Japan Geoscience Union Meeting 2013
(May 19-24 2013 at Makuhari, Chiba, Japan)

MGI34-08 会場:301A 時間:5 月 19 日 14:30-14:45

Hadoop による時系列衛星画像からの分散データマイニング -GIMMS を用いた植生指標の年間変動モデリングへの適用
Distributed data mining from time-series satellite imagery using Hadoop -experimental study for modeling of GIMMS NDVI-

西前 光 [1*], 本田 理恵 [1]
Kou Nishimae[1*], Rie Honda[1]

[1] 高知大学
[1]Kochi University

衛星による地球・惑星の観測では，大量の時系列画像が時々刻々と蓄積されている。こうした大量のデータに機械学習やデータマイニング手法を適用する事によって，有用な時空間変動パターンを発見（マイニング）する事が期待されている。そのためにはデータから時間，空間などの必要な断面を柔軟に抽出し，パターン抽出やモデリングを高速に行うことのできるシステムが必要である。これには大量のストレージだけでなく，高速に大量のデータを処理できるシステムが必要となる。

こうしたシステムを実現するため，近年クラウドやグリッドなどの分散データ処理のフレームワークとして注目されている Hadoop, MapReduce もコモディティなコンピュータの集合体に実装して一般のユーザーにも使用しやすい時空間データマイニングの分散システムを構築することを検討している。テストベッドデータとしては，25 年分の NOAA 衛星のデータを修正して作成された GIMMS の植生指標指数データ (NDVI) を用い，ここから植生を反映した年間変動のパターンをロジスティック関数でモデリングすることを目的としている。また，得られたパラメータから植生の時空間分布の変動や green up などの変化点などの高次の知識を抽出するものとしている。

今回は，予備実験として,（1）参照点の時系列データ抽出と（2）抽出した時系列データに対する最尤法によるロジスティク関数のモデリングの 2 つのプロセスを MapReduce で実装し，前回の発表 (西前，本田，2012) よりも多い最大 50 台の iMac を用いて分散処理のスケーリング効果を検証した。ネットブートシステムの利用によりクライアントのセットアップ・管理は簡素化させ，各マシンのローカル HDD 未使用部分を利用して 23TB の仮想ファイルシステムを構築した。実験の結果，時系列データ抽出の場合は，30 台程度で処理速度の効率化が進まなくなり，コア数を増加させても処理時間に変化が見られなかったが，モデリングについては反復によるローカルでの計算時間が長いため，コア数，計算機数に対して計算効率はほぼ線形に増加する事が確認できた。前者はファイルアクセスがボトルネックと考えられるが，データ抽出に要する時間はモデリングに比べて 2 桁小さいため，全体システムとしては分散化による十分な高速化実現可能であると考えられる。

また，Hadoop 上で実行できる機械学習ライブラリである mahout を利用して高次の知識発見を試みた。Mahout の利用により, クラスタリング, 分類, レコメンド等を Hadoop 上で分散処理することができる。 今回はロジスティク関数のモデリングで得られたモデルパラメータを mahout の機械学習ライブラリを用いてクラスタリングし，その時空間分布を可視化することでの変動発見を試みた。この内容についても発表時に報告する。

キーワード: Hadoop, 分散, 時空間, データマイニング, 植生指標
Keywords: Hadoop, distributed, spatio-temporal, data mining, vegetation index