大規模台風時系列画像コレクションのためのマイニングとサーチング*

北本 朝展†
国立情報学研究所‡

1 はじめに

自然科学のさまざまな分野において、センシング技術の発達が大量の観測データを生み出し、それとともに大量のデータを網羅的かつ汎用的に収集した大規模データコレクションを基盤とする研究スタイルが芽生えてくることとなった。例えば生物情報学や宇宙科学の分野では、ヒューマンゲノム計画や宇宙サーベイ計画などのプロジェクトにおいてその傾向が顕著に現れているが、同様の傾向は地球科学においても進行している [1]。

地球科学における主要なセンシング技術の一つは地球観測衛星であるが、それが高解像度・多頻度・多波長のセンサを備えるようになるにつれて、地球観測衛星が送信する地球環境データの量も加速度的に増加した。例えば気象現象の観測においても、気象衛星「ひまわり」のような雲の分布を観測する衛星に加え、降雨の立体構造や風の強さと方向を観測可能な新しいタイプの地球観測衛星が登場することにより、大気のメカニズムはより明確に理解できるようになった。

ただし、こうして得られた大量の地球環境データも、ただ人間が眺めてテープに保管し棚に並べておくだけでは機動的な活用が不可能である。そこで我々は、地球上に生じる重大な現象を観測した地球環境データを網羅的かつ汎用的に収集した高品質な大規模データコレクションを作り上げることで、そのような重大現象に対して、後述するような「データコレクションを基盤とする研究スタイル」を適用してみることを考えた。その重大現象として本論文で選んだのが「台風」である。本論文は、このような動機のもとで構築した大規模台風時系列画像データコレクションに対する、サーチングおよびマイニングに関する研究の課題および現状について述べる。

*Mining and Searching for the Large-scale Collection of Typhoon Image Sequences
†Asanobu KITAMOTO
‡National Institute of Informatics, 2–1–2, Hitotsubashi, Chiyoda-ku, Tokyo 101–8430, Japan

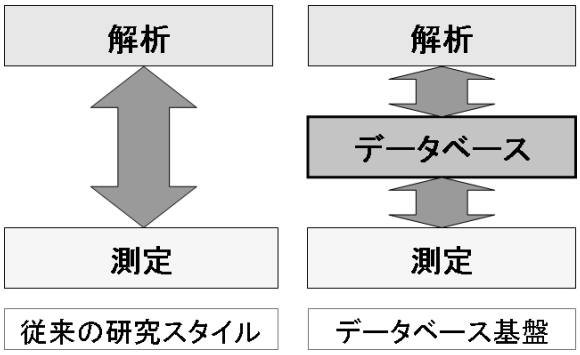


図 1: データコレクションを基盤とする研究スタイルの階層モデル。

2 データコレクションを基盤とする研究スタイル

最初に「データコレクションを基盤とする研究スタイル」について、図 1 の階層モデルの模式図を用いて説明する。まず従来のモデル（2 層モデル）では、データ測定とデータ解析を同じグループでおこない、両者が密接に関連するというスタイルのもとで研究が進んでいた。しかしデータコレクションを基盤とするモデル（3 層モデル）においては、データ収集者に相当するデータベース層を新たに加えることにより、データ測定層とデータ解析層とを分離することが可能になる。その利点は以下のようにまとめることができる。

1. 「測定」と「解析」という 2 つの作業を分業できるため、データを作る人とデータを使う人が密接に関わる必要がない。
2. すべての測定を一旦データベースに登録することで、同一の測定を複数人が無駄に反復せずにすむ一方で、複数人による繰り返し測定がデータの信頼性を高めることもある。
3. すべての測定結果がデータベースを経由することで、データの形式化が可能になるとともに、データ管理・操作のための見通しのよいアーキテクチャを実現できる。

さらにこの中間層に着目すると、データベース層には2つのインタフェースが存在することがわかる。まず、データ測定層へのインタフェースでは、データを収集してデータコレクションを構築するための設計が課題となる。一方、データ解析層へのインタフェースでは、データを世界中の研究者あるいは一般の人々に幅広く公開するための公共データベースの設計が課題となる。これら2つのインタフェースに求められるのは以下のような要件である。

1. 網羅性：多くの種類の解析に必要となりそうなデータを事前に測定しておくこと。
2. 一貫性：すべてのデータをできるだけ同等の品質および方法で測定すること。
3. 汎用性：特定のデータ解析方法に依存しないようなデータ定義方法を提供すること。
4. 多様性：特定のデータ検索方法に依存しないような柔軟なデータ操作方法を提供すること。

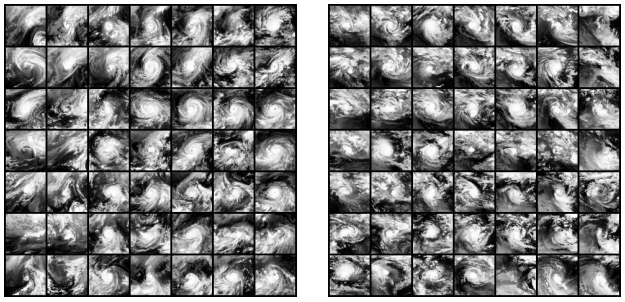
上記の課題は、上の2点が主にデータコレクションに関する課題、下の2点が主に公共データベースに関する課題である。これらの要件を満たすようなデータベース層の設計が、さまざまな挑戦的課題を含む重要な研究課題である。

では、本論文の主題であるサーチングおよびマイニングは、図1のどこの層に相当するのだろうか。図1では、データ解析層とデータベース層との区別が明確ではないが、本論文ではデータベース層は、形式化が可能なデータ操作を扱う層であると考える。ゆえに、サーチングの大部分はデータベース層、そしてマイニングについては、データベース層に含めるもの、およびデータ解析層に含めるもの、の両者が存在することになる。

以上のような研究課題を踏まえたうえで、我々のプロジェクト「デジタル台風」では、大規模な台風時系列画像コレクションの構築を継続している。以下ではこのコレクションの概要を紹介する。

3　台風時系列画像コレクション

本研究で構築している大規模台風時系列画像コレクションは、気象衛星「ひまわり」の衛星画像を対象とするものである。これは、西太平洋に発生する南北両半球の台風をすべて追跡して画像を蓄積するという意味で網羅的であり、またすべてのデータを同一センサ* ・同一縮尺で同一処理するという意味で、一貫性を保つものである。



(a) 北西太平洋　(b) 南西太平洋

図 2: 台風時系列画像コレクションの多様性。それぞれの画像コレクションに k-平均法を適用してクラスタリングした結果を示す。クラスタの代表画像は多次元尺度構成法を用いて 2 次元平面に配置されている。

コレクションの規模は、2003 年 1 月現在で、1995 年から 2002 年までの台風を対象におよそ 42,200 件（北西太平洋 31,200 件および南西太平洋 10,800 件）、台風系列数では 253 系列（北西太平洋が 183 系列、南西太平洋が 70 系列）に達している。その特徴は、時系列台風画像を表現するために、地球に固定した座標系から眺めるオイラー的表現ではなく、台風と共に動く座標系から眺めるラグランジュ的表現を用いた点にある。そして、台風中心が常に画像中心と一致するように台風画像を作成することで、台風雲システム全体の動きと台風雲パターン固有の動きとを分離して扱うことができる。その詳細については、他の文献を参照されたい [2]。

一方、この画像コレクションのための公共データベースについては、すべてのデータを原則公開するという方針でその準備を進めている段階であるが、原稿執筆時点でも、すでにデータ公開やデータ検索などが一部実現している（http://www.digital-typhoon.org/）。このデータに興味をもつ研究者がデータベースにアクセスし共通データを用いることにより、研究結果の共有や比較が進めば理想的である。このような公共データベースの重要性は、WWW (World Wide Web) の発達とともにますます高まると予想される。

*例えばこのような気象情報処理では、可視センサのデータをそれが有効な昼間だけ用いることで、昼間の解析精度を向上させるという処理が一般的であるが、データコレクションという観点からは、昼間と夜間で精度が異なるような処理は不適切だと判断する。

4 台風時系列画像コレクションのサーチング

台風時系列画像コレクションのサーチングとしては、類似（時系列）パターンのサーチングが最も重要な機能である。その理由は以下の通りである。

台風の強さや大きさを知るのに最も確実な方法は、飛行機などで台風中心まで飛行して必要な情報を実測する方法であるが、この方法は日本付近では現在用いられておらず、実際には気象衛星画像の台風雲パターンを人間が認識し解析するという方法（ドボラック法）が用いられている。この方法は、類似した雲パターンは類似した内部状態を表すという仮定に基づき、過去の類似パターンから得られた経験則を現在の状態推定に適用する、という枠組みである。この枠組みを参考にするならば、台風を解析するためには、類似（時系列）パターンのサーチングが重要な機能になる。

このような機能は、画像類似検索あるいは映像類似検索に関する過去の研究で頻繁に論じられたテーマであり、特に顔画像検索など、同一種類の画像を蓄積したデータベースに関するサーチングと共通する部分が大きい。ただし台風画像検索の場合、同一の雲パターンが出現する可能性はゼロであるため、同一人検索ではなく、「他人の空似」検索、つまり過去に出現した台風パターンの中から類似したものを検索する必要がある。ゆえにパターン間の類似度の定義が大きな問題となる。

類似度定義に関する第一の論点は、個々の観測パターンごとの類似度の定義である。理想的にはここに専門家の類似判定基準、つまり専門家がドボラック法を適用する際に着目する画像特徴を含めた類似度を定式化すべきである。そのような画像特徴には、目の有無や中心付近の雲の大きさ・厚さなどがあるが、これらは必ずしも定量的な判断基準ではなく、人間が認識しやすい（ゆえにコンピュータが認識しやすいとは限らない）画像特徴が選ばれる傾向がある。また類似性の判定は、熟練した解析者の主観的な作業に依存する部分が大きい [3]。そこで本研究では、台風雲パターンの統計的特徴からそのような画像特徴を見出すために、現在のところは統計的特徴の最も基本的なモデルである固有空間法を用いている [2]。ただし、この方法はパターンを一様に扱うため、微細な特徴（台風の目など）に対する選択性が低いという欠点がある。そこで複数スケールの画像特徴も表現可能なモデルについて研究を進めている。

第二の論点は、観測時系列パターンごとの類似度の定義である。すなわち、個々の観測パターンごとの類似度を拡張し、連続する観測パターンの集合どうしの類似度を定義することで、類似する時間変化をサーチングする機能を提供する。専門家にとっても、台風の 1 日程度の時間変化を動画で眺めることは、台風の時間変化の傾向を知るために有効な方法である。そこで本研究ではこのような時系列間の類似度を、Dynamic Time Warping (DTW)、つまり 2 つの時系列信号の時間軸上の伸縮を考慮した最適なアラインメント（整列）法によって計算する [4]。

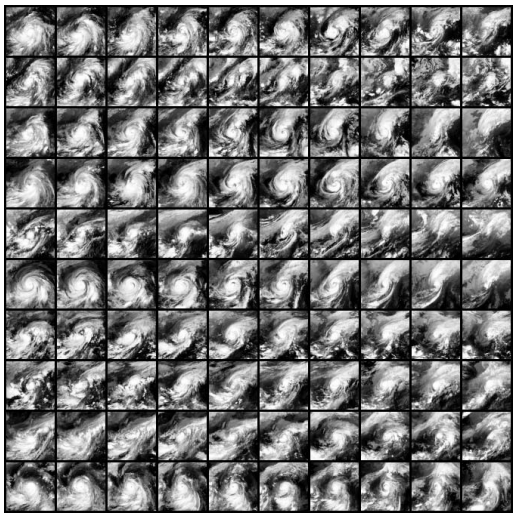


図 3: 複数台風時系列画像のアラインメント。

以上の論点に加えて重要なのが、画像コレクションをサーチングするためのデータ操作言語 (Data Manipulation Language : DML) の開発である。もちろん、これまでのデータベース研究で多数の DML が提案されており、SQL を代表として実用化された言語も多いが、それらは画像コレクションのサーチングという目的に必ずしも適合するわけではない。例えば台風時系列画像コレクションのサーチングにおいては、画像集合を一旦複数のグループに分割し、グループごとに類似画像を検索し、その上位の検索結果を収集して応答するという操作が必要になることがあるが、このような機能を関係データベースの操作言語で記述することは難しい。そこで、このような入れ子の問い合わせをきちんと記述するために、我々の目的に必要なデータ操作の体系化および DML の定義が必要である。このような、サーチングを主要機能とする簡易な DML およびその基盤となるデータモデルについて我々は研究を開始しており [5]、グループを対象としたデータ操作結果が再びグルー

プになるという意味で閉じているような、グルーピングを基本とする体系を構想している。このような DML の定義は、次章のマイニングにも関連する課題である。

5 台風時系列画像コレクションのマイニング

台風時系列画像コレクションのマイニングの目的は、台風の解析あるいは予測に有用な規則性あるいは不規則性を時空間データの中から発見することにある。そのような法則を発見するための方法として、気象学の標準的な方法である大気力学的アプローチによる物理過程のモデル化ではなく、確率的アプローチによるデータ生成過程のモデル化に基づいてマイニングすることを考える。

最も基本的な方法は、データのクラスタリングや分類によって、台風雲パターンの基本的な傾向を見る方法である。クラスタリングでは特に画像の変異や時間的推移を 2 次元平面上に並べて眺めることにより、その傾向を大まかにつかむことが可能である [6]。また、台風が発達するか衰弱するかを台風雲パターンという情報のみを用いることで分類することが可能であれば、台風雲パターンにはその傾向を見極めるだけの情報が出現していることになる。サポートベクトルマシンを用いることにより、実際には 73%程度の精度で、そのような分類が可能であることがわかった [7]。

さらに挑戦的な課題として、急速に発達する台風の予兆を発見するという課題がある [8]。急速発達する台風は、急速に強まる風や高波のために大きな災害につながることがあるため、このような現象が始まる前にその予兆を発見したいというニーズは大きく、それをもし雲パターンから発見できれば実用的価値は高い。この課題についてはまだマイニングに成功はしていないが、今後の課題として取り組んでいく予定である。

最後に前章で述べた DML の問題について、マイニングに関するデータ操作もこのような形式化を進めていく必要があると考える。というのも、サーチングがマイニングの一ステップである場合に、これらを統合して記述することの利点が大きいためである。先述の DML では、マイニングの多くの操作をグルーピング操作として捉えることにより、それらを言語として記述するための体系を提供することを目論んでいる。

6 おわりに

本論文では、我々が構築を続けている「台風時系列画像コレクション」を対象とするサーチングとマイニングについて、研究の課題と現状をやや総論的にまとめた。このような自然科学データを対象とする大規模公共データベースの構築、およびサーチング＆マイニング機能の提供という潮流は、今後も様々な分野でますます高まってくるに違いない。そのような状況に対して情報学は何が貢献できるのか。特にデータモデルやデータの形式化について貢献できることは多いはずであり、そうした観点からの研究を今後も続けていく計画である。

謝辞 気象衛星「ひまわり」画像を研究目的にご提供下さる、東京大学生産技術研究所の安岡教授、喜連川教授、ならびに根本助手に感謝いたします。なお、台風中心位置は気象庁の観測によるベストトラックデータに基づくものです。

参考文献

[1] Han, J., Altman, R.B., Kumar, V., Mannila, H., and Pregibon, D. Emerging Scientific Applications in Data Mining. *Communications of the ACM*, Vol. 45, No. 8, pp. 54–58, 2002.

[2] Kitamoto, A. Spatio-Temporal Data Mining for Typhoon Image Collection. *Journal of Intelligent Information Systems*, Vol. 19, No. 1,, 2002. 25–41.

[3] 鈴木和史, 元木敏博 (編). 台風 – 解析と予報 –, Vol. 197, **気象研究ノート**. 日本気象学会, 2000.

[4] 北本朝展. 台風時系列画像のマルチプルアラインメントに基づくデータマイニング. **電子情報通信学会技術報告**, Vol. PRMU2002-159, pp. 79–84, 2002.

[5] Kitamoto, A. IMET: Image Mining Environment for Typhoon Analysis and Prediction. In Djeraba, C., editor, *Multimedia Mining*, pp. 7–24. Kluwer Academic Publishers, 2002.

[6] Kitamoto, A. Evolution Map: Modeling State Transition of Typhoon Image Sequences by Spatio-Temporal Clustering. In *Discovery Science (DS2002)*, Vol. 2534, *Lecture Notes on Computer Science*, pp. 283–290. Springer, 2002.

[7] Kitamoto, A. Typhoon Analysis and Data Mining with Kernel Methods. In *Pattern Recognition with Support Vector Machines (SVM2002)*, Vol. 2388, *Lecture Notes on Computer Science*, pp. 237–248. Springer, 2002.

[8] 北本朝展. 台風画像コレクションからの予兆発見. **人工知能学会研究会資料**, Vol. SIG-FAI-A103, pp. 19–26, 2002.