ASF) が開発・公開しているオープンソース・モデルで実装したソフトウェアが、“Hadoop”である。Hadoop は、テラバイト、ペタバイトといった巨大なデータを処理するための分散処理基盤として、数多くの企業や研究組織

が、実際に活用している。Hadoop を用いたビッグデータの機械学習のためのライブラリが “Mahout” である。

分散データの分析ツールとして最も注目されているのは Hadoop であるが、これよりも優れた機能を持つ別のツールもある。“Spark” はカリフォルニア大学バークレー校 AMP Lab (Algorithms, Machines, and People Lab) で、大規模で低遅延のデータ分析アプリケーションを作成するために開発された。インメモリー・コンピューティングの基本要素を備えたスケーラブルなデータ分析プラットフォームであるため、Hadoop に優るパフォーマンスを発揮する場合がある。対話型のクエリのみならず反復的なワークロードも最適化することができる。また、Hadoop ファイルシステムと並行して実行することができる [8]。

Hadoop では、バッチ処理にフォーカスして処理を行う。この処理モデルは、多くのケース（Web の索引付けなど）に十分対処できるが、極めて動的なソースからリアルタイムの情報が必要となるケースには対応できない。ビッグデータのリアルタイム処理とは、絶え間なく発生することが見込まれるストリーミング処理（メッセージ処理やデータベースのリアルタイム更新）をすることができるということを意味する。また、連続的に発生するデータを、継続的に処理（連続的なクエリの実行や、クライアントへ処理結果をストリーミングで表示すること）をし続ける必要もある。これを実現した技術が、Twitter 社が開発した “Storm” である [9]。

分散処理環境での機械学習の実現には、深い分析とスケーラビリティを分離する必要がある。深い分析とは、機械学習に代表される集計を超えた複雑な解析（分類、回帰、近傍探索、推薦、異常検知、クラスタリング、グラフマイニングなど）を意味する。またバッチ型機械学習と同程度の精度を実現する必要もある。これを、コモディティサーバを並べた分散処理で行わなくてはならない。従来の機械学習エンジン単体では、正確な分析は行えるものの、データサイズは小～中規模で、バッチ処理、個別の開発が必要なケースが一般的である。これに対し日本発のビッグデータ分散処理環境である “Jubatus” では、誤差を許容範囲内に抑えつつ、データを高速処理する仕組みを実現している [10]。ビッグデータとその分析手法に対して、不確実な多くの課題が偏在する昨今の外部環境において、曖昧ながらも意思決定を下すというニーズに対し、リアルタイム分析を対象とし、さらにオンライン機械学習による深い分析という付加価値を提供するフレームワークである。株式会社 Preferred Infrastructure と NTT ソフトウェアイノベーションセンタにより、開発された。「深い分析」と「スケーラビリティ」の両立は、すべてのサーバで同時刻に同じ結果が返される保証をせず、緩やかにモデル情報が共有され、時間が経つにつれ、同じ結果が返されるような仕組みで実現されている [11]。

分散処理の代表的技術を、大規模データ蓄積、バッチ機械学習、ストリーム処理、分散機械学習の観点から比較し表 1 にまとめた。

| | Hadoop | Spark | Storm | Jubatus |
|---|---|---|---|---|
| 大規模データ蓄積 | ◎ HDFS | ◎ | ○ | 対象外 |
| バッチ機械学習 | ○ | △ | × | ○ |
| ストリーム処理 | × | × | ◎ | ○ |
| 分散機械学習 | ○ Mahout | △ | × | ◎ |

表 1. 分散処理の代表的技術の比較

## 5. 分散処理の事例

Yahoo! はビッグデータ処理に、Hadoop を活用している。Yahoo! JAPAN に蓄積されるビッグデータは多岐に渡る。コンテンツのカテゴライズやページビュー（PV）、クリック履歴のデータ、オークションやショッピングの商品の情報や販売履歴、Twitter からのつぶやきなどもある。月間ページビューが 507 億 PV、ピーク時のアクセス数が秒間 5 万件あり、そうした規模のデータが蓄積され、それを処理するために、国内最大級という 3500 台の Hadoop クラスタを運用している [12]。

## 6. まとめ

ビッグデータの活用でこれまで利用できなかった情報を引き出すことができるようになってきた。しかし、データの量が増えても、背後に潜むパターンを見出すことができなければ、大量のデータから知見を得ることはできない。やみくもにデータを解析すれば、ビジネスの競争力が向上するというわけではない。

データアナリティクスの分野で、2012 年、大きなブレークスルーとなったディープラーニング（深層学習。Deep Learning.）という新しい技術もある。ディープラーニングは、様々なデータの組合せにより表層のデータには見えない、データの背後の隠れた組み合わせで生じる概念を含む階層的構造を意味する。この手法は、テキストマイニング、音声認識、画像認識、化合物活性予測など多くの分野のコンテストで今までの手法を大きく凌駕し、中には人間の認識率を超える結果を出しているものもある。様々なデータの組合せから生まれるビッグデータならではの分析と言える。今後、本稿で解説したビッグデータ処理手法を皮切りに、新たなスタンダードとなる技術の開発が求められていくであろう。

## 7. 参考文献


[1] 総務省　情報通信白書　第 1 部　特集　ICT が導く震災復興・日本再生の道筋　第 2 章　「スマート革命」が促す ICT 産業・社会の変革

[2] 福井武弘　「標本調査の理論と実際」日本統計協会 (2013)

[3] 吉永恵一　リクルート流ビッグデータ活用術　「“ビッグデータ分析” は本当に必要か？」

[4] ウィキペディア「分散コンピューティング」

[5] ウィキペディア　「CAP 定理」

[6] S. Ghemawat, H. Gobioff and Shun-Tak Leung , SOSP '03 Proceedings of the nineteenth ACM symposium on Operating systems principles, pp.29-43 (2003).

[7] J. Dean and S. Ghemawat, Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation, pp.10-10 (2004).

[8] M. Tim Jones, 「Spark: 高速なデータ分析のための新たな手段」, developerWorks, IBM (2011)

[9] M. Tim Jones, 「Twitter Storm でビッグ・データをリアルタイムに処理する」, developerWorks, IBM (2012)

[10] 堀川ら, 「オンライン機械学習並列分散処理フレームワーク Jubatus」, NTT 技術ジャーナル 24(10), 30-35, 2012-10-00, 電気通信協会.

[11] 岡野原ら, 「大規模リアルタイム解析エンジン Jubatus の創り方」, 情報処理学会デジタルプラクティス 4 (1) 20-28 (2013).

[12] インプレス, インターネット・ウォッチ, 「ヤフーが日々蓄積するビッグデータの塊、3500 台の Hadoop で処理し地道に活用」