D-033

データ解析における並列分散処理基盤 Hadoop の利用
Using Hadoop for Data Analysis

大石 哲也† 橋本 司‡ 長谷川 隆三†† 藤田 博†† 越村 三幸††
Tetsuya Oishi Osamu Hashimoto Ryuzo Hasegawa Hiroshi Fujita Miyuki Koshimura

1. はじめに

インターネットの普及により，利用可能なデータの量は増大し，世の中には情報があふれている．これらの大規模データを管理し，高速に処理することはこれまでより一層難しくなってきている．そのため，スケールアップによって一台のコンピュータの性能を向上させる手法や，MPI という分散インターフェースを利用する手法で性能向上を行ってきた (e.g.[2])．しかし，コンピュータの性能を向上させるには限界があり，性能向上にかかるコストも甚大になってくる．また，従来の分散インターフェースでは，開発コスト及び実行コストの面で十分な成果が得られていなかった．

そこで，本研究では，並列分散処理基盤 Hadoop に着目し，複数のコンピュータを利用した分散環境を構築することで大規模データへの対応を可能にし，処理速度の向上を考える．Hadoop はオープンソースで公開されている Java で実装されたソフトウェアであり，一般に入手可能なコモディティ・コンピュータを利用して容易に分散環境を拡張することが可能である．(e.g.[3]) 更に，MapReduce という並列分散のモデルを用いることで，高い開発効率を維持しつつアプリケーションの開発が行える．

本稿では，Hadoop を Wikipedia からのデータ抽出，Twitter のデータ解析，SAT 問題へ利用する方法を提案する．以下，2 章でシステムの概要について述べ，3 章でシステムが妥当であるかを調査するための予備実験について述べる．最後に，4 章で結論を述べる．

2. システム概要

2.1 コンセプト

本研究では，Twitter や Wikipedia 等の大規模データ解析の処理基盤として，Hadoop を応用する手法の開発と，SAT 問題へ Hadoop を適用することにより，従来のアプローチでは達成困難であった課題への解決方法の開発を目的とする．

アプリケーションは，利用されるシチュエーションによって，処理に与えられるデータの形やデータサイズが異なる．また，求められる処理性能も，データの規模や処理の対象によって変わってくることが予想される．つまり，データ解析を行うシステムでは，インプットとしてのデータの形式やサイズに柔軟に対応する必要があり，要求される解法によって最適な性能で処理を実行するスケーリング技術も求められる．

データの特徴を挙げると，Twitter のデータは，その一つ一つはサイズの小さなデータであるが，毎時 1000 万投稿(e.g.[4])とも言われる膨大な量のデータが発生し，そのデータが時系列で並んでいる．一方 Wikipedia では，数 GB～数十 GB のファイルサイズを一つのファイルで持つデータの集合であり，Twitter と比較するとデータ更新の少ない静的なデータという特徴を有する．また，SAT 問題のデータでは，取り扱う問題によってデータの規模が変わる場合があり，かつ，解法のアルゴリズム次第で複雑さが高度になるという特徴がある．

このような，求められるデータ形式やサイズ，要求される処理能力が，取り扱う分野に依存するという課題に対して，Hadoop の持つプログラミングモデルやスケールアウトの特性を利用することで解決を図る．

2.2 概要

本件研究における Hadoop を用いたシステムでの，Hadoop クラスターの構成と，その上層のアプリケーションの関係について図 1 に示した．Hadoop は OS 上に構築される並列分散処理基盤層に位置し，複数のアプリケーションを実行させることが可能なインフラストラクチャとして機能する．その Hadoop 層の上層で，本研究のそれぞれのアプリケーションが実行される．

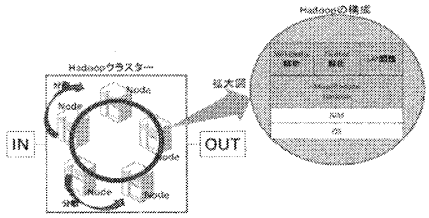

図 1 Hadoop のクラスターと Hadoop 内の構成

以降，実現を目指すシステムについて，想定している機能毎に詳細を述べる．

2.3 Wikipedia からのデータ抽出

近年，インターネットの普及に伴って，膨大な文書を閲覧することが可能となり，適切な文書を探すために検索エンジンを利用する機会が多くなっている．しかし，検索エンジンを利用しても，求める情報を得ることが難しい場合も多い．そこで，Web 百科辞書である Wikipedia を利用することで，適切な文書を探し出す手法が提案されてきている．(e.g.[1])

本研究では，検索単語の属する領域を Wikipedia から抽出する際，膨大なデータ量を扱う処理が必要になる

† 九州大学情報基盤研究開発センター Research Institute for Information Technology, Kyushu University

‡ 大亜電子株式会社 dia-electron Co., Ltd.

†† 九州大学大学院システム情報科学研究院 Faculty of Information Science and Electrical Engineering, Kyushu Universit

が，この領域において，Hadoop の並列分散処理を活用する手法を研究開発する．

2.4 Twitterのデータ解析

近年，急激に普及している Twitter では，ユーザー間で流される情報量も膨大になった．しかし，膨大な量の情報の中から，リアルタイムの高い有用な情報を得ることは容易ではない．興味のある分野の有用な情報だけを抽出することはさらに難しい．

これを解決する手法として，Twitter のユーザーを分野別(政治・IT・芸能・スポーツなど)にクラスタリングし，さらに，その中でも有用かつリアルタイム性の高い情報(流行)を多く提供しているユーザーを抽出することを考えた．

これらを実行する際，膨大な計算時間を要することが予想されるため，本研究では，大規模データを高速に処理する機能に Hadoop を用いる．

2.5 SAT 問題への Hadoop 利用

これまで本研究室では，モデル生成法に基づく独自の自動推論システム MGTP を開発し，準群の未解決問題の解決に成功する等により，評価を得てきた．しかし，検証対象の情報システムの複雑さや規模の増大も著しく，並列分散による自動推論のさらなる高性能化が必要とされている．実際に，回路検証・組み合わせ最適化問題等，記述量が大規模となる問題や探索空間が膨大となる問題が多数存在し，並列分散実行による効果が期待できる．

本研究では，並列分散処理基盤として Hadoop を利用し，従来のアプローチでは達成困難な課題への打開策を講じる．

3. 予備実験

3.1 実験方法

Hadoop 利用の有効性を確かめるため，いくつかの予備実験を行っていく予定である．本章では，その実験の概要について述べる．

以下のような観点で予備実験を行う．

(1) 実効性の検証

Hadoop クラスター上でのアプリケーション実行速度向上の台数効果や，大規模データを処理した際の性能劣化耐性等を検証する．

(2) 適用性の検証

データ形式やデータ特性などの観点から，MapReduce モデルで実装することが適しているかどうか，あるいはより高位の枠組みを利用すべきかどうかを検証する．

(3) 開発容易性の検証

開発効率を著しく落とすことなく，研究開発を進められるかを検証する．

3.2 考察

分散処理を使わずにデータ解析を行う場合，基のデータが大規模になると，現実的な時間では処理が困難になることが予想される．分散処理基盤の Hadoop を使うことで，処理時間を大幅に短縮することが可能になる．

また，MPI 等既存の分散インターフェースでは十分な成果が得られていない開発コストの面でも，Hadoop ではコモディティ・コンピュータを用いた分散環境の構築が容易であり，大規模データの処理や速度向上が達成できる点において優位である．

4. 結論

2 章で述べた課題の共通点は，以下である．

・形式の異なるデータの処理
・大規模なデータの処理
・高速な処理

形式の異なるデータを処理する点は，MapReduce を利用することで解決を目指す．MapReduce では，map 関数と reduce 関数を記述することでアプリケーションを開発できるが，この二つの関数を組み合わせると，Twitter 解析で求められる時系列処理や Wikipedia 解析で求められる文字列マッチング処理など複雑な処理も柔軟に記述できる．これにより，種々のインプットに合わせたアルゴリズムの実装が可能になる．

大規模データの処理および高速な処理の 2 点は，処理能力の向上を，コモディティ・コンピュータをシステムに追加するだけで実現可能という Hadoop の特徴を利用して解決できる．

今後は，Hadoop と上層のアプリケーションとを結ぶフレームワーク研究も視野に入れ，大量データの処理と高速性が要求される SAT、知的 Web 検索、化学組成分析等へ応用できる技術の研究も進めたい．

謝辞

本研究は科研費（21500102）の助成を受けたものである．

本論文は九州大学 システム情報科学研究院 長谷川・藤田研究室の研究成果と大亜電子株式会社の開発実績をまとめたものである．本研究の第 2 章の概要では，倉門氏および白木原氏に資料を提供して戴くとともに有益なご助言を戴いた．ここに両氏に対して感謝の意を表する．大亜電子株式会社の各位には研究遂行にあたり日頃より有益なご討論ご助言を戴いた．ここに感謝の意を表する．

参考文献

[1] Masada, T., Kanazawa, T., Takasu A., and Adachi, J., “Improving Web Search by Query Expansion with a Small Number of Terms”, NTCIR-5 Workshop Meeting, (2005).

[2] 太田寛，西谷康仁，“データ並列言語の通信生成方式とマルチグリッド法での最適化評価”，情報処理学会論文誌，Vol.42, (2001).

[3] 水野謙，菅沼俊夫，石崎一明，古関聰，上田陽平，小松秀昭，“並列トランザクショナルアプリケーションのためのプログラミングフレームワーク”，情報処理学会論文誌，Vol.48, (2009)

[4] “Google による統計データ” https://www.google.com/adplanner/planning/site_profile#siteDetails?identifier=twitter.com&geo=JP,(2010)