

Hadoop による時系列画像のための 汎用分散データマイニングシステムの構築

Costruction of distributed data mining system for time-series imagery using Hadoop MapReduce

西前光^{*1} 三好智也^{*2} 森 啓太^{*3} 本田理恵^{*4}
Kou Nishimae Tomoya Miyoshi Keita Mori Rie Honda

^{*1*}^{*2*}^{*3*}^{*4}高知大学
^{*1*}^{*2*}^{*3*}^{*4}Kochi University

The prototype of distributed data mining system for time-series imagery is designed and developed by using Hadoop and MapReduce. In extracting time-series data from those images, data transferred between Map and Reduce is decreased to about 3% of those in the conventional method by distributing the block of images instead of pixel and using compression. The scalability up to 50 nodes is confirmed for analysis of 1200 images with total amount of 2.4GB. The system is then applied to spatio-temporal correlation analysis of weather satellite images and its effectiveness is confirmed.

1. はじめに

蓄積された大量のデータ、いわゆるビッグデータからの知識発見が注目されている。ビッグデータとしては、一般的には、e-コマースのデータ、SNS、tweet などのテキストデータからのトレンド分析等がよく取り上げられるが、これらのデータには時々刻々と蓄積される画像も含まれ、その実例としては地球観測衛星の画像、様々な目的で所々に設置されたモニタカメラ、セキュリティカメラの画像があげられる。このような時間順に並んだ画像、すなわち時系列画像から特徴を抽出し、さらに時間、空間変動パターンに関する知識を取得することができれば、科学的発見や環境変動の予測のみならず、特定の環境での人々の行動パターンの理解、予測などの様々な分野に応用することができるであろう。また、時系列画像は大量のファイルとして蓄積されていることが多いため、その大規模な処理には、分散処理が有効と考えられる。本発表では、ビッグデータの解析環境として注目されている Hadoop を用いた大規模時系列画像の時空間分散データマイニングシステムの構築にむけた検討結果について報告する。

2. Hadoop, MapReduce

分散処理のフレームワークには Xgrid, Gfarm, Apache Hadoop などがあるが、近年注目されているのが Apache Hadoop [White 13] である。Apache Hadoop プロジェクトは、信頼性が高くスケラブルな分散コンピューティングのためのオープンソースソフトウェアを開発している。このソフトウェアには、共通モジュールである Hadoop common, 高スループットのアクセスを実現可能な分散ファイルシステムを構築する Hadoop Distributed File System (HDFS), ジョブスケジューラとクラスタのリソース管理を行う Hadoop YARN, そして YARN に基づいて分散処理をおこなう Hadoop MapReduce から構成される。また関連したプロジェクトとして分散データベースの HBase, データウェアハウスのインフラストラクチャである Hive, 機械学習ライブラリの Mahout などがある。これらによって大規模データの効率よい分散ストレージ、

分散処理、分散データベース、問い合わせなどの環境の実現が目指される。

図 1 に Hadoop で用いられる MapReduce の概念図を示す。MapReduce の処理は Map と Reduce の 2 つに分けられ、各データは <key, value> のペアで構成される。Map では key, value の組を作成し、Shuffle 処理で key 毎にデータをソートして複数のスレーブノードに割り当てる。Reduce 処理では受け取った <key, value> のペアに対して集約処理を行う。Map, Reduce にそれぞれどのような処理を割り当て、またどのような値を <key, value> に割り当てるかが、効率的なシステムを構成する鍵になる。

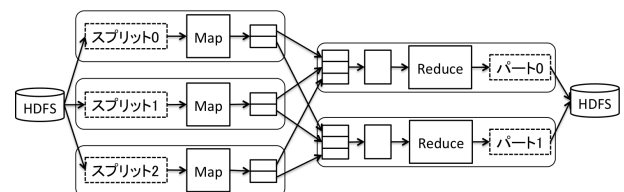


図 1: MapReduce の概念図

3. 想定するタスクとシステム

時系列画像からの知識発見には、(1) 全画像、全地点からの時系列データの抽出とその解析 (セグメンテーション, 相関分析, イベント検出, 予測), (2) 異なる時間の画像の特徴抽出とその分類, クラスタリング, (3) 画像からのオブジェクトの抽出と特徴記述, ならびにその追跡, といったタスクが含まれる。このようなタスクを大量の画像に対して実行するため、図 2 のようなシステムを構築する。ここでは数台のマスタと数十台のスレーブノードに対して Hadoop を実装し、HDFS による分散ストレージ, MapReduce による分散処理環境を実現する。マスタは HDFS を制御する NameNode と, MapReduce による分散処理を制御する JobTracker から構成される。入力には時系列画像を与えるが、簡単のため、空間的な配置はあらかじめそろえられ、等時間でサンプリングされているものとする。また時刻情報は画像名に含まれるものとする。

連絡先: 本田理恵, 高知大学 理学部 情報科学教室,
高知市曙町 2-5-1, 088-844-0111(1160), 088-844-8361,
honda@is.kochi-u.ac.jp

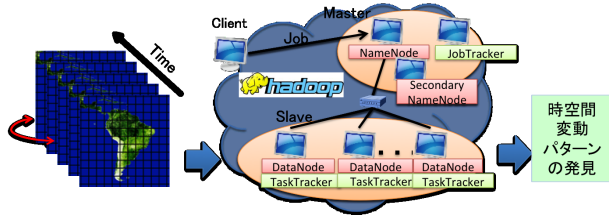


図 2: 想定する時系列画像の分散データマイニングシステム

タスク 1 の時系列解析を例にとると、まず画像を画素毎に分解して画素毎の時系列を抽出し、各時系列に対して処理を行うことになる。この処理は、1 段の Map, Reduce で実装することができるが、単純に $\langle \text{key}, \text{value} \rangle$ を $\langle \text{座標}, \{ \text{時間}, \text{画素輝度値} \} \rangle$ として、座標を key にして画素毎の時系列への集約を行うと、Map と Reduce 間のデータ流通量が膨大になり、スケーラビリティが確保できなくなる。画像は規則的な配置で配列されていることを考えると、分散の単位を画素とすることは冗長な処理といえる。

そこで時系列画像に適した分散化の手法として、画像をブロックに分割し、 $\langle \text{key}, \text{value} \rangle$ を $\langle \text{ブロック ID}, \{ \text{先頭座標}, \text{時刻}, \text{ブロックの輝度値野ベクトル} \} \rangle$ に割り当てて、さらに中間データを Hadoop 標準の圧縮アルゴリズム DEFLATE で圧縮することとした。これによって、Map, Reduce 間におけるデータ伝送の飽和によるボトルネックを改善することが期待できる（詳細は [西前ほか 14] 参照）。

4. 実験結果

実験には高知大学情報科学教室計算機システムの 53 台の iMac (intel Core 2 Duo 3.06GHz, 2 コア中 1 コアのみ使用, メモリ 4GB) を用いた。使用した Hadoop は version 1.2.1 で、レプリケーション数は 3 とした。地球観測衛星画像（植生指標画像）1152x1152 画素、各 16 ビット（実質 10bit）、1200 枚（時間方法のサンプリング点数）に対して、時系列を抽出し、平均、分散を計算する問題を扱った。この際、比較のために、(1) 画素を分散単位とする場合、(2) 画像のブロックを分散単位とする場合、(3)(2) にさらに DEFLATE で圧縮を行った場合について実験を行った。

実験の結果、MapReduce 間のデータ伝送量は (1) から (2) への変更で約 1/6 に、(2) から (3) の変更でさらに約 1/6 となり、合計で約 1/35 まで削減できた。さらに図 3 に示す通り、1-50 台のノード数で計算速度向上比（1 ノードに対する計算時間の比の逆数）を調べたところ、画素単位分散では 10 台程度で計算速度の向上比が頭打ちになっていたのに対して、ブロック分散と圧縮の併用により約 50 台までスケーラビリティを拡張できた。なお、ブロック分散圧縮ありのケースでの計算時間は 100 秒程度であった。これによって、Map, Reduce 間におけるデータ伝送の飽和によるボトルネックを改善し、スケーラビリティを回復することができることを確認した。

5. 気象画像への応用例

この結果を基に、気象衛星画像からの時空間相関分析にこのシステムを応用した。先行研究 [坂口, 本田 09] での同様のシステムの概念図を図 4 に示す。これは時系列画像中で、注目すべきある地点の部分時系列（基準時系列）があたえられたと

表 1: 使用計算機（マスター 3 台、スレーブ 50 台共通）の性能

諸元	値
プロセッサ	Intel Core 2 Duo (3.06GHz)
メモリ, HDD	4GB, 500GB
HDFS 用 HDD 容量	455GB(全システムで 23TB)
OS	Mac OS X 10.6.8
Hadoop version	1.2.1
Java version	1.6.0_45
ネットワーク	1000BaseT

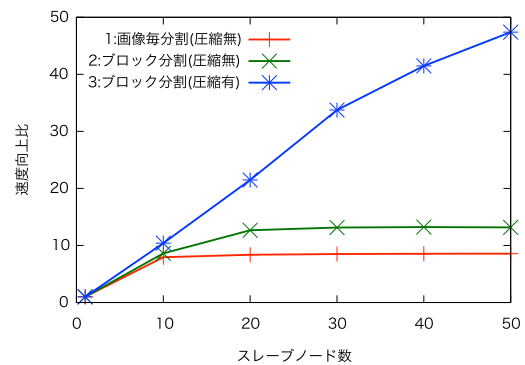


図 3: Map-Reduce 間データ削減時のノード数に対する計算速度向上比の変化

き、これと強い正または負の相関を持つ地点を網羅的に調べるものとする。この際、基準点と参照点の時間遅れも考慮する。これより、A 地点で特徴的な変動が起こったとき、数日後に数百 km はなれた B 地点で同様あるいは正反対の変化が起こるといったタイプのパターンの発見を行う事ができる。

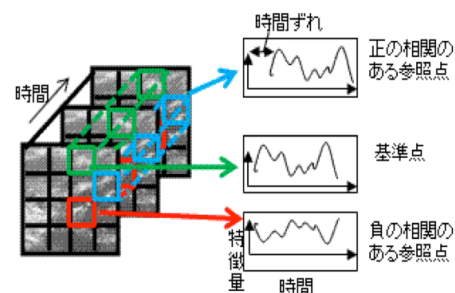


図 4: 衛星画像からの時空間相関分析概念図 [坂口, 本田 09]

高知大学気象情報頁 [高知大学 2014] にアーカイブされた運輸多目的衛星ひまわり MTSAT-1,2（ひまわり 6, 7）の IR1 画像を 2012 年 9 月から 12 月までの 2914 枚を用いた。画像からは 2 週間の基準時系列を取得し、時系列画像から抽出した様々な位置、開始時間の時系列から取得した部分時系列を参照時系列とし、その相関係数を計算し、正の相関、負の相関の高い箇所を可視化するものとした。なお、可視化の部分は現在は Hadoop の処理の対象外としている。

時系列抽出は前節と同様に1段の Map, Reduce で実施し、その後時系列画像を中間ファイルとして HDFS に保存し、その後、基準時系列を Distributed キャッシュに指定して、第2段目の Map で時系列毎の分散処理を行うものとした。なお画素を単位とするとノイズの影響を大きく受けたため、5x5 画素であらかじめ平均化を行う前処理をおこなった。抽出された相関係数の高いスナップショットの例を図5に示す。ここでは、画像中程に、正、負の相関係数の高い部分(それぞれ赤、青)が対になって現れていることがわかる。

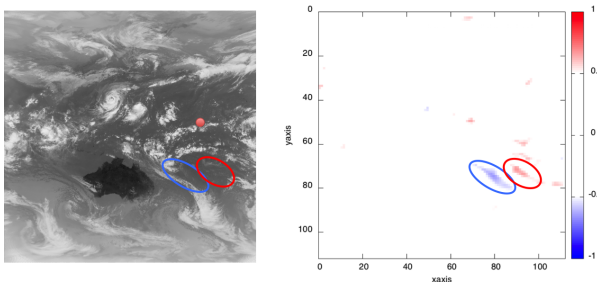


図 5: 相関係数の計算結果例。左側は時系列始点の画像(2012年10月13日02時、基準点より3日後、赤丸は基準点の位置)、右は基準時系列との相関係数(青が-1, 赤が1)。

図6に、基準点と図5で示した正負の相関係数が高い箇所の代表点の時系列(値の大小は画像の輝度:雲量を反映)をそれぞれ、緑、赤、青のグラフで示したものである。ここでは時間遅れは修正し、それぞれの時系列の始点を時間の減点に合わせている。これより正の相関の大きい箇所では基準点とほぼ同期した変動を示し、負の相関の大きい箇所では増減のパターンが逆転している事がわかる。このように、分散処理のフレームワークを利用することによって大量のファイルからも、複数の計算機をもちいてスケーラビリティを確保しながら、柔軟なデータの要約や特徴抽出を行い、時空間の知識発見を行うための基盤を形成することが可能になった。

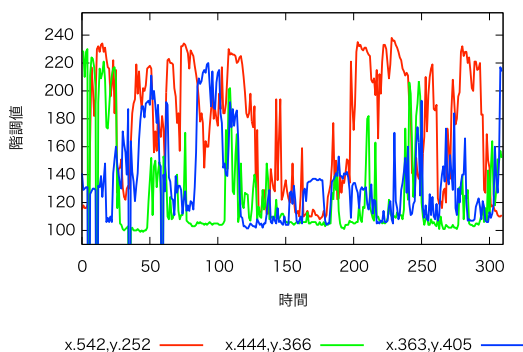


図 6: 基準地点(緑)と、図5の青、赤の囲み領域中心の時系列。

6. 汎用システムのデザイン

以上の検討を基に、想定する時系列画像からの汎用時空間データマイニングシステムのユースケース図を図7に示す。汎

用システムでは、様々な事例において様々な時空間の側面に注目した処理が行えるモジュールを備える必要がある。基本的には、タスクに必要なデータ抽出、前処理、特徴抽出、学習やモデリング、評価・可視化といったプロセスから構成される。今回のプロトタイプシステムでは、前処理、データ抽出、統計量の計算といったコアとなる部分を Hadoop で効率的に実装できることを確認した。しかし実際には試行錯誤によって適切なアルゴリズムを検討したり、前処理の部分にまでさかのぼって、ビンニングやサンプリングを行うことや、抽出した知識の一覧表示や時空間データの可視化も必要である。今後は汎用性に留意して全体的なシステムの詳細設計をすすめ、ユーザーインターフェースにも配慮したより汎用的なシステムとして構築していく予定である。

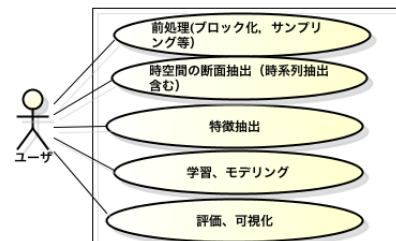


図 7: 汎用システムのユースケース図

7. まとめ

本研究では分散処理のフレームワークである Hadoop, MapReduce を用いて、大量の時系列画像から時空間の知識発見を行う分散データマイニングシステムの基礎として、時系列抽出とその処理について検討した。時系列画像に適した分散化の手法としてブロック分散を実装することにより、Map, Reduce 間のデータ伝送量をおさえ、スケーラビリティを改善することができた。また気象画像を対象にして時系列画像としての相関分析とその可視化によるインタラクティブな知識発見システムとしての実装を行った。今後は、より実践的な知識発見のアルゴリズムを実装し、可視化などのユーザーインターフェースも開発することにより、汎用的にインタラクティブに試行錯誤を通して知識発見を支援するシステムとしての開発を行うことが有望である。

参考文献

- [White 13] Hadoop, Tom White, オライリージャパン; 第3版 (2013)
- [西前ほか 14] Hadoop による時系列画像からの時空間画像分散データマイニングシステムの検討-気象衛星画像への応用-, 西前光, 三好智也, 本田理恵, DEIM Forum 2014 D1-6 (2014).
- [坂口, 本田 09] 坂口 祥太, 本田 理恵, “気象画像を用いた時空間変動における相関性マイニング”, 第23回人工知能学会全国大会論文集 (2009).
- [高知大 14] 高知大学気象管理グループ, 高知大学気象情報頁, <http://weather.is.kochi-u.ac.jp> (2014).