ネットワークに基づく

分散型地球環境データベースの構築

に関する研究開発

最終成果報告書

国立情報学研究所

小野欽司

計　宇生

フレデリック・アンドレス

北本　朝展

平成14年5月31日

目次

第1章 序章

1.1 研究プロジェクトの背景・目標

宇宙から観測する地球環境データ，つまり地球観測衛星データの現状をたとえてみると，まずは華々しいデビューを飾って将来にわたる活躍が嘱望されたのに，その後の活躍は本来の実力に比べれば今ひとつという，悩めるスター選手にたとえることができるのではないだろうか．現状の活躍は確かに可能性の片鱗を見せてはいるが，その本来の実力を存分に発揮しているとは言えない，というのが我々の印象である．

地球観測衛星データは確かに様々な分野で大きな功績を残しており，社会的にも大きな影響を与えてきた．例えばオゾンホールを観測した地球観測衛星データがフロンガスの廃止に結びついたり，アマゾンの森林破壊を表す地球観測衛星データが熱帯林破壊の衝撃を人々に伝えたりと，地球環境データは地球の現状を伝え，それに対して人々が感じ判断するための重要な素材を提供している．このように地球観測衛星データが有用なのは，人間が地上を歩き見聞きしながら情報を収集するのとは全く異なるスケールと新しい視点で，地球そのものを一望のもとに眺めることができるからである．またこのような観測を継続的におこなうここと で地球の変化も知ることができるため，地球の変調を警告するという意味でも，社会的に大きな影響を与えうるメディアである．また近年においては，米国同時テロによって破壊された世界貿易センター跡地を捉えた超高解像度画像や，その後の戦いにおける超高解像度衛星画像の活用など，やや異なる意味でその有用性が証明されたのも記憶に新しい．

にもかかわらず，多くの人々が地球観測衛星データに対して抱く思いというのは，「どこか馴染みがなく実感が湧かない」というものではなかろうか．例えば地球環境データとして我々にとって最もなじみが深いものの一つに，気象衛星「ひまわり」による気象衛星画像がある．この気象衛星画像は地球の大気中の雲の動き，あるいはその背景にある地球の大気の流れや地球の温度分布も表している．このような情報が天気予報の高精度化に与えた影響は非常に大きい．例えば気象衛星がない時代には，太平洋上の台風がある日突如として日本に接近し，対策を施す間もなく大きな被害を出すことがしばしばあった．気象衛星画像によって地球規模で大気の観測ができる今日では，少なくともこのように突如として台風が接近するという失敗はもはや起こりえないものとなった．そして台風の現在位置や勢力を手に取るように確認することができ，その情報をもとに災害に対する警報を事前に流すことが可能になるなど，気象衛星画像が台風観測に与えたインパクトは実に大きなものがある．しかし天気予報を見る多くの人々にとっては，ひまわり画像は「雲のような白いものが流れていく動画」として天気予報番組で眺める画像としてはなじみがあるとしても，これが我々の身の回りで起こっている実際の大気をまさに表現したものであり，そこには多くの情報が含まれているのだ，という実感はそれほど強くないようにも思える．

一方研究者の間では，地球観測衛星データの活用は着実に広がっている．究極的には地球に関係す

る研究はすべて地球観測衛星データを活用する余地があり，その応用範囲は非常に広いと考えられるにもかかわらず，研究者の間でも現状の活用範囲はまだ狭いように思える．その主要な原因はアクセシビリティの悪さ，データ解析技術およびデータベースの不十分さにあると考える．まずアクセシビリティについて言えば，欲しい地球観測衛星データを見つけるだけでも大変な労力がかかると知れば，多くの人はデータを探す段階でデータを使うことを断念してしまうだろう．またこのようなデータがインターネットで簡単に検索・入手できなければ，現在ではその利便性に不満をもっても不思議ではない．また，地球観測衛星データの専門家でない限り，地球観測衛星データの処理プログラムやデータ解析アルゴリズムを自分で作りたいとは思わないだろうが，現状では研究者ごとに独自の処理プログラムやデータ解析アルゴリズムを開発する場合が多く，それがソフトウェア資源の共有には大きな障害となっている．また近年はいくつかの商用ソフトウェアが比較的充実した地球観測衛星データ処理環境を提供しているが，これらの商用ソフトウェアはいずれもかなり高価であり，本格的な業務をおこなう企業や大学でない限り，購入することは現実的でない．

そこで本研究は，ネットワーク上に分散した地球観測衛星データを，簡単に素早く，集め，調べ，判断するための情報技術について，特にネットワーク技術とデータベース技術の観点から研究することを目的とする．このような複合的な作業を支援するためには，複数の要素技術の組み合わせとそれらの融合を目指していくことが必要である．例えば分散型地球環境データベースの基盤がネットワーク技術とデータベース技術にあるとしても，これらを別々に研究していたのでは，分散型地球環境データベースという総合的な視点はなおざりにされてしまうことになる．ゆえに，地球環境データベースという観点から，各種の要素技術をコーディネートし，それらの間の結合や統合までを対象とするような，総合的な研究プログラムが不可欠である．本研究はそのような総合的なプログラムを枠組みとして，地球環境データベースに必要な情報技術の研究開発に取り組む研究プロジェクトである．

ここで問題となるのは，このような要素技術の独立性を高めるためのインタフェースの設計および個々の要素技術との関連性の定義である．この問題に対して本研究が提案する枠組みは，要素技術の関係を階層構造として捉える階層モデルの方法論である．また本研究では，地球環境データに関するデータベースの一般論を最初から展開するのではなく，むしろある特定の対象に関する研究をケーススタディ的に深めながら，地球環境データベースの本質を探っていくことを目論んでいる．そのような豊富な情報内容をもつ研究対象として，本研究では気象現象としての「台風」を選び，「台風データベース」を地球環境データベースのケーススタディとして研究することとした．7.2章で後述するように，台風という研究対象は情報学的観点から見ても本質的に興味深い研究課題を数多く含んでいる．そこでこの台風データベースを一種のテストベッドとして活用することにより，この研究成果を究極的な目標である「地球環境データベース」実現への第一歩とする計画である．

1.2 研究プロジェクトの計画

本研究プロジェクトは平成9年度から平成13年度までの5ヵ年にわたる計画であり，以下のようなプランで研究を進めた．

平成９年度

1．タイのアジア工科大学(AIT)に気象衛星NOAAの受信局が完成するのを機に，2Mbpsの国際回線を活用して，国際的規模の地球環境データ共有の可能性について探る(第3章)．

2. 地球環境衛星データベースの基本的な要素技術の研究をはじめるために，まず画像内容に基づく検索を可能とするような強力な画像データベースのアーキテクチャについて研究する（第6章）．

平成１０年度

1. タイとの地球環境データ共有を実現するためのステップとして，タイ国内のネットワークインフラストラクチャの整備を支援するとともに，ネットワークを経由した地球環境データの共有に関する問題点を検討する（第3章）．

2. 地球環境衛星データベースの階層的アーキテクチャに基づき，各レイヤにプラグインするような要素技術の研究を開始する．具体的には，画素レイヤにおいては，地球環境衛星データに特有の統計的性質に基づく画像分類法を提案し，また意味レイヤにおいては，利用者とのインタラクションを重視した進化的画像検索インタフェースを提案する（第6章）．

平成１１年度

1. 地球環境データ共有のための技術的課題として，大規模地球環境データのキャッシング技術について研究を進める．具体的には，地球環境データへの利用者のアクセスパターンに応じたキャッシング技術を研究する（第5章）．またタイとの地球環境データ共有についても，タイ国内のネットワークインフラストラクチャの整備を引き続き進め，これが実現可能なレベルとする（第3章）．

2. 地球環境データベースについては，引き続き階層的アーキテクチャの各レイヤの研究を進める．昨年度までのテーマに加え，領域レイヤにおいては形状や配置に基づく画像表現モデルを新たに開発し，これらを統合した画像データベースのプロトタイプの実現まで結びつける（第6章）．

平成１２年度

1. ウェーブレット変換(wavelet transform)に基づく完全再構成なフィルタバンク構造（S+P変換など）を用いて，地球環境衛星データの多重解像度表現を得る．これをキャッシング技術と組み合わせ，キャッシングの数学的理論について検討する（第5章）．またタイとの地球環境データ共有について，具体的な問題点を洗い出し実現に近づける（第3章）．

2. 前年度の後半から地球環境データの対象を台風に絞り込み，地球環境データベースの具体例として台風データベースを研究することに方針を変更した．画像データベースの階層的アーキテクチャの利点は，このように対象を変更しても全体的なフレームワークは同一で済むという点にある．そこで今までの研究成果を活用しつつ，台風画像コレクションの生成法や台風雲形状の表現法など，より台風という対象に適した手法を追究する（第7章）．

平成１３年度

1. タイとの間で地球環境データを共有するための実証実験をスタートする（第3章）．また地球環境データのためのマークアップ言語，すなわちデータ定義・操作のための言語を設計し，これをXML(eXtensible Markup Language)構文で記述することで，相互運用性に優れた地球環境データベースのメタサーバを実現する（第4章）．

2. 台風画像データベースの研究をさらに進め，データマイニングの方向性から台風データを徹底的に研究する．ここでは，統計的な機械学習・データマイニング手法を導入し，大量の台風画像コレクションの中から規則性・不規則性を発見することを目標とする（第7章）．同時に上述のメタサーバにデータベース機能を組み合わせ，最終的には研究成果はWWWを用いて広く一般に公開する（第8章）．

1.3 研究開発の体制

小野　欽司	研究の取りまとめ役として，研究プロジェクト全体に目を配る．また特に国際的な規模での地球環境データ共有を実現するために，ネットワーク環境の整備と定常的な運用に関してアドバイスを与える．
計　宇生	ネットワーキング技術に関して，特に地球環境データのような大容量のデータを，必ずしも帯域が十分には確保されていない国際的なネットワーク上で，適切に交換するためのQoS(Quality of Service)技術について研究する．
フレデリック・アンドレス	地球環境データのためのデータベース技術について研究する．マルチメディアデータベースのために開発したデータベースエンジンのノウハウを地球環境データに応用し，データベース技術の革新を目指す．
北本　朝展	地球環境データを直接的に扱うための画像処理技術や画像解析技術を画像データベース技術に応用し，さらにこれを分散ネットワーキング環境に拡大するための研究を進める．

1.4 導入設備

本研究で導入した主な設備は以下の通りである．

平成9年度	PCワークステーション（50万円×2）
	ノートPC（20万円）
平成10年度	PCワークステーション＋ビデオ画像入力装置＋ハードディスク（160万円）
平成11年度	PCワークステーション（50万円）
	ディスクアレイ装置（90万円）
平成12年度	ディスクアレイ装置（80万円）
平成13年度	ノートPC（20万円）

第2章　研究開発の概要

2.1 個別研究テーマの階層構造

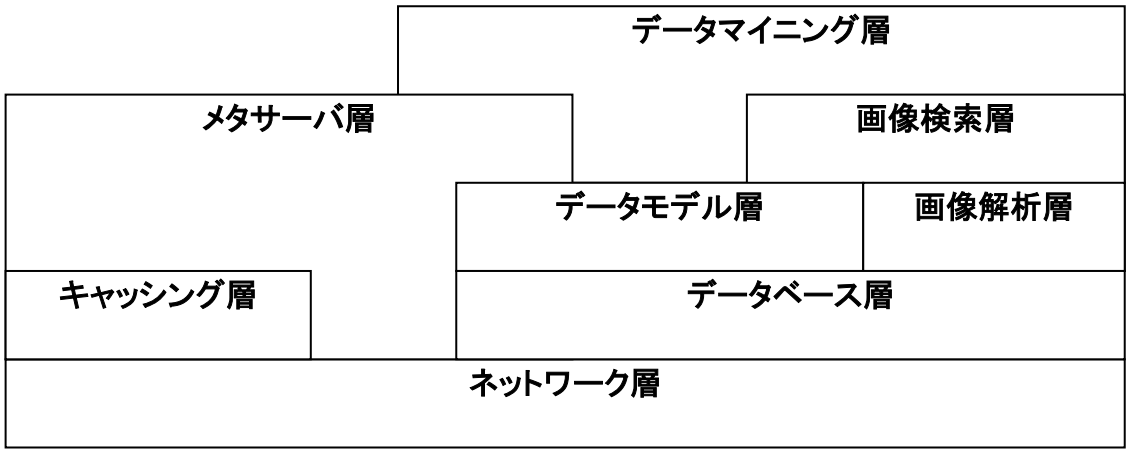


図 1 本研究の様々な個別研究テーマの階層構造.

研究プロジェクトの計画にも述べたように，本研究プロジェクトはさまざまな研究テーマを含む総合的な研究プログラムであるため，これらを一つの視点から統一的に概観するためのフレームワークが必要である．本研究でそのようなフレームワークに相当するのが図 1に示す階層モデルである．これは各研究テーマの関係をレイヤ構造で模式的に表現したものであり，上位層での情報表現が下位層での情報表現をカプセル化し，個々の層は自身が担当する範囲内で情報を変換または付加するという関係を表現するものである．また下位層の方がより基本的な処理を担当すると大まかに考えてもよい．このような階層構造を用いることで，複雑な問題を部分問題に切り分け，全体の見通しをよくすることが可能となる．そして個々のレイヤにおける問題を解決するとともに，複数レイヤのインタフェースにも同時に配慮することで，全体として整合性のとれた一つのシステムが完成してくることになる．これが本研究プロジェクトの研究方針である．

このような階層型のモデルは，OSI（open system interconnection)モデルを代表とする通信分野の基本的モデルに類似したものである．またデータベースの分野においても概念的なモデルとして多用され，内部レベル(internal level)，概念レベル(conceptual level)，外部レベル(external level)の3階層からなるANSI/SPARCモデルが代表的なモデルとなっている[32]．ここで，内部レベルとは物理的なデータの格納のレベル，概念レベルはデータベース全体を論理的に記述するレベル，そして外部レベルは個々の

アプリケーションごとのデータベースに対する視点である. このモデルに代わって, 最近では3層アプリケーションアーキテクチャという階層モデルが唱えられており, ユーザインタフェース層がクライアント機, ビジネスロジック層がサーバ機, データベース層が大規模データベースサーバ機に対応するという, よりネットワーク環境を重視したモデルとなっている. 本研究で提案する地球環境データベースは, 実質的にはこの3層アプリケーションアーキテクチャを踏襲する構成となる.

またこれらとは別に, 6.3章で述べる画像データベースのアーキテクチャも, このような階層モデルの一種である. これは図 1の階層モデルにおける画像解析層と画像検索層の2層を, 改めてさらに細かい階層構造に細分化しなおすことに相当する. 階層型のモデルはこのように画像データベースのアーキテクチャをはじめとする様々な側面において基本となるモデルであり, 実世界のデータを単なるビット列からデータ, さらには情報・知識へと構造化・抽象化していくためには, このような階層的思考が不可欠であると考えている.

2.2 各階層の研究成果

本報告書では, これから各階層の研究成果を順次述べていこうと思うが, ここでまず各階層の研究成果の概要をまとめてみたい. これらの研究成果は, インタフェースを介してお互いに関連してはいるが, 独立した研究成果でもある.

ネットワーク層（第3章）

ネットワーク層においては, 既存のTCP/IP層の上で大規模な地球環境データを共有するための実証的な実験が主な研究成果であり, 技術的に新しい点はあまりない. しかし実用的な面では, この実証実験の成功によって, 日本・タイ間で地球環境データを実際に準リアルタイムで共有できるようになったことは地球環境研究者にとって大きな前進である. この成果は現在のところ研究者コミュニティにおいて高い評価を受けており, 将来的に地球環境データを一般に公開できれば地球環境データの流通という面で大きな波及効果をもたらす部分となる.

キャッシング層（第5章）

キャッシング層においては, 地球環境データの実際のアクセスパターンからその性質を導いたという点に新規性がある. キャッシング技術自体の研究は数多くあるが, 地球環境データの特性に応じたキャッシング技術についてはまだ研究はそれほど進んでいない. ここで新たな方法論を生み出せれば, 一般のキャッシング技術にも影響をもたらす可能性はあるが, あくまで地球環境データの性質ということにこだわるなら, その効果は地球環境データにしか及ばないということになる.

データベース層（第7章）

データベース層においては地球環境データのフォーマット変換および集約・統合などを通して, データベースとして利用可能な形に膨大なデータを前処理し, これをデータベース化する作業をおこなった. このような作業は新規性・独創性に欠け, しかも膨大な手間を要する地味な作業ということでとかく軽視される傾向はあるが, 実際のところこの地味な作業なくしては, 実用的かつ大規模なデータベースは実

現し得ないものである. また最終的によい成果が得られるか否かは, 実際のところこのような前処理をどこまで真剣におこなうかに大きく影響されることも多い. したがってこの階層の研究で得られた実用的ノウハウは貴重な情報であり, 最終的な成果を支えるものである.

データモデル層（第4章）

データモデル層においては, 地球環境データベースのためのデータ定義の形式化, およびデータ検索の形式化について研究した. この種の研究は少なくとも伝統的な関係データベースに関してはほぼ完成の域に達しているが, 地球環境データのようにデータの種類や用途が多様な場合や, データマイニングのように操作が無数に考えられるような場合については, 形式化が困難であることから未だ研究は成熟したレベルに達していない. 本研究の試みはこれらに代わるほどの美しい理論を打ち立てるまでには到達していないが, 地球環境データのデータマイニングという視点から必要な機能を形式化したという点で, 新規性および実用的価値は高いと考える.

メタサーバ層（第4章・第8章）

メタサーバ層においては, データモデル層およびネットワーク層などの成果を統合し, 分散データベースにおいて重要な役割を果たすメタサーバを部分的に実現した. 特に本研究の独自性は, 地球環境データに適したデータモデルに基づいたデータ定義および操作言語の開発にある. 特に本報告ではデータ操作言語として提案するGRQL(Grouping and Ranking Query Language)を中心に説明する. これらの言語の構文にはXML構文を用いているため, 他のサーバとの相互運用性に優れた検索エンジンを構築することができる.

画像解析層（第6章）

画像解析層では地球環境衛星データを対象とする画像解析技術について研究した. 具体的には6.3章で述べる画像データベースアーキテクチャに沿って, 統計的画像分類手法, 雲パターンの形状分解表現などについて研究を進めた. これらのテーマのうち, 特に統計的画像分類手法については, 我々が知る限りこのようなモデルは過去に提案されておらず, しかも地球環境データの場合に本質的な問題となる混合信号の問題に正面から取り組んだ方法として, 独創性の面でも高い評価を受けている.

画像検索層（第6章）

ここでは地球環境衛星データベースの検索方法について, 特に対話的な研究方法という観点から研究を進めた. このような対話的な研究方法自体は, 特にテキスト検索の分野で長い歴史があるが, このような対話的な検索に進化的計算論を導入するというアイデアは過去にあまり研究されたことがなく, 新しい方向性を示すものとして注目されている. またこの画像検索の背後でやりとりされるメッセージは, 先述のデータ操作言語GRQLであり, このような言語の開発も画像検索層を支える技術となっている. 対話的な検索は下のデータマイニング技術の発展に伴って今後ますます主流になると考えられ, その意味で将来に期待をつなぐ研究である.

データマイニング層（第7章）

データマイニング層においては，大量の台風データコレクションを対象とした知識発見手法を中心に研究した．このようなデータマイニング手法を用いて，台風の解析や予測などに有用な情報を発見することができれば，災害の面などから社会的インパクトは非常に大きい．また台風の大規模画像データコレクション自体が，アメリカでもたった1箇所でしか構築していないほどの新規性の高いデータベースである．ゆえにこの階層の研究成果は本研究プロジェクトを代表する成果といってもよい．

以上のように，本研究ではすべての要素技術に対して，地球環境データベースという応用を意識した定式化をおこなっている．例えば，画像解析層においては台風雲パターンの解析，データモデル層においては台風画像検索のモデル化，データマイニング層では台風画像データマイニングのためのアルゴリズムの研究といったように，地球環境データベースとの関連を意識した．

また具体的な対象として台風を選ぶことにより，7.2章で後述するような様々な興味深い研究課題が各階層それぞれに生じる．また，台風画像系列は空間パターンの典型例であるのと同時に時系列信号の典型例でもあることから，時間の取り扱いも各階層における重要な研究課題となる．つまり，台風の雲パターンは，空間パターンとして捉えると，柔軟かつ生成・消滅が起こる物体のモデリングという画像解析分野でニーズの高い問題に直結するし，一方で台風は誕生から消滅まで個体のライフサイクルが完全に追跡できるパターンであることを考えると，時系列解析の対象としても挑戦的なテーマとなる．

以上の側面を考え合わせることで，時空間パターンからのデータマイニングという，データマイニングの中でも現在ホットな話題が浮上してくる．このように大容量で構造化されていないデータからのデータマイニングという観点から研究を進め，大規模なデータ処理に耐えうる高速でロバストな技術を開発することを研究の目標とする．

2.3 ネットワーキング技術の研究成果

さて本研究の研究計画では，もともと研究テーマをネットワーキング技術およびデータベース技術の2つに分類し，その方針に沿って計画を立てていた．そこでこのような分類に基づく研究成果についてもここでまとめておきたい．主要な研究成果は以下のようにまとめることができる．

1．SINETの国際回線を用いた大規模衛星データの準リアルタイム交換実証実験

2．分散地球環境データベースのためのメタサーバおよびXMLに基づくデータ定義・検索言語

3．地球環境データへのアクセスパターンに基づくキャッシング

これらはまったく独立した個別のテーマではなく，後述するように相互に関連をもつテーマである．まず第1項（第3章）では，国際的規模での地球環境データ共有実験をおこない，準リアルタイムに大規模衛星データを日本とタイの間で共有することができることを示した．次に第2項（第4章）では，分散地球環境データベースにおいて重要な存在となるメタサーバに関する研究，および地球環境データのためのデータ定義・操作言語について研究した．ここで，本研究で提案するデータ操作言語GRQLは，グルーピング操作を基本とする言語であり，画像データの類似性などに関する問合せにも対応できる言語となっている．またこの言語は，TCP/IP上のプロトコルに基づき，クライアントとメタサーバとの間でXML構文のメッセージをやり取りするという方式で構成されており，ネットワーク環境での使用が想定されたもので

ある. その意味でこの研究はネットワーキング技術とデータベース技術とにまたがる研究成果である. 最後に第3項(第5章)は地球環境データへのアクセスパターンの特性を活用して, アクセス頻度の高いデータをキャッシングすることによりメタサーバを経由するデータ転送を効率的にするための研究である. このような性質を地球環境データについて調べた例はまだほとんどないため, この実験結果はキャッシング技術に新たな知見を与える成果である.

2.4 データベース技術の研究成果

本研究の重点はネットワーキング技術よりもデータベース技術にあった. このデータベース技術に分類できる研究成果は以下のようにまとめることができる.

1. 地球環境データベースの内容検索技術

2. 地球環境データベースのデータマイニング技術

3. 台風画像データマイニングシステム(IMET)の構築

まず第1項第6章)では, 階層型の画像データベースアーキテクチャを議論し, 個々の階層の役割煮応じた画像解析技術を研究する. ここでは具体的には様々な画像解析技術を研究しており, その研究成果は多岐にわたる. 第2項(第7章)は個別のデータというよりも大量のデータコレクションの解析そのものに興味があり, 大量の地球環境データの中に埋もれた(統計的あるいは関係的な)性質を発見する技術について研究を進める. 特に本研究では, 気象学的および社会的に特別な重要性をもつ気象現象「台風」を対象として, 台風解析や台風予測に有用な規則性あるいは不規則性を発見することを目標とし, 機械学習技術やデータマイニング技術を画像解析技術と融合した技術を開発した. 最後に第3項(第8章)は以上のように得られた技術を総合し, 一つの台風画像データマイニングシステム(IMET)として構築したという研究成果である. このシステムはWWWでも公開されており, 一般の人がアクセスし使用することが可能となっている.

2.5 本報告書の構成

本報告書は以下の章において, 本研究プロジェクトの成果を順次述べていく.

まず第3章「地球環境データの国際的共有のための実証実験」では, 日本とタイとの間で国際的な規模での準リアルタイムデータ共有を実現するための様々な試みについて述べる. この実証実験のためのネットワーク環境について述べた後, 具体的なケーススタディとして台風衛星画像データの共有とその効果について述べる.

次の第4章は「地球環境データの定義・操作のためのマークアップ言語」を中心に述べる. ここで提案するデータ操作言語GRQL(Grouping and Ranking Query Language)は地球環境データのデータマイニングにはグルーピング操作がしばしば必要になることにヒントを得て, グルーピング操作を基本操作として設計するものである. グルーピング操作によって得られるデータ集合が再びグループとなるようにすれば, 操作を入れ子にして適用することも可能となり, 柔軟で強力な操作を実現することができる. またXML構文を用いることで相互運用性が高いマークアップ言語としている.

第5章では「地球環境データへのアクセスパターンのモデル」として，実際の地球環境データベースへのアクセスログを解析することで，地球環境データへの利用者のアクセスパターンを特徴付ける試みである．特に本研究では4つの観点からアクセスパターンの統計的性質を分析し，これをモデル化することで地球環境データへのアクセスパターンをモデル化した．

第6章の「地球環境データの画像内容検索モデル」は，本研究で提案する画像データベースアーキテクチャおよびその各レイヤを構成する種々の画像解析手法をまとめて述べる部分であり，多様な研究成果の概要をまとめている．

第7章の「地球環境データ（台風データ）のマイニング」は地球環境データマイニング，特に台風画像データマイニングの研究成果について述べるものであり，第6章の研究成果を受け継ぎ，個々のデータの検索にとどまらず，台風画像コレクション全体としての性質を明らかにするための機械学習，データマイニング手法の適用についてその成果をまとめる．

第8章の「IMET: Image Mining Environment for the Typhoon」は具体的なシステムとして構築した台風画像データマイニングシステム(IMET)に関する概要の説明であり，特にビューの具体的な設計に関するより細かい部分の議論についてまとめる．

最後に第9章は結論として，本研究プロジェクトで得られた研究成果をまとめ，今後の研究の展望を探ることとする．

第3章 地球環境データの国際的共有のための実証実験

3.1 はじめに

地球環境データにおいては，データを共有するための仕組みが本質的に重要である．その理由は，地球環境データがまさに地球規模の広がりをもつデータであり，しかも単一の衛星と単一の受信局では地球全域の観測をカバーすることが不可能に近いためである．例えば，静止気象衛星「ひまわり」はかなり観測範囲の広い衛星であり，日本の受信局で受信可能な衛星観測データだけでも，地球表面の1/3弱はカバーすることができる（解像度の問題を除けば）．しかし，例えば地球の赤道をすべてカバーするような観測データを得るためには，アメリカやヨーロッパなど他国が運用する静止気象衛星の観測データをつなぎあわせないと「地球の裏側」の観測データを得ることはできない．さらに，極地方の観測まで考えれば，赤道上空に静止する衛星からこの地方は見えないので原理的に観測不可能である．このような極地方の問題は地球を周回するタイプの衛星でカバーすることは可能であるが，これらの衛星は一般的に観測範囲がさらに狭く，しかも地上から衛星を見通せる範囲でないと衛星観測データを受信できないので，地球各地に受信局を設置し地球各地の観測データを受信しておく必要がある．

もちろん本来は，このような受信に関する問題を解決した上で，それらを交換・収集するメカニズムまで考えておかねばならない．ところが世界中の地球観測衛星データを収集するメカニズムを整備することは，簡単な作業ではない．世界各地に高速な通信ネットワークが張り巡らされていれば，これらの受信データを準リアルタイムで収集しつなぎ合わせることも可能であるが，世界各地ではむしろこのような高速な通信ネットワークが到達していない場所も多く，そのような場所では郵便ネットワークに頼ったデータ収集メカニズムに頼らざるを得ない．これでは準リアルタイムでのデータ収集は不可能である．

そこで本研究では，各地の地球観測データを地球規模で共有するための仕組みとして，世界各地の受信局を接続するための広帯域ネットワーキング技術や，それに付随する各種のネットワーキング技術について研究する．特に本章では，日本とタイとの間で地球環境データを共有するための実証的な実験について，その概要と成果をまとめる．大規模な地球環境データの交換に耐えうるような広帯域ネットワーク環境は，数年前には実現することがまだ難しかったが，近年は徐々に各地のネットワークがアップグレードされ，中にはギガビットクラスの帯域をもつネットワークも出現している．このような進歩は，地球環境データの共有には有利な状況である．

ただしこのような広帯域のネットワークは日本と米国間などの主要国間に限られるのが実情であり，東南アジアのように通信需要がそれほど多くない地域への帯域は依然として限られている．ただし地球環境データの重要性に関しては，これらの地域の観測データは決して主要国の観測データに劣るものでは

ない．そこで本研究では，大規模地球環境データベースの国際的共有を実現するために，タイ王国のアジア工科大学および日本の国立情報学研究所 / 東京大学との間の準リアルタイム地球環境データ共有を実現するためのネットワークを実現するための実証実験を進める．ここで準リアルタイムというのは，1秒単位ではなく，1時間や1日単位で見れば最新のデータを入手できる，ということを指す．つまり地球環境データに対して要求される「リアルタイム性」とは，時間を単位とするようなスケールであり，これが映像などのストリーム系データの「リアルタイム性」との相違点である．

3.2 日本とタイとの間での国際的共有

この実証実験は，国立情報学研究所とタイ王国・アジア工科大学Asian Institute of Technology (AIT)との国際共同研究に基づくものである．具体的に共有を目指すのは，アジア工科大学で受信するTERRA衛星MODISセンサおよびNOAA衛星AVHRRセンサからの受信データであり，これらのデータを国立情報学研究所や東京大学，AITが，インターネット経由で準リアルタイムに共有できるような情報基盤を確立することが目的である．TERRA衛星およびNOAA衛星のデータは東京大学でも受信しているが，両国で受信できる観測範囲が異なるため，観測規模をアジア全域に拡大するためには，アジア工科大学と東京大学で受信する観測データを共有し重ね合わせる必要がある．

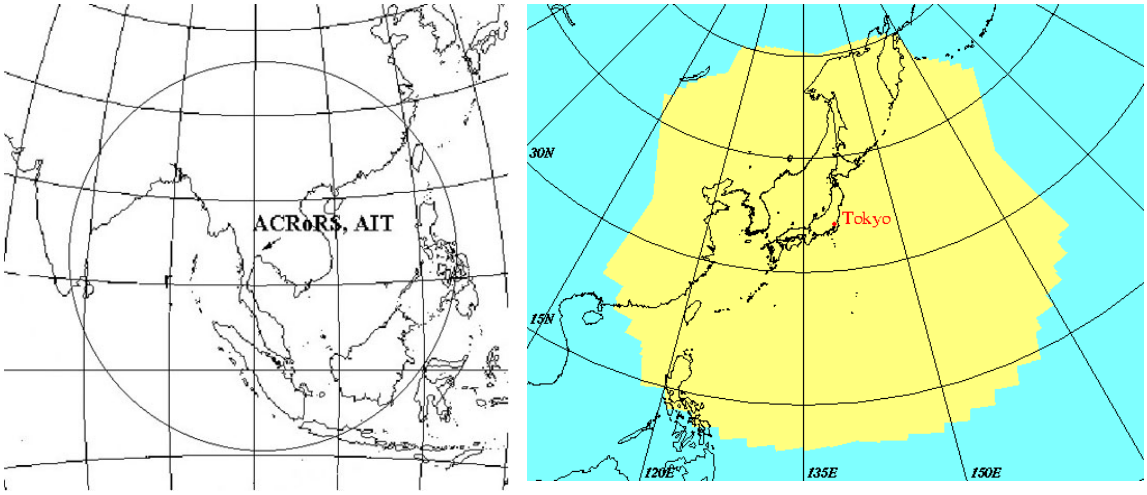


図 2　日本・東京(右)とタイ・バンコク(左)で受信できる衛星データの観測範囲の違い．右図は東京大学生産技術研究所提供，左図はアジア工科大学アジアリモートセンシング研究センター提供．

例えば日本・東京とタイ・バンコクで受信可能なNOAA衛星の観測範囲の相違を図 2に示す．この地球観測衛星は地球を周回する軌道(軌道高度平均約850km)を動きながら，衛星直下点を中心に走査することで地球を観測する．その観測データはその後地球観測衛星から発信されるが，地球表面が球面であることから，その送信電波を受信可能な衛星受信局は地理的に限られてしまう．逆に考えれば，ある地上の受信局は，そこから見通すことができる衛星位置の範囲によって，その観測範囲が制限を受けることになる．このような状況では，例えば東南アジアを上空から観測した地球観測衛星データを日本の受信局では直接受信することができないため，このような地球観測衛星データを日本で入手するためには，東南アジアでいったん受信した上でこれらのデータを日本まで転送する必要がある．

このようなデータ転送処理は，従来は磁気テープの郵送によりおこなっていたが，これではリアルタイム性および利便性に劣ることは明らかであり，この部分を通信ネットワークによって実現したい，との発想が生じるのは当然の流れである．つまり高速ネットワークにより地球観測衛星データを高速に転送できれば，複数地域あるいは全地球の地球観測衛星データを，準リアルタイムで多くの研究者が共有することが可能となるのである．ここに広帯域ネットワークの真価がある．

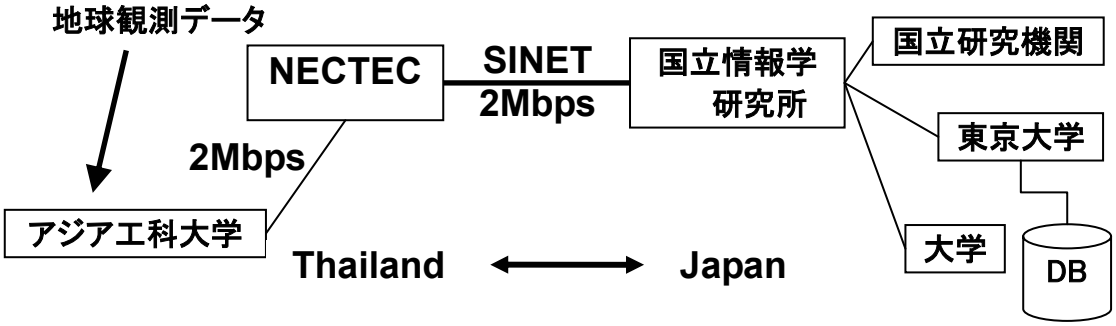


図 3　SINETの国際回線を用いた大規模衛星データ準リアルタイム交換実証実験の構成図。

そこで本研究では，国立情報学研究所が提供するSINET（Science Information Network)のタイ国際回線(2Mbps)をネットワーク研究基盤として活用し，この国際回線を用いて大規模地球環境データの共有を実現する．具体的な対象は図 2のNOAA衛星データ，およびその後継機種にあたるMODIS衛星データであり，これらの大量の衛星データを国際的に共有するためのネットワーク構成として，図 3のような実証実験ネットワークを構築した．

この実証実験ネットワークは，国立情報学研究所とタイのNECTEC（National Electronics and Computer Technology Center)との間で運用するSINETの国際回線を中心に，日本国内の学術ネットワークSINETおよびタイ国内の学術ネットワークThaiSARNとを活用して，日本とタイの各機関を結ぶものである．ここでSINETの国際回線の帯域は2Mbpsである．AITとNECTECとを結ぶThaiSARNは，実験開始当初は256kbps程度の帯域しかなく，大規模地球環境データの共有は難しかったが，現在ではSINETの帯域に合わせて2Mbpsに増強されており，さらに100Mbps以上への高速化も計画に上っている．SINETの日本国内回線はこの5年間で着実に増強されており，こちらはタイ国内ネットワークに比較すれば帯域に余裕があり，ボトルネックにはならないと考えられる．

以上のように本研究では，大規模地球環境データ共有を実現するためのネットワークとして，日本国内，タイ国内，およびそれらを結ぶ国際回線を増強し広帯域ネットワーク環境を実現した．次にこのネットワークを用いておこなった実証実験について述べる．

3.3 実証実験

今回の実証実験は，アジア工科大学で受信するTERRA衛星MODISセンサおよびNOAA衛星AVHRRセンサからの受信データを，アジア工科大学および国立情報学研究所の間で共有することを目的とする．前者のTERRA衛星MODISセンサデータは，平均して1日3回～4回の受信があり，データサイズは1回の受信ごとに700MByteから1400MByte程度と幅があるが，これらのデータはgzipという可逆圧縮ツールを用いて圧縮することで，1/3から1/10程度のデータサイズに圧縮することができる．したがって，1日で送信する必要のある実質的なデータ量は，およそ2GByte～2.5GByteに達する．

そこで図 3の実証実験ネットワークを用いて，実際に大規模地球環境データの送信実験をおこなった．実験では主にFTP (File Transfer Protocol)を用いた簡単なファイル転送を行い，これらの地球観測衛星データをアジア工科大学から国立情報学研究所まで転送するのに要する時間およびデータサイズから，単位時間あたりの転送速度を計算した．この方法で算出できるのは，実効的な転送速度，すなわち通信に必要なオーバヘッドを除いた実質的な速度となる．例えばTCP/IPでは送信するデータサイズに比べてヘッダのサイズが10%程度となるため，この部分がオーバヘッドとなり，たとえネットワークの帯域が2Mbpsであっても実際にはそこまで速度は向上しない．またFTPによるオーバヘッドも若干は生じる．

その結果，この実証実験ネットワークを用いて，図 4に示すような安定した転送速度(図ではOut方向)を達成した．図に示すように1週間を通してほぼ1．6Mbpsの速度が達成できており，最大では約1．8Mbpsの速度に達している．ここで1．6Mbpsという速度は約17．2GByte/dayに相当する速度であり，これは1日で送信すべき衛星受信データ量，約3GByteを上回る帯域である．ゆえに，MODIS衛星およびNOAA衛星という大量の地球環境データを，日本とタイの間で共有することが，少なくともネットワーク的には実現可能であることを示した．

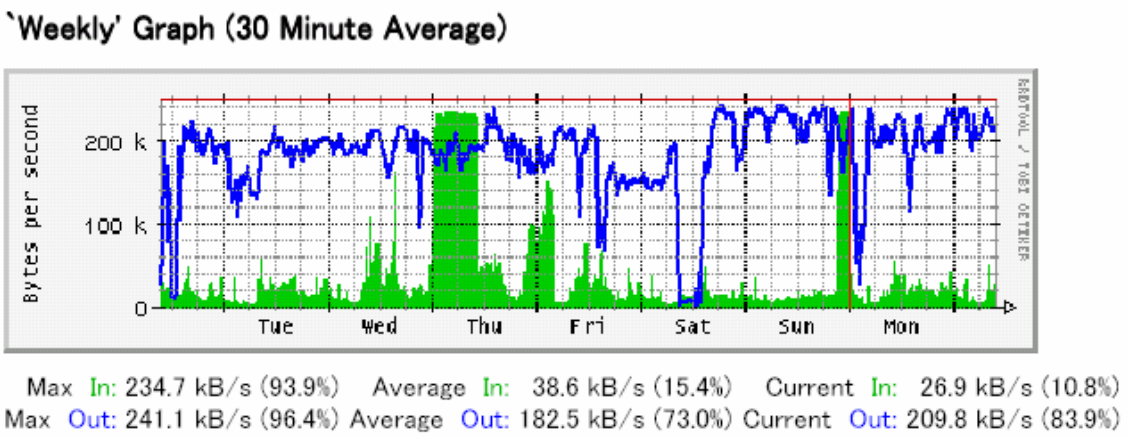


図 4 AITとNIIを結ぶネットワークの1週間のトラフィックの変動．このトラフィックのほとんどは今回の実証実験のトラフィックである．

3.4 ケーススタディ：台風画像データの共有

ここで，このようなネットワークを用いて地球環境データを共有することの意義を，台風画像データの共有というケーススタディを用いて説明する[22][25]．第7章で紹介するように，我々は地球環境データの中でも特に台風画像データの収集とデータベース化に関心をもっている．この場合に問題となるのが，東南アジア地域での台風画像データの入手である．というのも，我々が主に用いる気象衛星「ひまわり」は，東経140度上空35790kmに静止する衛星であるため，衛星直下点での地上視野(ground field of view : GFOV)はおよそ5kmであるものの，緯度経度が衛星直下点から遠ざかるにしたがって実質的な解像度が低下するためである．特にタイ(バンコク)は東経100度と衛星直下点から遠いため，解像度の低下によって良質な台風画像データが得られないおそれがある．

それに対し，アジア工科大学で受信するNOAA衛星データは地球周回衛星とよばれる種類の衛星であるため，地球上の任意の地点を衛星直下点に近い地点として精度よく観測することができる．また平均

高度が833kmと低いため，地上解像度が衛星直下点で1.1kmと，もともと「ひまわり」衛星と比べても優れた性能をもっている．ゆえに，この衛星がたまたま東南アジア上空を通過し，しかもその時に台風をうまく捉えることができれば，「ひまわり」衛星画像よりも高品質の画像データを入手できるのである．ただしこの衛星は観測頻度の面に難があり，ひまわり衛星が毎時間観測なのに比べて，NOAA衛星の場合1日最大数回に限られる．しかもいつも台風の上空を通過するわけではない．

そこで我々は，この2種類の衛星データを組み合わせて台風画像コレクションを構築することを考えた．つまり東南アジア地域で「ひまわり」衛星画像を補完するものとして，アジア工科大学で受信するNOAA衛星データを活用することにより，解像度が低下する東南アジア地域での台風画像データを補完しようという戦略である．

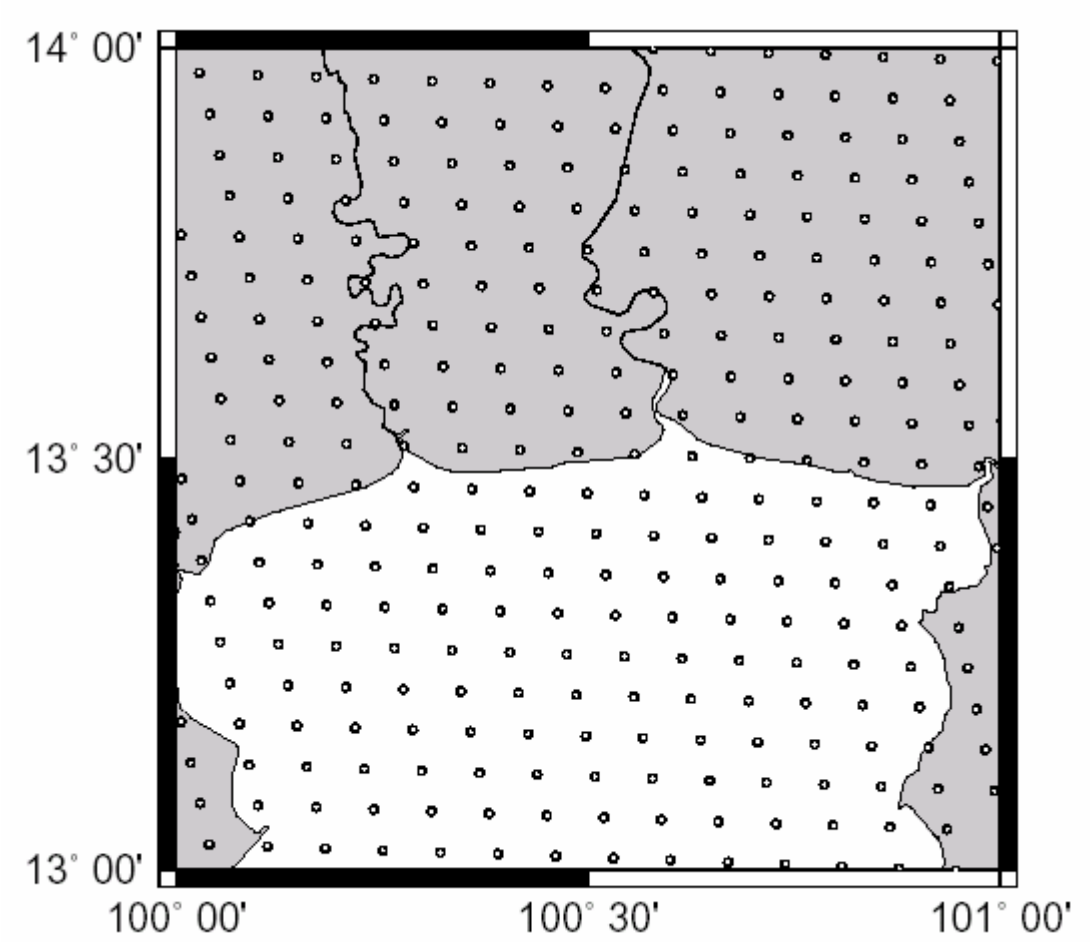


図 5 バンコク地域におけるひまわりVISSR画像のスキャンスポットの間隔（0000 UTC December 12, 1998）．白い円の中心はスキャンスポットの位置を示し，円の大きさはNOAA AVHRR画像の地上視野の大きさ（約1.1km）を示す．つまりNOAA AVHRR画像の瞬時視野はひまわり画像のスキャンスポットの間隔よりかなり小さくなる．

図 5ではバンコク付近での，二つの衛星の瞬時視野を比較している．ひまわり画像の場合はスキャンスポット（衛星画像上で単一の画素として現れる地上領域）の間隔はおよそ7.5kmとなっているのに対し，NOAA衛星がバンコク上空を通過すればスキャンスポットは1.1kmである．このようにNOAA衛星は地上解像度の面で大幅に優れていることが確認できた．

もちろんNOAA衛星にも欠点はある．具体的には観測頻度が小さいことと，観測範囲が狭いことが問題である．例えば1回の観測で幅2800kmの領域しか観測できないというのは，巨大な台風の雲パターンの直径がほぼ同じサイズであることを考えれば大きな制約である．したがって衛星が台風中心の上空を通過しない限り，台風雲パターンの全体像を捉えることは難しい．

実際にひまわり画像とNOAA画像で台風の雲パターンがどのように表現されるのかを示すのが図 6である．この図はほぼ同時刻に同地域を撮影したふたつの衛星画像を並べたものである．この場合，NOAA衛星画像で台風が画像の右側に位置していたため，中心付近はしっかりと捉えられたものの，台風雲パターンの全貌は捉えられていない．また地上解像度に関しては，この図でははっきりとはわからないが，ひまわり画像よりも解像度が高いデータになっている．それに対してひまわり画像は台風雲パターンの全貌は捉えているものの，ベトナムは比較的画像の端に近い場所であるため，実質的な解像度が多少低下しぼやけた画像となっている．また両者とも観測波長が非常に近いため，雲パターンも非常に近いパターンが現れている．

ここで，「ひまわり」画像に関しては日本で受信するデータもタイで受信するデータも同一のデータであるが，NOAA画像の場合は両者が異なるということに注意しておきたい．また先述したように，このような東南アジア地域の観測画像は日本で直接受信することが原理的に不可能である（日本からはタイ上空を飛ぶNOAA衛星を見通せないため）．ゆえに，地球環境データを国際的に共有する価値が，ここに生まれるのである．そしてこれら2種の地球観測衛星データを補完的に用いることによって，高解像度の台風画像データを得ることが可能となるのである．

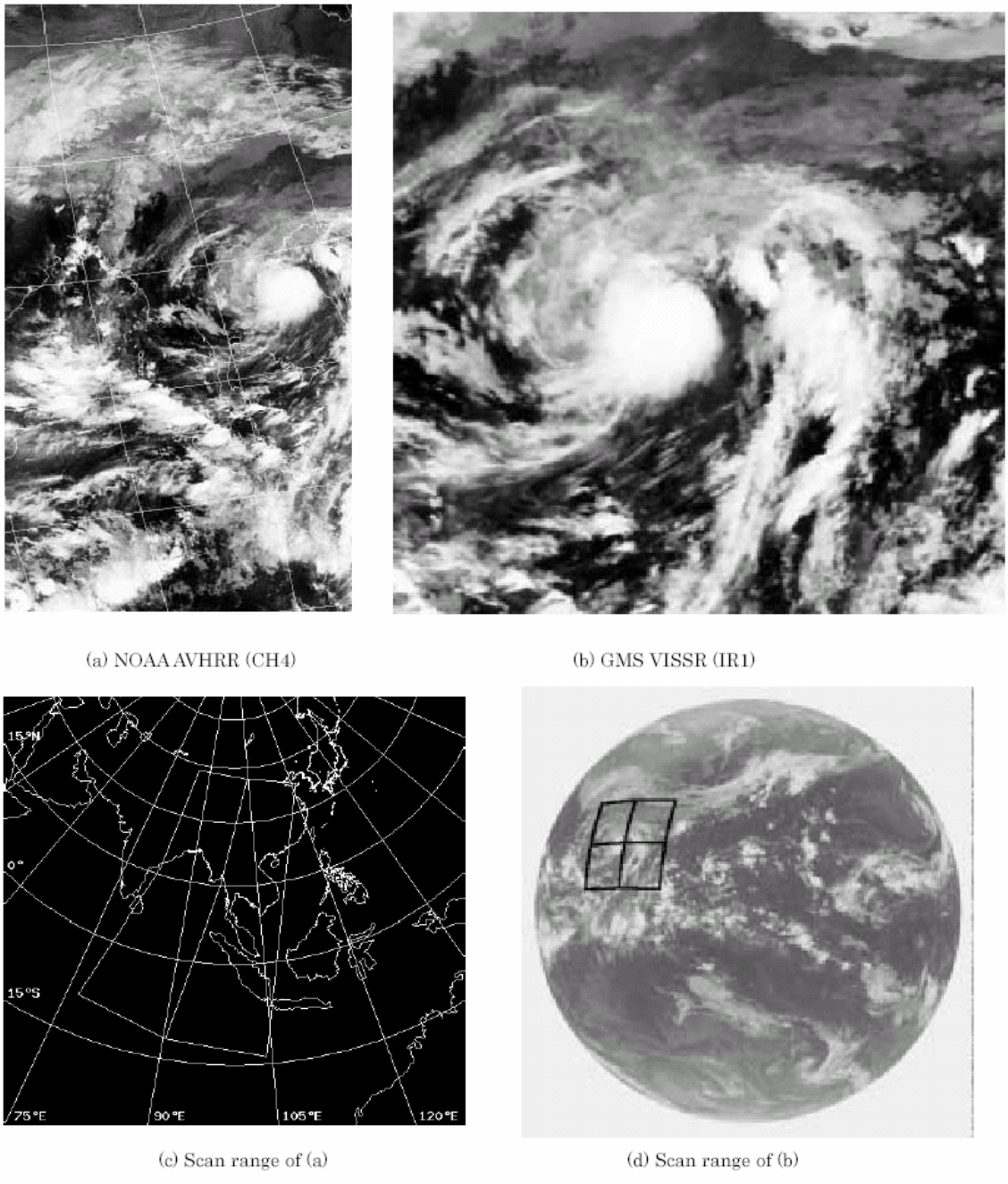


図 6　NOAA AVHRRデータとひまわり画像との比較. 画像は台風199921号の画像であり, 撮影日時は両者ともほぼ1999年10月19日2100 UTCである. 台風の位置はおよそ(17. 1°N, 107. 3°E)でありベトナムのフエ市の近くに位置している. このときの台風の勢力(内挿値)は, 中心気圧996 hPa, 最大風速35ノットである. (a) と (b)のスキャン幅は(c)と(d)に示す.

3.5 今後の課題

現在のところSINETの国際回線およびタイ国内のネットワークとも安定した転送速度を維持できている．したがって，現状の地球環境データの量であれば，インターネットを用いた準リアルタイムの大規模地球環境データ共有は定常的に運用することが可能である．ただし，地球環境データは常に高解像度，多波長，高頻度という方向に向かっていることから，地球環境データの量は，今後も単調に増加しこそすれ，減少することはあり得ない．また地球環境研究者は，地球環境データの解析に関してそれぞれ独自のノウハウを保有しているので，これらの処理済データを準リアルタイムで共有することまで考えを広げると，現状よりもさらに広帯域のネットワークが必要になるのは確実である．

ゆえに現状の帯域が十分であるわけでは決してなく，今後も継続的にネットワーク基盤を増強していくことは重要な課題である．特にSINETのタイ回線はこのように東南アジアと大規模地球環境データを共有するために不可欠のネットワークであり，その役割はきわめて大きいものがある．

またSINETの国内回線を利用することで，今後はこのような地球環境データを国内の研究機関とも準リアルタイムに共有していく計画である．その際に国立情報学研究所は主に地球環境データベースおよび地球環境データの検索機能を提供し，地球環境データのメタサーバとしての役割を果たす計画である．このようなデータベース技術については，次章以降で順次述べていくことになる．

将来的には，今回の実証実験で用いたようなFTPによる単純なファイル転送ではなく，さらに高度なファイル共有メカニズムの適用を実証することが課題である．複数の衛星データ受信局で受信したデータを共有するためには，データの所在および登録に関する情報を自律的に収集しデータを取得するようなメカニズムが必要だからである．この点については本研究では十分な検討をおこなうことができなかったため，今後の課題としたい．

第4章 地球環境データの定義・操作のためのマークアップ言語

4.1 はじめに

地球環境データを相互に共有し検索することを可能にするためには，すべてのデータを共通の形式で記述するか，複数の形式の間で相互に変換するための規則を定めておく必要がある．このような形式の必要性はかなり以前から繰り返し議論されてはきたが，地球環境データには複数の地球観測衛星，それも異なる機関から打ち上げられた地球観測衛星が含まれることから，実際のところ地球観測衛星のデータフォーマットを統一することさえ部分的にしか実現できていない．また，地球環境データの利用者は，少数の地球環境データを対象とし自分の関心に沿った独自の処理をおこなうことが多いため，他の利用者の利便性までを考慮した統一的な形式の定義に対する関心がそれほど高くなかったという事情も否定できない．しかし，地球環境データを用いた研究が広がりや深みを見せるにつれて，利用者が多くの種類の地球環境データにアクセスする必要性が増しており，複数の地球環境データを検索し，組み合わせ，加工するという作業を支援する統一的なフォーマットに関するニーズも高まりつつある．ゆえに，データに関する記述（メタデータ）を統一することと，複数の地球環境データを効率的に検索することという，二つの作業を支援するための研究が重要な課題となってきた．

このような長年の課題に対して，近年の大きな変化が影響を与えつつある．まずデータの記述フォーマットとして，XML[34]がその存在感を増してきた．XMLの利点は少なくとも構文(syntax)の面ではデータを統一的に記述できるという点にある．したがって，XML文書のためのパーサやコンバータなどの既存のツールをそのまま活用でき，しかも互換性にも優れている．第二に，デジタルアース計画や電子政府計画などに端を発する，大規模地理情報空間の構築へと向かう世の中の流れである．すべての地理的情報を統一的な座標系のもとで記述するという大規模なプロジェクトに起因して，少なくとも地理的情報の記述に関しては統一的なフォーマットが実現しつつある．これらを背景として，本研究では地球環境データを定義するための言語および操作するための言語を定めることとした．

ここでデータ定義言語は，地球観測データの記述だけではなく，これらを処理して得られる2次データや3次データの記述にも用いる．また処理手順とデータとを関連させて記述することにより，自己記述的な地球環境データを生成できる．また地球環境データから抽出した画像特徴などに関してもデータ定義言語を用いて記述することにより，このメタデータを対象にした地球環境データの検索を統合的におこなうことも可能となる．

一方，データを操作するための言語を設計するには，データを定義するための言語とは異なる設計が必要である．データの操作方法はほとんど無限に存在し，これらをすべて言語仕様に取り込むことは不

可能であるから，データ操作方法の形式化から基本的な操作を抽出し，その他の多くの操作を基本操作の組み合わせとして実現する，ということに関する考察が必須である．そこで地球観測衛星データを中心とする地球環境データの検索操作に着目し，このような検索操作に適したデータ操作言語の設計を研究の目的とする．この種の言語として現在最も広く使われているデータ操作言語SQL(Structured Query Language)と比較すると，本研究で提案する言語GRQLはSQLのような理論的な美しさと完全性は備えていない．しかし我々の目的は，そのような理論的な美しさからスタートするのではなく，現実問題からの要求を分析し設計するというデータベース本来の方法論に沿って，地球環境データ検索に適した言語を設計することにある．

4.2 データ定義言語

一般的にデータベースの特徴は，個別のプログラムやデータの物理的論理的条件に依存せずに情報の操作ができることにある．このような独立性を実現するためには，情報の分類方法，すなわちデータベースのスキーマを定義し，個々の情報はスキーマに沿うようにデータベースに投入する．その他にもデータベースの設計には考慮すべき点が多数存在し，それらを段階にまとめてみると，典型的には以下のように表すことができる[32][33]．

1．要求の収集と分析

2．スキーマ設計

3．データベースシステムの選定

4．データモデル変換

5．物理データベース設計

6．データベース実現

このサイクルは順序立てて進むわけではなく，数多くの修正やフィードバックを繰り返す試行錯誤の結果，妥当なスキーマに収束するまで繰り返す．このようなデータベースの設計は実際には多くの研究においてすでに研究・議論されたテーマであり，その方法論もこれまでに多く提案されている．しかし，データベースに関する要求は応用分野によって様々に異なり，そのような要求を抽象化，形式化しスキーマに変換していく作業は，決して容易ではない．

データ定義言語(data definition language)は情報を記述する言語であり，情報の利用の仕方とは独立に，情報を宣言的なデータとして表現するための言語である．一般的にデータ定義言語では，情報内容の記述に加え，情報内容の無矛盾性を保証するための一貫性制約も記述できることが求められる．そして，情報内容を系統的に記述できるような見通しのよい言語を得るためには，対象とする問題領域に関する深く体系的な知識が究極的には必要となる．このような知識体系を一般的に構築するのは困難であるため，問題領域ごとの知識体系を反映したデータ定義言語を定め，情報の相互運用性を高めることを目的とするプロジェクトが，近年は非常に活発化してきている．

その背景にインターネットの普及に伴うデータ交換の活発化があるのは確かだが，近年の活発な動きを刺激した最大の理由が，XML[34]の登場にあるのはおそらく間違いないだろう．XMLの登場によって，データ定義言語を書くための「構文」に関する有力な標準が固まり，少なくとも構文レベルでは多様なデ

ータ定義言語が同一のフォーマットで記述できるようになった. このことが, あらゆる分野において情報記述をXML化する試みを刺激し, XML構文を用いたマークアップ言語が各分野で一気に花開くこととなった. さらに, 複数のマークアップ言語を名前空間メカニズムによって共存させることもできるため, 既存のマークアップ言語を取り入れ部分的に拡張することで, 相互運用性に優れた新たなマークアップ言語を定義することも可能となった.

ただし当然のことながら, XMLはデータ定義言語の構文を定めたに過ぎず, これですべての問題が解決するわけではない. むしろ, 実際のところ, 従来からの本質的な問題はそのまま残されていると言ってもよい.

1. 問題領域において記述すべき情報内容を選び出す方法はXMLが扱う範囲外であり, また表現したい情報内容に最適な要素が既存のデータ定義言語にいつも用意されているとは限らない.

2. 同一あるいは非常に類似した情報内容に対して, 複数のデータ定義言語がある場合に, どちらを使えばよいかがわからない. このようなことが起こるのは, 隣接した研究領域における知識体系の相違が原因であることもあるし, あるいは同一の研究領域における専門家間の意見の相違などを反映していることもある.

このような状況に対しては, Semantic Web[41]の試みを代表的な例として, 意味的に同一の情報内容を共有するための推論規則を定める方法が盛んに研究されている. 例えばXMLデータの抽象的な意味を, メタデータを記述するための規格であるRDF (resource description framework)[40]で記述し, さらに抽象的な意味を定義するオントロジーを用いて, 抽象的な意味を定義し共有する仕組みを提供する. ただしこのようなシステムはたとえ実現するとしても遠い将来のことであり, 現実的なレベルで成功を収めた方法はまだ出現していない.

本研究で定めるデータ定義言語は対象をあらかじめ地球環境データに限定したものであり, しかも標準化を目指すような大規模なプロジェクトでもない. むしろ本研究で用いる地球環境データを統一的に記述するためのスキーマを発見的に見出すための試みであり, その意味では現在も進化しつつあるスキーマである. ゆえに現段階のデータ定義言語は, 地球環境データを統一的に記述可能な(仮想的な)データ定義言語から考えれば, はるかに未熟な段階であると言わざるを得ない.

4.3 データ操作言語

データ操作言語(data manipulation language)とは, データの検索や登録など, 様々なデータベース操作をサポートするための言語である. データ操作言語は概念スキーマを用いて記述され, データを定義するための言語とは異なるスキーマをもつ. データ操作言語で用意すべき命令とは, 情報構造の利用方法やデータベースの成長に伴って変化するため, 原理的には無数の命令が必要となり, あらかじめすべての処理命令を特定しておくことはできない.

この種の言語で最大の成功を収めているのはSQL(Structured Query Language)である. これは関係代数という数学的理論を背景とすることが大きな特徴である. また, 選択(selection), 射影(projection), 結合(join)などごく少数の演算子の組み合わせで複雑な操作を記述できることも魅力的であり, 実際にこの言語は関係データベースの検索には非常に強力である. しかし関係代数の範疇に入らない情報構造に関する検索操作は難しく, 特にデータマイニングのように多様な操作の組み合わせからデータを特徴付けるための機能に対しては, 関係代数の枠組みが適しているとは必ずしもいえない.

このような問題は広く認識されており，それらの問題点に対する関係代数の拡張，新しい数学的構造，あるいはデータ操作に関する新しいスキーマの提案などを目的とする研究は，実際のところレビューするのが難しいくらいに多数存在する．最近ではXML文書への問合せに特化した言語としてXML Query[44]などの言語も提案されている．しかしこれらの既存の言語は，それぞれの目的に最適なようにデザインされたものであるため，我々が地球環境データベースにおいて実現したい機能という観点では，必ずしも最適なデザインとはなっていない．

そこで本研究では，地球環境データベースに必要な操作を発見法的に形式化する方法により，データ検索言語を定めた．このような方法は，数学的理論基盤からトップダウン的に定めるデータ定義言語よりも見通しが悪いのは確かだが，最初に数学的基盤を確立することは容易ではなく，研究を進めながら重要な操作を順次形式化していくというボトムアップのアプローチも現実的には有効であると考える．

本研究で提案する言語(Grouping and Ranking Query Language : GRQL)は，XML構文に基づく言語である．またスキーマにはSQLやXML Queryなどの既存のデータ操作言語から語彙を借用している．SQLとの最大の違いは，グルーピングやランキングといった操作を基本的な操作として定義した点にある．グルーピングは共通の特徴をもつデータをまとめる操作，ランキングはある特徴を基準にしてデータを並べ替える操作であり，これらの操作を重視することで，データを探索する能力が大幅に向上する．また並べ替えに関しては，データ間の類似度あるいは距離といったものを基準にできることも，特に類似画像検索などを考えた場合に有効な特徴である．

4.4 関連規格

4.4.1 データ定義言語

地球環境データの定義には，これまでHDF (Hierarchical Data Format)やnetCDF等の，画素値配列定義に関する規格，あるいは個別の衛星データのヘッダに含まれる，極めて汎用性の乏しいフォーマットを用いるしかなかった．それに対し，近年のXMLの普及に伴って，地球環境データの定義にむけたマークアップ言語が部分的ではあるが提案されるようになってきている．

Earth Science Markup Language (ESML)

ESML (Earth Science Markup Language)[35]は，地球環境データの相互運用性を重視したマークアップ言語であり，異種フォーマットのデータセットを統一的に扱うためのメタデータを提供することを目指している．ESMLファイルはデータファイルの内容，構造，または意味を記述するものである．ESMLスキーマはESMLファイルを生成するための規則を定義する．ESMLファイルはデータファイルの外部に生成されるので，既存のデータファイルを変更することなくESMLを用いることができる． ESMLのメタデータは以下の3種類からなる．

1．構文的(構造的)メタデータは，データをビットまたはバイトの観点から記述する．これらのメタデータは，データファイルのビット列に構造を与えるためにESMLパーサが利用する．例えば，構文的メタデータはESMLパーサに，次の32ビットを2の補数表現のビッグエンディアン32ビット整数と解釈せよ，といった指示を与えることができる．

2. 意味的メタデータは，構文的メタデータによって記述される要素に意味を与えるものである．これらのメタデータは，パーサがデータの意味を「理解」できるようにする．例えば意味的メタデータは，構文的メタデータによって"Longitude"と名前付けられた項目が，10000分の1ラジアンを単位として表された経度の値を記述していることをパーサに指示することができる．

3. 内容的メタデータは，データセットを人間が読めるような用語で記述する．これらはESMLパーサで解釈されることもあるかもしれないが，コンピュータに可読な形で記述されていることは必ずしも必要ない．典型的な例として，例えば内容的メタデータがデータの由来や系統を記述することがあるが，これらはデータの自動処理に不可欠な情報ではない．

Geography Markup Language (GML) / G-XML

いずれも地理情報を表現するマークアップ言語であり，今後は大規模地理情報基盤において用いられることになる言語である．特にGML(Geography Markup Language)[36]は地理情報の記述言語として，Open GIS Consortium Incを中心に世界の多くの組織が参加する大規模な規格となっているのに対し，G-XML(Geography XML)[37]は主に日本の組織が中心となって規格化したマークアップ言語である．ただし類似した内容をもつマークアップ言語であり，両者の統合も検討されているようである．

例えばG-XML文書は，計量地理空間（MetricGeospace)，位相地理空間(TopologicalGeospace)，関心地点(POI)，移動体(Mover)，経路(Route)，画像(Picture)，描画スタイル(RenderingRule，RenderingRuleList)の組合せによって構成されている．これら計量地理空間，位相地理空間，関心地点，移動体，経路，画像及び描画スタイルは，ほかのG-XMLの構造で指定され，さらにその構造は，G-XMLの部品で指定される．また，記述されたG-XML文書データ全体にメタデータ(Metadata)を関連付けることも可能であり，メタデータはG-XMLの部品を用いて指定することができる．

MPEG-7

MPEG-7は「マルチメディアコンテンツ記述インタフェース(multimedia contents description interface)」ともよばれ，マルチメディア環境での視覚・聴覚データの記述を可能とするための標準的な技術を確立することを目的としている．MPEG-7は視覚・聴覚データの特徴を，以下のコンテキストで標準化する．

1. **記述子(Descriptors : D)** それぞれの特徴表現の構文と意味とを定義する．

2. **記述スキーマ(Description Schemes : DS)** 構成部品間の関係の構造と意味を特定する．

3. **記述定義言語(Description Definition Language : DLL)** 新しいDSあるいはおそらくDを作り出すことを可能とし，またすでに存在するDSの拡張や修正を可能とする．

4. 記述の多重化や同期の問題，伝送の機構，ファイルフォーマットなどに関するシステムツール

ここで記述定義言語は，DS内のあるいはDS間の空間的，時間的，構造的，概念的な関係を表現できなければならない．またこの言語は，ひとつあるいはそれ以上の記述子と，記述子が記述するデータ間にリンクあるいは参照を張るためのモデルを提供しなければならない．現在のところこの言語の設計は，XML Schema[39]およびRDF (Resource Description Framework)[40]に大きく影響されている．

MPEG-7の中で特にマルチメディア情報の記述に用いられる規格はMDS (Multimedia Description Scheme)と名付けられ，その中で映像信号の特性を記述するためのメタデータとしてMPEG-7/Visualが

定められている. このメタデータは, 領域記述や輪郭線記述のための記述子を提供しており, 第7章で述べる台風雲パターンの記述などに使える可能性がある. しかしそれ以外の面では, 今のところMPEG-7を用いる利点はそれほど大きくない.

Weather Observation Definition Format (OMF)

OMF[42]はアメリカ海軍の気象用XMLマークアップ言語であり, 気象観測レポートという特定の文書を記述することを目的とする言語である. OMFは, 地上あるいは海上の気象観測, レーウィンゾンデとパイバル観測を組み合わせた気象観測, そして海温や海流などの海洋プロファイルデータを記述するためのものである. また重大な気象警報や気象予測を記述するための要素も含んでいる. このフォーマットは気象の基礎的な観測データを記述するには適しているが, 本研究のように気象衛星画像も含む記述を得るという目的にはあまり適していない.

4.4.2 データ操作言語

データ操作言語ではSQLが突出した地位を占めているが, それ以外にもいろいろな言語が, 特にデータ検索を意図して提案されている. また近年の大きな流れは, XML文書のような木構造のデータに関するデータ検索言語や, ネットワーク環境でのプロトコルまでを含めるようなデータ検索言語である.

Structured Query Language (SQL)

SQL(Structured Query Language)は関係データモデル上のデータ操作言語として著名な言語であり, 関係データベースにおいては事実上の標準言語となっている. その起源はE. F. Coddが提案した関係論理および関係代数であるが, SQLはこれらの論理および代数を具体的な構文規則によって記述することを可能としている. ただし実用的な側面をより重視して開発したため, 関係論理および関係代数に見られる美しい数学的構造は多少犠牲にされている. そこでSQLは誕生以来さまざまな改良を経ており, 最近はマルチメディアデータのための検索言語SQL/MMなど, 従来のような表形式のデータではない非構造的なデータへも適用範囲を広げてきている[43].

SQLの問合せは「何を対象とし(FROM), どのような条件を満たすもののうち(WHERE), この属性を示せ(SELECT)」という形で問合せを記述し, このような単純な構造にもかかわらず多様な検索問合せを記述できるという柔軟性が大きな特徴である. その他にもグルーピングのための操作(GROUP BY)や並べ替えのための操作(SORT BY)などの操作もSQLに備わっている. その他の拡張機能までを含めればきわめて実用的かつ確立された言語となってはいるが, 実際はSQLでは記述できない問合せもよく知られており, 決して万能のデータ操作言語ではない.

XML Query

XML Query[44]はXML文書に対する問合せ言語であり, XML文書の階層構造の中から問合せの対象となる部分を選び出し, その部分をフィルタリングし, その結果をツリー構造にするという操作を処理できる言語である. これらの処理はFLWR式とよばれ, FOR節, LET節, WHERE節, RETURN節から成る. 構文としてはSQLに類似した構文を持つXQueryと, XML構文を用いるXQueryXとがあり, XMLの普及とともに今後の発展が期待できる言語である.

Multimedia Retrieval Markup Language (MRML)

MRML(Multimedia Retrieval Markup Language)[45]は，本研究で提案するデータ操作言語とかなり近いアイデアに基づくマークアップ言語である．基本的にはマルチメディアデータの検索のための言語であり，検索に必要な操作をマークアップ言語として形式化している点に特徴がある．またネットワーク環境におけるクライアントとサーバ間の通信プロトコルとしての側面も明確に意識されており，ユーザインタフェースとデータベースサーバとを明確に分離することで，マルチメディア検索のための共通基盤を作り出すことを目的としている．このような視点は我々の研究と非常に近いものである．なお問合せの記述にはXML構文を用いており，その点も我々の研究と共通している．ただし，彼らが提案する言語ではマルチメディア検索の形式化がまだ未成熟であり，直交性に優れた操作を提供することには成功していない．また構造化が不十分なため，操作の指定とパラメータの指定が平坦に入り混じってしまっており，問合せがどのような操作を意図しているのかを読み取りにくい構造となっている．

4.5 地球環境データのためのデータ定義・検索言語の提案

本研究では以上に述べた既存のマークアップ言語を参考にしながら，独自のデータ定義・検索言語を提案する．またこれらの言語のXML構文を用いたマークアップについても研究する[31]．このうちデータ定義言語については，このような問題を一般的な枠組みで考えることも不可能ではないが（例えばRDF），その水準はいまだ初歩的水準にとどまっており，現状では問題領域に適切な文書型定義を個別に定めていくという，地道で発見法的な方法しかないのが実情である．となると問題は「標準化」の問題，つまり情報記述の整理・記述に関する教科書的な視点を見出す問題に帰着してしまい，そこからは有用な結論を導き出すことが難しくなる．

そこで本研究では主にデータ検索言語について議論し，データ検索言語としてGRQL（Grouping and Ranking Query Language)を提案する．これはその名前の通り，グルーピングおよびランキングを基本操作とする言語である．ただしこの言語におけるグルーピングとは，いわゆるデータ抽象化におけるグルーピングとは異なり，SQLにおけるGROUP BY演算子に近い意味で用いている．またランキングもSQLにおけるORDER BY演算子に近い意味をもつ．であれば，わざわざ新しいデータ検索言語を提案しなくても，既存の言語でこれらの機能は実現できるのではないだろうか，との疑問が生じるだろう．それでも本研究では，新しい言語を開発する価値はあると考えている．その理由を以下に述べる．

グルーピング

グルーピングとはある基準にしたがってデータを複数のグループに分割する操作である．このようにグルーピングを重視する理由は，本研究のようにデータマイニング目的にデータベースを利用する場合，データをグループに分割し，グループごとに処理を加えていく操作が多いためである．ただし本研究で用いるグルーピングは，既存の概念を大幅に拡張しており，ただ単に属性値の一致性のみでデータを分割する手法を意図するものではなく，データを分割する一般的な手法である種々のクラスタリング手法も広義のグルーピング手法とみなすものである．また概念としては，統計学における層化(stratification)，すなわち母集団をいくつかの層に分割するという操作の概念も含めたものである．

このようなグルーピング（分割計算）の機能は標準的データ操作言語であるSQLにも含まれている．しか

しこの機能には不十分な点がある. というのも, 現状のSQLにおけるグルーピングの機能は, グループごとに1組の値を対応させる計算(平均値や合計など)のために用意されており, 分割したグループごとにさらに操作を施すような操作は想定されていないのである. これは非常に大きな制約である. またデータ(あるいはデータベース用語でいえば実体)をグループに分割した後, 個々のグループに対してさらにグルーピング操作を加えるという入れ子操作についてもSQLでは実現できない. そこで本研究で提案する言語ではグルーピング操作の結果を再びグループと定義することにより, グルーピング操作を入れ子にすることが可能となり, グルーピング操作の柔軟性を大幅に向上させることができる. このようにグループを基本とするアイデアは, アレイ言語(配列を基本とする言語)にも類似した考え方である.

ランキング

ランキングとはある基準にしたがってデータを並べ替える操作である. これは画像類似検索などにおいて, 問合せに対するデータの類似度に応じてデータを並べ替えるという操作において必須の機能であるため, 本研究ではグルーピングと並ぶ基本操作として重要視している. ただし, 似たような操作はSQLでもORDER BYとして用意されているが, その違いは何だろうか.

まず関係データベースにおいては, 表の要素であるタプルの並びに特別な意味を見出さない. というのが基本的な方針であるため, ORDER BYは出力結果を整えるという, あくまで付加的な機能にとどまっている. したがって, 例えばある基準で並べ替えた上位5件のデータを対象に新たなグループを形成し, それに対して副問合せをおこなう, といった操作が実現できない[1]. それに対して本研究で提案するGRQLではこのような操作が可能であり, ランキングの結果をもとにグルーピングし, その結果のグループにまたランキングを適用し, といったように連鎖的にこれらの操作を適用することができる.

それに加え, 画像は一般に構造をもつデータ(抽象データ型)であり, しかも類似度の定義を画像表現モデルごとに独自の尺度を用意しなければならない. また並べ替えの際の基準となる数値は動的に定まるものであり, あらかじめデータベースに投入しておくことはできない. このような機能は既存のデータベースの拡張機能やオブジェクト指向データベースを用いれば実現できないことはないが, GRQLではこれらの機能を最初からできるだけ考慮した言語を設計する計画である.

なおこの言語はSQLと同様に, 問合せは宣言的に記述することを基本とする. つまり利用者は具体的な操作手順を記述せず, 何の情報を得たいかだけを問合せに記述する. また本研究では, データ操作言語のうちデータ検索にかかわる部分のみを提案する. というのも, 地球環境データの性質を考慮すると, データの削除や更新はほとんど生じず, したがってデータの登録・削除・更新などについて考慮する必要が薄いためである.

4.6 データ操作言語GRQLの概要

以上にまとめたように, 本研究で提案するデータ操作言語GRQLはグルーピングとランキングを基本とした言語である. GRQL (Grouping and Ranking Query Language)という名称は, その基本操作に着目した命名であるが, もちろんこの言語は他の側面も持っている. 例えばこのデータ操作言語の対象となるデ

[1] 我々の不勉強のため、すべての関係データベースを網羅的な調査したわけではないので、あるいはデータベースによっては可能なものがあるかもしれない。

ータは主に画像データや地球環境データなどのマルチメディアデータであり，データの種類に着目すればマルチメディア検索言語MDQL(Multimedia Retrieval Query Language)と命名することもできる．また使用環境としてはメタサーバ環境，あるいはインターネット上の3層アプリケーションアーキテクチャを想定しているため，ネットワーク環境で用いるプロトコルに着目することもできる．さらにXML構文を用いるマークアップ言語であることも一つの重要な側面である．

このように，この言語の命名には種々の観点が考えられるが，本研究ではその操作が最も本質的な特徴を表していると考え，この側面から言語を命名した．それは，SQLのようなデータ操作言語の本質が，対象とするデータにあるのではなく，データ操作の背景にある論理構造―関係代数や関係論理―にある，という考えを踏まえた命名である．GRQLの背景にある論理構造についてはまだ数学を用いた形式化が十分に進んでいないが，将来はそのような論理構造を明らかにすることで，この言語の存在基盤をよりくっきりと明らかにしていきたいと考えている．ただし実用面を考えれば，対象とするデータに特有の操作や，ネットワーク環境に対応するプロトコルなどの付加的な機能を実現していくことも，非常に重要な課題であることは間違いない．

そこで以下では本研究で提案するデータ操作言語GRQLをより具体的に見ていきたい．表 1はGRQLの基本的な操作および指示をまとめたものである．いずれもSQLやXML Queryなどの言語から語彙を借用している．上記の操作で最も特徴的なものはグルーピングに相当する操作GROUP-BY，およびランキングに相当する操作RANK-BYである．以下ではこれらの操作および記述の概略を説明する．

表 1 データ検索言語GRQLの基本操作および記述．

FOR-GROUP	問合せを適用するグループの選択操作．
WHERE	グループ内で検索条件を満たすデータを選択する操作．
TASK	グループ内のデータに対する演算や操作の指示内容の記述．問題領域に依存．
RANK-BY	グループ内のデータの並べ替え操作．
FETCH	グループ内から抽出するデータ個数の指定であり，特にRANK-BYと共に用いる．
RETURN	データの中から結果として取り出す属性の記述．射影に相当する．
GROUP-BY	グループ内のデータをさらに分割し，子グループを得るための操作．

FOR-GROUP

問合せQUERYを適用するグループを選択する．最上位の問合せではグループは1個しか存在しないため，この操作は効果を何ももたらさない．しかしこの操作が副問合せ(sub-query)に現れる場合には，上位の問合せでグルーピングされた個々の子グループに対する問合せを指定するという意味をもつ．

WHERE

グループ内で条件を満たすデータを選択する操作である．この条件を指定するためにFILTERという操

作を用意し，データ属性と問合せ指定との一致性に基づいて選択する機能などを用意している．またFILTERをANDやORを用いて結合し，複雑な条件を指定することもできる．

TASK

グループ内のデータに対する演算や操作の指示内容を記述する．これは画像のような複合オブジェクトに対する操作に必要なパラメータを記述する部分であり，オブジェクト指向データベースにおいては，オブジェクトに対するメソッドとして実装する部分に相当する．このような部分の操作を形式化することは非常に困難であり，操作方法が無数に考えられる中では，形式化できる部分は単純で典型的な操作に限られると考えるべきであろう．そこでTASKに関しては問題領域に依存した実装を積極的に進める．

RANK-BY

グループ内のデータを整列基準にしたがって並べ替える操作である．データ間の大小関係にしたがってデータを昇順あるいは降順に並べ替えるという単純な操作であるが，大小関係を形式的に定義することは意外に難しく，現状では属性に応じて適切な関数を用意する方法をとっている．

FETCH

グループ内から取り出すデータ個数を指定するものであり，RANK-BYと組み合わせることで検索を効率的に実行すること，および新たなグループを形成することに活用する．

RETURN

データの中から結果として取り出す属性を記述するものである．関係代数では射影に相当し，普通は最上位の問合せのみに適用する．

GROUP-BY

あるグループ内のデータをさらに子グループに分割する操作である．属性値が離散的であれば，属性値が一致するデータごとにグルーピングすればよいが，属性値が実数値であればこれを離散的な値に変換する必要がある．また本研究ではグルーピングの概念を拡張して，クラスタリングや層別化によるデータの分割もグルーピングの一種とみなしているため，通常のグルーピングよりは非常に多様なグルーピング操作を適用することが可能である．

以上の操作の組み合わせが問合せQUERYを構成する．このQUERYは，グルーピング後に生じるグループそれぞれにさらにQUERYを適用するという形で入れ子の問合せとすることも可能である．

本研究では以上の問合せをXML構文によって記述する．その文書型定義(Document Type Definition : DTD)を表 2に示す．この文書型定義はまだ不完全であり重要な部分のみを抜き出したものであるが，上記の操作に加えいくつかの属性の定義や出現順序なども定義している．要素queryの子要素として出現するfor-group要素は，問合せqueryへの副問合せとなる．さらに問合せXML文書のルートノードENVELOPEはheader要素およびbody要素から構成されており，全体としてheader要素はセッション情報やトランザクション情報の記述，そしてbody要素は問合せの中身の記述という構成になっている．

表 2 データ操作言語GRQLの文書型定義(DTD)の一部.

```
<!DOCTYPE envelope [
<!ELEMENT envelope (header, body) >
<!-- header -->
<!ELEMENT header (server?|transaction?) >
<!ELEMENT server (#PCDATA) >
<!ATTLIST server port NMTOKEN #IMPLIED>
<!ELEMENT transaction (#PCDATA) >
<!-- body -->
<!ELEMENT body (for-group)+ >
<!ELEMENT for-group (query)+ >
<!ELEMENT query (where?|target?|task?|rank-by?|return?|fetch?|group-by?|for-group*) >
<!-- target -->
<!ELEMENT target (select|match) >
<!ELEMENT select (#PCDATA) >
<!ELEMENT match (#PCDATA) >
<!-- where -->
<!ELEMENT where (and?|or?|filter*) >
<!ELEMENT and (and?|or?|filter*) >
<!ELEMENT or (and?|or?|filter*) >
<!ELEMENT filter (#PCDATA) >
<!ATTLIST filter select CDATA #REQUIRED >
<!ATTLIST filter type CDATA #REQUIRED >
<!-- task -->
<!ELEMENT task (example?|let?) >
<!ELEMENT example (dynamic|constant) >
<!ATTLIST example type CDATA #IMPLIED >
<!ELEMENT dynamic (#PCDATA) >
<!ELEMENT constant (#PCDATA) >
<!ELEMENT let (function?|measure?) >
<!ELEMENT function (#PCDATA) >
<!ELEMENT measure (parameters)? >
<!-- parameter is dependent on the type of measure, so here omitted -->
<!-- rank-by -->
<!ELEMENT rank-by (#PCDATA) >
<!ATTLIST rank-by order (ascending|descending) #IMPLIED >
<!-- return -->
<!ELEMENT return (select)+ >
<!-- fetch -->
<!ELEMENT fetch (from?|size?|to?) >
<!ELEMENT from (#PCDATA) >
<!ELEMENT to (#PCDATA) >
<!ELEMENT size (#PCDATA) >
<!-- group-by -->
<!ELEMENT group-by (select)+  >
<!ATTLIST select discretize-type CDATA #IMPLIED >
<!ATTLIST select discretize-unit CDATA #IMPLIED >
] >
```

4.7 ケーススタディ：台風画像検索

表 3は，GRQLを用いて記述したXML検索要求メッセージ，およびこのデータ検索要求に対するXML検索応答メッセージの例を示す．検索応答に関する文書型定義は省略するが，検索要求に対応するデータのリストを応答するのが基本的な動作であり，この場合は問合せ画像への類似度で整列した順序つきリストを応答メッセージとしている．またGRQLにおいて@マークで始まっている文字列は，データ検索エンジンで特別に解釈する文字列である．例えば@exampleは，問合せデータを意味する文字列で

ある[1]が，この検索要求例では問合せ画像をデータベースエンジン自身がランダムに選ぶように指示しているため，この値を利用者側が事前に知ることはできない．このように，データベースエンジン側で動的に検索要求を解釈する必要が生じるような場合に，このような一種のメタ文字を用いた指示をおこなう．

[1] このような意味そのものは、あらかじめ検索エンジンへ作り込むという形で実現するしかない。ただし「作り込む」とは、操作に対応する処理をあらかじめ検索エンジンに作り込む、というのと同等の意味においてである。

表 3 具体的な問合せに対応するXML構文によるGRQLのエンコーディング.

問合せ

1. 台風9903号から一つの問合せ画像をランダムに選ぶ. そしてこの台風系列に属する他の画像は検索対象から除く.
2. データベース中の画像を台風系列ごとにグルーピングし, グループごとに, グループ中のデータと問合せ画像との距離を計算し距離が小さい順にデータを整列する.
3. それぞれのグループから最大2件の画像を取り出して親グループに集約し, 再び親グループ内で距離の小さい順に整列する. その後親グループから最大5件のデータを取り出す. 結局これら一連の操作によって, 一つの系列から最大2件, 全体で5件の, 問合せ画像に類似した画像を検索することができる.
4. 検索した5件のデータに対して, 台風系列の名前と, 画像の名前, そして問合せ画像への距離を応答する. そして最後に問合せ画像に関する情報を応答する.

XML検索要求メッセージ

```
<?xml version="1. 0" encoding="UTF-8"?>
<envelope>
<header>
<server port="59300">localhost</server>
<session user="kitamoto" id="1">
<transaction>1</transaction>
</session>
</header>
<body>
<for-group>
<query>
<task>
<example type="single">
<constant select="folder">9903</constant>
<dynamic select="name">@random</dynamic>
</example>
</task>
<where>
<filter select="folder" type="equals"
   not="1">@example</filter>
</where>
<rank-by order="ascending">value</rank-by>
<fetch>
<from>0</from>
<size>5</size>
</fetch>
<return>
<select>folder</select>
<select>name</select>
<select>value</select>
<select>example</select>
</return>
<group-by>
<select>folder</select>
</group-by>
<for-group>
<query>
<task>
<let variable="value">
<function target="example">distance</function>
<measure type="euclid" option="squared">
<parameters>
<min>0</min>
<max>30</max>
</parameters>
</measure>
</let>
</task>
<rank-by order="ascending">value</rank-by>
<fetch>
<from>0</from>
<size>2</size>
</fetch>
</query>
</for-group>
</query>
</for-group>
</body>
</envelope>
```

XML検索応答メッセージ

```
<?xml version="1. 0" encoding="UTF-8"?>
<envelope>
<header>
<session user="kitamoto" id="1">
<transaction>1</transaction>
<matching>24300</matching>
<elapsed>0. 000000e+00</elapsed>
</session>
</header>
<body>
<example>
<folder>9903</folder>
<name>GMS599060113</name>
</example>
<list number="5">
<item order="0" id="0">
<folder>9902</folder>
<name>GMS599042809</name>
<value>1. 962298e+00</value>
</item>
<item order="1" id="1">
<folder>9514</folder>
<name>GMS595091908</name>
<value>3. 230482e+00</value>
</item>
<item order="2" id="2">
<folder>9915</folder>
<name>GMS599091606</name>
<value>3. 372034e+00</value>
</item>
<item order="3" id="3">
<folder>9509</folder>
<name>GMS595082415</name>
<value>4. 487362e+00</value>
</item>
<item order="4" id="4">
<folder>0003</folder>
<name>GMS500070219</name>
<value>5. 203874e+00</value>
</item>
</list>
</body>
</envelope>
```

4.8 メタサーバ環境への拡張

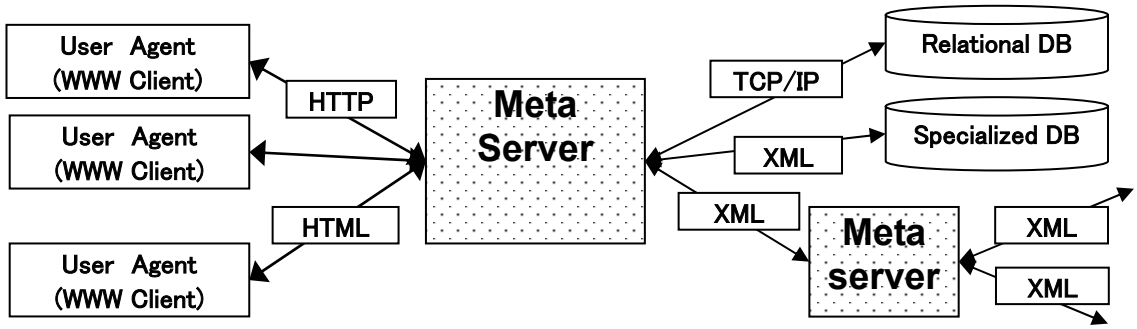


図 7 XML に基づく分散衛星画像データベースシステムの構成図。

本研究で提案するデータ操作言語GRQLは，さらにメタサーバ環境での活用も念頭に置いている[12]. ここでのメタサーバ環境とは図 7に示すように，クライアントとメタサーバ，データベースがネットワーク上で連係するという3層アプリケーションアーキテクチャの構成を踏襲するものである. そして下記のような動作を基本的な動作として想定する.

1. 利用者側のユーザエージェントは，適当なWWWインタフェース(FORMなど)を使って入力された問合せを，HTTPリクエストに載せてメタサーバに送信する.

2. メタサーバ側ではそれを適当な検索要求に組み立て，それぞれの問合せを送信すべきデータベースを選択して，検索要求メッセージを送信する. このときの問合せ言語として，バックエンドデータベースが関係データベースであればSQLを用いるが，8.2節で述べる特徴空間探索エンジン(FSE)に対しては，データ操作言語GRQLを用いてXML検索要求メッセージをエンコードし，TCPコネクション上で送信する.

3. 検索要求を受け取ったデータベースサーバは，検索結果を検索応答メッセージに組み立て，メタサーバに送信する.

4. 検索応答をすべて受け取ったメタサーバは情報を適切に加工し，利用者側のユーザエージェントに応答する.

このように，データベースへの直接的な問合せをメタサーバに担当させることで，柔軟な問合せ機構をサポートする階層的な構成が実現する. ただし，このような通信プロトコルに関するサポートは現在のところ不十分であり，(1)TCPコネクションを確立し，(2)メッセージを送信し，(3)コネクションを閉じる，といったきわめて基礎的なプロトコル部分しか実現していない. しかし将来的には，このようなプロトコルをさらに充実させていく必要がある. その最大の理由は，6.7節に述べる対話型検索のサポートである. このようなタイプの検索は第7章で述べるデータマイニングにおいても重要である. というのも，人間とコンピュータとの相互作用に基づくデータマイニングの実現は，この分野においても大きな課題と指摘されているからである. このような状況では，個々の検索トランザクションの管理にとどまらず，複数のトランザクションをまたがり利用者状態を保持することのできる，セッション管理をサポートする仕組みが必要となる.

4.9 今後の課題

本研究で提案するデータ操作言語GRQLは，一般的なデータ操作言語を研究する過程で生まれたわけではない．むしろ，第8章で述べるデータマイニングシステムIMET，あるいはこのシステムが動作するWWWサイト「デジタル台風」http://www．digital-typhoon．org/を構築する過程で，そこに必要となる機能をリストアップしながら既存のデータ検索言語に不満を感じたことがきっかけで生まれた言語である．このように経験的な試行錯誤から誕生したために，言語の形式化についてはまだあまり洗練されたデザインとはなっていない．

しかし本研究の問題意識は，関係データベースあるいは既存のデータモデルやスキーマに地球環境データの構造を無理やり押し込めるのではなく，地球環境データを適切に表現し効果的に検索するためにはどのようなデータモデルやスキーマが適当かを考えるべきである，というところにある．このように，そもそもデータベースの設計に戻って考え直すというアプローチは，現在のように多種多様なマルチメディア情報を扱う必要性が生じている時代では重要な問題意識であると考える．もちろんそこで問題となるのは，アドホックな形式化をできるだけ避け，常に数学的な形式化を意識する，という態度だろう．

最後にデータ定義言語およびデータ操作言語に関する今後の課題をまとめる．

データ定義言語

本研究では地球環境データに適したデータ定義言語を構築することを目指したが，現段階ではまだ発見法的な段階にとどまっている．ただし自分流の言語がたとえ標準規格と異なっても，単なるタグ名や情報内容の機械的な変換だけでこと足りるならば，XSLT（XML Stylesheet Transform)[46]を用いた変換によって互換性を容易に確保できる．ゆえに，このような意味では，自分流の言語であることは大きな問題ではないようにも思える．しかし，これはやや楽観的過ぎる見通しであるのも確かであり，本質的には情報内容が世界をみる見方に依存しており，その意味で互換性を確保するためにより高度で微妙な変換が必要になる，というところに問題がある．この問題については本質的な意味での解決策は現時点では存在せず，当面は実用的な観点から議論していくしかない．

データ操作言語

データ操作の方法もデータ定義と同様に無限の操作が考えられるという意味で形式化は困難であるが，関係データモデルにおける関係代数や関係論理のように，数学的構造を発見できる可能性があるという点で，データ定義言語とは事情がやや異なる．したがって独自に考案したデータ操作言語に対応する美しい数学的構造を発見できれば大きな成果であるが，画像検索のような問題領域でこのような美しい数学的構造を発見できるのかはまだよくわかっていない．ただし，このような形式化は確かに美しいものの，現実問題を解決するのに必要な操作を具体化し，抽象化し，洗練させ，形式化することで言語を組み上げていくという現実的試みも，数学的論理と同様に重要であると考える．またこのような現実的試みの先に，新たな形式化のパラダイムが見えてくることにも期待している．

一つの興味深い方向性は，利用者プレファレンスを考慮したデータ操作言語である．これは6.7節で述べるような対話型検索システムを念頭においたものであり，利用者フィードバックを用いて利用者のプレファレンスを学習する過程をサポートするものである．先述のMRMLにおいても個人化された検索への対応は考慮されており，GRQLも通信プロトコルをより強化することによって，このようなパーソナルなデータ操作言語へ発展させていくことも可能であると考えている．

第5章 地球環境データへのアクセスパターンのモデル

5.1 アクセスパターンを用いたキャッシング

地球環境データベースの基盤技術の一つに，大容量の地球環境データを効率的にネットワーク上で入手するための技術がある．このような技術の一つにキャッシング技術，すなわち1度アクセスされたデータを次のアクセスに備えて保持しておく技術がある．もし次回のアクセスがキャッシュに蓄積されたデータにヒットすれば，ネットワーク経由でのダウンロード，あるいは同一ホスト内のディスクI/Oやテープ I/Oといったボトルネックを回避できるため，全体としての利便性は大幅に向上する．したがって，効率的なキャッシングを達成するためには，データへのアクセス確率の偏りを発見し，アクセス確率が高そうなデータを優先的に保持する仕組みが不可欠である．

ところで，この「アクセス確率」とは，そもそもどのような値なのだろうか．実際のところ，キャッシングするかどうかを意思決定する時点では，個々のデータのアクセス確率は未知である．より正確に述べれば，真のアクセス確率は時間とともに変動するために，ある時点での真のアクセス確率は永久に未知であるかもしれない．しかし本研究ではキャッシングの問題を，あくまでアクセス確率の推定という文脈で捉えることを目指しており，真のアクセス確率を何らかの付加情報から推定するという枠組みに基づく．またこのとき，アクセス確率の推定値が真のアクセス確率に近いほどキャッシングが効率的になることも言える．

ただし，たとえ真のアクセス確率に近い値が推定できたとしても，そもそもアクセスパターンの分布が一様分布であれば，理論的にキャッシングの効果がないことが証明されてしまう．したがってアクセスパターンを用いたキャッシングにおいては，以下の2点に関する解析を進めていく必要がある．

1．アクセス確率分布に関する議論

2．個々のデータのアクセス確率の推定に関する議論

本研究では，地球環境データの特質を考慮した地球環境データへのアクセスパターンの解析，および個々のデータへのアクセス確率を推定する方法について研究する[16]．本研究で特に地球環境データに対するアクセスパターンを議論するのは，地球環境データへのアクセスパターンが，例えば一般のWWWデータのように様々な種類のデータを含むデータへのアクセスパターンとは，かなり異なるのではないかと予想するからである．またデータ更新や削除などの点で地球環境データには独特の性質があり，このような問題領域に依存した事前知識を活用した効率的なキャッシングについても検討したい．

5.2 地球環境データへのアクセスパターン

地球環境データへのアクセスパターンには以下のような特徴がある.

地球環境データは新鮮さが重要である

地球環境データ, 例えば気象衛星「ひまわり」画像などでは, 最新データを入手するためのアクセスが急増する場合がある. これは現在の天気を知りたいというニーズが大きいことを考えれば当然であるが, 継続的にデータを収集する人々からのアクセスも最新のデータに集中することが多い. そして, ある一定期間が過ぎた後はアクセス頻度が徐々に低下し, データを大量処理する研究者からのアクセスなど, ごく限られた利用者からのアクセスに限られてくる. ゆえに地球環境データに関しては, 「現在」から計測してデータがどのくらい「新鮮」であるかが, アクセス確率に決定的な影響を及ぼすと仮定できる. 似たような傾向は一般のデータウェアハウスについても指摘されており, これは経時的な大規模データベースにおいて利用者は新鮮なデータに興味をもつ, という一般的な傾向を示していると考えられる.

地球環境データは更新や削除がまれである

地球環境データは基本的に更新や削除がない. 更新や削除がもしあるとすれば, 測定後に判明した誤りの訂正, 不正データの削除, あるいは記憶領域の不足による過去データの削除, などの場合にしか発生しない. このような更新や削除に関する性質は, 一般のデータウェアハウスのように経時的にデータを蓄積する場合に共通した特質であるともいえる. このことはキャッシング効率にある種の効果を及ぼす. 例えばデータをキャッシュすべきかどうかを決定するのに, データ更新や削除を検査するために必要なコストや, キャッシュされたデータの一貫性を確保するためのコストを考慮する必要がなくなる. このことは, キャッシングに関係する問題を, 一般の場合に比べて単純化してもよいことを示している.

本研究ではこのようなキャッシング技術を, 分散型地球環境データベースの中のメタサーバで活用する計画である. 本研究で想定する分散型地球環境データベースのネットワーク構成を図 7に示す. ここでは利用者は以下のような方法でメタサーバを利用する.

1. 利用者はメタサーバにアクセスし, 検索要求メッセージをメタサーバに対して送信する.

2. メタサーバは検索要求メッセージを解析し, もしその検索要求がメタサーバ内に保持されているデータに対する検索要求であれば, メタサーバのデータを利用者に送信して検索を終了する.

3. もしそれがメタサーバ内に保持されていないデータであれば, メタサーバは個々のデータベースサーバに対して, 条件を満たすデータをメタサーバに送信するように要求する.

4. 個々のデータベースサーバはメタサーバに対して, 条件を満たすデータを送信するが, 条件指定の粒度によっては, 本来の検索要求に比べて多めのデータが返される場合もある.

5. メタサーバはデータを受け取ると, 検索要求に適合したデータを利用者に送信し, それ以外のデータは自身のキャッシュ領域に保持する.

6. メタサーバは定期的にキャッシュ領域を確認し, キャッシングポリシーにしたがってキャッシュデータを整理する.

このようにメタサーバにキャッシュ領域を確保すれば，地球環境データのオリジナルデータが蓄積されているデータベースサーバに向けて利用者が直接アクセスする必要性が減少するため，メタサーバが利用者に近い場所に存在する限り，全体的なデータ転送効率を向上させることができる．

さて本研究では，地球環境データのアクセスパターンを以下の4つの観点から解析しモデル化する．

1．アクセス頻度の偏り

2．アクセス時刻間隔の偏り

3．観測時刻とアクセス時刻の間隔の偏り

4．解像度に対するアクセス頻度の偏り

これらの観点は，網羅的なリストでもなければ，演繹的に導かれた指標でもないが，地球環境データのアクセスパターンに関する我々の仮説を反映したものである．また上記の4に関する研究は十分に進めることができず，結果としてまだ仮説にとどまっている．ゆえに本報告では，主に上記の1～3に関して実際の地球環境データのアクセスパターンを解析した結果を述べ，4については関連する研究に関連する初歩的な実験結果および今後の研究の方向性についてまとめる．

5.3 キャッシング技術

キャッシングという考え方自体は以前から情報技術のあらゆる分野で研究されてきた研究課題であり，その有効性も広く実証されている．そしてこのキャッシング技術が，ネットワーキング技術という文脈において注目を集めたのは，WWWの爆発的な発展に対応するための現実的な解決策の一つ，という理由が大きい．ネットワーク的に利用者により近い場所に，アクセス頻度の高いデータのみを保持するキャッシングサーバを設置すれば，もとのサーバへのアクセスを減少させる効果があり，結果的にネットワーク全体でのデータ配信の効率性と信頼性を高めることができる．本研究はこのような観点からのキャッシング技術を，特に地球環境データという文脈で考え直してみたい．

キャッシング技術を特徴付けるものは，キャッシュ管理のための削除ポリシーである[47][48]．キャッシングサーバにおいてデータを保持しておくことができるメモリ領域あるいはディスク領域は有限であることから，重要度の低いデータを効果的に発見し削除するための戦略を発見することが，キャッシング性能の向上の鍵を握ることになる．このような戦略に関する提案は既に膨大な数に上っているが，中でも最も基本的な戦略と考えられているのが，最もアクセス間隔が長いデータを削除するLRU（Least Recently Used)や，最もアクセス頻度が低いデータを削除するLFU（Least Frequently Used)などの戦略である．そこで，これらの基本的な戦略に加えて，いかに問題領域の特性を加えた戦略を考案できるかが．重要な研究課題になる．またネットワーク環境を考えれば，オリジナルデータを取得するのに要するネットワーク転送時間を考慮したキャッシングポリシーなども考慮すべき対象である．

以下ではキャッシングポリシーを最適化問題として定式化するため，いくつか必要な定義と目的関数を導入する．まず地球環境データの集合 D を $D=\{D_1,\ldots,D_N\}$ とする．地球環境データの集合は，あらかじめすべて把握しておくことが可能な場合も多いため，集合の大きさ N は既知であると仮定する．次に個々のデータに対してキャッシュ決定関数 $x_i \in \{0,1\}$ を定め，データ D_i がキャッシュされていれば $x_i=1$，キャッシュされていなければ $x_i=0$ となるような変数として定義する．またデータ D_i のサイズを

s_i，このデータの入手に要するデータ転送速度をv_iとし，キャッシュ領域のサイズをSとする．

次に，キャッシングポリシーの「良さ」を表す指標となる目的関数を定義する．まず最も広く使われる指標であるキャッシュヒット率を定義する．時刻tにおけるアクセスに関するキャッシュヒット率を表す目的関数$Q(t)$およびそれに関連する最適化問題は，以下のように表すことができる．

$$\max_x \quad Q(t) = \sum_{i=1}^{N} x_i(t) q_i(t) \tag{1}$$

$$\text{subject to} \sum_{i=1}^{N} x_i(t) s_i \le S \qquad \sum_{i=1}^{N} q_i(t) = 1 \tag{2}$$

ただし$q_i(t)$は時刻tにおけるデータD_iへのアクセス確率である．この目的関数は個々のデータのサイズを考慮していないが，データサイズによる重み付けを考えれば，以下のサイズ加重キャッシュヒット率といった目的関数を考えることもできるだろう．

$$\max_x Z(t) = \sum_{i=1}^{N} x_i(t) s_i q_i(t) \tag{3}$$

さらに，上記の目的関数とは逆の観点から，キャッシュでヒットしなかったデータを改めて取得するために要するコストを最小化することを考えれば，以下のような目的関数にたどりつく．

$$\min_x \quad V(t) = \sum_{i=1}^{N} (1 - x_i(t)) v_i s_i q_i(t) \tag{4}$$

$$\text{subject to} \sum_{i=1}^{N} x_i(t) s_i \le S \qquad \sum_{i=1}^{N} q_i(t) = 1 \tag{5}$$

ここでは問題を単純化するために，データ転送速度v_iは時刻によって変動しないと仮定した．以上のように種々の目的関数に応じて，これを最大化するようなキャッシュ決定関数x_iを定めることができる．このキャッシュ決定関数を適当な時点で計算し，$x_i = 1$となるようなデータのみをキャッシュ領域に残せば，目的関数を最大化するという意味での最適なキャッシングが実現する．

上記の問題は，一般に組み合わせ最適化問題とよばれる問題に属する．特に上記の問題の場合は0－1ナップザック問題という，よく研究された問題に帰着させることができるため，この問題に対して考案された種々の高速アルゴリズム（分枝限定法）などを活用できる．ただしこの問題は理論的にはNP困難な問題であることから，たとえ効率的なアルゴリズムがあるとしても，データ数が巨大な場合は何らかの

近似計算が必要になる.[1]

さて本研究では, このような最適化問題そのものよりは, むしろ時刻 t におけるデータ D_i へのアクセス確率 $q_i(t)$ をどのように推定するかという問題に関心がある. これは時刻 t のアクセスがデータ D_i となる確率であるため, 当然のことながら未知の値である. ゆえにデータに関する事前情報を活用してこの値を推定することを考える.

こうして推定したアクセス確率を $\hat{q}_i(t)$ とする. 例えばキャッシュヒット率を最大化する場合には, 推定したアクセス確率に対して最適化されたキャッシュ決定関数 $\hat{x}_i(t)$ を得ることができる. ここで, このように推定したキャッシュ決定関数を用いて, 真のアクセス確率 $\tilde{q}_i(t)$ に対するキャッシュヒット率を考えてみると, 推定キャッシュヒット率は真のアクセス確率と推定キャッシュ決定関数の積, すなわち $Q^*(t)=\sum_{i=1}^{N}\tilde{q}_i(t)\hat{x}_i(t)$ となり, これは真のキャッシュヒット率 $\tilde{Q}(t)=\sum_{i=1}^{N}\tilde{q}_i(t)\tilde{x}_i(t)$ を用いて計算したキャッシュ決定関数を用いた場合のキャッシュヒット率よりも小さな値となる.

また上記の関係を別の関係から分析してみる. 推定アクセス確率で最大化した目的関数を $\hat{Q}(t)=\sum_{i=1}^{N}\hat{q}_i(t)\hat{x}_i(t)$ とすると, 真のキャッシュヒット確率と推定キャッシュヒット確率との差は以下のように表すことができる.

$$\begin{aligned}\tilde{Q}(t)-\hat{Q}(t)&=\sum_{i=1}^{N}\tilde{q}_i(t)\tilde{x}_i(t)-\hat{q}_i(t)\hat{x}_i(t)\\&=\sum_{i=1}^{N}\tilde{q}_i(t)\left(\tilde{x}_i(t)-\hat{x}_i(t)\right)+\left(\tilde{q}_i(t)-\hat{q}_i(t)\right)\hat{x}_i(t)\end{aligned}\tag{6}$$

つまり真のキャッシュヒット率と, 推定したキャッシュヒット率との差は, 上式のように2つの項に分割して表現できることがわかる. このうち第1項はキャッシュ決定関数が真のアクセス確率に応じて最適化されていないことに起因する差, また第2項はアクセス確率の推定値が真のアクセス確率と一致していないことに起因する差, に相当する. ゆえに推定アクセス確率を真のアクセス確率にできるだけ近づけることが, 重要な問題となるのである.

5.4 アクセス確率の推定

そこで本研究では, 地球環境データのアクセスパターンに基づき, 個々のデータのアクセス確率を推定するためのモデルを確立する. 以下ではそれぞれの要素について, 実際に地球環境データのアクセスログを解析しながら検討する. 対象とするアクセスログは, 東京大学生産技術研究所が提供する地球環境データベース[49], 特にひまわり衛星画像へのアクセスログである.

[1] このようなパフォーマンスの問題から、実際のキャッシングサーバにおいてはLRUなどの初歩的なアルゴリズムが多用されるようである。

このアクセスログに対して，具体的には先述の4つの要素からアクセス確率を推定するモデルを追究する．理想的には4つの要素が独立であれば，アクセス確率はこれら4要素について独立に求めた確率の積として求めることができる．このような独立性の仮定は厳密には成立しないと考えられるが，本研究では簡単のため独立であると仮定する．また個々の利用者ごとのアクセスパターンは考慮の対象から外し，アクセスごとの独立性，つまり時刻 t のアクセス確率と時刻 $t+1$ のアクセス確率が独立であると仮定する．むろん同一の利用者が連続してアクセスすればそこには強い相関が生じるが，かといってマルコフモデルなどの履歴を考慮するモデルを用いると，モデル構築や統計情報収集が困難となる．ゆえに本研究では利用者ごとの履歴を収集せず，すべてのアクセスを独立と仮定する．

次に個々の要素のモデル化については，一般にWWWのアクセスパターンがジップ法則(Zipf's law)あるいはジップのような分布(Zipf-like distribution)にしたがうことが多いとの報告を重視し[50]，ここではジップ法則にあるようなべき乗型の分布 $P \sim \rho^{-\beta}$ を用いる．そしてグラフの傾きからべき乗の指数を求め，その指数の値からアクセスパターンの偏りについて議論する．なおこの部分は予備的な実験のため，最小二乗法のような厳密なフィッテイングはおこなっていない．

5.4.1 アクセス頻度の偏り

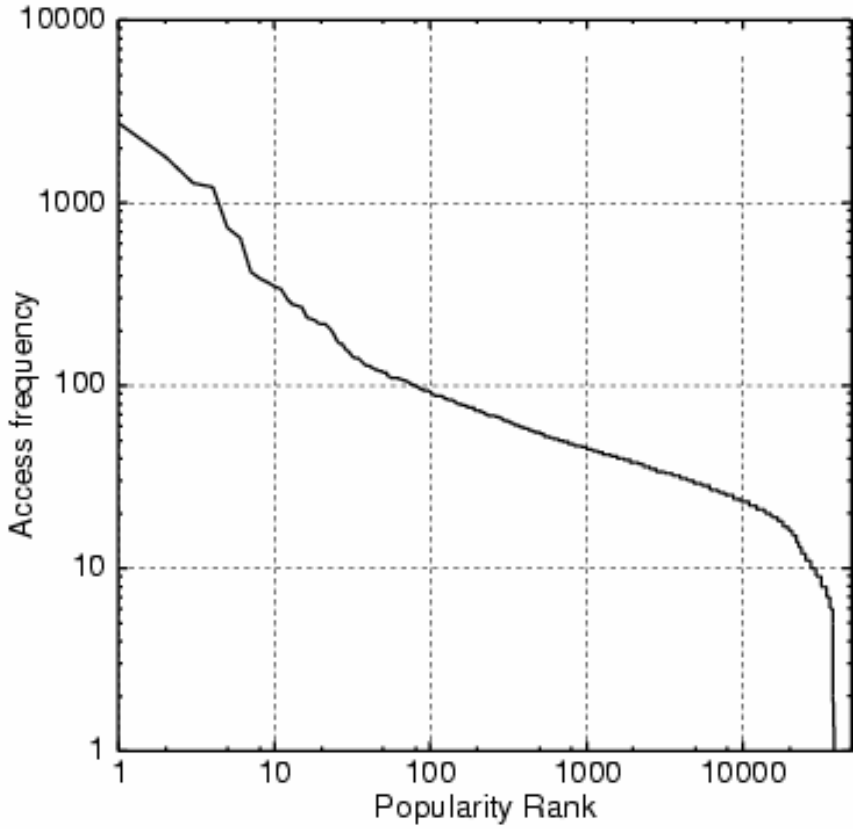


図 8 データに対するアクセス頻度の偏り.

アクセス確率を考えるために，まず個々のデータに対するアクセス頻度の偏りを考える．ここでは個々の

データD_iごとにアクセス回数A_iを調べ，アクセス回数の多い順にデータを並べ替える．ここからアクセスランキングごとのアクセス確率を推定したい．このようなデータの「人気度」を活用したキャッシングポリシーとしては，代表的な例としてLFU(Least Frequently Used)が知られており，最もアクセス頻度が少ないデータから順番に削除するというポリシーに基づき，アクセス頻度の多いデータは今後もアクセスされる確率が高いという経験則を利用している．そこで図 8では，データに対するアクセス頻度の偏り，つまりデータの人気度を調べてみた．このグラフは38631件の衛星画像に対する728697件のアクセスから計算したものである．

図 8のグラフを眺めると，およそ上位100位を境にして，アクセス頻度が高いグループとアクセス頻度が低いグループとに分割できることがわかる．その理由をモデル化するのは困難であるが，例えば台風が接近するとアクセスが増加するなどの現象を考慮すると，例えば災害などの現象に関連するデータが前者のグループに属するのではないかと推測できる．一方後者はランダムなアクセスが卓越するような，あまり重要ではないデータに対応すると考えられる．

このような知見を数学的にモデル化するために，データに対するアクセス頻度の偏りを2つのべき乗側の混合として表現することを考える．すると，大まかなフィッテイングの結果より，ランキングが高い（100位以内）部分は$\beta = 0.8$のべき乗則，またランキングが中程度（100位から10000位）の部分は$\beta = 0.3$のべき乗則で近似できることがわかる．また，さらにランキングが下位（10000位）の部分については，そもそも頻度が小さすぎてデータの信頼性に欠けるため，この部分はモデル化しない．

このようなべき乗則を用いてアクセス確率を推定するには，まずあるデータのランキングを調べ，その値を上で求めたべき乗則の式に代入すればよい．なお確率を計算するためには，確率の総和が1になるような正規化定数が必要である．そのためにまず，ランキング上位のべき乗則の方から正規化する．これはそれぞれのべき乗則が当てはまる最小の値から最大の値までの値を足し合わせ，おのおので比例配分することによって計算できる．この方法は厳密な数学的方法ではないが，これを厳密におこなう方法については今後の課題とする．

5.4.2 アクセス時刻間隔の偏り

次にアクセス時刻間隔について調べる．アクセス時刻間隔とは，あるデータD_iに対するj回目のアクセス時刻$T_i(j)$から，$j+1$回目のアクセス時刻$T_i(j+1)$までの差$T_i(j+1)-T_i(j)$を指すものであり，これをすべてのデータに対して調べて分布をモデル化し，そこからアクセス確率の推定に有用な知見を得ようというのがここでの目的である．このようにアクセス時刻間隔に関連するキャッシングポリシーとして代表的なのがLRU(Least Recently Used)，つまりアクセス時刻間隔が大きいデータから削除していくというポリシーである．図 9ではデータに対するアクセス時刻間隔の偏りを，アクセス時刻間隔を計算できた595789件から計算したものである．

図 9によると，アクセス時刻間隔はほとんどが2日あるいは3日以内に集中している．これは，最新のデータへのアクセスがほとんど3日以内に集中しており，この間にはアクセス時刻間隔が非常に短いという性質に対応している．その後はアクセス時刻間隔が長くなるたびに，そのような事例が発生した件数は単調に減少していき，例えば前回のアクセスから100日経過後にもう一度アクセスされるという事例は，全体595789件のうち，1日あたり200件以下ほどであることがわかる．つまり，アクセス時刻間隔が空いてしまったデータに再度アクセスがある可能性は高くないと言えるのである．

この性質を数学的にモデル化するため，ここではグラフ全体を単一のべき乗則で近似する．本来のジップ法則の意味から考えれば，このように時間間隔を横軸としたグラフにべき乗則を当てはめることに多

少の違和感はあるが，単なる関数当てはめとしてここではべき乗則の関数を用いる．すると係数はおよそ $\beta = 1.3$ と，かなり速やかに減少する関数が得られた．以上の関数に対する正規化定数の計算は前項と同様である．

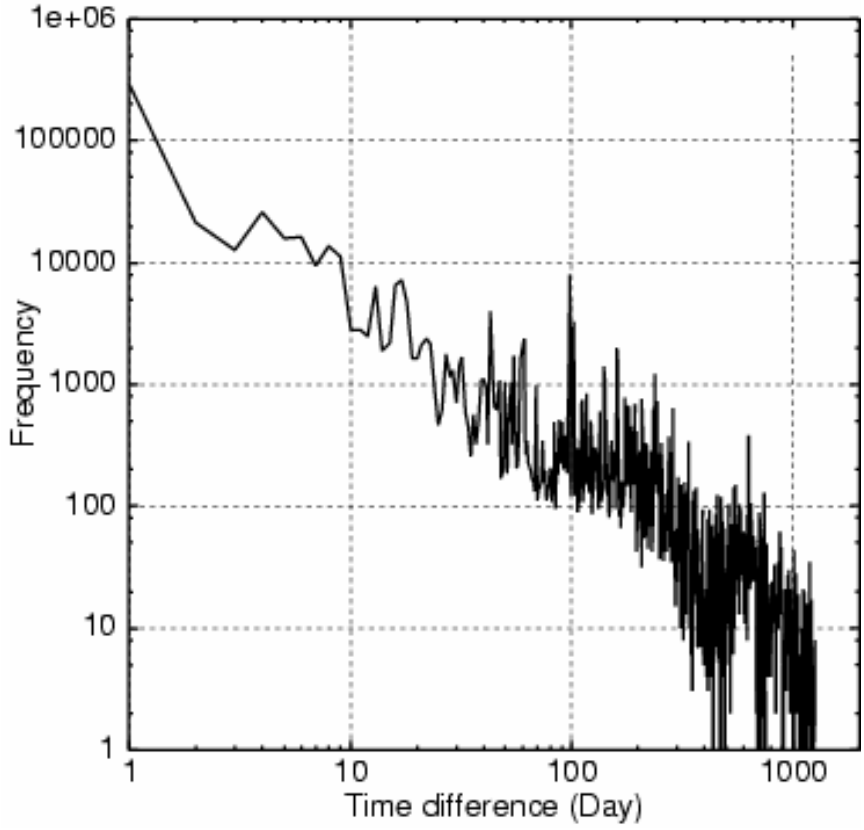


図 9 データに対するアクセス時刻間隔の偏り.

5.4.3 観測時刻とアクセス時刻の間隔の偏り

データの「新鮮さ」に関するアクセスの偏りは，経時的にデータを蓄積していく場合に特に顕著な現象である．すなわち，データ D_i には観測時刻 R_i という時間が関連づけられており，観測後の更新や削除がめったに起きない，という場合である．地球環境データは一つの典型的な例であり，またデータウェアハウスでもこのような性質はしばしば観察されている．つまり，できるだけ新しい情報を得たいという利用者の要求を考慮すると，データの「新鮮さ」に基づくキャッシングポリシーが有効であると考えられるのである．そこで，観測時刻とアクセス時刻の間隔の偏りを調べてみる．具体的にはデータ D_i への j 回目のアクセス時刻 $T_i(j)$ について，$T_i(j) - R_i$ を観測時刻とアクセス時刻の間隔と定義する．これをすべての i および j について調べると，観測時刻とアクセス時刻の間隔に関する偏りの分布を得ることができる．これを計算可能な671013例について計算した結果が図 10である．

図 10のグラフは，おおよそ3個の部分に分割することができる．まず観測時刻直後には，非常にアクセス頻度の高い期間が1日から3日程度持続する．この部分は，新鮮なデータを入手したいという，利用者の強い要求を反映したものである．このピークの後には，時間の経過とともに頻度が緩やかに減少す

る期間が，観測後3日から100日程度続く．最後に受信後100日以後の期間は，頻度がほぼ一定値に落ち着き，データに対するランダムなアクセスが卓越していることを示唆している．この100日という数字には具体的な根拠[1]は見つかっていないが，約3ヶ月で新鮮なデータに対する需要がほぼゼロになることは，これ自体興味深い知見であると考える．

以上の性質を3個のべき乗則の混合として近似する．まず初期の3日程度は $\beta = 4.7$ のべき乗則となる．次にその後100日までは頻度は緩やかに減少し，この部分はおおよそ $\beta = 1.0$ のべき乗則で記述できる．最後に100日以後はおおよそ $\beta \sim 0.0$ に近くなっており，この期間は完全にランダムなアクセスとみなしてもよいと考える．

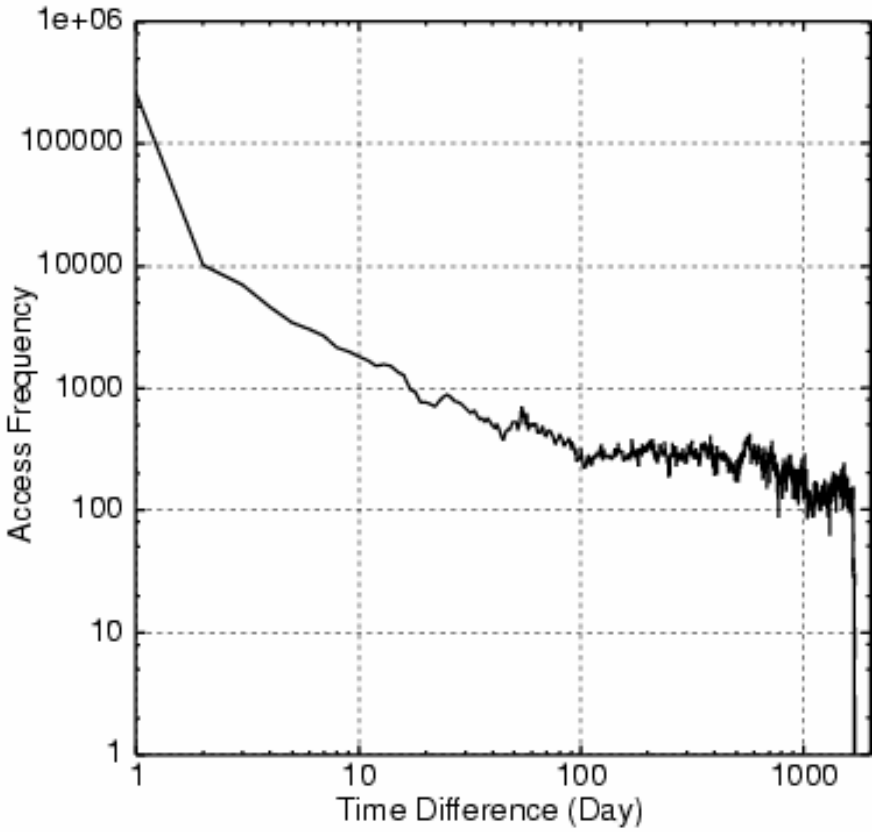


図 10 データ受信時刻とアクセス時刻の間隔の偏り.

5.4.4 解像度に対するアクセス頻度の偏り

地球環境データの中でも，特に地球観測衛星データのような画像データを対象とするならば，データの解像度を活用したキャッシングも有効であると考えられる．というのも，地球観測衛星データへのアクセスは，そのほとんどが低解像度のデータ，あるいは概要画像へのアクセスに集中しているからである．利

[1] 「人のうわさは75日」ということわざは，情報の新鮮さに対する人々の意識を表現していると考えられる．

用者の通常のアクセスパターンとは以下のようなものである.

1. 低解像度のデータを用いて, データ内容を確認したり比較したりしながら, 最終的に必要なデータを絞り込む.

2. 高解像度のデータについて, 必要なデータのみをダウンロードする(ただし専門家のみ).

特に高解像度絵のデータはサイズが巨大であり, しかもより高度な処理技術が必要となることから, これらのデータへのアクセスはほとんど専門家に限定されている. したがって低解像度のデータを優先的にキャッシングすることで, ほとんどの人にとってヒット率が向上する, という効果が期待できる. またキャッシングされた低解像度データが高解像度データの復元に使えるような画像符号化法を用いれば, キャッシュされた低解像度データは無駄とはならず, より効率的なキャッシングが実現できる.

この方法は, キャッシュ決定関数を $x_i \in \{0,1\}$ ではなく $x_i \in [0,1]$ のような実数値とみなし, この実数値が解像度に対応するようなソフトなキャッシングポリシーともみなせる. つまり, 低解像度のデータを多数キャッシングすることにより $x_i = 0$ というケースが減少し, 低解像度データへの要求を満たすことでヒット率が大きく向上するのである. しかし本研究では, この仮説を裏付けるような実際のアクセスログが入手できなかったため, このような予想を定量的に証明することはできなかった. したがって以下では, このようなキャッシングの基礎となる画像の多重解像度符号化について簡単に紹介する.

5.4.5 ウェーブレット変換を用いた画像の多重解像度符号化

このような解像度に応じた符号化法として, 利用者の要求に応じて部分的なデータを任意の解像度で符号化し送信するという, 段階的な符号化方式が有用であることが知られている[51]. この方式は以下のようなシナリオの場合に有効である.

1. 利用者は概要画像を探索しながら, 必要な画像およびその画像中の必要な領域を決定する.

2. 必要な領域についてさらに高解像度の画像をダウンロードして探索する.

3. 以上の操作をオリジナルデータの解像度に到達するまで繰り返す.

このとき, 概要画像という低周波領域の画像を, 高解像度画像を復元する際のデータの一部として使うことができる, というのが段階的な符号化方式の中心的アイデアである. そのため低周波領域は改めて送りなおす必要がなく, 高周波領域のデータだけを送信することで, 高解像度画像を完全に復元することが可能となる. ゆえにこの方法は, 大規模な地球観測衛星データを対象に, 解像度を変化させながら探索的に検索する場合には有効な方法となる.

このような符号化方式は, ウェーブレット変換[53]との関連で近年注目を集めるようになってきた. これは, ウェーブレット変換を用いて画像信号を複数のスケールに分割して表現(多重解像度表現)し, 完全再構成なフィルタバンク構造を用いることにより, 可逆な符号化を実現するという方法である. 本研究ではこのように, ウェーブレット変換を用いた無損失の(可逆の)符号化方式のみに関心がある. その理由は, 地球環境データの場合は個々の画素値ごとに意味があり, 見た目に変化がないから多少の損失があってもよい, という評価基準は無意味なためである.

このような可逆のウェーブレット変換のひとつにS+P変換がある[52]. この変換はHaar変換と類似した変換であるが, 特に観測信号が整数値となるデータに対するウェーブレット変換に適したものである. この変換を繰り返し低周波画像に適用することによって, Mallatの多重解像度ピラミッド表現を得る[53]. その概念図を図 11に示す.

これは階層的ピラミッド構造による空間的周波数の階層的分解を示しており, 左上のブロックは最も低周波(最も低解像度)の信号に対応するブロック(これをスケール L_0 と表記する)である. このブロックは画像の大まかな構造を保持しており, この部分だけで概要(クイックルック)画像として用いることができる. これに対して, このブロックに隣接する3個のブロックは, これに続く低周波の信号を表現しており, 主に低周波領域で表現しきれなかったエッジ部分などに対応する信号が現れてくる. これらを同様にスケール L_1 とする. 次のスケールは L_2 であるが, ここで注意すべきことは, ピラミッドが間引き構造になっているために, スケールがあがるごとにブロックの大きさが大きくなる点である. そこでスケール L_2 のブロックは図 11に示すように 4×4 のサブブロックに分割し, L_0 のブロックと同じ大きさになるようにする. この操作を同様に繰り返すことで, 大きさが同じで種々の空間周波数と位置に対応するブロックが誕生する. 図 11の場合は L_3 まで想定しているので, $2^3\times2^3=64$ 個のブロックが生まれる. これらを別々に管理することで, 段階的な符号化法を実現する. これを具体的なシナリオに沿って説明する.

1. クライアントはまず概要画像を要求する. これは L_0 のブロック0に対応する. この概要画像を見ながら, 利用者は画像の右上方に興味深い対象物を見つける.

2. 次にクライアントはそれをもう少し拡大して眺めるために, 1段階高い解像度の画像を要求する. このとき L_1 スケールのブロックをすべて送信するのではなく, 画像の右上方を復元するのに必要なデータのみを送信すればよい. したがってクライアントはブロック1を受信し, これと既に受信したブロック0を組み合わせてクライアントは画像を復元する.

3. 同様にクライアントからのさらに高解像度な画像への要求に応じて, ブロック3 (L_2)およびブロック6 (L_3)を送信し, クライアントはこの時点でオリジナルデータの部分画像を完全に復元することができる.

このような画像符号化法によって, 画像全体のデータを送信するよりも少ないデータ量でオリジナルデータの部分画像を完全に復元することが可能である. もちろん, 衛星画像全体の中で利用者が必要とする部分や解像度が既知であれば, そのデータのみを符号化し送信したほうが効率的である. しかし利用者がデータを眺めながら対話的に検索し, かつ最終的にはオリジナルデータの一部のみを必要とするならば, このような方法は有効であると考える.

本研究ではウェーブレット変換に基づくこのような階層的符号化法を気象衛星画像に対して適用し, 解像度による圧縮効率の違いなどを調べた. 実験の対象は気象衛星NOAAの衛星画像であり, この画像には第3章で述べたようにタイで受信する気象衛星NOAAの衛星画像を用いている. その理由は, タイで受信したデータを日本のサーバにキャッシングし, 日本の利用者に対して提供する, といったサービスを念頭に置いているためである. この画像は 2048×5296 画素という巨大な画像であるが, これに対してウェーブレット変換を適用し図 11の分割を計算してみる.

まずウェーブレット変換の結果を図 12に示す. ウェーブレット変換としてはHaarウェーブレットを用いている. この結果は気象衛星NOAAの画像をスケール L_1 まで分割した画像である. 右上の画像は大きさが4分の1の概要画像となっており, それ以外の領域は概要画像では記述しきれないエッジ情報を示し

ている. これら4領域の情報を再び組み合わせることでオリジナル画像が完全に復元できることに注意する. つまりこの変換は可逆な変換である.

次にこのウェーブレット変換をさらに適用して得られるブロック分割に対して, それぞれのブロックのエントロピーを計算した結果を図 13に示す. これはもともとの画像を$16\times16=256$ブロックに分割して計算したものであり, おのおののブロックは128×331画素のブロックとなっている. その結果, 画像の一般的な性質から予測できるように, 最も低周波領域の画像が符号量最大となる一方, 高周波領域の部分画像の符号量は前者と比べて25%から75%少なくなることが判明した. これは, 低周波領域に加えて高周波領域を再現するために必要となる付加的なデータ量は, データ全体をそのまま符号化するよりは, かなり小さくなることを示唆している.

このようなウェーブレット変換に基づく画像符号化法は, 実際には研究プロジェクト期間中に標準化が進み, 上記の類似アルゴリズムがJPEG (Joint Photographic Experts Group) 2000[54]という新しい規格に実装されることとなった. この符号化は可逆画像符号化および段階的符号化がシームレスに統合されているため, 本研究のように可逆符号化が必要な用途にも十分に活用できる規格である. したがってJPEG2000の普及に伴って, このような解像度に応じたキャッシング技術に注目が集まることがあるかもしれない. 本研究では実際にJPEG2000のコーデックを改良してこのようなキャッシングアルゴリズムを実装する段階には至らなかったが, 今後はさまざまな形でこのようなキャッシング技術が普及する可能性はある.

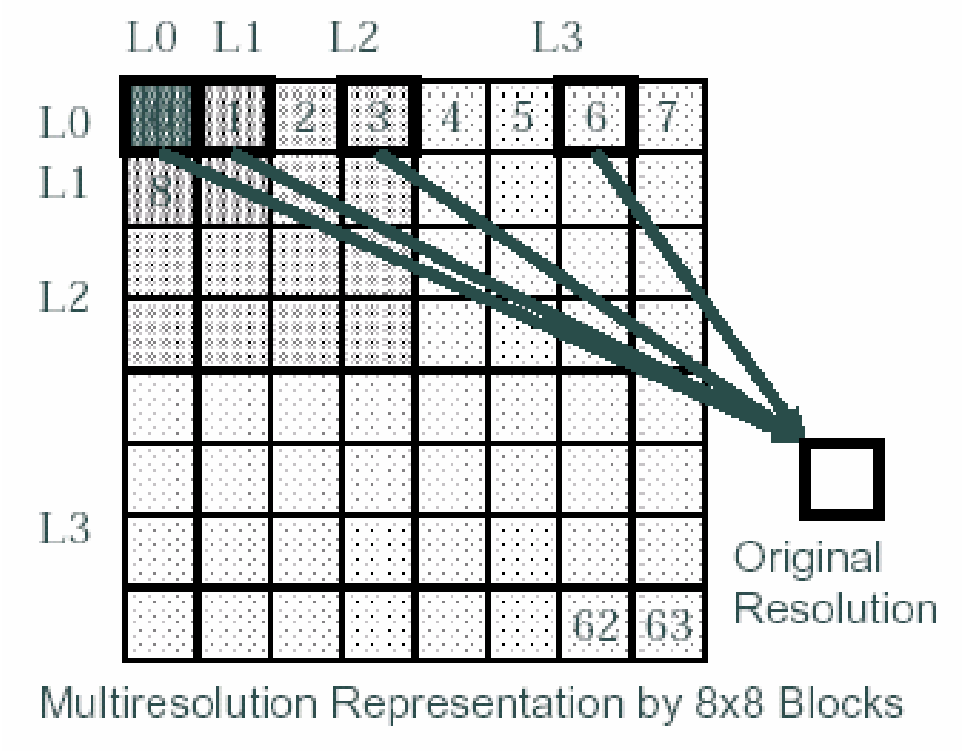


図 11 ピラミッド構造に基づく多重解像度画像表現とブロック分割の概念図.

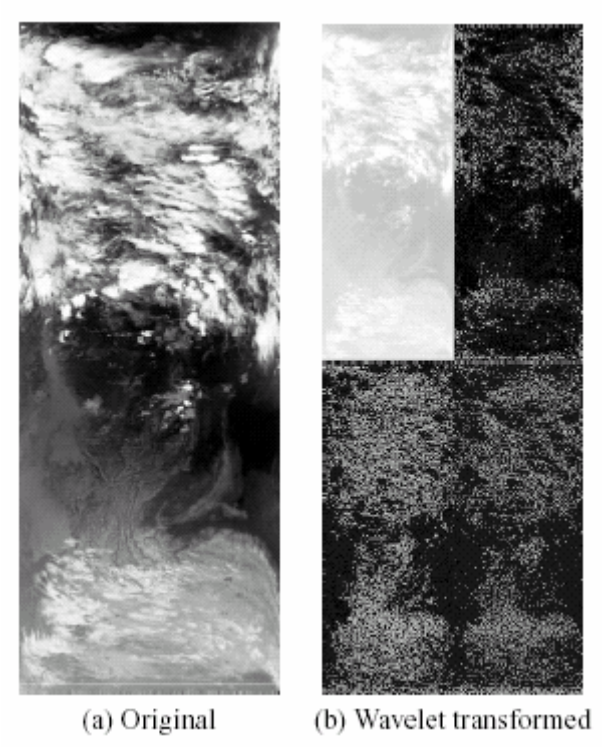


図 12　気象衛星NOAA画像に対するウェーブレット変換の適用.

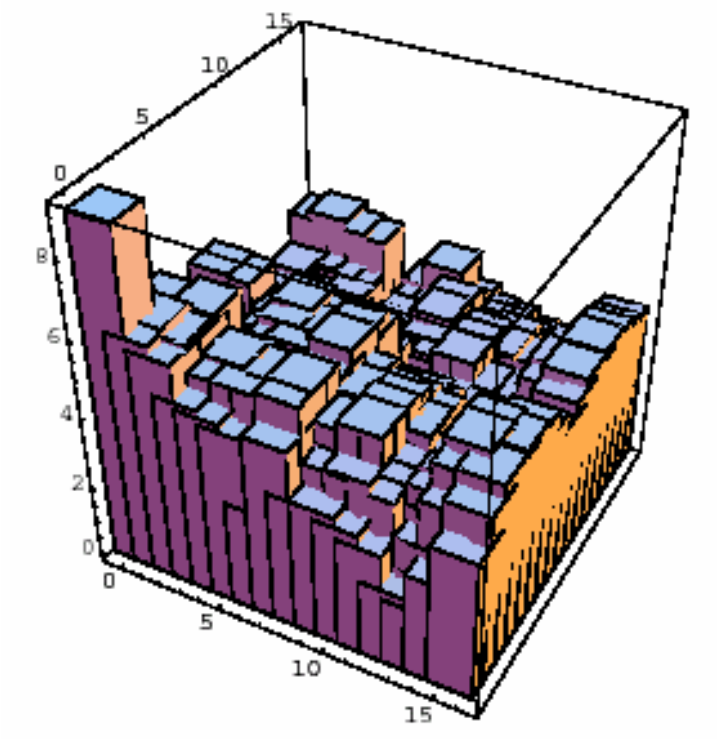


図 13　NOAA衛星画像にウェーブレット変換を適用し分割した場合の, ブロックごとのエントロピー. もともとの衛星画像は10ビットであることに注意.

5.5 地球環境データキャッシングの有効性

以上の結果から本研究では，地球環境データへのアクセスパターンの解析，およびアクセス確率の推定について述べた．本来であればこれにシミュレーション結果を付してキャッシングポリシーの優劣について論じるべきであるが，本研究では時間の都合からそこまで達することができなかった．この点については今後の課題としたい．ただし図 8のグラフを見る限り，地球環境データに関しては，ランキング上位のデータへのアクセスは比較的集中度が低いと考えられる．このことは指数 β の値が1よりも小さいことからわかる．逆に言えば，ランキング下位のデータへも一定のアクセスがあり，この部分が無視できない割合を占めているということになり，これはアクセス頻度のランキングに応じたキャッシングポリシーが有効でないことを示唆している．

一方で図 9や図 10のグラフは，比較的よい集中度を示している．これらは時刻間隔を表現するグラフであり，むしろこちらの情報を用いる方が有効なキャッシングポリシーを考案することができそうである．しかし図 10のグラフから導けるポリシーは，要するに最新のデータをキャッシングし，古くなるごとに削除していくという，強いて名づければLRO(Least Recently Obtained)とでも呼べそうな，あまり芸のないポリシーである．これを実現するためには，単純に時間とともに減衰するようなパラメータをキャッシュ決定関数に対応させるのが最も簡単であり，さらに5.4.4節でも述べたように，実数値のキャッシュ決定関数を解像度に対応させるとすれば，最終的には時間とともに解像度が減衰するようなキャッシングポリシーになる．これにLRA(Least Recently Accessed)ポリシーである図 9を組み合わせると，地球環境データに適したキャッシングポリシーを作り出すことができる．

以上のキャッシングポリシーは，逆の側面から眺めれば，プリフェッチの手法とも似ているところがある．例えば最新のデータへのアクセスがあらかじめ多いことがわかっているので，これを実際のアクセスに先立ってあらかじめダウンロードして保持しておく戦略がプリフェッチである．そして，単にデータの新鮮度でキャッシングするというポリシーならば，これはプリフェッチ的な手法でも実現可能である．これを発展させて，利用者の好みを先読みしてデータをプリフェッチすることが可能であれば，それもネットワークの利便性を高めることにつながるだろう．例えばある地域のあるデータを見ている利用者が，次に隣接地域のデータにアクセスするのか，それとも同一地域の別の種類のデータにアクセスするのか，それとも同一地域の別の日時のデータにアクセスするのか，といった行動パターンを予測するのである．しかしこのようなアクセスパターンに関してはまだ全くモデルはできておらず，プリフェッチ的な手法で解決できるのは，現在のところ最新データのプリフェッチ程度しかないと考えられる．

第6章 地球環境データの画像内容検索モデル

6.1 はじめに

画像データベースを有効に活用するためには強力な画像検索機能が不可欠であるが，中でも類似画像の検索機能はその重要な機能の一つである．そのためこれまでにも多くの研究が類似画像検索の問題を扱ってきた．しかし類似画像検索という問題はいろいろな要素が複雑に絡み合った問題であるため，画像データベースの構成も必然的に複数の処理手法を組み合わせる構成とならざるを得ない．ではこのとき，各処理手法はいかなる処理目的のもとで選択され，また相互にどのような関係を持っているのだろうか．このような問題意識が従来の研究では十分に意識されずに明確に整理されていなかったと考える．つまり従来の研究の問題点は，これらの点を明確にせずにシステムを構築するために，画像データベースとしてのパースペクティブに欠けていたことにあると考える．

そこで本研究では，類似画像検索システムを構築するためのフレームワークとなる「階層モデル」を提案する[1]．このモデルでは「画像内容」というものを意味的な観点から分割された複数の階層で表現する．階層的な構造を基礎とすることにより，類似画像検索システムの全体的な構造がより明確となり，また各レイヤに適切な処理手法を独立に「プラグイン」することにより，システムのフレームワークを崩すことなしに幅広い問題領域への応用が可能となる．本研究は地球環境衛星データのための類似画像検索システムを階層モデルに基づき構成し，各階層にプラグインするための手法についてもそれぞれ研究成果を得た．

6.2 従来の研究

類似画像検索の代表的な研究としては，電総研で構築された画像データベースである，商標・意匠データベースTRADEMARKや電子美術館ART MUSEUM[55]，またIBMにおいて研究されるQBIC[56]などがある．前者の研究では，画像特徴量として比較的処理が簡単でロバストな特徴量を用い，画像から抽出された画像特徴量ベクトルを多変量解析の手法を用いて適切な空間に写像することにより，人間の感性をも考慮した類似画像検索を可能としている．また後者の研究では，自動または半自動的な方法で抽出された画像特徴量を画像検索のためのインデックスとして用い，大規模画像データベースに対しても実用レベルの高速な画像検索を実現している．他にも，風景画像から画素単位対象物ラベル付け手法を用いたシステムや，プリミティブ分解に基づくシステムなど，数多くの研究がなされてきた．

これらのアプローチに対し，キーワードを検索インデックスとして用いるという伝統的なアプローチの研究も依然として盛んである．例えば放送用画像データベースにおいて，木構造化されたシソーラスに基づいて画像をキーワード付与ベクトルで表現し，このベクトル同士の内積によって画像間の類似度を計算するシステム，状態遷移モデルを用いて画像に自動的にキーワードを付与するシステムなどがその例である．しかしこれらの研究では，システムごとに検索対象となる画像の種類をかなり限定しており，そのため各システムの総合的な評価を相互に比較することは難しい．このとき，各システムの内部で組み合われ用いられている各処理手法が，類似画像検索システムという全体構造の中でどのような位置を占めているか，またシステムが全体としてどのような構成になっているか，といった点を把握するのが困難となるところが問題である．本研究が提案する階層モデルは，このような問題点に対する一つの解決策として提示するものである．

6.3 階層モデルという方法論

階層モデルという用語は二つの異なる意味に使われる．まずは階層という用語を「解像度の観点」から捉える用法がある．この場合は，ピラミッド構造や四分木などの階層的なデータ構造によって画像が表現され，このときの階層モデルの使用目的は種々の画像処理の高速化である．しかし本研究ではもう一つの意味& 全体としては複雑な問題を複数の階層ー部分問題--に分割して考えることにより問題構造の把握を容易とする方法論を指す用法として用いる．このように「意味的な観点」から画像内容を複数の階層に分割して表現することで，個々の階層で処理すべき内容も明確にしやすくなり，また個々の階層での処理手法の実装とシステムの全体的な構造とを分離して考えやすくなる．このような階層モデルという方法論は決して新しいものではなく，すでにさまざまな分野で使用されている．通信分野での中心的な概念であるISO/OSIの参照モデル[57]はその代表的な例であり，またコンピュータビジョンで用いられる代表的なモデル[58]や動画像の表現モデルにも同様の階層表現が用いられている．そこで本研究では，この階層モデルという方法論を類似画像検索システムに適用する．

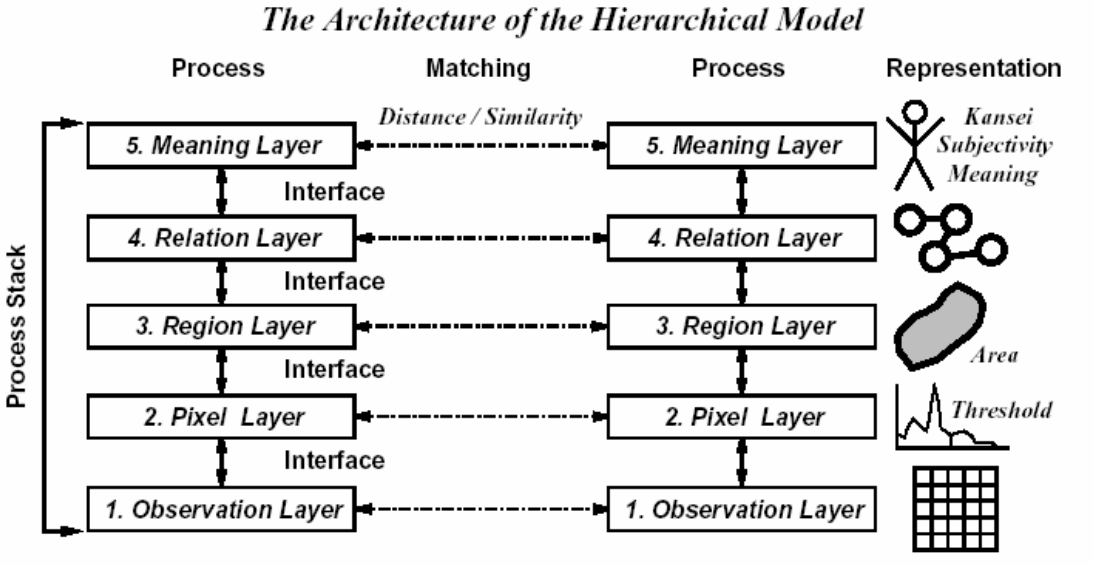


図 14 階層モデルの概念図．垂直方向に5個のレイヤ，水平方向に3個の要素から構成されている．

まず階層モデルの概念図を図 14に示す. 本研究で用いる階層モデルは, 垂直方向から眺めると, 観測レイヤ・画素レイヤ・領域レイヤ・関係レイヤ・意味レイヤの5層(レイヤ)から構成され, また水平方向から眺めるとプロセス・表現・マッチングの3個の要素から構成されるというモデルである. このモデル全体を「アーキテクチャ」と呼ぶ. 以下では, 垂直方向から眺めた5レイヤの役割, および水平方向から眺めた3要素の役割について説明する.

6.3.1　垂直方向から眺めた5レイヤの役割

観測レイヤ (Observation Layer)

センサを通して実世界を観測した結果を画素値配列という形に記録し, さらに画素値配列を前処理し以後の処理に使える形式の画像に変換する方法を定めるレイヤである. したがってこのレイヤでは, どのようなセンサを用いるかという問題だけではなく, センサを通して観測された画素値配列に対して画像の再生・復元・補正などの処理を行なう問題や, 雑音除去などの前処理を行なう問題などがこのレイヤに含まれる. 具体的には, 衛星観測画像の幾何補正処理や雑音除去のための平滑化処理などが含まれる.

画素レイヤ (Pixel Layer)

画像の最小単位である, 画素を単位とした処理方法を定めるレイヤである. このレイヤは, 注目画素と非注目画素とを分離する問題や, 各画素を分類クラスに分類する問題など, 以後のレイヤの基礎となる処理を行なうことを目的としている. 具体的には, 画像ヒストグラムなどの画素を単位とした統計量を計算する処理や, 統計量を用いた画素単位の画像分類, 画像空間上の局所的な演算を用いたエッジ検出などが含まれる.

領域レイヤ (Region Layer)

画素レイヤで抽出された画素単位の処理結果を用いて, 同一クラスに属し相互に隣接する画素の集合である領域という概念を考え, 領域を単位とした処理方法を定めるレイヤである. 具体的な処理では, 領域に対して形状特徴量やテクスチャ特徴量を計算する処理が一般的に用いられるが, さらに進んで形状分解(領域に基づいた方法)や多角形近似(輪郭線に基づいた手法)などの領域構造を抽出できる複雑な処理も用いられる.

関係レイヤ (Relation Layer)

領域レイヤで抽出された領域間の関係を計算する方法を定めるレイヤである. 本研究で用いる関係という用語は, 最も単純でほぼ自明な関係である隣接関係から, 隣接していない領域間に何らかの遠隔作用を仮定する関係までを含む幅広い用語として捉える. 例えば, 上下左右などの位置関係, 物理法則のアナロジーに基づいた関係, 主観的輪郭など知覚的な計算モデルに基づいた関係などを定義することができる. なお, このレイヤまでに抽出された情報を総合し, 画像表現モデルを構築する処理も含まれることがある.

意味レイヤ (Meaning Layer / Semantic Layer)

画像の意味や人間の感性・主観などを計算論に基づいて処理する方法を定めるレイヤである．つまりこのレイヤは，「人間」という要素を含めた柔軟な検索を実現するために重要な役割を果たすものとなる．具体的には，画像検索に対する利用者の観点を反映する問題や，画像に付与されたキーワードの意味的な類似性を計算する問題，さらには人間の画像に対する印象などを用いた感性情報処理などが問題となる．これらの処理手法は，人間の認知モデルや言語モデルなどから演繹的に導くことができれば理想的だが，現実には人間が与えた例示データなどから帰納的に学習する処理が必要となることが多い．

6.3.2　水平方向から眺めた3要素の役割

プロセス (Process)

階層モデルのアーキテクチャを決定するためには，まず各レイヤに対して，下位レイヤから受けとった情報を処理して上位レイヤへと渡すために必要な処理手法を定める必要がある．このように各レイヤに実装する処理手法のことを本研究では「プロセス」と呼ぶ．各レイヤに実装するプロセスの垂直的な集合（スタック）は問題領域に応じて適切に定める必要があり，そのよしあしが類似画像検索システム全体の総合的な性能を決定する．

表現 (Representation)

各レイヤのプロセスによって処理された後，データベースに蓄積され検索インデックスに用いられる情報の表現形態を指す用語である．表現は各レイヤに実装されるプロセスに強く制約されるが一意には定まらない．また各レイヤの間に設けられた「インタフェース」は，下のレイヤの表現を出力として受け取り，上のレイヤの入力に渡す役割を果たす．表現によって記述された情報のみが検索インデックスとして使用されることを考えると，検索性能を向上させるためには画像検索の目的に応じて適切な表現を採用することが重要である．

マッチング (Matching)

類似画像検索とは，利用者によって与えられた検索キーとデータベースに蓄積されたインデックスとの間で距離(類似度)を次々と計算し，距離が小さい(類似度が大きい)順に蓄積画像を出力する処理である．この中で距離を計算する処理を指す用語がマッチングである．マッチングも各レイヤに実装されるプロセスに強く制約されるが一意には定まらない．さてマッチングでは，検索キーと蓄積インデックスの間で「同位レイヤ」の記述を比較して距離(類似度)が算出される．このとき，一致に基づくマッチング(exact matching)ではマッチングのアルゴリズムは比較的単純となるものの，非一致に基づくマッチング(inexact matching)ではマッチングのアルゴリズムによって検索結果が大きく変動しうる．そのためアルゴリズムの選択は検索性能の向上に重要な意味をもつ．また，グラフ構造などの構造をもつ表現のマッチングはアルゴリズムが複雑になり計算コストが増大しがちであるため，アルゴリズムの選択は検索時間の高速化にも重要な問題となる．

6.3.3 スタックの設計

表 4 階層モデルにおける各レイヤで用いられるプロセスおよびその目的

レイヤ	スタックで用いられるプロセス	プロセスの目的
意味レイヤ	多変量解析・ニューラルネットワーク・遺伝的アルゴリズムなどの教師つき学習アルゴリズム	画像の意味や人間の感性・主観に基づいた処理
関係レイヤ	物理法則のアナロジーに基づいたモデル，ルールに基づいたモデル，ゲシュタルト法則など人間の視覚特性に基づいたモデル	領域間の関係の構築および画像表現モデルへの構造化
領域レイヤ	形状特徴量・テクスチャ特徴量などの計算，形状分解・骨格線抽出・多角形近似などの構造抽出	領域の形状的特徴や構造的特徴の抽出
画素レイヤ	物理世界や画素値空間の統計的性質を用いたヒストグラム解析や統計的分類手法，人間の視覚特性や近傍画素の画素値分布などを用いたエッジ抽出など	画素単位での注目画素の分離や分類クラスへの画素分類
観測レイヤ	物理世界の観測に最適なセンサの選択，センサの補正，得られた画像の幾何学的な補正など	センサを通した物理世界の観測および種々の画像補正

画像検索システムを構築するためには，まず検索対象画像や検索目的を考慮して適切なスタックを決定する作業が必要となる．そこでまず，各レイヤで用いられるプロセスの一例を表 4に示す．このようなプロセス候補の中から，構築するシステムに適切なプロセスをレイヤごとに選択し実装するという流れで類似画像検索システムが完成する．ここで，通信分野でも通信ネットワークの用途に応じて多くの規格が提唱され，複数のプロトコル・スタックが使用され続けていることを考えると，類似画像検索の場合にも検索対象画像や検索目的に応じて異なるスタックが用いられるのはむしろ当然のことと考えてよい．

また階層モデルでは，すべてのレイヤが完全に独立し任意のプロセスを各レイヤに組み込める設計が理想的である．ただし現実には，各レイヤのプロセスがこのように自由にプラグインできるという設計は難しい．なぜなら，各プロセスはインタフェースに依存しており，なおかつプロセスから独立した形で各レイヤ間のインタフェースを定めるのが困難であるからである．そこで，レイヤ間の相互作用を弱めることと，普遍的でありかつ有用な情報を失わないというインタフェースの定義とが，スタックの設計に必要な指針となる．こうして有用なプロセスを複数のスタックで共用できれば，システムの構築はより簡単になろう．

6.4 本研究で用いるアーキテクチャ

本研究では気象衛星画像に対する画像検索に関心がある．特に気象衛星画像としてはひまわり衛星画像を用いており，本研究ではひまわり衛星画像に適した構造の画像データベースを構築する．実際に用いたアーキテクチャの一部を示す．以下では各レイヤのプロセスについて概説する．

観測レイヤ

衛星画像の場合には，このレイヤで幾何補正処理および地図化処理が必要である．これらの処理をおこなうには画素の性質は必要ではなく，衛星画像の各画素および地図化画像の各画素の緯度経度，およびそれらの対応のみが必要である．本来は地上基準点(Ground Control Point)を用いたさらに高精度の処理が必要であるが，本研究ではこれを省略した．それでも，気象衛星が送信してくるパラメータを用いることで，ほとんどの場合においてほぼ正確な幾何補正処理をおこなうことが可能である．

画素レイヤ

画素値という確率変数が従う統計的性質に基づくプロセスを提案した．この手法は比率成分解析と名づけており，特に衛星画像において特徴的な混合画素（ミクセル）の処理に威力を発揮する手法である．これによって衛星画像の各画素を雲（複数の種類）や海，陸などに分類し，その後の処理に用いる．

領域レイヤ

初期の研究では，このレイヤに変形モデルに基づいた形状分解プロセスを用いていた．その後，今後分布推定に基づく正規混合分布モデルの当てはめによる楕円領域抽出を研究した．このように画像中に含まれる領域を簡単な領域の集合として表現することが目的であるが，現実的にはこのようなプロセスを雲のような不定・柔軟形状に対してロバストに計算することは難しい．そこで本研究の後半ではあえて画像の領域分割をおこなわず，画像全体を1領域とみなして画像全体の特徴ベクトルを計算する方法を用いた．

関係レイヤ

本研究の前半では，物理法則のアナロジーに基づいた引力モデルを領域間に定義し，これによって画像中の領域の関係を階層化属性つき関係グラフ(Hierarchical Attributed Relational Graph)で表現するという方法を用いた．本研究の後半では，画像全体を1つの領域とみなしているため，階層化属性つき関係グラフとしてはちょうどノードが一つだけのグラフと等しくなり，グラフを導入する意味がなくなった．そこで研究の後半ではこのレイヤは空となっている．

意味レイヤ

本研究では適合度フィードバック(relevance feedback)に対応する方法として，利用者の反応を学習しながら画像検索を進めていくための対話型進化的機構を提案した．これは利用者の判定を遺伝的アルゴリズムの適応度に反映させることで，利用者の好みをより強く反映するような個体を繁栄させる（あるいは利用者の好みを反映しない個体を淘汰する）という形で学習を進めていく方法であり，これを画像検索の問題に応用したものである．

以下ではこれらのレイヤに関連して研究した成果の概要を順番に紹介する．なお，以下に述べる研究成果がすべての研究を尽くしているわけではなく，例えば領域レイヤなどなどに関する研究成果は一部省略しているものがある[15][21]．これは研究プロジェクト終了時の相対的重要度に基づく取捨選択である．

1．6.5章「比率成分解析に基づく画像分類法」では，「画素レイヤ」にプラグインするための統計的画像分類法について概要を説明する．この手法は仮想分布という確率モデルを導入するところは独創的なアイデアであり，特に衛星画像中に出現する雲の分類を根本的に考え直すところから発生した技術である．しかしその背景から説明するのはそれだけで多くのページを必要とするため，本報告では主要部分のみを箇条書きに近い形で述べる．

2．6.6章「楕円形状分解による台風雲パターンの表現」では，「領域レイヤ」にプラグインするための画像処理手法について概要を説明する．この手法は，目的関数の最小化により物体形状抽出をしようというエネルギ最小化の概念と，複雑な物体形状を単純な部品の組み合わせとして理解しようという形状分解の概念とが，その背景にある．具体的には台風雲パターンを楕円の集合体として表現するための方法を追究し，その結果を台風の気象学的解析に適用した結果，それらが従来知られていた知見を再現することを確認した．

3． 6.7章「対話型進化的計算論に基づく画像検索」では，「意味レイヤ」にプラグインするための対話的な検索方法について概要を説明する．対話的な検索過程の方法論である適合度フィードバック(relevance feedback)の対話過程を特に進化的計算論を用いてモデル化することにより，適応的な画像検索法を実現することを目指す．また画像検索を目的とする場合に重要となる改良点などについてもまとめる．

6.5 比率成分解析に基づく画像分類法

6.5.1 はじめに

本研究は比率成分解析(FCA)，すなわち混合した状態で観測された信号から成分信号の混合比率を復元する，という逆問題を解くことを目的とする解析手法を提案する[27]．この種の問題に対しては多くの解析手法が過去に提案されているが，本研究で提案するモデルの最大の特徴は，成分信号の混合比率に対する3種の制約および仮定--正値制約，総和制約，確率性--に基づき，成分信号の混合過程に起因する確率分布，すなわち「仮想分布(virtual PDF)」という新しい概念を提案する点にある．また正規分布とディリクレ分布を各信号分布のモデルとする場合には，この仮想分布のモーメントやキュムラントの厳密解を解析的に求めることができ，これを用いて確率分布を近似表現できることも示す．さらに，実際の観測信号から成分信号を推定するという逆問題についても，仮想分布を含む混合分布モデルに対するEM (Expectation-Maximization)アルゴリズム風の解法を提案し，このモデルが既存の正規混合分布モデルなどにわずかな計算コストで追加できることを述べる．

本研究において「信号の混ざり方」のモデルとなるのは，数学的に扱いやすい線型混合過程とする．

$$x(t) = \sum_{m=1}^{M} a_m(t) s_m(t) + n(t) = S(t)a(t) + n(t) \tag{7}$$

ここでtは観測信号の座標を表し，観測信号$x(t) \in R^N$は二つの種類の隠れ変数の線形和，すなわち純粋信号からの成分信号$s_m(t) \in R^N$およびその重み付けとなる混合比率$a(t) \in R^M$の線形和として観測されると仮定する．ここで観測信号はN個の観測チャネルからの信号を表すものであり，また混合比率はM個の成分信号$C_m \in C = \{C_1, ..., C_M$について以下の制約を満たすものとする．

$$a_m(t) \in F = \{a_i(t) \mid a_i(t) > 0 \quad and \quad \sum_{i=1}^{M} a_i(t) = 1, \quad for \quad \forall t \} \tag{8}$$

また$n(t) \in R^N$は加法的ノイズを表現する確率変数である. 以上に述べた一般的な線形混合過程の中でも, 我々は特に以下のような問題設定に興味がある.

1. 正値制約 $a_m(t) > 0 \quad for \quad \forall m, \forall t$

2. 総和制約 $\sum_{m=1}^{M} a_m(t) = 1 \quad for \quad \forall t$

3. 確率性 $a_m(t)$と$s_m(t)$はどちらも確率変数A_mとS_mの実現値である.

このような問題設定のもとで, この線形混合過程が作り出す確率分布を考慮したうえで, 混合分布推定および混合比率推定をおこなうことを目的とするデータ解析手法のことを, 本研究では比率成分解析(FCA)と命名し, このデータ解析手法の性質を確率論および学習理論の枠組で記述する.

6.5.2 問題の背景

そもそもこのような問題意識を抱えるに至った背景には, 画像解析における画素分解(pixel unmixing)という問題がある[4][5][6][10][13][14][18]. この問題では, 1画素に対応する観測信号が実際には複数の成分信号の混合であると考え, その比率を観測信号から推定することを目標とする. ここで実世界の観測対象が連続確率場であると考えれば, 画像すなわち画素配列という表現は, 連続確率場を離散的な(重なりのある)区画に分割し, その区画ごとに信号を平滑化した表現であるとみなすことができる. 個々の区画の大きさは観測手段に応じて異なるが, 例えばリモートセンシング衛星画像などの場合には, 1区画の大きさが実世界の数キロメートル四方に対応する場合があり, したがって複数種類の成分信号が区画内において混合した状態で観測されることも珍しくない. この現象は本質的にはスケールの問題[59], つまり連続確率場に特徴的なスケールに比べて, 観測手段によって規定されるスケールがはるかに大きい場合に顕著な影響が生じる. また半透明物体を通過する信号にも同様の混合現象が生じる.

画像表現に不可避であるこのような問題をきちんとモデル化することで, 観測信号から成分信号の混合比率を復元するための方法を得たい, というのがこの研究の動機となっている. 例えば気象衛星画像において, 雲と海が1画素内に混合して観測される場合に, 雲と海の混合比率を気象衛星画像の画素値から復元したいという要求がある. ただしこの問題には以下のような困難が潜んでいる.

1. 空間的には画像信号の相関がある程度は観察されるが, たとえ隣接した画素どうしであっても, 混合比率が同一となることは期待できない.

2. 気象衛星画像においては, 晴天領域は混じり気の少ない均質な領域とみなせるが, 一方で画素内に雲と晴天領域が混ざって観測される不均質な領域もある. したがって信号全体でみれば, 単一の信号源から発生する純粋な信号と, 複数の信号源から発生する成分信号が混合した信号とが共存することになる. これはすべての信号が混合信号であるとのパラダイムでは扱うことができない.

類似した問題を解くための枠組としては, 後述するような多くの手法がこれまでに提案されてきた.

6.5.3 関連手法

独立成分解析(independent component analysis : ICA)

独立成分解析[60][61]が対象とするのは未知情報源分離(blind source separation)の問題であり，特に未知の情報源の独立性を利用して信号を復元する点に特徴がある．本研究で提案するFCAの目的はICAに近いが，基本的な問題設定にはそもそも大きな違いがある．

1．**混合比率** ICAは混合比率に何の制約も課さないが，FCAは正値制約および総和制約を課す．

2．**混合比率の確率性** ICAはすべての座標で混合比率が同一であるのに対し，FCAは座標ごとに変動する確率変数である．ただし最近はICAでも，確率的な問題設定を扱う研究が増えつつある．

非負行列分解(nonnegative matrix factorization)

特に画素値の非負性に着目し，観測行列を非負の成分に分解することで，画像の疎(sparse)な表現を得ることを目的とする方法である[62]．しかし画素値に対応する物理量を考えると，この零点とは例えば物理学における絶対零度のような本質的な零点というよりは，観測手段および確率変数の実現値によって規定される，物理的意味の希薄な値に対応することも多い．それに対してFCAは，混合比率という物理的に意味のある量を扱う点が異なる．

混合分布モデル(mixture density model)

このような複数の成分信号の混合を扱うための代表的モデルの一つに，混合分布モデル[63]がある．このモデルでは，個々の観測値は成分情報源のどれか一つから生起する信号であると仮定する．それに対し本研究のモデルは，個々の観測値自体が複数の成分信号の混合であると考える．つまり本研究のモデルは，確率変数の実現値のレベルで混合が生じる，換言すれば確率変数の観測過程において混合が生じる状況をモデル化するものである．ゆえに通常の意味での混合分布モデルはFCAが扱う問題を適切に表現できない．

6.5.4 仮想分布の導出

まず成分信号の混合が瞬時混合として起こること，すなわち他の座標の信号が畳み込まれることはないと仮定する．次に瞬時混合において同じ成分の組み合わせが関わる観測信号の実現値だけを集めることが可能であると仮定しよう．すると，これらの観測信号は，成分信号のある特定の組み合わせに依存する分布にしたがうと仮定できる．本研究ではこの分布を仮想分布(virtual PDF)と命名する．その理由は，実世界にはこの分布に対応するものが存在せず，センサで観測した信号のみに一種の「幻影」のように出現するからである．

このような仮想分布の性質を調べるため，まず観測信号中に出現する成分の集合を$C=\{C_1,...,C_M\}$とする．そしてi番目の成分からの成分信号をC，そして成分信号が従うPDFを$S_i \sim p_i(s_i \mid \Phi_i)$と定義する．ここで$\Phi_i$はパラメータベクトルである．次に2個の成分信号の混合に関してあらゆる組み合わ

せを列挙するため，集合 C の部分集合で大きさが2であるものを

$$\Omega_2 = \{\omega_{(i,j)}^{(2)} = \omega_k^{(2)} \mid C_i \in C, C_j \in C, i \neq j \tag{9}$$

と表す．ただし $k = 1,...,\begin{pmatrix} M \\ 2 \end{pmatrix}$ である．同様の方法で部分集合 $\Omega_i, (i = 2,,M)$ を決めることができる．ただし集合の要素数は $\begin{pmatrix} M \\ i \end{pmatrix}$ である．

次に混合信号 $\omega_k^{(M)}$ の確率分布を表現する．一般性を失うことなく，この集合に含まれる成分の添え字を $\omega_k^{(M)} = \{C_1,...,C_M$ と付け直す．するとこの成分の混合信号に対応する確率分布 $p(x \mid \omega_k^{(M)})$ は，混合比率 a の周辺分布として以下のように計算できる．

$$p(x \mid \omega_k^{(M)}) = \int p(x, a \mid \omega_k^{(M)}) da = \int p(x \mid a, \omega_k^{(M)}) p(a) da \tag{10}$$

ここで $p(a)$ は混合比率の事前分布である．以下では記法を簡略化するために $\omega_k^{(M)}$ という記述は省略する．ここで次に上の式の $p(x \mid a)$ という項に着目する．ここに(7)式の関係を代入すると，

$$p(x \mid a) = p(a_1 s_1 + + a_M s_M \mid a) = p(s_1,...,s_M \mid a) \tag{11}$$

が得られる．ここで雑音項 n は簡単のために省略しているが，この項を含めた形で議論することも可能である．

さらにここで，各成分信号が独立であるという仮定を導入すると，(11)式は以下のように書き換えることができる．

$$p(x \mid a) = \frac{1}{a_1 \cdots a_M} p_1\left(\frac{s_1}{a_1}\right) * \cdots * p_M\left(\frac{s_M}{a_M}\right) \tag{12}$$

ここで $*$ は畳み込み演算を指す．定義より，混合に関わっている成分信号のみが部分集合に含まれているため，すべての混合比率は $a_i > 0$ である．

ここで畳み込みを扱うのに便利な関数として，特性関数を導入する．これは虚数単位 j および転置 T を用いて $\varphi(w) = \int e^{jw^T x} p(x) dx$ と表される関数であり，これを用いて確率分布 $p(x \mid a)$ の特性関数 $\varphi_{X|A}(w)$ は以下のように書き換えることができる．

$$\varphi_{X|A}(w) = \frac{1}{a_1 \cdots a_M} \prod_{m=1}^{M} a_m \varphi_m(a_m w) = \prod_{m=1}^{M} \varphi_m(a_m w). \tag{13}$$

最後に求めたかった $p(x \mid a)$ は，(13)に対する逆変換の式を用いて

$$p(x \mid a) = \frac{1}{(2\pi)^N} \int_w e^{-jw^T x} \varphi_{X|A}(w) dw \tag{14}$$

となり，ここから最終的に $p(x)$ を(10)式を用いて求めることができる

6.5.5 混合分布モデル

本研究で用いる混合分布モデルの特徴は，以下のような関係式で表すことができる．

$$\text{Virtual PDFs} + \text{Real PDFs} = \text{Mixture Density} \tag{15}$$

ここでReal PDFは各成分信号の分布を表現するものであり，いわゆる伝統的な混合分布モデルはすべての分布がReal PDFに属するモデルである．それに対し本研究のモデルは仮想分布を含んだ形で混合分布モデルを定式化する．すなわち(15)式を正式に書くと，

$$p(x \mid \Phi, \Theta) = \sum_{l=1}^{M} \pi_l p_l(x \mid \Phi_l) + \sum_{\omega_k^{(l)} \in \Omega_l}^{|\Omega_l|} \pi_{\omega_k^{(l)}} p_{\omega_k^{(l)}}(x \mid \Phi, \Theta) \tag{16}$$

$$\sum_{l=1}^{M} \pi_l + \sum_{l=2}^{M} \sum_{\omega_k^{(l)} \in \Omega_l}^{|\Omega_l|} \pi_{\omega_k^{(l)}} = 1 \tag{17}$$

となる．ここで Φ は成分信号分布のパラメータ，Θ は混合比率分布のパラメータであり，$\pi_{\omega_k^{(l)}}$ は仮想分布の混合確率を表現するものである．このような混合分布モデルをデータから学習し隠れ変数などを推定することが，我々が提案する手法「比率成分解析」(Fractional Component Analysis : FCA)の計算の目的である．その解法にはEMアルゴリズムを改良した手法を用いる[63][64]．

さて仮想分布は理論的には(14)式から導くことができるが，ごく特殊な場合を除けばこの仮想分布は閉じた形で求めることはできず，実際に分布の形状を計算するには膨大な数値積分計算が必要となる．ゆえに本研究ではこの分布の近似により，仮想分布の近似計算を高速化する方法を提案する．このような近似方法として，本研究では分布のモーメントおよびとキュムラントに基づく近似方法である，グラム・シャリエ展開を用いる[60][61]．具体的に仮想分布のモーメントは，(10)の両辺をモーメント展開し，べき乗が同じ項を両辺で比較することで，以下のような関係式を得ることができる．

$$m_X^{(l_1,\ldots,l_N)} = \int_F p(a \mid \Theta) m_{X|A}^{(l_1,\ldots,l_N)} da \tag{18}$$

つまり仮想分布のモーメント $m_X^{(l_1,\ldots,l_N)}$ は，混合比率の事前分布 $p(a \mid \Theta)$，および混合比率の条件付き信号分布 $p(x \mid a)$ のモーメント $m_{X|A}^{(l_1,\ldots,l_N)}$ から計算できることがわかる．ここで $m_{X|A}^{(l_1,\ldots,l_N)}$ の方が $m_X^{(l_1,\ldots,l_N)}$ よりも計算が容易であり，かつ(18)が閉じた形で計算できるようなモデルがあれば理想的である．なぜなら上記の条件が満たされる場合，仮想分布のモーメントを閉じた形で求めることが可能となるからである．そのような例の一つとして，以下のような強力なモデルが実際に存在する．

$$p_m(x \mid \Phi_m) = \frac{1}{(2\pi)^{N/2} |\Sigma_m|^{1/2}} e^{-\frac{1}{2}(x-\mu_m)^T \Sigma_m^{-1} (x-\mu_m)} \tag{19}$$

$$p(a \mid \Theta) = \frac{\Gamma\left(\sum_{m=1}^{M} \theta_m\right)}{\prod_{m=1}^{M} \Gamma(\theta_m)} \prod_{m=1}^{M} a_m^{\theta_m - 1} \tag{20}$$

ここで(19)は各成分信号の確率分布を表す正規分布であり，(20)は混合比率の事前分布を表すディリクレ分布である．これらの分布族を導入することで，実際に(18)を閉じた形で求めることができる．しかもこれらのモデルは決してアドホックなモデルではない．正規分布はこの種のモデルとしては最も基本的なモデルであるし，またディリクレ分布もこれが妥当なモデルであるとの部分的な証拠をシミュレーションにより既に得ている[10]．

$$m_X^{(1)} = \frac{\sum_{m=1}^{M} \theta_m \mu_m}{\theta}, \qquad m_X^{(2)} = \frac{\sum_{m=1}^{M} (\theta_m)_1 \left(\mu_m^2 + \sigma_m^2\right) + \sum_{\substack{m,n=1 \\ m \neq n}}^{M} \theta_m \theta_n \mu_m \mu_n}{(\theta)_1}$$

$$m_X^{(3)} = \frac{\sum_{m=1}^{M} (\theta_m)_2 \left(\mu_m^3 + 3\mu_m \sigma_m^2\right) + \sum_{\substack{m,n=1 \\ m \neq n}}^{M} \theta_m (\theta_n)_1 \left(\mu_m \mu_n^2 + 3\mu_m \sigma_n^2\right) + \sum_{\substack{m,n,o=1 \\ m \neq n \neq o}}^{M} \theta_m \theta_n \theta_o \mu_m \mu_n \mu_o}{(\theta)_2}$$

$$m_X^{(4)} = \frac{\sum_{m=1}^{M} (\theta_m)_3 \left(\mu_m^4 + 6\mu_m^2 \sigma_m^2 + 3\sigma_m^4\right) + \sum_{\substack{m,n=1 \\ m \neq n}}^{M} (\theta_m)_1 (\theta_n)_1 \left(\mu_m^2 \mu_n^2 + 6\mu_m^2 \sigma_n^2 + 3\sigma_m^2 \sigma_n^2\right)}{(\theta)_3} +$$

$$\frac{\sum_{\substack{m,n=1 \\ m \neq n}}^{M} \theta_m (\theta_n)_2 \left(\mu_m \mu_n^3 + 6\mu_m \mu_n \sigma_n^2\right) + \sum_{\substack{m,n,o=1 \\ m \neq n \neq o}}^{M} \theta_m \theta_n (\theta_o)_1 \left(\mu_m \mu_n \mu_o^2 + 6\mu_m \mu_n \sigma_o^2\right) + \sum_{\substack{m,n,o,p=1 \\ m \neq n \neq o \neq p}}^{M} \theta_m \theta_n \theta_o \theta_p \mu_m \mu_n \mu_o \mu_p}{(\theta)_3}$$

図 15　1次元仮想分布の場合のオーダーk のモーメント$m_X^{(k)}$．$m \neq n \neq o \neq p$ はどれも異なる値をとるという意味であり，また$(\theta)_n = \theta(\theta+1)\cdots(\theta+n)$ である．

このような分布を導入した場合のモーメントの具体的な式を図 15に示す[27]．ここでは1次元信号の場合の4次までのモーメントしか示していないが，少なくともこの範囲では，成分信号分布のパラメータおよび混合比率分布のパラメータを用いて，仮想分布のモーメントが表現できていることがわかる．これらのモーメントからキュムラントを導出する公式は既によく知られており，例えば低次のキュムラントは$c_X^{(1)} = m_X^{(1)}$ や$c_X^{(2)} = m_X^{(2)} - m_X^{(1)} m_X^{(1)^T}$ のように簡単に導出できる．

最後にこれらのキュムラントを用いて，仮想分布をグラム・シャリエ展開によって近似する．この展開は正規分布 $N(c_X^{(1)}, c_X^{(2)})$ からの隔たりの大きさをキュムラントの積で表現するため，キュムラントが閉じた形で得られていれば近似確率分布も容易に計算できる．またこの時の近似精度は，キュムラントを高次まで計算すれば向上させることができるが，計算誤差や計算量などの現実的な面を考えれば，1次元信号分布の場合で4次キュムラント程度，2次元以上の信号分布では2次キュムラント程度（共分散行列）までを用いた近似が，妥当な選択であろう．

6.5.6　今後の展開

本報告書では仮想分布の導出の部分について説明をおこなったが，実際にはこのような各種の分布パラメータをデータから学習するための機構であるEMアルゴリズムに関する議論や，混合比率の推定問題など，さらに進んだ議論が必要である．

これらの研究成果は，最終的には気象衛星の雲分類処理，すなわち気象衛星「ひまわり」の各画素を

雲や海，陸などに分類するために用いるものである．気象衛星画像には雲と海が混ざったような画素が多数出現するが，比率成分解析を用いるとこれらの画素のモデルを確率論的にきちんと定義することができるため，このような画素から「逃げずに」雲分類処理をおこなうことができる．雲分類処理アルゴリズムへの比率成分解析の適用についてはまだ中途半端なレベルにとどまっているが，このようなアルゴリズムをまずは早急に完成させたい．なぜなら，信頼性の高い雲分類アルゴリズムは第7章で述べる台風データマイニングにおいても重要な要素技術となるからである．

6.6 楕円形状分解による台風雲パターンの表現

6.6.1 形状分解

本研究の具体的な対象である台風画像データベースにおいては，台風雲パターンの表現法が重要な研究課題である．台風雲パターンの形態学的特徴については7.3.1節でも後述するが，楕円形とらせん形が基本構成要素になると考えてよい．まず楕円形は台風中心部の雲の表現に適しており，その物理的根拠はコリオリ力＋遠心力と気圧傾度力とのつりあいで生じる傾度風にあると考える．一方らせん形は台風中心から伸びるレインバンドの表現に適した形状であり，その物理的根拠は摩擦力の影響で等圧線と一定の角度で吹き込む摩擦収束にあると考える．

本研究では，特に前者の円形領域の表現という問題を考え，変形楕円を用いた形状分解法を提案する[19][21][23]．ここで形状分解(shape decomposition)とは，物体の形状を複数のより基本的な部分に分解し，物体を「部分」の集合として表現する方法である．特に本研究では，楕円を基本構成要素として，台風雲パターンを楕円の集合として表現する方法を提案する．

その方法の概略を以下で説明する．まず7.4.3節で述べるように，本研究では画像分類アルゴリズムを適用することで各画素に対して雲種のラベルを付与する．その後形状分解の目的に応じて，関心のある分類クラスに正のスコア，関心のない分類クラスに負のスコアなどの値を割り当てる．次に楕円内部に含まれる画素のスコアの和(エネルギ)を計算し，楕円をパラメータで変形させながらエネルギを最大化するようなパラメータを求める．正負のスコアを定義しているために楕円要素のエネルギには極大値が存在する．こうして決定された楕円要素を取り除き再び同様の最適化を繰り返す．これをまとめると下記のようなアルゴリズムとなる．

1．分類対象画像の画素すべてに，分類クラスに対応したスコアを割り当てる．

2．スコアが正の画素の分布状況から楕円要素の初期パラメータを設定する．

3．楕円内部に含まれる画素のスコアの和をエネルギとし，このエネルギが最大となるように楕円パラメータを最適化する．

4．最適化された楕円要素のエネルギがしきい値よりも大きければ，形状分解要素として採用する．

5．円要素を複数回発見した後に終了する．

6．形状分解要素の内部に含まれる画素のスコアを負の重複ペナルティに置換し，2に戻る．

6.6.2 楕円要素のパラメータ最適化

ここで最適化すべきエネルギ関数に着目する. これを本研究では, 楕円要素の内部に含まれる画素のスコアの総和, と定義する. そしてエネルギ関数が最大となるパラメータを最適解とし, そのパラメータに対応する楕円を形状分解要素とする. このような枠組はエネルギ最小化に基づく変形モデル[66][67]に触発されたものであり, エネルギ関数の定義が形状分解結果を間接的に制御する点が特徴である.

ここで楕円要素のパラメータ表現について述べる. 先述のようにラグランジュ的表現を用いる場合, 画像中心を原点とするのが自然な座標系である. また台風中心を座標原点とした極座標 (r,θ) を楕円のパラメータベクトルを w として, 以下の表現を用いる.

$$f(w) = f(r,\theta,\phi,p,a)$$

ここで $a = q/p$ は縦横比である. こうして楕円形状分解問題は, 5次元のパラメータベクトル w の最適化問題に帰着し, その最適化にはPowell法を用いる. その理由は, 本研究のように関数の微分を正確に求めるのが困難な場合は, 準ニュートン法などのように微分を用いる方法よりは, 微分を必要としないPowell法などの方が, 安定して高速に解を求めることができるためである. ただしPowell法では探索方法の特徴から, 探索方向の初期設定におけるパラメータの並び順によって, たどりつく局所最適解が異なる場合がある. そこで簡単なテストの結果, パラメータの並び順を上記の式のように決定した. また各パラメータの変動範囲がほぼ同じになるようにスケーリング処理も施している.

なお, 具体的なスコアの設定や, 細かいヒューリスティクスについての説明などについては, 本報告では省略する.

6.6.3 時系列解析

台風の時系列解析としてよく取り上げられるのは, 台風の雲の動きを利用した風ベクトルの推定という問題である. これはブロック単位での雲の動きを, 前後の画像の相関などを用いて推定するものであり, コンピュータビジョンにおけるオプティカルフローの問題とも共通する部分が多い. しかしこのような方法では通常の衛星観測の間隔よりも短い観測間隔でないと, 風ベクトルを安定して推定するのが難しい. 一方ここでは, 時系列解析を形状分解要素のパラメータの予測という文脈で議論する. ある時刻 t での楕円パラメータを w^t とし, 時刻 T での w^T については過去のパラメータ $w^t,(t=1,\ldots,T-1)$ から予測可能と仮定する. そのための方法として最も単純な方式である持続予測 $w^T = w^{T-1}$, および1次自己回帰過程(AR)や移動平均過程(MA)などのモデルに基づくカルマンフィルタなどがある[68][69]. 以下では単純な持続予測を用いた結果を示す.

6.6.4 実験結果

楕円形状分解

本研究の提案手法を用いた形状分解結果を図 16に示す. 灰色で示された楕円が形状分解要素として採用された楕円である. 積乱雲は白色で表されており, 白い画素領域に正のスコアが割り当てられて

いる. 台風の中心雲領域が比較的まとまっている場合, 具体的には図 16の5行目や6行目の画像に対しては, 中心雲領域をうまく抽出するような形状分解結果が得られた. さらに5行目の例では, 中心雲領域とバンド状雲領域に異なる楕円が当てはめられており, 両者を異なる部分として表現することが可能となっている. それに対し, 2行目や3行目のように, 台風の中心雲領域から南の方向へ太いバンド状雲が伸びている場合は, それらをまとめて大きな楕円で表現しており, 中心雲領域とバンド状雲領域という異なる性質の雲領域が一体の雲システムとみなされてしまっている. この問題を解決するために, 新たなペナルティを導入すべきか, あるいは新たな形状分解要素を導入すべきか, などの点については今後の検討課題である.

台風雲パターンのサイズの日変化

次にこのような形状分解結果を用いて, 台風の形状変化を統計的に解析する試みについて述べる. 例えば台風の雲パターンのサイズの日変化などは, 楕円の大きさを用いて自動的に計測できる.

熱帯海洋上の対流活動には明らかな日変化があり, 早朝に対流活動のピーク, 午後から夕刻に対流活動のボトムがあることが知られている[70]. この対流活動の変化にしたがって, 台風のサイズにも日変化が観察される. 本研究で関心がある「積乱雲」に相当する雲領域は, およそ1500LSTから1800LST(地方時)に最大になる, と気象学の文献には記されている. そこで本研究では, 同様の日変化現象が本研究で提案する手法でも確認できるかどうかを検証する[23]. 具体的には, すべての形状分解要素系列の中から, 以下の条件を満たす時系列を探索する.

1. 発生から消滅までに24時間以上追跡可能である(ある程度安定した要素のみを対象とする).

2. 個々の形状分解要素の面積が S_{th} 以上である(ある程度大きな要素のみを対象とする).

3. 画像中心と楕円中心の距離が D_{th} 以下である(中心付近の雲領域のみを対象とする).

以上の形状分解系列に対して, さらに以下の計算をおこなう.

1. 抽出された時系列に対して平均面積を求め, 各時刻の面積を平均面積で正規化する.

2. 各時点の観測時刻を地方時に変換したのち, グラフにプロットする.

1995年から1998年の台風91個の時系列画像を対象とした実験では, 上記の条件を満たす楕円要素時系列は60個検出できた. これらの楕円要素時系列を対象として計測した, 台風サイズの日変化を示すグラフが図 17である. この図には, 0000LSTがボトムで1700LSTがピーク, その間は直線的にサイズが増加 / 減少するという, 明瞭な日変化パターンが現われている. この変化パターンは, 追跡初日の0000LSTから2400LSTのみならず, 2日目や3日目にも繰り返されているため, かなり安定したパターンであると考えられる.

図 16　台風199624号に対する楕円形状分解の結果. 灰色の楕円が形状分解構成要素となる楕円である. 4 観測画像ごとにサンプリングした. 画像の下の96110318というインデックスは 1996年11月3日 1800 UTCの衛星観測画像を意味する.

この結果は，気象学の文献[70]で報告されているピークの時刻とボトムの時刻にもほぼ一致している．ゆえに本研究で提案する形状分解手法は，気象学的にも意味のある性質を抽出する能力があると評価できる．なお図 17では，平均的に3日目が面積最大となっている．これは形状分解要素には，初日から3日目ごろまで発達を続け以後は緩やかに衰える，という性質があることを示唆している．

台風の眼の検出確率の日変化

同様に台風の眼の日変化についても計測してみた．台風の眼を抽出する場合にも，中心雲領域を抽出する場合と同一の形状分解アルゴリズムを用い，スコアの設定のみを変更する．しかし中心雲領域の場合とは異なり，眼のサイズは小さいものでは数画素幅しかなく，楕円パラメータの推定誤差が大きくなるのは避けられない．そこで眼の日変化に関しては，各画像で眼が検出されたか否かのみを調べることで，地方時と眼の検出確率との関係を調べることとした．

まず本研究の手法による眼の検出確率は約16%となり，気象学の文献[70]で言及されている値23%よりもやや低い値となった．その理由は小さな眼の検出に漏れが多いためと推測される．次に図 18に眼の検出確率の日変化を示す．眼の検出確率は0700LSTに最低となり，およそ2100LSTに最高となった．気象学の文献では，眼の検出確率ではなく眼の大きさに関する調査結果として，眼の直径は早朝0600LSTから0730LSTに最小となり1500LSTから2100LSTに最大となる，との記述がある．そこで，眼のサイズと眼の検出確率とが正の相関関係にあると仮定すれば，本研究の結果は従来の知見と同じ結論を示唆していると考えることができる．このように眼の検出確率に関しても，従来の知見を裏付ける結果が得られた．

以上のように，本研究で提案する楕円形状分解に基づく台風雲パターンの表現は，エネルギ関数に基づくパラメータ最適化という枠組みに基づき，気象学的にも意味のある時間的変化パターンを抽出することができた．この意味で本手法は台風雲パターンの表現に有効であると結論する．

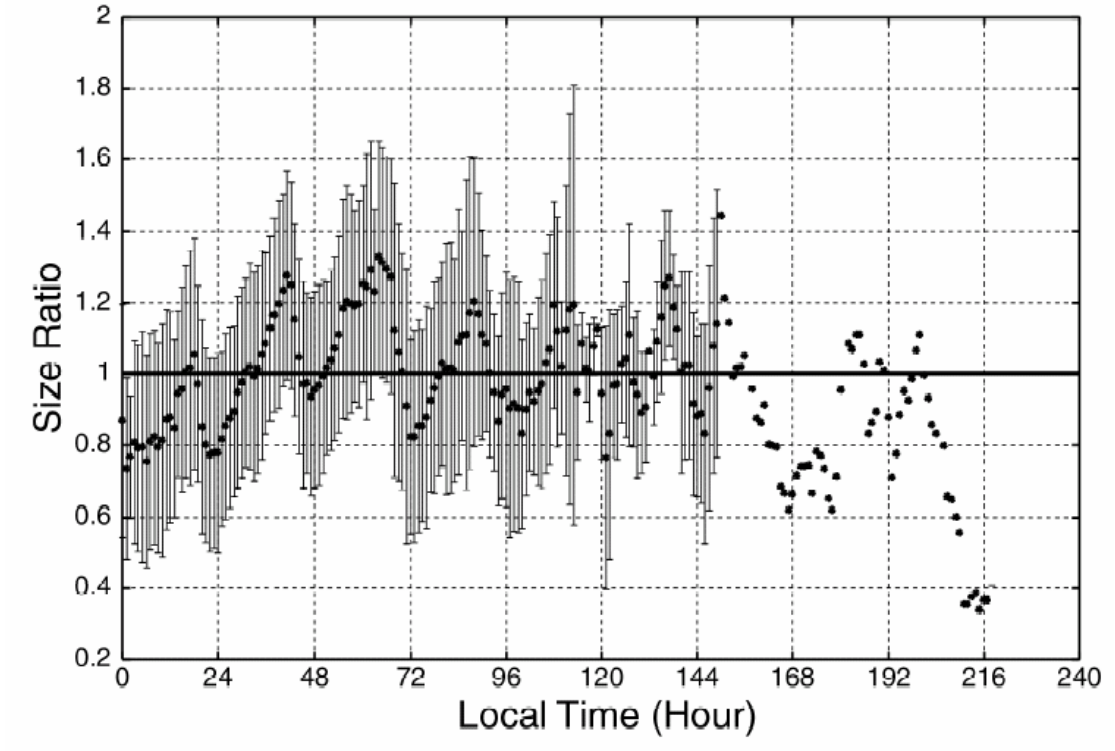


図 17　台風サイズの日変化．長期間の追跡に成功した楕円要素を対象に面積の日変化を計測した．横軸は台風中心の地方時(1時間未満は四捨五入)，縦軸は各楕円要素の追跡期間中の平均面積を1として正規化した大きさ．エラーバーは標準偏差を表す．ベストトラックの階級が「台風」である観測画像のみを対象とした．

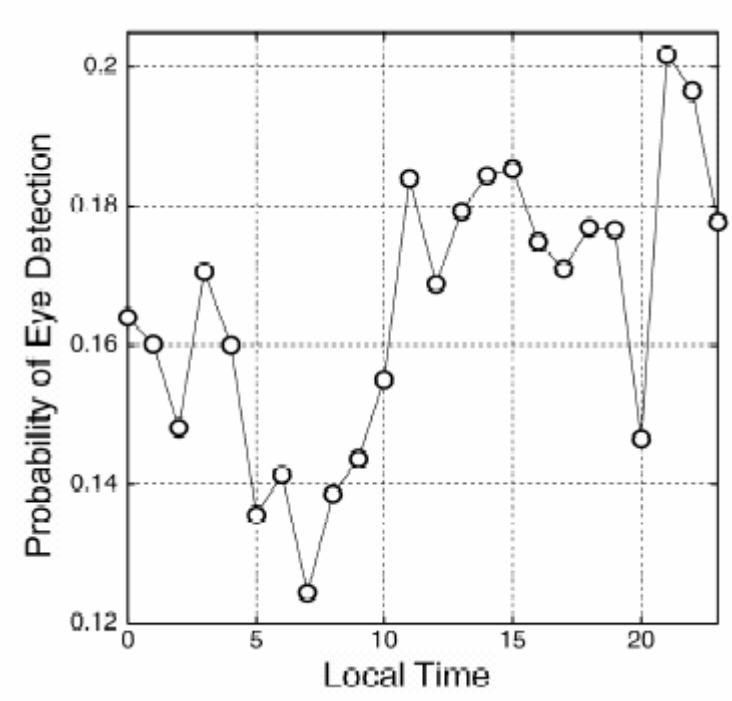


図 18 台風の眼の検出確率と地方時との関係. ベストトラックの階級が「台風」である観測画像のみを対象とした.

6.7 対話型進化的計算論に基づく画像検索

6.7.1 はじめに

利用者とコンピュータとのインタラクションによる発見的な探索方法を重視する画像探索法である「画像散策」の方法論を提案する[2][3][7][9][11]. 関連手法と比較した場合の本手法の特徴は, 進化的計算論[71]の枠組みを用いて類似尺度を対話的に最適化できるという点にある. 本研究で用いる進化的計算論は, 無世代型で非同期的な遺伝的操作という, 対話型進化的計算論[72]に適した特徴をもつ「待ち行列型アルゴリズム」であり, 利用者からの適合度フィードバックの数値を各個体の適応度と関連付けることによって類似尺度を柔軟に適応させるというのが基本的なアイデアである. 気象衛星画像を対象とした画像散策実験を画像散策履歴などの観点から分析した結果, 複数の類似尺度を同時に試す進化的計算論の集団的な性質が, 画像散策に有効であることがわかった.

画像データベースを有効に活用するためには, 利用者にとって重要な内容を持つ画像を, 画像データベースの中から素早く効率的に探し出す内容検索機能が必須である. しかしその研究はいくつかの大きな課題を抱えており, あるデータの適合度が状況に依存して多様に変化するという「検索の個別性」はその課題の一つである. すなわち画像検索とは, 画像データベースに蓄積された個々の画像に対して適合度を付与した後に整列し, 適合度の大きい順番に出力する処理であると考えることができる. ところが実際のところ, この「適合度」とは利用者の検索目的や興味・知識・主観などに依存する尺度であ

り，個々の検索によって変化するものである．つまり，このような検索の個別性に対処できるような柔軟な検索方法の実現が重要な問題なのである．

そこで本研究では，利用者とコンピュータとのインタラクションに基づく対話型処理を用いて検索の個別性に対応する方法を提案する．ただし「対話型処理」とは，単に検索 / 提示のループを繰り返す検索システムを指すのではなく，利用者とコンピュータとの明示的 / 暗黙的なインタラクションを通じて，柔軟で適応的な振る舞い(学習)や検索支援を可能とするような，高度な検索システムを目標とするものである．

6.7.2 画像散策とは

本研究ではまず以下のような用語を定義する．ここで本研究の中心は「画像散策」である．

画像検索	検索キーに対して類似度の大きい画像を効率的に探し出すプロセス
画像散策	利用者の検索目標に対して適合度の大きい画像を対話的・発見的に探し出すプロセス
画像探索	像検索と画像散策とを合わせた総称

「画像散策」は，従来からブラウジングやナビゲーションなどと呼ばれてきた手法と類似した考え方であり，利用者とコンピュータとのインタラクションによる発見的な探索方法を重要視する手法である．従って，検索目標が不明瞭にしか想起できない場合，あるいは画像データベースを対話的に散策する過程を通して徐々に検索目標を形成する場合などで，画像散策の考え方が有効になると考えられる．

さて画像散策というアイデアを実現するための探索方法(シナリオ)について次に考察する．基本的には，以下の二種類の情報を利用者とのインタラクションを通じて収集することが，本質的な課題になると考える．

1．画像特徴空間のどこを重視するのか(範囲指定)

2．画像特徴空間のどれ(どの軸)を重視するのか(重み指定)

この二種類の情報は利用者の検索目的や興味・知識・主観などに依存する情報である．これらに関する基礎的な情報は，事前の調査によって収集することは可能である．しかしその情報を検索の個別性に合わせてチューニングするためには，利用者との明示的 / 暗黙的なインタラクションを通した情報収集が不可欠であると考える．

さてここで，仮想的な多次元画像特徴空間上に画像データベースに蓄積された全画像をプロットした概念図である図 19を用いて，画像散策のシナリオについて説明する．利用者が目標画像(図 19の●)に到達するまでの過程は，以下のようなステップに分解できよう．

１．検索キーの明示的な指定による基準点の設定

まず利用者が最初にすべきアクションは，最初の検索キー(図 19の○)の設定である．この検索キーは，利用者が「画像特徴空間のどこ」を重視するかという範囲指定に関する情報を，システム側に明示的に指示する役割を果たす．そして，画像型の検索では，例示画やスケッチ画を画像特徴空間に写像

した点を検索キーとするのが一般的な方法である. しかし, 利用者が頭の中に想起している「真の」検索目的を完璧に表現する検索キーを実現するのは不可能であり, 不完全な検索キーからいかにして検索目標に到達するかが問題である.

２．類似画像検索を用いた検索キー近傍の探索

検索キーと蓄積画像との間で類似画像検索を実行する. これは画像特徴空間内で, 検索キーの近傍に存在する画像を検索する処理に対応する(図 19の楕円状等類似度線[1]). この際に複数の類似尺度を併用できれば, いろいろな観点のもとで定義された「近傍」を検索することが可能となる. ただし, 検索キーと検索目標画像が画像特徴空間内で非常に離れている場合には, この操作だけでは目標に到達できない.

３．検索キーの更新による探索空間内の移動

画像散策過程では, 現時点で用いている検索キーよりも類似検索上位画像の方が, 検索目標画像により近い画像内容を持つ場合がある. このような場合には利用者の指示により検索キーを更新する. これは画像特徴空間上で, 検索目標画像により近い空間に移動することに相当する(図 19での×の移動). また検索キーのランダムな更新は, 画像散策過程が袋小路に入り停滞してしまった場合に有効である.

４．フィードバックに基づく類似尺度の適応

「画像特徴空間のどれ(どの軸)」を重視するかという重み指定に関する情報を考える. 本研究では画像検索の目標を, アルゴリズム的に計算された類似度によって順位付けられた検索結果と, 人間が想起している「真の」検索目的に対する各画像の主観的適合度の順位付けとの整合性を最大化すること, と考えている. このような目標を達成するために, アルゴリズム的類似度に複数のパラメータを導入しておき, このパラメータを適切に調整することによって利用者の検索目的や主観に適合した柔軟な検索を実現する方法を考える. 具体的には, 利用者に上位検索画像に対する主観的適合度を入力してもらい, この数値を類似尺度の調整機構にフィードバックする方式を用いる(図 19での楕円状等類似度線の変形).

このように本研究が提案する画像散策シナリオでは, 2から4を繰り返すことによって, 範囲指定と重み指定の両方を適応的に更新する方法を念頭に置いている. このように利用者とコンピュータとのインタラクションを活用する対話的な検索方法は, 情報検索の分野で主にテキスト検索を対象に研究が進められてきたテーマであり, 利用者が検索結果に与える適合度フィードバックに基づき検索キーや重み付けを更新する方法に関しては, 既に多くの研究の蓄積がある[73]. 最近は本研究のように, 類似画像検索にこのアイデアを適用しようとの試みも, 数多くおこなわれるようになってきた[74]. ただし, フィードバック作業に要求される利用者の労力を可能な限り削減することが重要な研究課題であり, そのためには利用者のあらゆる反応を追跡し, そこから最大限の情報を効率的に抽出する機構を考える必要がある. このような点についてはまだ研究が十分に成熟していない.

---

[1] もちろん、重み付きユークリッド距離などの単純な類似尺度を用いない限り、等類似度線は楕円とはならない。

本研究ではこのような対話的な検索過程には進化的計算論を対話型に拡張した方法論である「模擬育種法」が適していると考え，これを適用する方法で研究を進める．さて類似画像検索に適した対話型進化的計算論とはどのようなものだろうか．

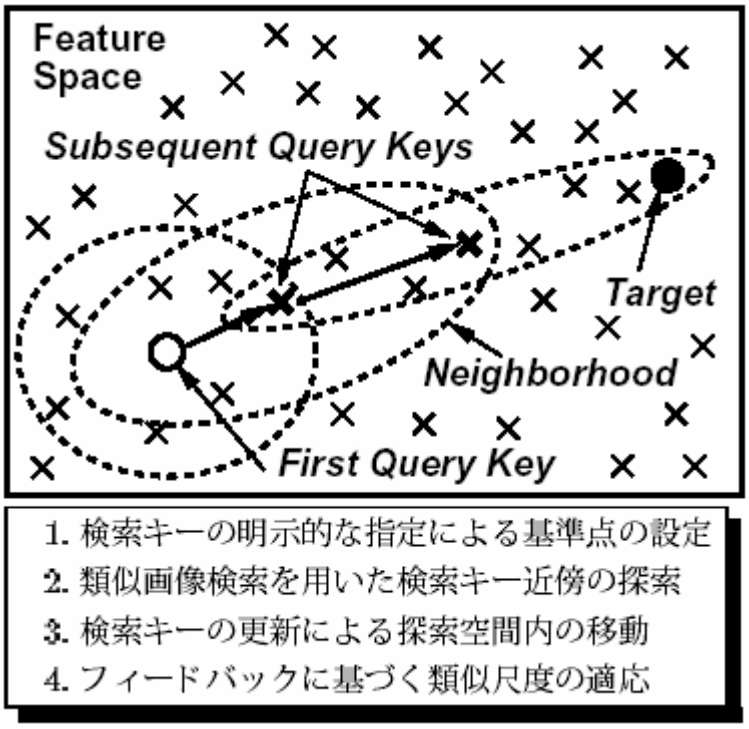


図 19　仮想的な画像特徴空間の概念図と，画像散策のシナリオを構成する各ステップ．

6.7.3　対話型進化的計算論

先述したように本研究の大きな特徴は，適合度フィードバックによる類似尺度の適応に進化的計算論の枠組を用いる点にある．なぜ進化的計算論をこの問題に適用するのか，更にこの問題に適用するに当たってアルゴリズムをどのように改良すれば良いのか，などの点についてここで述べる．

「進化的計算論」とは，生物の進化過程に触発された確率的探索アルゴリズムを指す．中でも代表的なのは，遺伝的アルゴリズム(Genetic Algorithm: GA)[71]と，遺伝的プログラミング(Genetic Programming : GP)である．まず進化的計算論の基本構成は以下のようにまとめられるが，詳細については紙面の都合上省略する．

1．初期集団の生成

2．終了条件が満たされるまでループ

2－1．遺伝子型から表現型を生成

2－2．表現型に対して適応度を評価

2－3. 遺伝子型に対して遺伝的操作を適用

3. 適応度に基づく選択

4. 交叉

5. 突然変異

さて本研究で用いる「対話型進化的計算論」は，上に述べた「進化的計算論」を対話型に拡張したアルゴリズムである[72]. 具体的には上の(2-ア)において，各個体の表現型に対する「良さの評価」を人間自身が行うという部分が対話型への拡張部分である. こうして得られた評価値を各個体の適応度と関連付ける仕組みを備えることによって，進化的計算論の枠組を用いた対話的なパラメータ最適化が可能となる. このような対話型進化的計算論の枠組を用いる利点は以下のようにまとめられる.

> 「人間が高い評価値を与えた個体に高い適応度を付与する」という自然な規則で進化的計算論の適用が可能である. 次に進化的計算論は勾配情報などを使わないロバストな方法であるため，人間が与える「大ざっぱな」評価値に基づく探索が可能である.

進化的計算論のように集団に基づく探索方式では，複数の評価規準を同時に探索することも可能となる. このように対話型進化的計算論の枠組は，画像散策への応用に適した複数の長所を備えていると言える. 本研究では先述のように「重み指定」型の画像散策過程を用いるため，重み係数の役割を果たす検索パラメータを個体の遺伝子としてコーディングし最適化することを目標とする.

6.7.4　待ち行列型アルゴリズム

次にアルゴリズムの改良について，我々らは模擬育種法の欠点である「待ち時間の問題」を解消する方法として「待ち行列型アルゴリズム」を提案している. このアルゴリズムは，「世代」という概念に基づく逐次的な処理を解消し，無世代型のアルゴリズムとして構成されているのが大きな特徴である.　加えて待ち行列型アルゴリズムでは，「集団」よりも「待ち行列」がアルゴリズムの基本となるため，非同期的な遺伝的操作や個体数に因われない構成が可能となる. 紙面の都合上，本研究では待ち行列型アルゴリズムに特徴的な「遺伝的操作」に絞り簡単に紹介する. またモジュール構成については図 20を参照のこと.

適応度に基づく選択

選択には「先着順トーナメント選択」という方法を用いる. まず待ち行列の先頭から2個体を取り出して適応度を比較し，適応度が高い方を勝者，もう一方を敗者と決定する. そして勝者は一定の確率 r_w で2個体に増殖させる一方，敗者は一定の確率 r_l で死滅させる. このように2個体のみのローカルな比較で選択操作が行えることがこの方法の特長である. 選択操作を生き抜いた個体は，それぞれ勝者待ち行列(W-Queue/WQ)と敗者待ち行列(L-Queue/LQ)の2本の待ち行列に加えるが，勝者待ち行列を以後の遺伝的操作で優先させることで，有望な個体のライフサイクルを加速する戦略も併用する.

交叉

交叉としては「勝者待ち行列内の交叉（WQ2)」,「勝者待ち行列とエリート集団[1]との交叉（WQEP)」,「敗者待ち行列とエリート集団との交叉（LQEP)」の3種類を用いる. 各回の交叉ではこの中から任意の交叉を確率的に選び出すことになるが, このとき勝者待ち行列が関係する交叉を選ぶ確率を高めることで, 勝者待ち行列を優先する戦略を用いる. またエリート集団との交叉は, エリート集団にプールされた有望な遺伝子を待ち行列個体集団に定期的に注入する役割を果たす.

突然変異

突然変異にはある固定した確率で各遺伝子を変化させる方法を用いるが, 突然変異率の値自体は交叉のタイプに応じて変化させる. なぜなら, WQEP交叉は有望な遺伝子同士の交叉であるため高い突然変異率は有害となる可能性が高いが, LQEP交叉では突然変異率をむしろ高めて探索範囲を拡大する方が有利と予想できるからである.

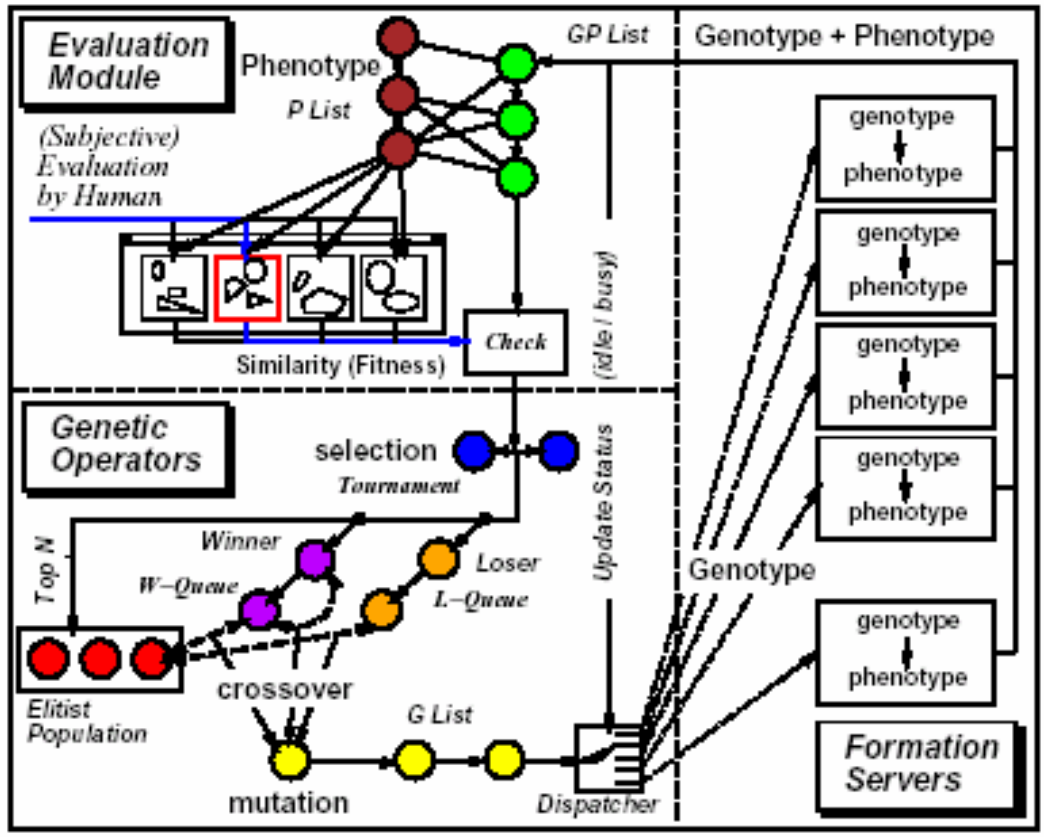


図 20　遺伝的操作モジュールの模式図.

[1]探索の過程で発見された N 個の最良個体（重複なし）をプールしておく集団であり, 固定サイズの個体集団として管理する.

6.7.5　適合度の入力

適合度の入力インタフェース[1] として，本研究では図 21に示す3種類のインタフェースを用意した．いずれも同時に$3\times3=9$件の画像を表示するインタフェースを用いており，この9枚は中央に表示される問合せ画像と，残り8件の評価画像からなる．いずれのインタフェースにおいても，すべての画像に対して評価を与える必要はなく，また任意のタイミングで問合せ画像を更新できる設計となっている．このインタフェースの背後では，評価対象画像は各個体の待ち行列と関連する待ち行列として管理しており，表示場所に空きができれば直ちに待ち行列の先頭の画像を表示するようにしている．そして利用者が適合度を入力すると，その画像を上位に検索した個体すべてに従って適合度が分配されるため，適合度の評価回数と個体数とは一致しない．

ここで各個体に分配する適合度について簡単に説明する．各個体の表現型は，遺伝子型（類似尺度の重みパラメータ）に基づき検索した場合の上位B件の画像であるとする．このとき各個体の適応度Fは，このB件の画像に対して利用者が入力した適合度の重み付き和として以下を計算する．

$$F=\sum_{i=1}^{B}W\left(\frac{c_i}{c_1}\right)e_i$$

ここでe_iは上位i位の画像に対して利用者が入力した適合度であり，c_iは上位i位の画像と問合せ画像との類似度である．また関数$W(x)$は類似度を用いた重みを調整するための関数であり，類似度が大きい画像への重みほど大きくするための役割を果たすものである．

このように適応度を計算するため，各個体は上位検索画像すべてに対して評価が定まる（評価しないというのも一種の評価である）までは待ち行列中に待機する．そして評価が定まれば直ちに上記の2固体間遺伝的操作を受け，進化的計算論による最適化のステップに入る．以上がインタラクションのプロセスである．

---

[1]この章の成果は多少以前のものであるため、画像散策のためのGUI（Graphical User Interface)は、WWWベースではなく、C言語およびOSF / Motif 1.2によって構築している。今後このシステムを作り直すことがあれば、WWWベースのシステムとし第8章のシステムと統合する計画である。またそもそもこの手法は個体をネットワーク上に「ばらまく」方法であるため、分散コンピューティングあるいはクラスタ環境への拡張も容易である。

(a) グラフ尺度法 [18]	(b) 2 値法	(c) 選一法 [18]
いわゆる「グラフ尺度法」に基づき、各画像の下部に表示された数直線の 1 点をクリックして、[0,1] 区間の準連続数値としての重要度を入力する。もし 8 枚の評価画像すべてが重要でない場合は、"Refresh" ボタンによってすべての画像に重要度ゼロを付与する。	重要度を 2 値で、すなわち評価画像がユーザにとって正の (positive) 重要度か負の (negative) 重要度かを選択して入力する。しかし実際には積極的に評価を与えられない場合も多く、ユーザが何も評価を与えない画像に対しては消極的な重要度ゼロを付与する。	ウィンドウに表示された 8 枚の中から最も重要度の高い画像を 1 枚だけ選び出し、その画像のみに正の重要度を与える方法である。他の画像は重要度ゼロである。もし重要な画像が 1 枚もなければ "Refresh" ボタンによってすべての画像に重要度ゼロを付与する。

図 21　適合度入力インタフェースの3つの例.

6.7.6　画像散策実験

画像散策実験の手順

まず検索開始キー（例を右画像の左上に示す）、およびユーザが想起している到達目標画像（例を右画像の右上に示す）とを設定する．次に画像散策のシナリオに基づき，到達目標画像を常に想起しながら画像散策を進め，到達目標画像が評価用ウィンドウに表示された時点で実験を終了する．

評価方法

画像散策過程の進行に伴って，到達目標画像の類似検索順位が単調に向上すれば理想的な結果である．すなわち，検索キーの更新による探索空間の移動によって類似検索順位の大幅な向上が起こり，また進化的計算論による類似尺度の最適化によって類似検索順位の継続的な向上が起こると期待される．そこで、各個体に対応する類似尺度と、各時点での検索キーを用いて類似画像検索を行った場合の，到達目標画像の類似検索順位の変動を追跡し、これが時系列的に向上しているかどうかを確認する。

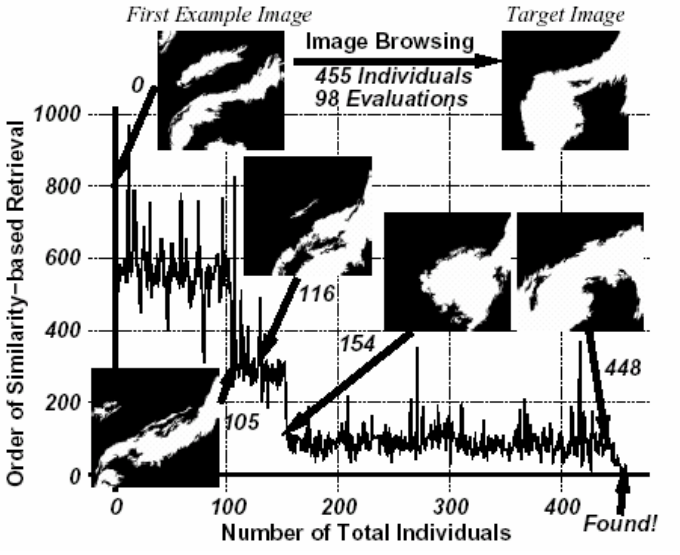


図 22　画像散策実験の手順と画像散策履歴の評価.

図 22に示す手順にしたがって画像散策の実験をおこなった．その結果，検索開始キーから検索キーを変更することなく到達目標画像に早々と到達できる簡単なケースもあれば，検索キーを何回変更して

も到達できない困難なケースもあることがわかった. そこでほぼ中程度の難易度に相当する場合の画像散策履歴を図 22に示す.

まず画像散策の初期段階では, 1027件の画像の中で到達目標画像の類似検索順位は約600位に過ぎず, このままでは目標画像が上位に検索される可能性はほぼゼロである. しかし, 105個体目, 116個体目, 154個体目に検索キーの更新により到達目標画像に近い探索空間に移動した結果, 類似検索順位は段階的に上昇して100位程度にまで達した. その後は順位が一時停滞するものの, 448個体目の検索キーの変更を経て, 最終的には98回の適合度入力と455個体の生成後に到達目標画像を4位に検索して画像散策実験は終了した.

この一方で, 進化的計算論を用いない場合, すなわち類似尺度を固定化して類似画像検索の繰り返しのみで画像散策する実験も行った. 定量的な比較はここでは述べられないが, 類似尺度を固定化しておくとすぐに袋小路にはまってしまい, 検索キーをランダムに更新しない限り画像散策が進まないことがわかった. この結果は, 複数の類似尺度を保持しつつ画像散策を進める方法の優位性を示している.

6.7.7 今後の課題

まず3種類のインタフェースを比較してみたい. 我々自身が使用した経験では, (c)の選一法が最も使いやすいインタフェースであると感じられた. その次が(a)のグラフ尺度法である. それに対して(b)のようにカテゴリを選ぶ方法はかなり面倒に感じられた. いずれのインタフェースでもすべての画像に対して適合度を付与する必要はなく, 結果的にはそれほど多くの評価回数を必要としない場合が多かった. むしろ利用者がフラストレーションを感じるのは, 画像散策がどのように進行し, 果たしてうまく進行しているのかが, 利用者にとってわかりにくい点にあるのではないかと考えられる. この問題を解決するためには, 利用者とのインタラクションを支援し, また現在の状況を適切に可視化する機能が必要である.

また進化的計算論の適用が真に有効であるのか, という点についても議論があろう. なぜなら, 進化的計算論はロバスト性に大きな特徴があるものの, どちらかというと「時間はかかっても良いからロバストな探索アルゴリズムが欲しい」という場合に適したアルゴリズムである. このようなアルゴリズムが人間とのインタラクションという「時間との勝負」が重要な問題に果たして有効なのだろうか. この場合は決定論的なアルゴリズムの方が有効である可能性もあり, これらは今後の課題として検討していきたい.

最後にこのような高度な検索機能を第8章で述べるデータマイニングシステムに組み込むという課題がある. データマイニングにおいて人間とのインタラクションが重要であるという意見は最近になって盛んに唱えられるようになってきており, そこからアクティブマイニングなどの新しい流れが生じてきている. すなわち, データマイニングのように本質的に探索的な処理においては, コンピュータが最初から最後まで自動的に処理し結果だけを表示するような計算パラダイムは適しておらず, むしろコンピュータが規則性あるいは不規則性を発見し, それを見て人間の考え方が変化して別の観点から物事を眺めるようになり, そこからまたコンピュータに新たな指示を与えるといったような, 人間とコンピュータが相互に影響を与え合うような関係が望ましいと考えるアプローチである.

本研究のアイデアは, そのようなプロセスを一種の進化と捉え, そこに生物学的メカニズムを導入することにより, より有効なインタラクションを見つけたいという動機で始まったものである. 本研究では画像検索というタスクに絞って研究を進めたが, 今後はデータマイニング分野への応用も目指していきたい. またこの際に, 構造をもつデータへのデータマイニングを可能とするための, 遺伝的プログラミングの適用も興味深い研究課題である.

第7章 地球環境データ（台風データ）のマイニング

7.1 はじめに

台風は地球上に出現する気象現象の中で最もドラマチックな現象の一つであり，その渦巻き型の形態の美しさは多くの人々の興味を引いてきた．その一方で，台風は最も破壊的な気象災害を引き起こす気象現象でもあり，その社会的な重大性から台風解析や台風予測には気象学を中心とした多くの研究者が多大な努力を費やしてきた．

これらの研究を通して台風の成因や構造の理解，および台風解析と予測の精度は着実に向上してきているものの，困難な課題が未だに多く残されているのも確かである．例えば現在の台風解析は，衛星画像上の台風雲パターンを熟練者が目視で解釈し判断する方法が主流であるが，この方法は過去の経験則を土台とした方法であり，数理的な根拠をもつ洗練された方法とはなっていない．また台風予測に関しても，台風雲パターンから台風急発達の予兆を発見するなどの観点からの予測は，現在の主流である数値予報にとっては困難な課題である．本研究はこのような台風雲パターンの解析や予測といった問題に対して，情報学的パラダイム，すなわち大量の台風画像コレクションを用いて，統計的パターン認識手法や機械学習手法に基づく，情報論的あるいは統計的な論理や根拠をもつ台風解析手法や台風に関する知識発見を目標とする[15][19][20][21][23][24][26][28][29][30][31]．

このような情報学的パラダイムの基盤となるのが大量のデータセット，すなわち34000件以上に達する台風気象衛星画像の大規模データコレクションである．このデータコレクションは，南北両半球の台風画像を，網羅的にかつ一貫性を保ちながら収集したものであり，気象庁などが発行するベストトラックデータに基づく高品質のデータコレクションである．このデータコレクションに対する空間的データマイニング，時間的データマイニング，さらには時空間的データマイニングの研究，つまり台風雲パターンというデータに対して，情報学的に何が言えるかを徹底的に探っていこう，というのが本研究の動機である．こうした研究から得られる結論は，気象学的な知見に一致するかもしれないし，あるいは全く新しい視点を与えるものかもしれない．おそらく新しい視点の方が，台風解析や台風予測，またはそれを通して社会に与えるインパクトは大きくなると考えられる．

例えば単に台風を予測するのではなく，まれにしか発生しないが重要な事象が起きそうだ，ということを事前に察知するという予兆発見技術を活用できれば，現在の気象予測とは違った意味で価値のある情報を与えることができるかもしれない．なぜなら，台風に起因する災害の防止や軽減などには早期警報が一定の効果をもたらすが，この早期警報は特にまれではあるが重要な事象の場合に最も効果が大きいためである．したがって，現在の気象予測とは全く異なるアプローチで問題に挑む価値はあると考え

ている. 本報告では現在までにおこなった研究の一部を紹介する.

7.2 なぜ台風を研究対象とするのか

社会的インパクト

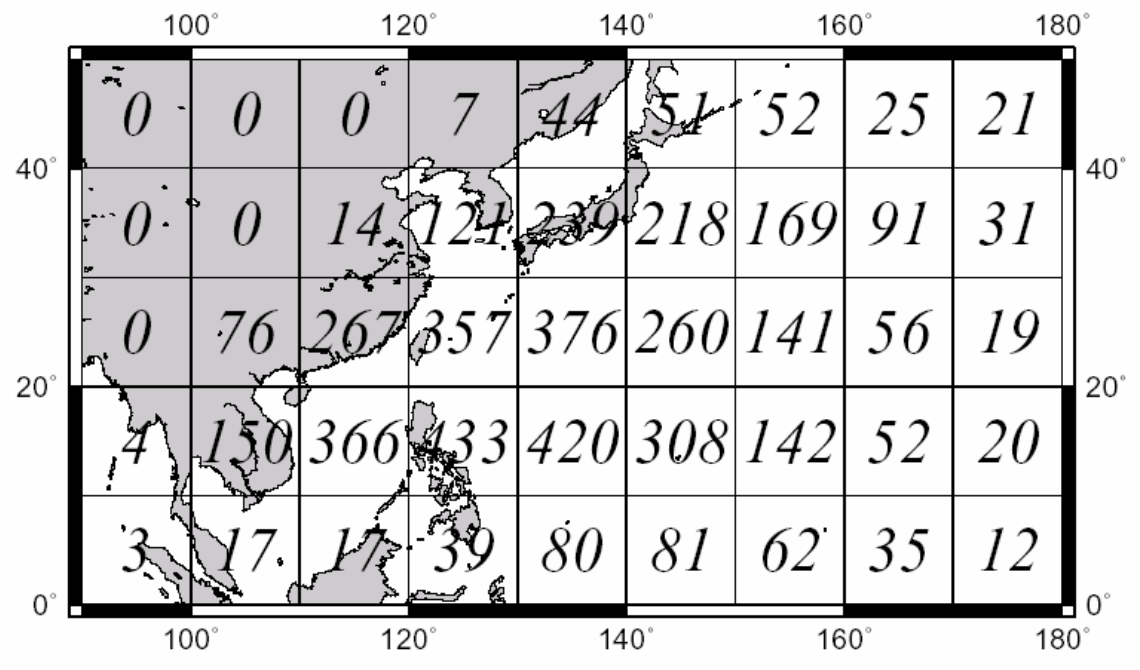


図 23　北西太平洋地域における台風通過地域の分布. 1951年から1999年までの1320個の台風について, ブロックごとに通過件数を数えた.

ここで改めて述べるまでもないが, 台風は日本に大きな被害を与える気象現象であり, 被害の軽減に結びつくような台風解析・台風予測技術への期待は常に大きい. そこでまず, 日本, あるいは東アジア地域にどの程度の台風が通過あるいは上陸するのかを見てみたい. 図 23は北西太平洋地域における台風通過個数の分布を示す. これによると, 台風が最も頻繁に通過する地域はフィリピン東方海上であり, 全台風のほぼ3分の1が通過するという高い頻度を示している. しかもこの地域では多くの台風は最盛期であり, その点からもこの地域が台風に関して最も危険地域であることは間違いない. 次に危険な地域となりそうなのは, 日本の南海上, 東シナ海および南シナ海, またこれらの地域に接する日本, 中国そしてベトナムあたりの国々である. これらの国々では, 台風によって引き起こされる災害, すなわち強風, 大雨, 高潮などによって過去に繰り返し大きな被害を受けているため, 台風による災害をいかに軽減するかということには大きな関心がある. このように, 台風に関する研究は社会的インパクトが大きい, というのが, この研究の一つの動機となっている.

情報学的な面白さ

気象衛星画像に現れる雲パターンの中でも，台風の渦巻き状の雲パターンはひときわ目を引く存在である．発生当初は小さな雲のクラスタであるものがしだいにまとまって渦巻き状に変形し，やがてくっきりした眼をもつ最盛期を迎える．しかし台風の北上とともに渦巻き状のパターンは急速に崩壊していき，やがて雲パターンの消滅をもって台風もその短い一生を終える．これは台風の「典型的な」ライフサイクルであるが，しかしどの台風のライフサイクルもこのような「典型的な」パターンをたどるわけではなく，台風ごとにそのライフサイクルは千差万別に異なる．また台風の雲パターンは短時間のうちに形態が爆発的に変化することも多いことが，長年の衛星観測を通して明らかとなってきた．このように，台風の雲パターンは，渦巻き状の雲パターンよりもはるかに豊かな多様性をもち，ライフサイクルも個性に富む存在，それが台風という気象現象である．

このような時空間的なパターンとしての多様性をモデリングするという研究課題は，特に画像解析技術として非常に挑戦的な課題である．また台風雲パターンを，発達しそうな台風やそうでない台風に分類する問題などは，まさに典型的なパターン認識問題であるし，台風の典型的なパターンをデータセットの中から見出す問題も，クラスタリングなどの典型的な機械学習問題として定式化することができる．つまり台風の雲パターンを研究対象とみた場合に，そこには情報学的に興味深い研究課題をいろいろと発見することができるのである．またこのような実世界の多様性に耐えうるロバストな手法は，情報学における長年の課題であり，この問題で培った技術は他の領域，例えば生物学のように時間変化する柔軟物体（個体，細胞など）を対象とする研究や，さらに他の時系列画像へも応用できる可能性がある．

研究素材としてのデータの特徴

上に述べたようなパターンとしての面白さは，他の地球環境データと比較するとより際立つものである．例えば陸地に関する地球環境データでは，季節ごと，あるいは1年ごとに，特定地点の状態がどのように変化するかに興味がある．そのような状態を推定するためには，陸面に関する専門知識に基づく精緻なモデルを作り込むことが重要な研究課題であり，そこには機械学習やパターン認識などの情報学的手法が活躍する余地は小さいように思える．それに比べると，台風には情報学的課題をより多く見出すことができる．

また台風を他の気象現象と比べても，台風の特殊性が際立ってくる．一般の気象データ，例えば気象測候所やアメダスにおける観測データは，固定地点での気象要素の時間変化を記録したものである．このような気象記録を，ここでは7.4.2章で述べるようにオイラー的視点と呼ぶ．それに対してテレビの天気予報などに登場する天気図は，高気圧や低気圧などの移動するものの特徴点の動きを追跡している．これをラグランジュ的視点と呼ぶ．台風も明らかに後者の視点に属するものである．しかし入手可能な気象データというのは，実は大部分がオイラー的視点に基づくものであり，低気圧の中心位置を1年分すべて記録したデータのようなラグランジュ的視点のデータセットはほとんど入手できない[1]．実際のところラグランジュ的視点のデータは，特徴点の移動にあわせて気象測器を移動させることができない限り，測定することが困難なのである．またオイラー的気象データからラグランジュ的気象データを推測するにしても，それを果たして精度よく推測できるのかという点に難問がある．ゆえに，ラグランジュ的視点に基づく気象データセットは，台風データセット（ベストトラック）がほとんど唯一の存在といってよい．このように台風に関してだけデータセットが存在するというのは，これが世界の気象機関にとって最重要の気象

[1] 新聞の天気図画像を毎日スキャンして画像解析すれば、1年分の低気圧軌跡データセットを作るのも不可能ではないが、そうやって作ったデータの精度は全く保証できず、しかもデータセットの網羅性にも疑問符がつく。

現象であり，特別の体制で観測に臨んでいるおかげである．このように台風データセットというのは，気象データの中でもかなり特殊な位置付けのデータセットと言える．

と同時に，台風の雲パターンの多様性を考えたとき，これが挑戦できるギリギリのレベルの多様性であることも指摘しておきたい．もし一般の（温帯）低気圧を対象とするならば，こちらは雲の多様性が大きすぎ，対象をモデル化するための手がかりをつかめない恐れがある．それに対して台風は基本形+変異といった枠組みでなんとか扱える（と思われる）程度の多様性であるため，ここに情報学的パラダイムで挑戦する余地が生まれるのである．

以上のように「台風画像データコレクション」は，研究素材としても他のデータに比べて研究に適した優れた特徴をもっており，データに基づく研究を進めるには適した研究素材であると評価する．

7.3 台風に関する気象学的課題

台風とは北西太平洋に起源をもつ成熟した熱帯低気圧(tropical cyclone)に与えられる地域的な用語である．成熟したかどうかは最大風速で判定され，日本の気象庁では最大風速が34ノット以上(17．2 m/s以上）の熱帯低気圧を「台風」と呼んでいる．世界の他の地域で発生する熱帯低気圧は，ハリケーンやサイクロンなどと呼ばれるが，いずれも物理的には同じ性質をもった大気擾乱である．

7.3.1　台風雲パターンの形態学的特徴

台風雲パターンの最大の形態学的特徴は「眼の壁雲」と「らせん状のレインバンド」にある[75]．眼の壁雲は台風中心付近の暖かい空気の核による強い上昇気流に対応しており，その中心に向かって下層では摩擦収束による吹き込み，上層では巻雲の吹き出しが発生している．また空気塊に働くコリオリ力の向きから，円形の等圧線に沿って吹く風は北半球では反時計回りに，南半球では時計回りに回転することになる．このようなコリオリ力や遠心力などの作用により，台風の中心雲領域の形態モデルとしては円形や楕円形が考えられる．一方らせん状のレインバンドは帯状に組織化された多くの対流雲に対応している．これがらせん状になる理由は完全に解明されていないものの，地表付近の風が等圧線とある角度をなして吹き込むことが重要な理由であると予想されることから，形態モデルとして等角らせん（ 対数らせん） $r = a^{\theta}$ を用いることができる．

このような円形の核とそこからのびるらせん状バンドという点で台風に類似した自然界のパターンには，宇宙に存在する銀河の形態がある[76]．銀河の形態学では，理想的な銀河の形態を表す「典型」と現実の銀河の形態とを比較し，現実の銀河と類似した「典型」に基づいて銀河を分類し解析することで，銀河の進化過程を把握する試みがなされている．この問題意識は，台風の形態学にもある程度は共通している．果たして台風には「典型」はあるのだろうか．また「典型」に基づいて解析することには利点があるだろうか．

7.3.2　台風解析技術の現状

気象学では一般に(1) 観測，(2) 解析，(3) 予報，の3点が問題とされているが，その事情は台風に関しても全く同じである．まず観測に関しては，台風を直接観測するための有力な手段となっていた飛行機

観測が廃止された現在では，気象衛星が熱帯低気圧監視の最大の武器となっており，台風解析においても気象衛星画像が大きな役割を果たしている[70]．この台風解析の結果は，台風予報の精度にも大きな影響を与える．というのも，台風モデルを気象予報モデルの初期場に設定する際に用いられる，擾乱を特徴づける少数のパラメータ(台風の中心位置，中心気圧，強風半径など)は，台風解析の結果に大きく依存しているためである．このように台風解析の善し悪しが台風予報の精度に大きく影響するという意味で，両者には密接な関係が存在している．

まず台風予測技術については，気象力学の運動方程式系に基づく数値予報モデル（気象シミュレーション技術）の進展によって，台風予測の精度は大幅に向上した．台風発生の予測など，まだ困難な問題も残っているが，進路予報は最も高い精度を達成しており，また強度や大きさなどに関する発達予報もそれに続く精度を達成している．ただし台風予測の精度が格段に向上した理由は，計算機性能の向上に伴うモデルの解像度の増強による部分が大きく，必ずしも台風の理解の進展に裏打ちされたものではない，との指摘もある[70]．実際のところ，台風は様々なスケールの激しい気象現象が複雑に絡みあうことから，そのメカニズムの理解には台風特有の困難さが存在し，シミュレーションによる台風の再現さえも決して容易ではない．ゆえに，確かに台風予測は進歩したものの，まだそのシミュレーション能力は万全には遠いのである．

一方の台風解析技術の現状については，専門家の文章をそのまま引用したい[70]．

> 衛星データによる熱帯低気圧の解析手法は，最初の極軌道気象衛星TIROS時代から米国を中心に精力的に開発が進められ，1960年代から70年代に大きな進展を見せた．この解析手法の特徴は，衛星画像から熱帯低気圧の「雲パターン」を認識することをベースにしているところにある．その成功した典型がDvorak法であろう．もちろんより客観的に熱帯低気圧の気象要素を抽出しようとする研究も数多く行われているが，現在も，少なくとも現業的に利用される解析手法は，この時代の延長上にあると言っても差し支えない．

このように台風解析は，「雲パターン」の認識という人間のパターン認識能力に依存した主観的な方法であり，1970年代以降は画期的なブレークスルーもなかったようである．確かに衛星観測データから実際の大気状態を復元する技術（データ同化技術）についても研究は進歩したが，しかし気象予報モデルと比較すればその進歩は地味なレベルにとどまっているように思える．その本質的な理由はおそらく，乏しい観測値から無限自由度力学系の状態を復元するという問題が，結局のところ解が一つに定まらない不良設定問題に帰着してしまい，大気力学理論とは別種の難しさを乗り越えられないことが原因であると考えられる．この困難を打ち破るためには，現状ではやはり，熟練した専門家のパターン認識能力，あるいは見えないものを見る「想像力」に頼らざるを得なくなるのである．

以上の議論をもとに，多少図式的に台風解析と台風予報とを比較してみると，両者は対照的な方法論に基づいていることがわかる．

1．台風予報― 気象力学の運動方程式系に基づく演繹的モデル

2．台風解析― 雲パターンに対する解釈と過去の観測記録との対応関係に基づく帰納的モデル

この両者を比較してみると，やはり弱点は台風解析の方にあると言えるだろう．しかし台風予報の精度向上のためには，数値モデルの高度化とともに台風初期値の精度向上，つまり台風解析の精度向上が不可欠であることを考えれば，台風解析に関しても画期的なブレークスルーが欲しいところである．そこでこれまで30年以上も使い続けられ，上記の文章中で「成功した典型」とも言及されているドボラック(Dvorak)法について，次節で検討してみよう．

7.3.3　ドボラック法

ドボラック法とは，衛星画像の雲パターンから熱帯低気圧の強度を推定する方法であり，アメリカ大気海洋局(NOAA) のハリケーン研究者V. F.　Dvorak により開発された[82]. これは衛星画像の雲パターンから熱帯低気圧の強度を推定することを目的とした手法であり，ドボラックはその強度推定の基準となるような台風雲パターンの「典型」を考案した. この典型に基づき，まず解析者は現実の台風雲パターンと最も類似した典型を探し，次に各典型に対して定められた分類木などの過去の経験則を適用することで熱帯低気圧の強度が簡便に推定できる，というのがドボラック法の仕組みである. この手法は，その簡便さから現在では世界中の熱帯低気圧解析センターで幅広く用いられており，日本でも気象衛星センターにおける熱帯低気圧の解析にはドボラック法に基づく方法が用いられている. 表 5にはドボラック法の処理手順をまとめた. この処理手順では，数箇所で雲パターンの認識処理が必要である. その多くの場合では，パターン認識は実際には手動的または半自動的であり，解析者の主観的な判断が入り込む余地がある. また示数の設定に関しても過去の事例から得られた経験則が多数組み込まれている.

さて，現在でも世界中の気象機関でドボラック法が活用されていることは，台風雲パターンという情報が台風の強度推定に必要な情報を含んでいることの強力な状況証拠になっている. またドボラック法はあくまで過去の観測例から人手により定式化された経験則であるため，気象力学モデルから演繹されるような確固とした理論的基盤を有するわけではない点に着目する. すなわち，大量のデータに基づく大規模機械学習手法や，人間の意思決定という問題を長年扱ってきた人工知能的手法を，このドボラック法という枠組みに注ぎ込むことによって，台風解析という問題に関する何か新しい展開が得られるのではないだろうか.

7.3.4　過去の情報学的アプローチ

気象学に対する情報学からの貢献は，残念ながらこれまでのところあまり大きくはない. とはいえ，情報学にとって気象学が目新しい応用分野というわけではなく，特に1980年代にエキスパートシステムが流行した時代には，気象予測をエキスパートシステムによって実現しようとの気運が盛り上がった時代があった. 例えばShootout-89[78]では人工知能システムの比較研究として，6種類のシステムが参加して，気象予測に関する大規模な比較実験がおこなわれた. システムの内訳は，伝統的なエキスパートシステムが3種，小規模エキスパートシステムと線形モデルのハイブリッドシステム，類推に基づくシステム，そして人間の認知判断に基づくシステムであった. しかし実験の結果，気象予測は予想されたよりもはるかに困難であることが判明した. というのも，これらのシステムはいずれも，持続予測あるいは気候予測[1]よりも有意に優れた結果を生み出せなかったからである. ゆえに最近では，情報提示の工夫により意思決定者を支援するエキスパートシステムの研究[79]などに，研究テーマの力点も移ったようである.

一方，台風画像解析は時空間系列パターン認識の典型的な問題であり，また台風画像予測は非線形時系列予測の典型的な問題であることから，台風データへの情報学的アプローチの適用例も少数ながら存在する. 例えば，流体力学に基づき時系列衛星画像から台風の雲の動きを解析する研究[80]や，衛星画像の台風雲パターンをdynamic link architectureと動的輪郭を用いて解釈する研究[81]などがある. また衛星画像を用いない研究には，ニューラルネットワークを用いた台風データマイニングの研究や

[1]持続予測では直前の観測値を予測値とし、気候予測ではある年月日の過去の平均値を予測値とする。これらは予測評価の最低水準である。

ファジー理論を用いた台風進路予測の研究などが散見される．しかしこれらの研究はいずれも，情報学的手法の一応用例として気象学的問題が取り上げられたという色彩が濃厚であり，両分野の橋渡しをして実際に気象学に何らかのインパクトを与えるという段階までには達していないように感じられる．

1．**CSC（Cloud System Center)の決定**　雲パターンから中心位置を推定する．眼が明瞭であれば比較的容易であるが，組織化された雲パターンでない場合には困難な作業である．

2．**DT数(Data T-number)の決定**　DT数とは「同じ雲パターンであっても，雲頂温度や雲バンドの長さなど画像上で測定できる雲の要素を測定し数値で表せば，その強弱を客観的に判断できる」との発想に基づく示数である．最初に衛星画像の雲パターンが，″Curved Band″，″Shear″，″Eye″，″Embedded Center″のどのパターンに最も近いかを判定する．次に，雲パターンごとに異なる雲の要素を測定する．測定対象は，雲システムの中心部に関する要素CF（Central Feature)，およびCFを取り巻く雲バンドに関する要素BF（Banding Feature)であり，これらの測定値を換算してDT数を求める．

3．**CCC（Central Cloud Cover）**　熱帯低気圧の雲システムの中心部が，冷たい丸みを帯びた雲に覆われた場合の補正をおこなう．

4．**前24時間の変化傾向**　前24時間のT数の変化量から，発達/衰弱/変化無しのいずれかを選択する．

5．**MET数(Model Expected T-number)の決定**　「平均的な台風の発達モデルでの発達/衰弱時の変化率を参考にすると，これまでの雲パターンの変化傾向から現在の強度を推定できる」との発想に基づく示数である．具体的には24時間前のT数，および変化傾向や過去の変化レートからT数を補正する．

6．**PT数(Pattern T-number)の決定**　PT数とは「現在の雲パターンを過去の数多くの台風(すでに強度がわかっている）から作成した平均的な雲パターンと比較することにより，現在の強度を推定できる」との発想に基づく示数である．これは台風の発達パターンの模式図から，最も類似しているパターンを選択することにより決定できる．

7．**T数(T-number)の選択**　DT数，MET数，PT数の3個のT数から適切なものを採用する．最も優先度の高い示数はDT数である．

8．**最終T数**　ある時間内のT数の変化量に制限を加え，その変化量を越えたときの補正処理をおこなう．

9．**CI数(Current Intensity number)の決定**　熱帯低気圧が発達期にあるか衰退期にあるかでT数を補正する．このCI数が熱帯低気圧の強度（最大風速）と関連づけられている．なお中心気圧は，最大風速との変換表に基づいて変換する．

10．**FI数(Forecast Intensity number)の決定**　24時間後のCI数を予測する．

表 5 ドボラック法の処理手順[70].

7.3.5　ヒューリスティックな台風モデル

そこで次に，台風雲パターンを表現するモデルとしてどのようなモデルが使われているのかをまとめてみる．まず気象学分野において用いられる台風モデルに着目してみると，それらが意外とヒューリスティックなモデルであることに気づく．

まず数値予報システム上では，台風は3次元格子網上の圧力場などの形で暗黙的に表現される，ということになっている．しかし現実的には，中心付近に集中した構造を格子網上に自然に再現するのは難しいため，「台風ボーガス」と呼ばれる人為的な構造を格子網に埋め込む方法が用いられている[70]．この構造はほぼ円形の圧力場で表現される簡素なモデルであり，実在の台風の非対称性などの形状特徴はほとんど無視されている．したがってこのような形状特徴パラメータを設定することにより，台風予測の制度が向上する可能性がある．

一方，専門家の目視判断による台風解析手法であるドボラック法においても，ヒューリスティックな台風モデルが用いられている．この手法ではあらかじめ「典型的な」台風雲パターンを線画で描いておき，各パターンに対して台風の勢力を推定するために用いる経験的なルールを調べておく．そして台風解析の段階では，専門家はこれら線画と実際の雲パターンとを見比べることにより，実際の台風に最も類似した線画が属するクラスに分類，そして各クラスの経験的ルールを適用することによって，台風の勢力を推定するという流れになる．この手法の鍵となっている「典型的な」台風雲パターンは，専門家の長年の経験に基づくものではあるが，それを選択する数理的な根拠は乏しく，類似したパターンを選択する際にも主観的判断が入り込む余地がある．

7.3.6　確率統計的な台風モデル

このようなヒューリスティックな台風表現モデルに対し，本研究では確率統計的モデルを基本とする．このようなモデルは，気象理論のよう理論的基盤にはやや欠けるものの，少なくともデータセットの統計的性質という数理的な根拠に基づく結論を導き出すことができる．この方法で最も基本的なモデルは，台風雲パターンを状態空間（特徴空間・位相空間）の一点として表現する方法，そして大気運動の時間発展を特徴空間中の1本の軌跡として表現するモデルである．

このようなモデルで表現したとき，空間中の点の出現確率に偏りがあるか，または軌跡が遷移する確率に偏りがあるか，という問題は興味深い．というのも，これらの偏りをデータの分布から学習できれば，その結果を台風解析と台風予測の双方に有効活用できるからである．この問題に関連して，気象予報官たちは以下のような印象をもっているようである．すなわちこれまでの経験によると，大規模大気運動の時間発展には，普段より変化の遅い準定常期と変化の早い遷移期とがあり，さらに前者の中にはいくつか再帰性の高い流れ型がある，というのである[83]．このような天気図型は位相空間でデータ密度の極大領域として認識されるはずであり，その領域をレジームとよぶ．いくつかの天候レジーム間の遷移をマルコフ連鎖として統計的に調べてみると，起こりやすい遷移とそうでないものの間に有意な差があった．

以上の知見によれば，おそらく台風雲パターンにもデータ密度の極大領域が存在するだろうし，起こりやすい状態遷移もあるはずである．ゆえにそのような規則性を学習し時系列のモデリングに活用できれば，台風解析や予測に有用な知識を発見することができるかもしれない．例えばクラスタリングはそのようなレジームを発見するための一方法であるし，先述のドボラック法は，レジームの図式化およびレジーム間遷移の規則性を，長年の経験に基づき直観的に導出した手法とみなすことも可能だろう．

一方，データセットから統計的に根拠のある「成分」を抽出し，台風雲パターンを「成分」への射影の結

合として表現する方法も一般的なアプローチである．例えば顔認識など多くの応用で有効性が確認されている主成分分析（Principal Component Analysis: PCA)[1]を用いて，台風雲パターンを固有ベクトル「固有台風」の結合として表現する方法がある．PCA は台風データに特別に適した数学モデルではないものの，多くの確率統計的モデルの基礎となるものであり，またデータの次元を削減するための標準的な手法としても位置づけられる[65][84]．そこで本研究でも，まずPCAから得られる固有成分を用いた台風雲パターンの表現について研究する．

7.3.7　本研究のアプローチ

本研究のアプローチでは，モデリングの対象となるのは「データそのもの」，あるいはデータ生成過程(data generating processes)である．例えば7.3.4章で述べたように，エキスパートシステムなどの人工知能的アプローチでは，主に専門家の認知過程(cognitive processes)や意思決定メカニズムをモデル化することを意図していたが，気象という物理過程および専門家の認知過程の双方が大規模な複雑系であったため，結果的にはこのような意図は失敗に終わった．

それに比べ，本研究のアプローチでも専門家は依然として重要な存在ではあるが，それはデータについて何を見るかを知っている，いわば熟練した「教師」としての存在であり，それをモデル化の直接的な対象とするわけではない．一方で本研究のアプローチは，物理過程(physical processes)のモデル化に関心を注ぐ気象学的アプローチとも異なる．やや図式的に比較すれば，気象学においてデータは（気象）理論を補強・改善するための補助的存在であるのに対し，本研究のアプローチではあくまでデータが中心であり，（気象）理論はデータから学習した結果を裏付けるもの，という位置づけになろう．このようなデータ中心のアプローチの中で，本研究では「類似性に基づく推論」を基本に据える．この方法は以下の2ステップを基本構成とする．

1．データベース中から過去の類似事例を検索する．

2．過去の類似事例に基づき現在の事例についての推論をおこなう

この論理を台風に適用するには，台風という複雑な時空間パターンに適した類似性の定義，類似事例の検索，類似事例からのルールの抽出，類推に基づく予測，などが具体的な研究課題となる．そのためには，時空間パターンという構造化されていないデータを適切に構造化するための数理モデルを見つけ出し，そのモデルを用いて台風雲パターンを表現していく必要がある

このとき，統一的かつ一貫した基準のもとに収集された高品質のデータとして，大量の台風データが使える状態を実現する必要がある．本研究ではまずそのようなデータコレクションの構築から始めた．これが「台風画像コレクション」である．

[1]気象学では、主成分分析（PCA)よりも経験直交関数(Empirical Orthogonal Function : EOF)という名称が一般的に使われる[85]。

7.4 台風画像コレクションの概要

7.4.1　ベストトラック

本研究では，北半球および南半球で発生した台風[1]の衛星画像を統一的な基準で収集したデータベースである台風画像コレクションを研究基盤とする．ここではきちんと枠取りされたwell-framedな台風画像を収集すること，つまり台風画像と台風中心とを正確に位置合わせすることが重要であり，その基礎データとしての役割を果たすのがベストトラック(最終解析結果)と呼ばれるデータセットである．ベストトラックとは，一定時間ごとの台風の位置や強度を事後の入念な検討を経て決定した記録のことを指し，1951年以降に発生したすべての台風に関して，一定時間おきの中心位置や中心気圧・最大風速などを気象庁が編纂している．

台風中心位置の決定という問題は，一見容易そうに思えるものの実際は困難な問題である．熟練者がある時点で下した決定も，台風のライフサイクルが確定した後で振り返ってみると，不適切な決定であったことが判明する場合も多い．そこで台風のライフサイクル全体を見直せるようになった段階で，すべてのデータを再検討し改めて総合的な判断のもとに台風の経路を決定することが必要となるのである．ゆえにベストトラックは，台風観測後にリアルタイムで推定した結果ではなく，台風の一生に関するすべての記録と関係する気象データが得られた後で専門家が検討し決定した結果である．ゆえにベストトラックは，中心位置，大きさ，強さに関する「正解データ」，あるいは近似的な「グランドトルース」とみなすことができる，信頼性の高いデータである．衛星画像のみから中心位置を自動的に決定するのは困難であることを考えれば，この「グランドトルース」の存在は重要である．また，これだけの継続性と一貫性をもって記録がまとめられている気象現象は，他には存在しないと思われる．

本研究では日本の気象庁（Japan Meteorological Agency : JMA)による北半球（北西太平洋・南シナ海）のベストトラックに加えて，オーストラリアの気象局（Bureau of Meteorology : BOM)による南半球（南西太平洋・南東インド洋）のベストトラックも併用する．これによって南北両半球の台風画像を作成でき，地域ごとの台風の性質の違いも比較することができるようになる．ただし両者は台風（中心付近の風速が 17．2m/s 以上の熱帯低気圧）の記録だけではなく，前駆期にあたる熱帯低圧部の記録や，台風から衰えた温帯低気圧の記録なども含んでおり，各気象要素観測における基準や測定法も同一ではないため，両者をそのまま比較する場合には注意を要する．

7.4.2　台風画像の生成

気象衛星画像としては気象衛星「ひまわり」 GMS-5 を用いる．この気象衛星は赤道上空35790km に位置する衛星であり，地上の観測者からは東経140度赤道上に静止しているように見える．北半球ばかりでなく南半球を含む幅広い領域を観測できることや，観測頻度が1時間に1回と比較的高いことが特徴である．この気象衛星から台風領域を切り出し地図投影することでwell-framedな台風画像を生成する．地図投影法にはランベルト等積天頂図法を用い，がそれぞれ2500kmとなるような縮尺を用いた．

[1]南半球の熱帯低気圧は「台風」ではなく「サイクロン」と呼ばれるが、両者は本質的に同一の気象現象である。

ラグランジュ的表現

従来の台風画像データベースは，そのほとんどが地球に固定された座標系を用いて台風を表現していたため，台風そのものの変化よりはむしろ台風システム全体の移動に注目が集まる傾向があった．しかし本研究のように台風雲パターンの時系列的な変化を表現したい場合，むしろ台風雲パターンに固有の変化に関心がある．そこで台風システム全体の移動の影響を排除し台風雲パターン固有の動きを分離して眺めることができれば便利である．

そこで本研究で作成する台風画像コレクションでは，台風中心が時系列画像において不動点となるように，台風中心と投影画像中心とを常に一致させながら，台風周辺領域を衛星受信画像から切り出す方式を用いる．これはテレビの天気予報番組のように地球に対する固定的な座標から台風を眺めるオイラー的表現ではなく，台風中心と共に動く座標系から台風を眺めるラグランジュ的表現を用いることを意味する．このような表現によって，台風雲パターンの動きから台風システム全体の移動ベクトルを分離することが可能となる．その例を図 24に示す．ラグランジュ的表現では台風の移動は直感的にはわからないが，台風そのものの変化を記述するには適した座標系であることがわかる．もちろん台風そのものの移動を見たければ，移動ベクトルを別に地図にマッピングすればよい．

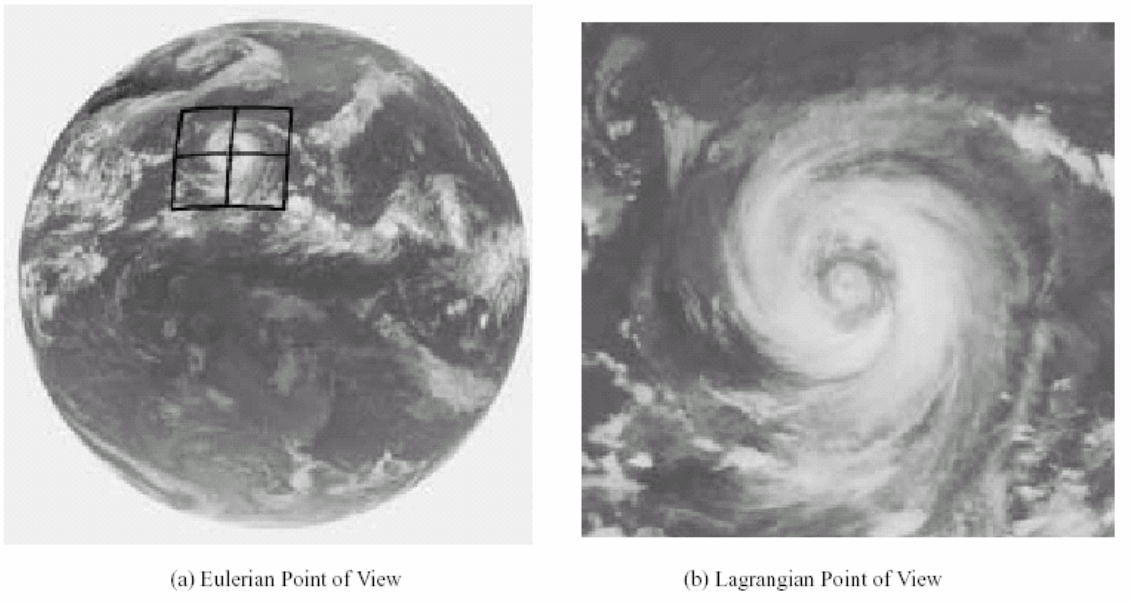


図 24　台風を眺める二つの視点である，(a)オイラー的視点．(b)ラグランジュ的視点．

地図投影法

台風画像の作成とは，衛星受信画像における台風周辺の雲パターンに対応する画素を，特定の地図投影法に基づいて地図投影画像の画像座標系に写像する処理のことを指す．このとき，台風雲パターンの時系列変化を適切に解析するためには，地図投影法に対して等積や等角などの好ましい性質を要求すべきである．そこで本節では地図投影法の性質や，画像の大きさなどのパラメータに関する検討をおこなう．

まず衛星受信画像側のパラメータとして，台風周辺域として取り出す範囲の大きさについて検討する．

その範囲は，台風の雲パターンを十分にカバーできるほどの大きさとする必要があるが，最も巨大な台風の大きさがおよそ直径2000kmであることを考慮し，画像中心を通る垂直線と水平線の長さが半径1250kmとなる範囲を台風周辺域と定めた．実際はこの範囲には収まりきらない台風が数年に1個出現するが，巨大な台風をすべてカバーする範囲を考えると，逆に小さな台風にとっては領域が大きくなりすぎてしまう．そのようなバランスも考え，本研究ではこの値を将来的には1300kmから1400km程度へと大きくする計画である．なお衛星受信画像の幾何補正については，S-VISSR（Stretched-VISSR）に含まれる衛星の軌道姿勢情報に基づく幾何補正のみを行っている．

次に投影画像側のパラメータを検討する．本研究で用いる地図投影法はランベルト等積天頂図法(Lambert azimuthal equal-area projection)とする．この図法には，地球球面上の原点からある角距離で囲んだ球面上の円形の表面積と，地図上の中心を原点として描いた円の面積が等しくなる，という性質がある．ゆえに本研究の用途には以下の利点がある．

1．台風中心と球面上の原点とを一致させることにより，台風の地理的な移動に関わらず地図上での台風の見かけの大きさを一定にできる．

2．歪みが投影画像中心からの半径にしたがって増加するため，台風のような円形に近い物体では比較的影響が小さい．

ここで，等緯度経度図法やポーラステレオ図法・ランベルト等角割円錐図法など，気象分野でよく用いられる地図投影法と，本研究で用いる地図投影法を比較する．

等緯度経度図法	緯度と経度が画像上で等間隔に直交し，任意の画素の緯度経度が計算しやすいという利点はあるものの，台風の移動に伴って投影画像上の見かけの大きさや形が大きく歪むため，異なる画像同士の比較が無意味となる．
ランベルト等角割円錐図法	数値予報の格子モデルとして用いられ，球面上と地図上の対応する点の近傍において任意の2方向の挟角が等しくなるという優れた性質(等角)を備えているが，本研究で関心がある等積性の性質は持たない．

特に等緯度経度図法では，台風が北に進んだときの形状の歪みがひどく，南北方向がつぶれた楕円形のような形状になってしまう．ゆえに地図投影法の選択は，台風の形状を正しく解析するためには不可欠の検討事項である．本研究では以上の考察からランベルト天頂等積図法を地図投影法として用いる．この投影法を用いて512×512画素の投影画像を作成する．この特定のサイズを選んだのは以下の理由による．

1．衛星直下点の赤外バンドの地上分解能5km/pixelを基準とすると，512画素幅が台風周辺域の範囲2500kmに相当すること．

2．画像の大きさが2のべき乗であると処理が効率的な場合があること．

7.4.3　画像分類

本研究の主題は台風雲パターンの解析であるため，以下の解析では衛星画像の観測画素値そのものではなく，画素ごとの雲の分類情報を扱いたい．そこで表 6に示すひまわりの赤外2チャネルと水蒸気1チャネルの画像データを組み合わせる雲分類アルゴリズムにより，雲画素の抽出および分類処理を前処理としておこなう．このアルゴリズムは気象庁で用いている雲分類アルゴリズムに独自の改良を加えた

ものである. その基本的な手順を以下に示す.

1. 各赤外画像を輝度温度に変換し, 輝度温度のチャネル間の差分も合わせて計算する.

2. 赤外の波長帯における水蒸気の吸収率の差から雲領域とそれ以外とを分類する.

3. 雲以外の画素から海面/地上温度をロバスト推定し, さらに標準大気の仮定に基づいて雲の輝度温度を雲の高度に換算し雲の種類を判別する.

表 6 気象衛星「ひまわり」搭載のセンサ VISSR の諸元.

観測チャネル	可視 (VIS)	赤外 1 (IR1)	赤外 2 (IR2)	水蒸気 (WV)
観測波長(μm)	0. 55-0. 9	10. 5-11. 5	11. 5-12. 5	6. 5-7. 0
画像サイズ(幅×高さ)	9216×9160	2560×2290	2560×2290	2560×2290
衛星直下点地上解像度(km)	1.25	5.0	5.0	5.0

7.4.4 台風画像コレクションの現状

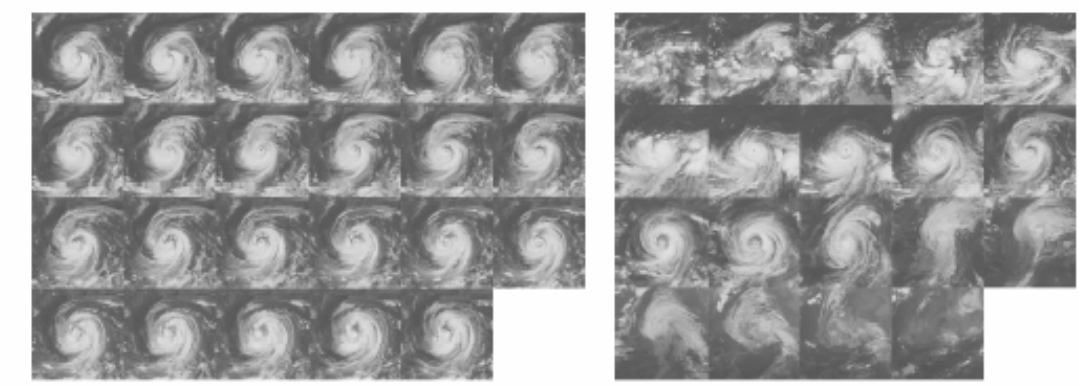
本研究で用いる台風画像コレクションの現状を表 7に示す. まず台風画像の数は北半球と南半球を合計して約34000件程度となっており, 件数自体はデータベースとして中規模であるものの, 時空間系列データとしては変化に富むデータセットであり, またデータマイニング研究に適するように一貫した基準のもとに前処理を施した, 高品質のデータセットである点が特徴である.

具体的に台風雲パターンの多様性を見るために, 図 25では台風雲パターンの変異度を, 複数の時間スケールで眺めた場合の例を示した. まず図(a)の方は1時間間隔で眺めたものであり, この場合は隣接する画像間の変化は小さい. これは台風雲パターンが, 全くランダムではなくある程度は連続的に変化すること, 言い換えれば連続する観測の相関はかなり高いことを示唆している. しかし図(b)のように24時間間隔で眺めれば, 隣接する画像においても台風雲パターンの変異はかなり大きく相関も小さくなっている. そして誕生から消滅までのライフサイクルを通してみると, 一生の間に大きな変化を示していることがわかる. ゆえに台風画像コレクションはパターンの時系列的な変動という意味で, 変化に富んだデータセットであり, このような柔軟な変化を数理的に表現可能なモデルの構築は, 画像認識・コンピュータビジョン分野における大きな課題の一つである.

さて本研究で構築する台風画像コレクションに類似のコレクションは, 我々が知る限り世界でも他に1つしかない. その一つとは, 米国NESDIS/CIRA (National Environmental Satellite Data and Information Service / Cooperative Institute for Research in the Atmosphere)が作成している「熱帯低気圧衛星画像アーカイブ」である[77]. こちらは大西洋および中央～東太平洋のハリケーンを対象としたもので, 本研究の台風画像コレクションの対象地域を, たまたま補完する形となっている(インド洋だけは両者の範囲に入らない). また主としてアーカイブ対象とする衛星の違いにより, 彼らの画像アーカイブにはより短い観測周期のデータも含まれている. 彼らの本職は気象学者であるが, 彼らもこのようなハリケーン画像アーカイブの有用性を, 気象学的立場から主張している. しかし基本的には, このプロジェクトは「気象学者による気象学者のための」プロジェクトであり, 本研究とは動機の面で大きく異なっている.

表 7　台風画像コレクションの現状.

	北半球	南半球
ベストトラック提供者	気象庁	オーストラリア気象局
緯度範囲	赤道より北	赤道より南
経度範囲	東経100度 -- 東経180度	東経90度 -- 東経170度
台風シーズン数	6	5
台風系列数	136	62
台風画像数	約24500	約9400
系列あたりの台風画像数	53－433	25－480



(a) 1997 年 8 月 15 日の台風 9713 号（1 時間ごと）　(b) 台風 9713 号の一生（24 時間ごと）

図 25　2種類の時間スケールで見た場合の台風雲パターンの変動.

7.5 台風画像コレクションに対するデータマイニング

7.5.1　固有台風表現

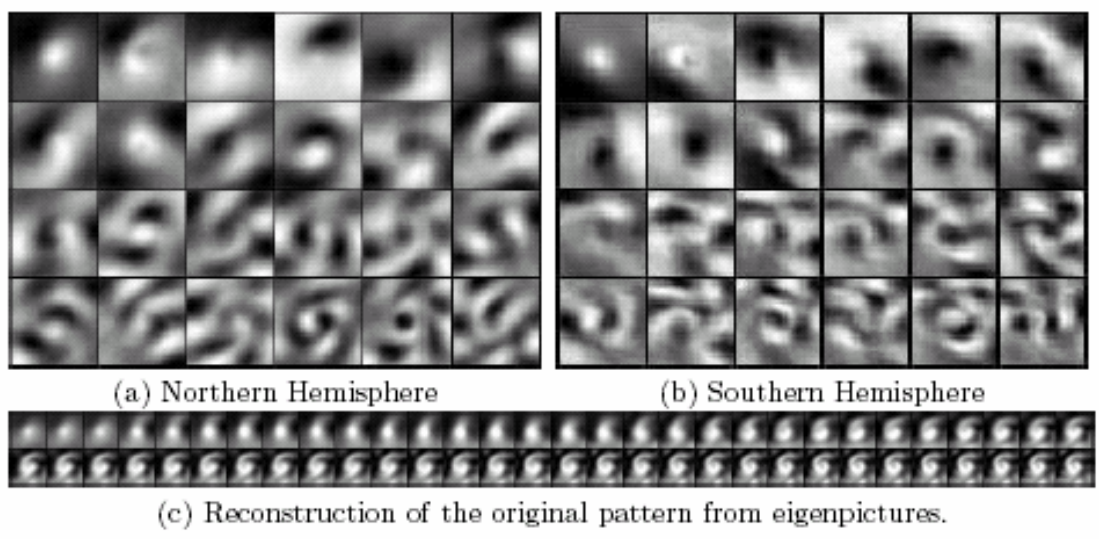


図 26　台風雲パターンの固有台風表現. (a)は北半球のデータ, (b)は南半球のデータ, そして(c)は任意の台風雲パターンが, 実際にこれらの固有ベクトルの線形和として表現できることを示している.

本研究では確率統計的な台風モデルを基本的なモデルとして用いるが，多くの確率統計的なモデルは，画素値配列あるいはそれに相当するような非常に低レベルの画像特徴量配列を特徴空間に射影し，高次元特徴空間における数学的操作に基づきパターン認識をおこなう方法を用いる場合が多い．このような方法は，画像の領域分割などの手法を用いずに，画像の画素値の分布そのものを問題とすることから，全体論的解析(holistic analysis)ともよばれる．これらの手法は，特徴に基づく解析(feature-based analysis)のように，対象とする問題領域に適した特徴抽出アルゴリズムの選択やロバスト性などに煩わされないことが少ないのが大きな利点である．また特徴空間への射影方法はその基準によって多様に考えることができ，たとえば判別分析(LFA)や独立成分解析(ICA) [60][61]などの手法がよく知られている．さらに画像特徴ベクトルを特徴空間内に射影した点が，物体の見かけの変化や時間的変化によって描く軌跡を解析することによって，特徴空間内での物体のパラメトリックな表現を得ることもできる．

本研究では台風雲パターンの最も基本的な表現法として，すでに顔画像認識などの分野で評価が確立している固有台風表現を用いる．そこでまず，このような固有台風表現を得るための計算手順について簡単にまとめておきたい．まず7.4.3節に述べた画像分類法を適用して，512×512画素の大きさの台風画像を画素ごとに雲（種類も含めて）と雲以外の海か陸に分類する．次にこの分類画像を32×32個のブロックに均等分割し，このブロック内の雲量を計算する．このとき個々のブロックは16×16画素の小ブロックとなり，ブロック内の雲画素の重みcの総和をとる．ここで重みとは雲の「重要度」を仮想的に考えたもので，例えば積乱雲は重み1（最大値）であるのに対し絹雲は重み0．5であるというように，それぞれの雲を台風の勢力決定に関する重要性に応じて区別して扱おうというものであるこのとき個々のブ

ロック内の加重雲量は $\frac{c}{16\times16}\in[0,1]$ と計算できる．よって，これらを画素値とする雲量画像を生成することで，元の台風画像から $32\times32=1024$ 画素の雲量画像を生成でき，これを台風雲パターンの特徴ベクトルとする．ここではブロック内の平均雲量を計算する操作が同時に次元削減にもなっている．

こうして計算した雲量画像の特徴を見るために，まず「平均台風」，つまりすべての台風画像を単純に平均した画像を図 26(a)および(b)の左上に示す．平均台風が示すように，台風雲パターンは中心付近に平均値が高い（雲量が平均的に大きい）コア領域があり，その他の部分は平均としてはそれほど高くない平均雲量となっていることがわかる．また北半球では北西側，南半球では南西側の平均雲量が小さい点が特徴的であり，これは対流圏の大気の大きな流れの反映であると考えられる．また平均台風の右横の画像は「分散台風」，すなわち雲量の分散を計算したものである．

次に主成分分析の結果として得られる固有台風画像についても図 26に示す．ここで固有値問題の計算には行列計算ライブラリのCLAPACKおよびATLASを用いた[88]．図は固有ベクトルに対応する固有値が大きい順に並べ替えたものであり，上段の第1固有台風から下段右隅の第22固有台風までを図示している．第1固有台風から第3固有台風までは主に南北方向の雲量傾度を表現しており，また曲率をもった雲のパターンも現われている．それに対して以下の第8主成分程度までは主に台風中心付近の雲塊を表現していることが読み取れる．さらに固有値の小さな主成分は主にスパイラルバンドのパターンを表現しており，固有値の現象に伴ってスパイラルバンドの分岐や数が増加する傾向があることがわかる．

また北半球と南半球の台風はおおむね同様の傾向を示しており，このことはこれらが同一の気象現象であることを示唆していると考える．ただしここで同一とは，パターンの南北（画像では上下）を反転させた上での同一性である．このように南北方向に反転させる必要があるのは，赤道をはさんで温度傾度やコリオリ力の向きが反転することが原因である．

そして固有台風は，もちろん台風画像コレクションの部分集合（例えば強い台風のみ）に対しても計算することができる．例えば強い台風（カテゴリ5）のみを対象に固有台風を計算すると，台風の眼の部分の微細構造がより強く出現する，などの興味深い成果が出ている．

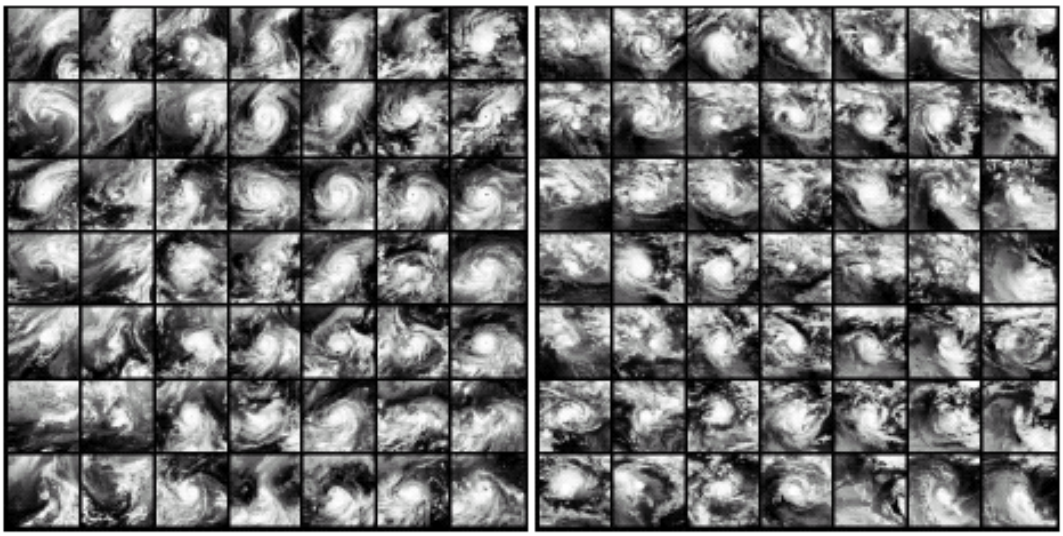
最後に主成分分析による次元削減について述べる．主成分分析で得られる固有ベクトルのうち，固有値が大きい順に固有ベクトルを選び，それらのベクトルが張る固有空間で台風雲パターンを表現することにより，特徴ベクトルの次元を2乗近似の意味で効率的に削減することができる．本研究では累積固有値が85%となるような次元数を選び，この固有空間に特徴ベクトルを射影することで台風雲パターンを表現する方法を以下では用いる．その具体的な次元数は，北半球台風が1024次元から71次元へ，南半球台風が1024次元から79次元となった．このように次元数を約15分の1にしても，およそ85%の信号成分は表現できている．この近似のよさが必ずしも分類のよさに直結するものではない点には注意すべきであるが，いずれにしろこのような次元削減は高次元特徴ベクトルに対して有効な前処理の一つである．

7.5.2　クラスタリング

K-平均クラスタリング

上記の固有台風では台風を表現する基本的な「成分」を抽出したが，今度はむしろパターンとして「典

型的な」ものを抽出することを考える．これは台風雲パターンの多様性を少数のパターンで代表させることに相当し，そのために必要なのは類似したパターン同士を結びつけグルーピングするような処理，すなわちグラスタリング処理をすることになる．クラスタリングのための手法はすでに多数提案されており，ここですべてをまとめて述べることは不可能であるが，本研究ではまず最も基本的な方法として，K-平均クラスタリング[65]を台風雲パターンに対して適用してみた．



(a) Northern hemisphere image collection (b) Southern hemisphere image collection

図 27 K-平均クラスタリングによる台風雲パターンのクラスタリング．(a)が北半球，(b)が南半球を表している．クラスタリング後にサモンのマップ化により大まかにパターンを整列させている．

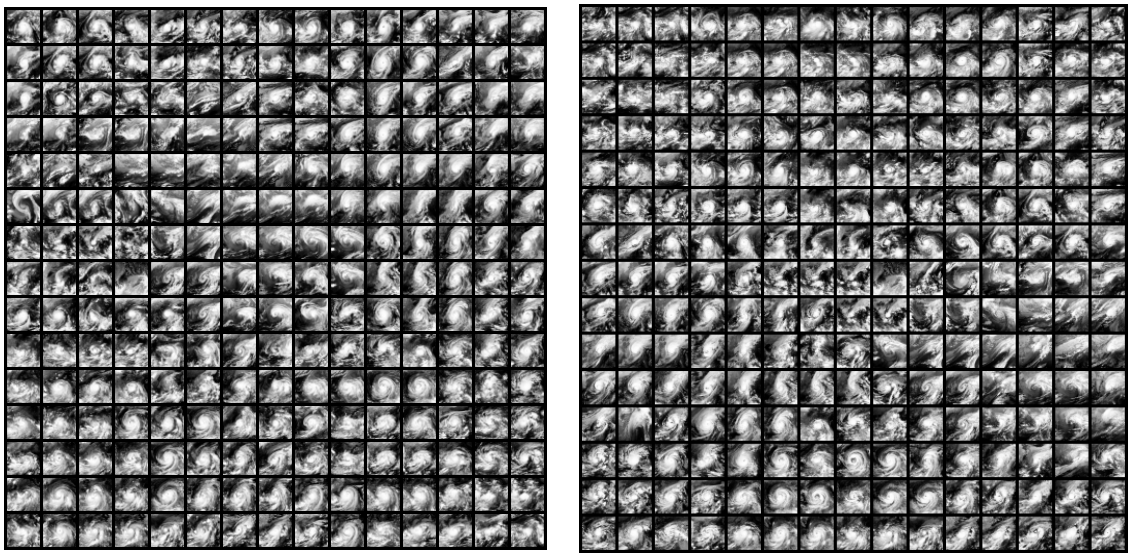
その結果が図 27である．これは北半球および南半球それぞれの台風画像コレクションに対してK-平均クラスタリングを適用し，49個のクラスタを抽出した結果である．その方法は以下の通りである．各画像の特徴ベクトルとしては，PCAで次元削減した特徴ベクトルを用い，また各クラスタの代表画像としてはクラスタ重心に最も近い画像を選ぶ．さらにクラスタリングの後処理として，多次元尺度構成法の一手法であるサモンのマップ化(Sammon's mapping)[90]をクラスタ代表点に適用し，クラスタ代表点を2次元平面上に位相的に配置する．つまり2次元配列上で隣接する画像は代表点どうしの距離も小さい．

また図 27は台風雲パターンの多様性を49個のパターンで代表したものと考えることができる．確かに渦巻状の雲パターンもあれば，東北方向に伸びた雲パターン，または形の定まらない雲パターンも現れている．しかしこれらはPCAで得た特徴ベクトルに対しユークリッド距離を類似尺度と定義した場合の代表パターンであり，異なる特徴ベクトルや異なる類似尺度を導入すれば異なる代表パターンが得られるという性質のものである．ゆえに，これらが本当に台風の雲パターンを代表するものと即断はできない．

その意味で，ドボラック法における代表的な雲パターンとは，専門家の長年の観察経験から形成されてきた抽象的なクラスタを具体的に表現するもの，と考えることができる．このときクラスタを形成する過程では，おそらく特徴ベクトルとしても類似尺度としても，非線形なアルゴリズムが作用しているはずである．我々が目標とするのはそのような代表的な雲パターンを，データマイニング的手法により導出することである．またそのような代表的パターンの集合にみられる規則性や，代表的パターンの間を遷移する際の規則性などをモデル化することにある．

自己組織化マップ

このようなクラスタリングの一種に自己組織化マップ(Self Organizing Map: SOM)[89]がある．自己組織化マップは，クラスタ代表点が2次元多様体上に拘束されたK-平均クラスタリングとみなすこともでき，クラスタ代表点の間に位相的構造が導入されるため，パラメータ学習後に得られるクラスタ代表点マップの上では，データはある種の順序を保ちながら並ぶことになる．また自己組織化マップは，多次元特徴空間を，2次元平面に非線形射影しているとみなすこともでき，このとき2次元多様体は多次元特徴空間のデータ分布を近似していることに相当する．いずれにしろ，隣接するニューロンド間では位相が導入されると自由に動けなくなるため，ただし多次元特徴空間の近似という意味では自己組織化マップはK-平均クラスタリングよりも劣ることになる．



(a) 自己組織化マップ（平面的位相）　　(b) 自己組織化マップ（トーラス的位相）

図 28　自己組織化マップによる北半球台風画像コレクションの2次元平面配置およびクラスタリング．(a)と(b)では位相構造が異なるが他のパラメータは同一である．

図 28は自己組織化マップによる北半球台風画像コレクションの2次元平面配置である．こちらは15×15のニューロンを用意し，近傍を適当に縮小しながら代表点の座標を学習した．図では(a)と(b)とでは使用した位相構造が異なり，前者は平面的な位相であるのに対し，後者では平面の右側（上側）と左側（下側）が位相的に接続しているので平面の端は定義できない．つまり高次元特徴空間中のデータの分布をトーラスの位相構造をもつ多様体で近似していることになる．

これらの結果は図 27に比べてより位相構造が強く浮きでた平面配置となっており，隣接した画像間では台風雲パターンが連続的に変化している．しかし位相的制約が強すぎるという印象も否めず，かえって台風雲パターンの多様性が図から失われているような印象さえある．例えばこの例の場合，代表点による近似誤差を平均2乗誤差で測定してみると，(a)の場合が41.2，(b)の場合が40.5なのに対し，同じノード数を用いたK-平均クラスタリング（結果の図は省略）では29.0であった．つまり平均2乗誤差が自己組織化マップでは4割ほど増加しているが，その原因は位相的制約の強さにある．しかしK-平均クラスタリングの場合でも緩やかな位相は多次元尺度構成法によって実現できることを考えれば，自己組織化マップの導入にはそれほど大きな利点はないようである．

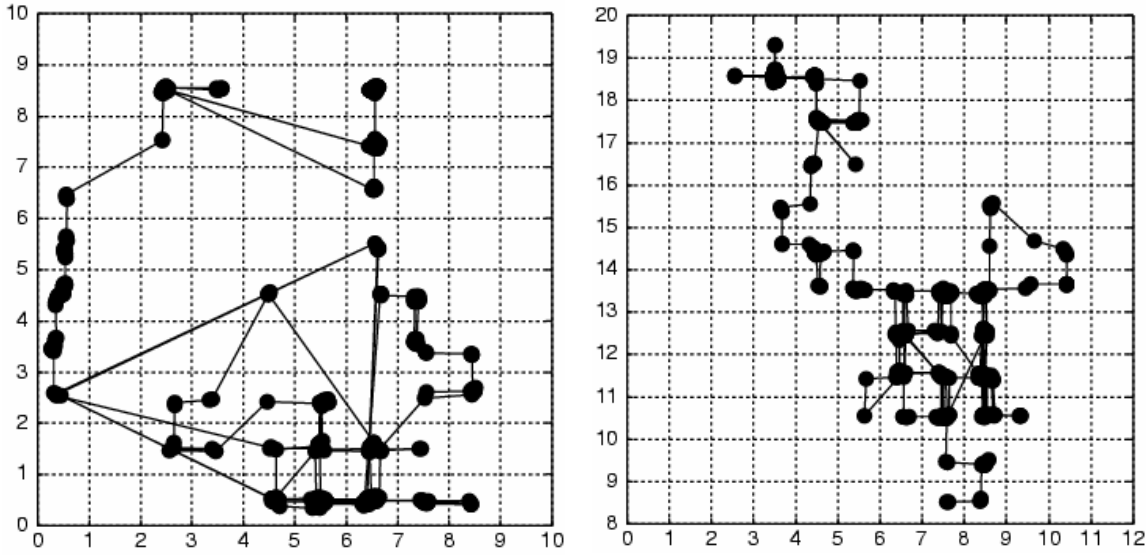
また自己組織化マップのもう一つの弱点は，確率論的枠組みを導入することが難しいという点にある．

例えば隠れ変数モデルなどを導入し，自己組織化マップの上で状態遷移の確率分布などを扱うことが直接的にはできない．このことも自己組織化マップの一つの限界を示しているといえる．

状態遷移マップ

このようにクラスタリングをおこなう理由は，一つには台風の典型的なパターンを発見してそれをリストアップすることにより，未知台風を過去の類似台風と関連付けて理解し，その勢力を推定するための規則を導く，という点にあった．もう一つの理由は，いったんこのように特徴空間を離散化し状態を作り出すことで，台風画像を時系列信号とみなしたときの状態間の遷移から規則性を発見する，という点にある．もちろん連続空間上でそのような性質が導き出せればよいのだが，今回の場合は，特徴空間のサイズに比べて標本数がそれほど多くないことを考えると，離散的にすることでモデルの複雑さを減らし，規則性を見つけやすくすることを狙うのが現実的な選択肢であると考える．

現在のところこのような状態遷移を隠れマルコフモデル(Hidden Markov Model)でモデル化する実験を進めているが，本報告ではその結果には触れず，状態遷移がどのような性質を持つかを図に示して説明したい．図 29は状態遷移マップ，つまり自己組織化マップのニューロン上で，ある台風のライフサイクルが示した状態遷移の挙動を示すものである．ここでは自己組織化マップの二つの位相構造に対して台風のライフサイクルがどのような挙動を示すかを比較したが，この場合では(a)では状態間に不自然な跳躍が見られるのに対して，(b)では不自然な跳躍が起きておらず，その意味で後者のほうが自然な結果が得られたと評価できる．なぜなら，台風の雲パターンは1時間単位ではほぼ連続的に変化すると仮定できるため，位相的なマップ上では距離的に近い隣接するニューロン間で跳躍が起こることが自然なためである．



(a) 自己組織化マップ（平面的位相）　　(b) 自己組織化マップ（トーラス的位相）

図 29 状態遷移マップ．自己組織化マップで学習したニューロン上で，ある台風のライフサイクルが示した状態遷移の挙動．なおここで用いた学習結果に対応する画像はここに載せていない．

上記の結果は単に台風系列を自己組織化マップ上にプロットしただけであるが，今後はこのような状態遷移に確率的モデルを対応させ，データからの遷移確率の学習を通して台風の状態変化の規則性あ

るいは不規則性を捉えることを狙っている. そのような研究プログラムを進めるためには, 自己組織化マップは確率的構造を導入しにくいという点に欠点がある. 我々は他の確率モデルに状態遷移モデルを組み込むことで, このような遷移確率の学習を進めていく計画である.

7.6 予兆発見にむけて

7.6.1 予兆発見のテーマ

表 8 予兆発見として重要と考えられる研究テーマと, それに対する気象学者の意見 (2001年9月現在).

	予兆発見の研究テーマ	気象学者の意見
発生予測	熱帯収束帯 (InterTropical Convergence Zone: ITCZ)の積乱雲塊の中から, 台風に発達しそうな積乱雲群の予兆を発見する.	台風発生の予兆発見は, 気象学でもメカニズムに未解明の部分が多いので, 興味深いテーマだと考える. またもし予兆を発見できるなら, その発見は段階が早ければ早いほど望ましい.
強度予測	短時間の間に急速に発達しそうな台風の予兆を発見する.	雲パターンの予測については, たとえ経験豊富な予報官であっても6時間後の状態でさえ想像がつかない場合も多く, このような方法で予兆が発見できるか疑問である.
進路予測	進路が定まらない迷走台風について, 今後の進路の予兆を発見する.	迷走台風の進路に対する, 雲パターンの非対称性など台風固有の原因はあまり寄与していないのではないか. やはり大きなスケールの大気の流れが主要因ではないか.

予兆発見には種々の定義があるが, 本研究では比較的まれではあるが重要な事象の発生を予測する技術, という定義を用いる. このような技術は, すべてのデータに関して平均的に予測能力を向上させる研究からは生まれず, むしろデータの規則性と不規則性を明確に意識する研究から生まれると考えている. あるいは. 損失関数に大きな非対称性がある場合(発生は稀ではあるが大きな損失が生じるような事象に対する損失関数)の最小化という問題に通じるかもしれない. このような予兆発見技術を台風の問題に適用することを考える[29]. そのためにはまず, 情報学的アプローチが貢献できそうな具体的テーマをあらかじめ見極めておきたい.

例えば, 台風予測に関しては, 気象学では特に進路予測を中心にした研究が進んでおり, その精度は着実な向上を見せている. それに加え, 台風は大きなスケールの大気の流れに乗る形で移動する場合がほとんどであるが, そのような大気の流れの計算は数値予報が得意とする分野である. したがって, 気象学的手法が有効と思われる進路予測などの問題では, 情報学的アプローチの貢献は限定的なものにとどまる可能性がある. その一方で, 気象学の枠内では扱いにくい問題もある. 前述の「台風雲パターンの認識」といった問題も, その代表的な例である. それに加え, 台風予測が難しい場合がいくつか知られているが, このような問題こそ, 気象学的アプローチが困難に直面しており, むしろ新しいアイデアを試すのにも適した問題なのではないかと考える.

そこで表 8には，現状では予測の困難な3種類の問題を取り上げ，それに対する気象学者の意見をまとめた．この中では，特に「台風発生の予兆発見」というテーマに大きな関心がありそうで，このテーマは今後の予兆発見研究において一つの有望な方向であると考えることができるだろう．しかしこの研究には，データ収集などに現状以上の計算資源を必要とするため，すぐに研究を開始することはできなかった．そこで本研究では2番目の研究課題である「台風の急速発達現象」という問題を取り上げる．この問題は文献[77]でも，画像コレクションに基づく台風研究の重要な応用の一つに記されている問題である．本報告では，その予備的実験結果について述べる．

7.6.2　予測可能性の問題

ここで予兆発見の問題に移る前に，気象現象において本質的な問題である予測可能性の問題に触れておきたい．これは端的には，長期予測不能性(long-term unpredictability)と短期予測可能性(short-term predictability)の問題を指す．大気力学系の非線形性により，たとえ大気が決定論的なシステムであっても長期的な予測は不可能であるというのが問題の本質であるが，一方で天気予報の成功に見られるように短期的な予測は十分に可能であり，したがってその「短期」が具体的にどのくらいの期間なのか，が議論の焦点となる．図 30でこの性質を検証してみた．

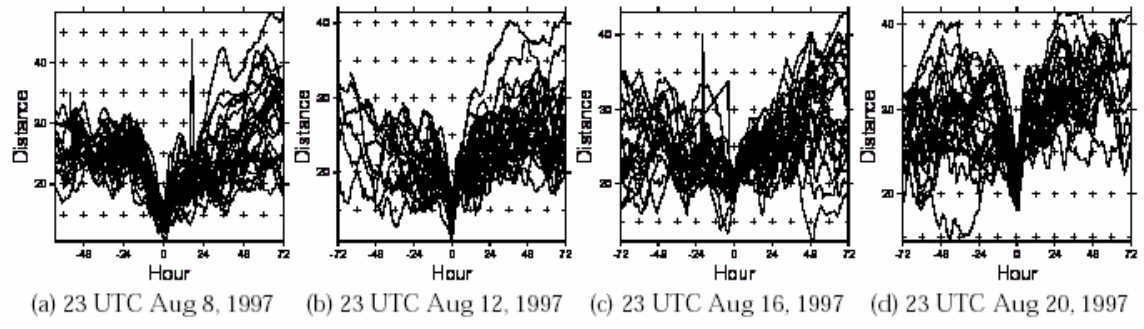


図 30 台風199713 号の予測可能性．(a) は発生期，(b) は発達期，(c) は最盛期，(d) は消滅期に対応する．

これは，時間ゼロの時点で，ある台風雲パターンに対する類似事例をデータベース中から検索し，これら類似事例の組が，その後の時間発展においてどのように類似性を失っていくのかを示した図である．いずれのグラフでも，基準とした台風199713号の事例とその他の台風系列の事例との距離が，時間の経過に伴って急速に拡大している．このことは，ある時点では類似事例であっても，両者のその後の時間発展は大きく異なることを示している．特にこの例の場合は，12時間以内での距離の増加割合が大きく，類似関係は非常に短時間しか成立していない可能性もある．なおこのグラフの平均的な傾きは，おおよそ最大リヤプノフ指数に対応する[86]．

このような性質は，類似性に基づく推論に対しては大きな障害として働く．たとえ過去の類似事例が発見できたとしても，両者のその後の時間発展は大きく異なりすぐに類似性を失う，というのがこの主張の帰結だからである．実際のところ，気象学者でカオスの発見者の一人でもあるE． Lorenzは，類似大気パターンに基づく気象予測である類推法(Method of Analogues)を30年以上前に提案した[87]．彼は2つの気象パターンの類似度（誤差）が時間とともに変化する様子に着目し，そのための研究対象として「高層大気の気圧（高度）分布」という気象パターンを選んだ．その結果，誤差が2倍に膨らむ時間は2．5日から8日であり，この数値がおおよそ大気の予測可能性に対応すること，また本当に類似したパターンを

過去のデータから見つけ出すのは困難なことを指摘した．このような予測限界が生じる本質的な理由は大気力学系の非線形性にあり，これこそ Lorenz が発見したカオスである．

起こりやすいパターンがあるとの主張と，過去には類似事例は見付からないとの主張と，相反するように見えるこれらの主張は，実は表現モデルに関わる問題ではないか，と我々は考えている．大気は無限自由度のシステムであるため，単純に考えれば類似事例が出現することはまず有り得ない．しかし表現モデルの工夫により適切な次元圧縮が可能となれば，次元の呪いの影響を避け，概念的に類似した事例を類似事例と認識できるようになる．つまり，変数の数を減らして少数の巨視的変数で記述することが予測能力を減少させるとは限らず，小さな内部構造にとらわれない巨視的な構成要素間の関係に基礎をおいた時空間構造の簡略な記述が重要となる．

7.6.3　急速発達現象の予兆発見

予兆発見というゴールへ向けての具体例として，本研究では台風の急速発達現象(rapid deepening, explosive intensification)について考察する．この現象が重要なのは，この現象の発生に伴って事前予測よりもはるかに強い風が台風周辺で吹くことになれば，船舶の遭難や避難の遅れなどに直結する危険な状況を引き起こす，という点にある．このような急速発達現象の定義としては以下のものが代表的であろう．

1．24時間で中心気圧が42hPa以上（または30hPa以上）低下する．

2．12時間で中心気圧が30hPa以上低下する．

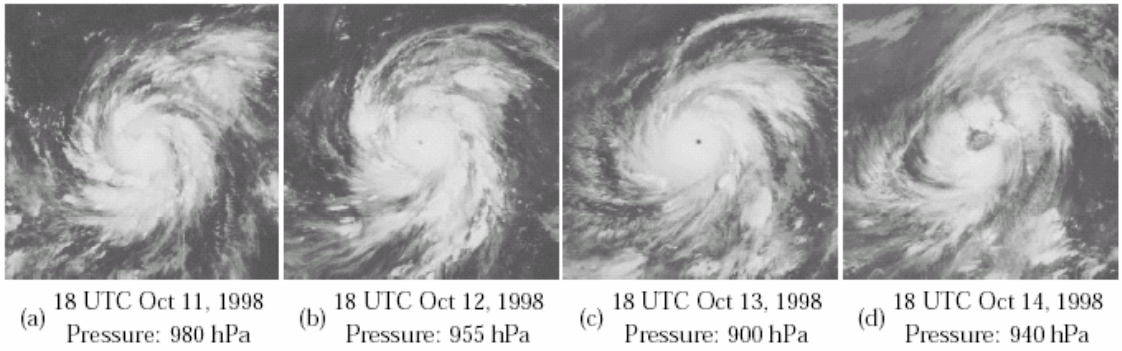


図 31 台風の急速発達の例. 18 UTC October 12, 1998から18 UTC October 13, 1998の24時間に中心気圧が55hPa低下した.

図 31は北半球台風画像コレクション中で最も急速な発達を示した事例であり，24時間で中心気圧は55hPa低下している．同様の急速発達現象は，北半球で136系列中4系列，南半球では62系列中3系列に発生していた（42hPa / 24hourの定義による）．すなわち急速発達現象とは，平均として5パーセント前後の台風系列に発生する，比較的稀な事象となる．

このような急速な発達は数値予報モデルでも予測できない場合が多い．その際に，特定の雲領域の拡大などに着目すれば，急速発達現象の予兆発見に有用であるとの研究も報告されているが，決定的な方法はまだ発見されていない．図 31の場合には，(b)あるいは(a)の雲パターンから，(c)のような急速な発達が起こりうることを察知することが課題となるが，これらの画像を実際に眺めれば，その困難さがうか

がえる.

我々もまず，特徴空間でそのような事象の集合が示す分布の可視化から始めた．そして図 32は，24時間の中心気圧変化$|\Delta p|$が10hPa以上の事例を，主成分で構成される空間にプロットしたものである．対角線上側が気圧低下の事例（台風は発達），下側が気圧上昇の事例（台風は衰弱）に対応し，大きな円に対応する事例ほど急速な発達（衰弱）を示す．これらの図では発達事例と衰弱事例の分布は大きく重なっているようだが，どちらかに特有な領域も図には現れている．このようにして，急速な発達に対応する雲パターンの特徴を学習していくことが，今後の研究課題である．

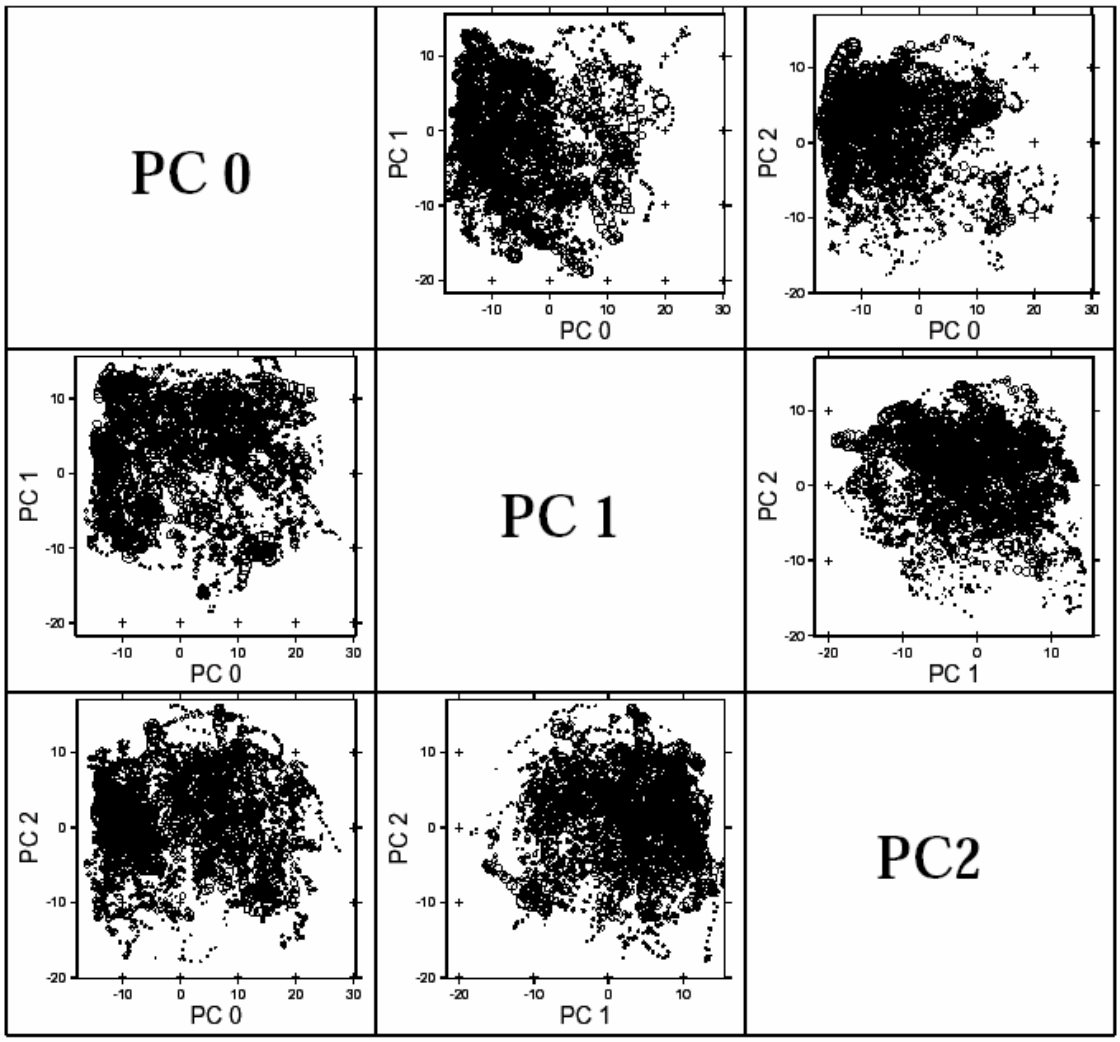


図 32　24時間の中心気圧変化$|\Delta p|$が10hPa以上の事例を，主成分(principal component：PC)が張る空間上へプロットした図．対角線の上側が気圧低下の例，対角線の下側が気圧上昇の事例であり，各円の面積は$|\Delta p|^{1.5}$に比例する．対角線をはさんで対応するグラフを比較することにより，特に気圧低下が顕著な領域を検出する．

7.6.4　予兆発見の可能性

本研究の結果はまだ予備的な段階であるが，大規模データからの学習という方向性は，情報学的アプローチのセールスポイントになると我々は考えている．確かに気象学には確固とした理論体系が備わっているが，現実の台風雲パターンに偏りがあるかどうか，などの疑問に対する解答を，理論から直接的に導き出すことが困難なのもまた確かである．また気象衛星データも，観測開始からすでに20年以上を経過して膨大な量のデータが既に蓄積されており，これらを人手で整理し検証していくことは，現実的には不可能に近い．よってそこには，現実世界に関する情報を集約し，学習し，提示することのできる情報学的手法が活躍する余地があると考える．特に予兆発見といった挑戦的課題では，そのメカニズムにも未知の部分が大きいため，単に気象現象を統計的に記述するにとどまらない，学習モデルの価値は大きいはずである．

最後に学習方法に関して言及しておきたい．台風に関しては気象現象としてはおそらく最も充実した，気圧・風速等のメタデータ（ベストトラック）が作成されている．しかもこれらは（気象庁の）専門家によって精査されたメタデータであるため，これらのメタデータを正解データと仮定することにより，教師あり学習の枠組を台風画像に対して適用できる．ただしここで注意すべき点は，専門家の精査を経ているとはいえこれらはあくまで推定値であり，決して真値ではないという点である．このことは，同一の台風に関するメタデータでも，各国の気象機関によって推定値にばらつきが生じることからも明らかである．

教師あり学習は，このようなメタデータを所与の真値として信用する立場，一方の教師なし学習は，モデルの学習結果からメタデータを修正するという目的も含む立場とみなすことができよう．あるいはメタデータを参考値として教師あり学習を適用し，その結果に応じてメタデータを修正し，再度教師あり学習を適用するという，データとモデルの相互作用に基づく立場もありうる．単に教師つき学習を適用する，あるいは気象学的手法を模擬するだけでは不十分であり，現在のメタデータを真値に近付けるという方向性も視野に含めるべきだ，というのが我々の主張である．

最後に，台風に関する研究は，それによって引き起こされる災害の防止や軽減に結びつけるというのが最終的な成果となるため，この方面に関する研究が重要となる．システムとしての視点で考えれば，おそらく最も重要となるのは地理情報システム(GIS)との連携であり，それによって降水・高潮など，それぞれの地域に特有の災害を防止することに威力を発揮できるだろう．一方人々への予報の伝達という視点で考えれば，確率論的台風表現モデルの学習結果は，現在気象予報の主流となりつつある，降水予測に代表される確率的表現[91]になじみやすい，という利点がある．こうして，例えば「現在の台風が1週間以内に非常に強くなる確率は20%です」といったように，現在では不可能な多様な種類の予報を実現することが，さまざまな視点からの意思決定を要求される人々に対する有用な情報源になりうると考えている．

第8章 IMET: Image Mining Environment for the Typhoon

8.1 はじめに

本研究では台風画像データマイニングのためのシステムIMET(Image Mining Environment for the Typhoon)を構築した. 本章ではこのシステムの概要を説明する. このシステムは, ネットワーク上に分散するデータベースサーバおよびそれらを束ねるメタサーバ, という構成になっており, 利用者はメタサーバにWWWクライアントからアクセスすることになる. すなわちこれは図 7の構成を踏襲したものであり, ネットワーク上で多種類のデータベースを動作させる環境へも容易に拡張することが可能である.

ネットワーク上のデータベースエンジンとしては, 従来型の関係データベースエンジン[1]に加え, 本研究では特徴空間探索エンジン(Feature Space Explorer : FSE)を独自に構築した. このエンジンは特に画像検索という用途を考慮して構築された画像検索エンジンであり, 特徴空間における種々のベクトル演算や, クラスタリングも含む広義のグルーピング操作をサポートし, また第4章で述べた検索言語による問合せを受け付けることが可能なシステムである. 一般に関係データベースは性能の面で問題が生じることが多いが, この特徴空間探索エンジンは画像特徴ベクトルの構造を考慮した設計となっているため, 類似画像検索などの処理をより高速に実行することができる.

本章ではこれらのデータベースエンジンに問合せを送信し, 結果をまとめて表示するメタサーバの役割, 特にメタサーバが提供するビュー(利用者視野)という視点[32]から, このデータマイニングシステムについて説明していきたい. ここでビューを用いて説明するのは, ビューが最終的には利用者へのインタフェースとなることが理由である. ビューの背後で動作するメタサーバ, および問合せ言語については既に第4章で説明済みである. そこで残された部分である利用者インタフェースの部分, つまりビューから眺めていくのが自然であると考える.

なお以下で述べるビューは, メタサーバが検索応答をまとめたXML文書を生成し, その文書をXSLTによってHTMLに変換する方法で, クライアントに送信している. これらのビューはWWWクライアントを経由して外部からもアクセス可能であり, 「デジタル台風」という名称でhttp://www. digital-typhoon. org/ において公開済みである. このサイトの背後では上に述べたバックエンドデータベースサーバが稼動し, メタサーバからの問合せに応じている.

[1] 本研究では関係データベースエンジンとしてPostgresQLを用いている。

8.2 特徴空間探索エンジン（FSE）

先述のように本研究では特徴空間探索エンジン(Feature Space Explorer : FSE)を独自に構築した．このエンジンは特に画像検索という用途を考慮して構築された画像検索エンジンであり，特徴空間における種々のベクトル演算や，クラスタリングも含む広義のグルーピング操作をサポートし，また第4章で述べた検索言語による問合せを受け付けることが可能なシステムである．現在のところは特徴ベクトル型の画像特徴のみをサポートしているが，将来的にはこのように画像表現に依存する部分を「プラグイン」するだけで，他の画像表現も取り入れることができるような仕組みにすることを考えている．

この特徴空間探索エンジンは，現在のところデータ件数がたかだか数万件程度であることを前提としているため，検索インデックス構造などにはほとんど手間をかけていない．というのも，台風画像データベースの場合，そもそも画像を無制限に増やすことが不可能であり，当面のところは件数が数万件程度にとどまることが確実なためである．したがって，数万件程度であれば1回の問合せには実用的な速度（例えば上位10件の類似検索等であればミリ秒単位）で終了するという性能を達成している．

このような高速性は，実は検索インデックスをメインメモリ上に展開することで達成しているが，現代のサーバ機のスペックを考えれば，数万件程度のそれほど大きくない検索インデックスをメインメモリ上に展開することは，十分現実的な選択肢になると考えてよい．またこのようにメモリ上に展開することで，ディスクアクセス等のI/Oボトルネックを回避し，容易に効率を向上させることができる．この方法は明らかに数十万件以上の大規模データにはそのまま適用できないが，地球環境データの増加ペースとハードウェアの性能向上ペースを考えれば，少なくとも台風画像データベースに限れば，現在の方式は将来にわたってもそのまま活用できそうである．

8.3 デジタル台風

先述のように，このデータマイニングシステムIMETはWWWで公開しているものである．そのトップページを図 33に示す．この図にあるように，本研究プロジェクトの愛称は「デジタル台風」であり，この命名はデジタルアース計画[93]に触発されたものである．本研究では台風に関するデータをすべて地理的時間的に結合して統一的な情報基盤上に展開していく計画であり，将来的には情報基盤上で統合した様々なデータの関連性を可視化し利用者に提示することで，利用者は情報環境の中を自由自在に歩き回りながら，台風あるいは地球に関する新しい発見ができるような情報環境を確立したいと考えている．このような環境が実現するのはかなり将来であるが，本研究はその第一歩であり，将来に向けて地球環境データの可視化についても常に気を配りながら研究を進めてきた．

図 33　デジタル台風のトップページ.

8.4 ビュー

8.4.1 単一観測ビュー

まず，単一観測とは個々の観測を指すものであり，気象衛星ひまわり画像から生成する台風画像と，その時刻に対応するベストトラックデータ，すなわち，台風中心位置，中心気圧，最大風速などから構成される．単一観測ビューの目的は，まず個別のデータの特徴を知るとともに，台風の雲パターンがどのようなパターンであるのか，といったように主に空間的な観点からデータを眺めることにある．

単一観測ビューの例を図 36に示す．これはある台風系列に対してある観測時刻に得られた種々のデータを統合して示すためのビューである．これは，観測時刻(UTC)と台風名をキーとして，様々なデータを結合したビューであると考えることもできる．ここに示すのは以下のような情報である．

1. ベストトラックデータ

2. 全体のひまわり画像のどの部分を切り取り台風画像としたかを示す画像

3. ひまわり画像で処理に用いた3種類の画像，すなわち赤外1チャネル(IR1)，赤外2チャネル(IR2)，それから近赤外水蒸気(WV)の各画像

4. 上記3種類の画像を処理して得た雲分類画像

5. 特徴ベクトルとして用いる固有台風画像

6. この単一観測を問合せとする高度なデータマイニング操作へのリンク

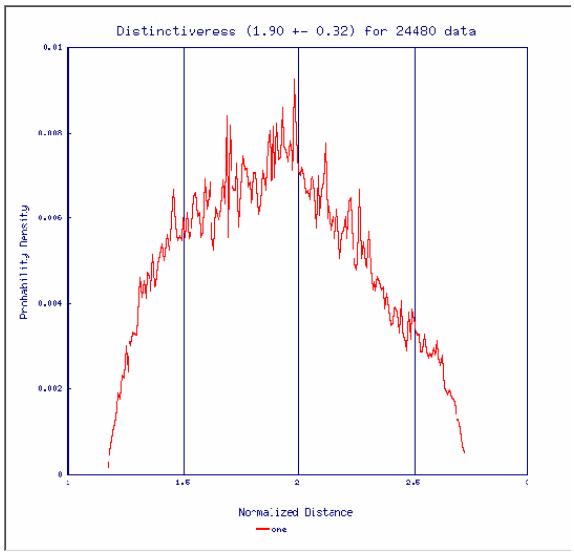


図 34 Distinctiveness を計算するための距離分布の図.

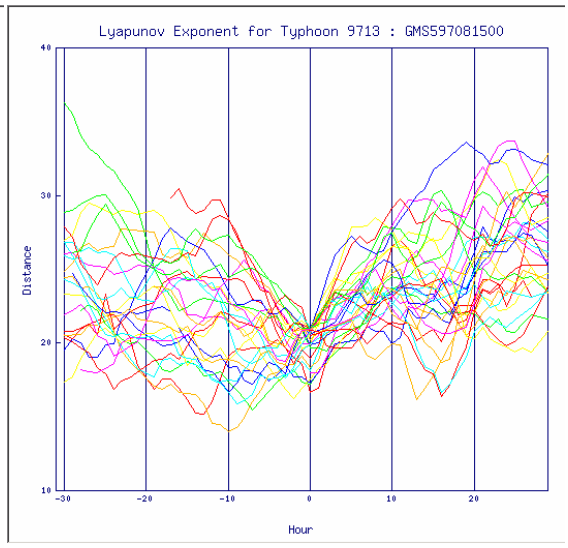


図 35 リヤプノフ指数を計算するための距離の時間変化の図.

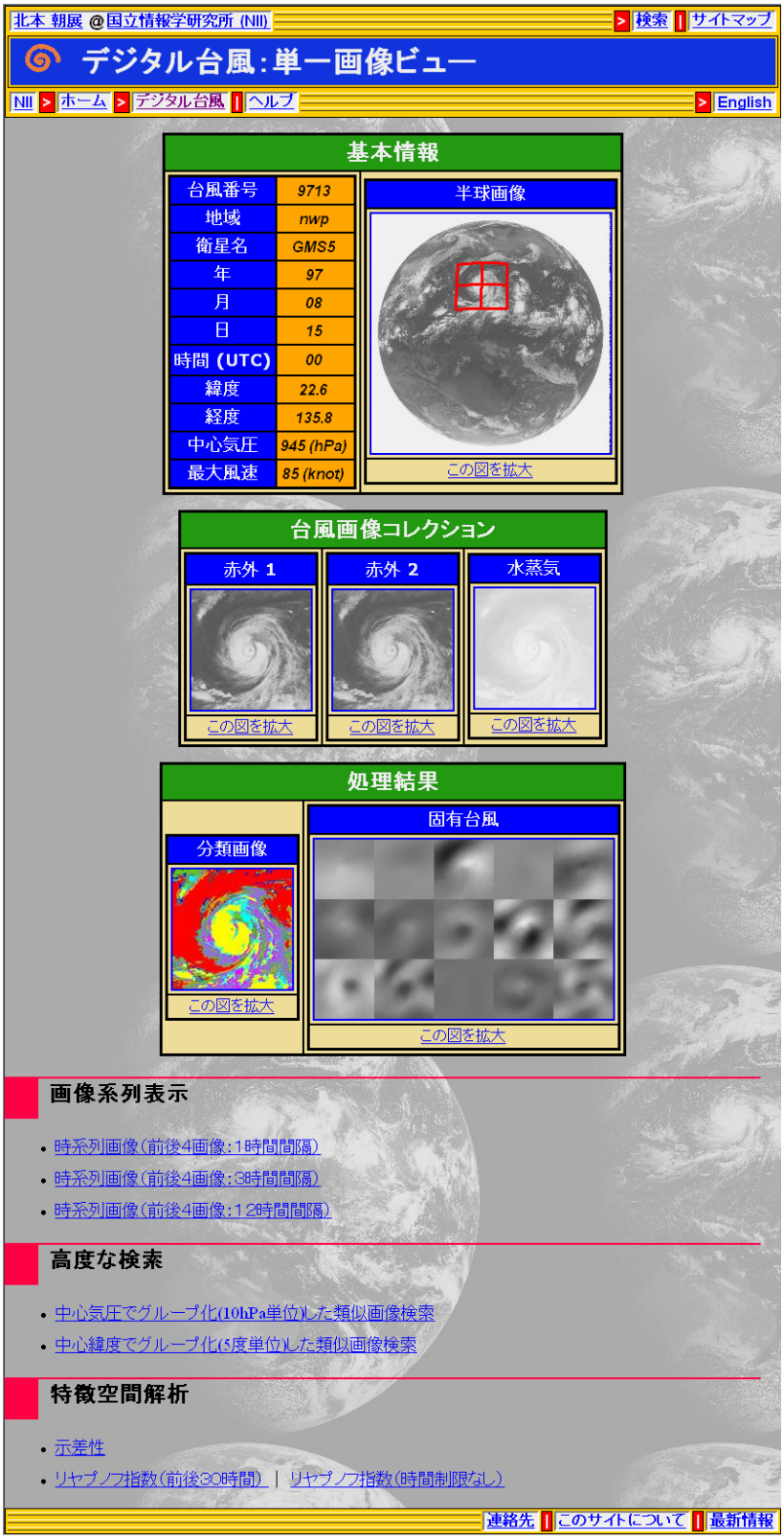


図 36 単一観測ビュー．台風に関するメタデータ(中心気圧など)に加え，衛星画像やそれから得た雲分類画像，また下部にはこのデータに関するさらに高度な問合せへのリンクが表示される．

上記6のデータマイニング操作については，現在様々な機能を実装中であり，今後さらに増やしていく計画である．そこで現在実装している機能の中から二つを紹介する．一つは特徴空間内において単一観測画像の近傍を解析するための機能である．具体的には，単一観測画像に対応する特徴空間上の点を原点とみなし，この原点の周囲に分布する他の観測点への距離の分布を計測する機能である．このような分布はDistinctivenessとも呼ばれ，特徴空間内で基準点の周囲にどのくらいの観測点が密集しているかを表す基準となる．これを計算すると，図 34に示すように，距離が小さい観測点は少ないが，距離が中程度の観測点が最も多く存在するという，一種の「次元の呪い」[65]に対応するような結果が得られる．

次に図 35では，台風系列間の類似度の変化を追跡するための機能を示している．これは，この単一観測画像の観測時刻を基準時刻ゼロとしたとき，そこから時間を進める・遡ると，異なる台風系列間で対応する画像間の類似度がどのように増加していくかを示している．このグラフの傾きはおおよそ最大のリヤプノフ指数に対応し，ある観測画像の近傍がカオス的にどのような性質をもっているのかを直感的に把握することができる．また類似観測間の距離が増大する様子から，アンサンブル的な予報をおこなったときの予測限界あるいは予測可能性についても，大まかな見積もりを得ることができる．このような時間変化を具体的な図で示すのが図 37である．

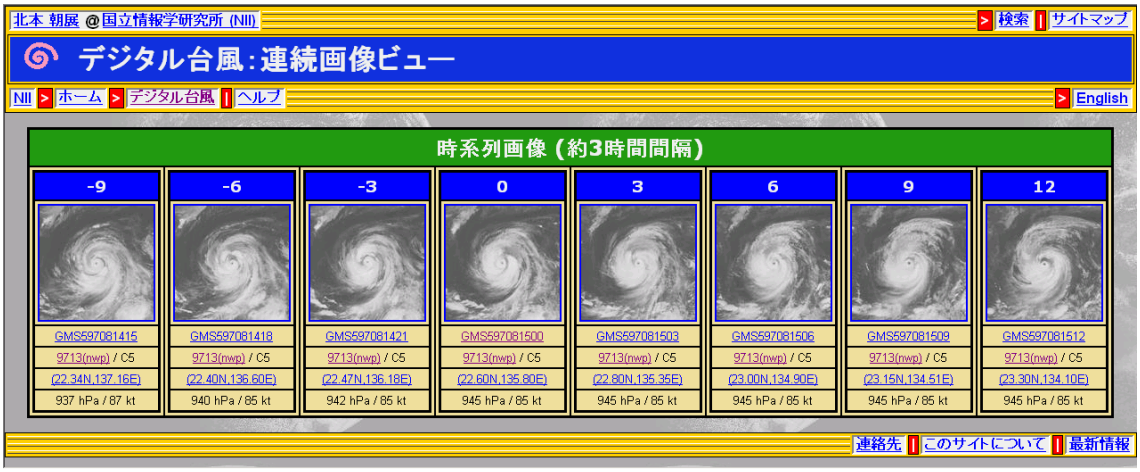


図 37　台風画像の連続画像ビュー．ある時刻を基準として，その前後の観測画像を表示する．

このような形で，単一観測をキーとしてデータマイニング操作をおこない，さらにこの単一観測に関連するデータを検索し，再びデータマイニング操作をおこなうというように，利用者が対話的に台風データベースを散策するような環境を実現する．ここで単一観測に関連するデータを検索する操作として，画像間，つまり台風雲パターン間の類似度を用いて，キーとなる画像に類似した画像を検索する機能を照会する．この機能を実現したビューを類似画像ビューと呼ぶ．その例を図 38に示す．ここでは問合せ画像に類似した上位14件の画像を検索し表示している．ここで検索を実行する際には，例示画とデータベースの画像の間ですべての組み合わせに対して類似度を計算する．しかし検索結果を得る際には，一つの台風系列から最大一件の台風画像しか表示しないという機能を用いている．これは，そのような制限なしでは，観測時刻が接近した台風画像が多数上位に食い込むためである．このような機能は，表3に示すデータ操作言語GRQLのグルーピング機能を用いることで実現でき，このことから，台風画像データベースに対してグルーピングが有用であることがわかる．

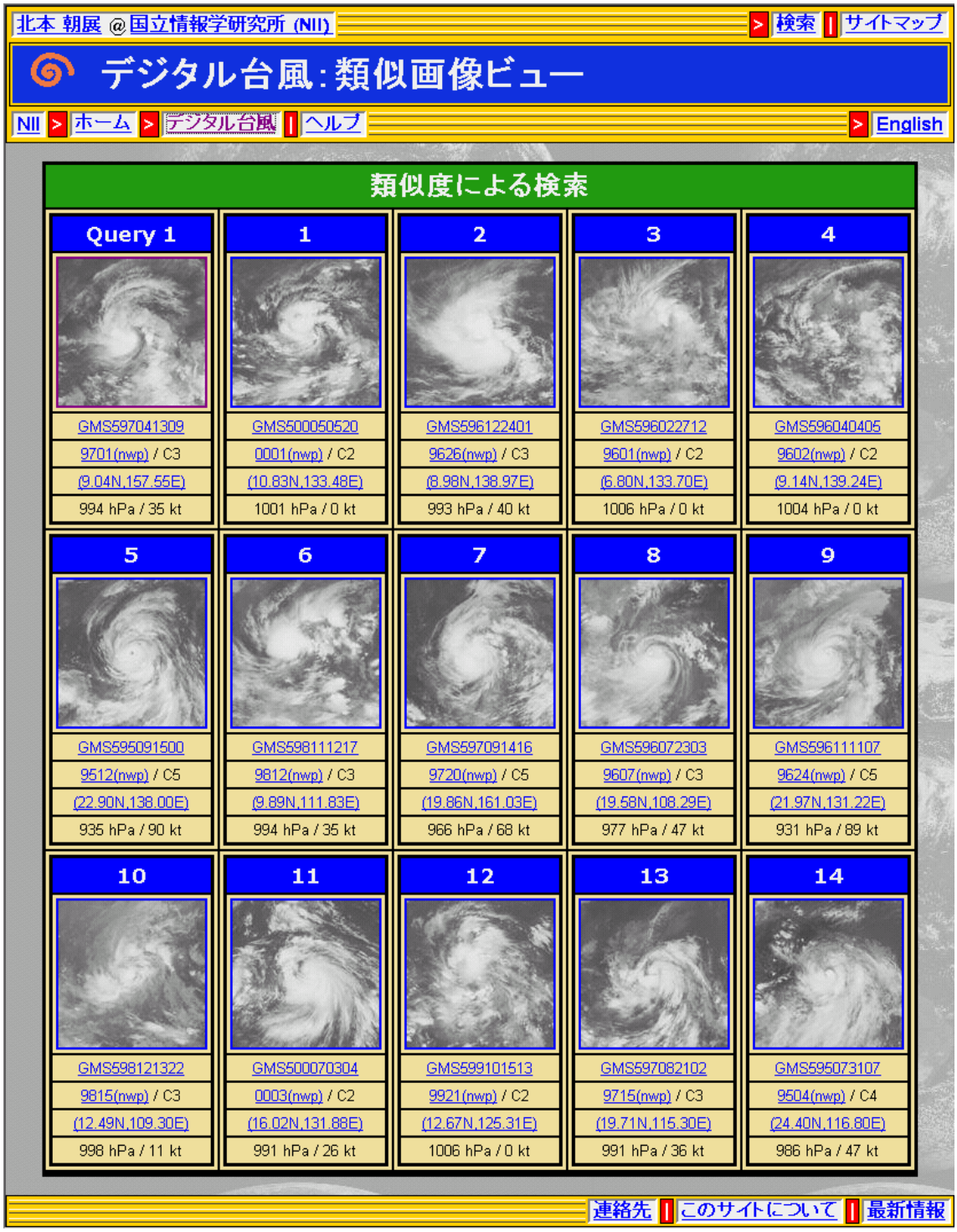


図 38　類似画像ビューの例. 左上が問合せ画像であり, それに類似した14件の画像を検索している. 同一観測系列からは1件しか検索されていないことに注意.

ここで類似画像ビューは，単一観測の順序つきリストとみなすこともできる. このビューは，単一観測ビューや単一系列ビュー，類似画像ビューへのリンクを備えており，これらのリンクをクリックすることで，台風データベースをさらにナビゲーションできる. そこで次に，単一系列ビューに関して説明する.

8.4.2 単一系列ビュー

単一系列ビューとは，同一の台風に関する情報を集約して表示するビューである．これは主に個々の台風の時間的な変化を見るためのビューであり，通常は観測時間の昇順に並べたリストを表示することで，同一台風の誕生から死滅までのライフサイクルの変化を眺める．図 39は北半球の台風199713号のライフサイクルを集約して表示するビューであり，以下の情報を示している．

1．台風の地理的移動経路

2．1日ごとにサンプリングした台風画像のリスト

3．台風雲パターンの変化と安定を直感的に把握するためのリカレンスプロット

4．この単一系列を問合せとする高度なデータマイニング操作へのリンク

ここでリカレンスプロットとは，時点iの信号と時点jの信号との相関関係$D(i,j)$を画素座標(i,j)に描画する方法であり，時系列信号のもつ非定常性の検出にも優れているとされるものである．特にカオス時系列解析ではアトラクタの構造を視覚化するツールとして用いる[92]．

また単一系列ごとの情報を集約するだけでなく，複数の台風系列の比較をおこなう機能も重要である．これは基本的にはtime warpingすなわち単一系列を時間軸方向に伸縮させつつ，複数の台風系列の間で最適な対応を求める問題に帰着させることができる．このような問題の解法として代表的なのは，動的計画法(dynamic programming)を用いたDPマッチングによるものであり，DTW(Dynamic Time Warping)とも呼ばれる手法である．これを用いると，台風ライフサイクルの時間軸を正規化しつつ，台風系列の比較をおこなうことも可能となる．本研究ではその初歩的な段階として，2個の台風系列の間で最も類似した対応を求める機能を実装した．その結果を図 40に示すが，基準時刻ゼロに類似していた台風が，時間の経過とともに異なるパターンに変化していく様子が表現できている．このように類似した事例が類似しない事例に離れていくという性質は，7.6.2章で述べたカオス的性質を表現していると考えることができる．

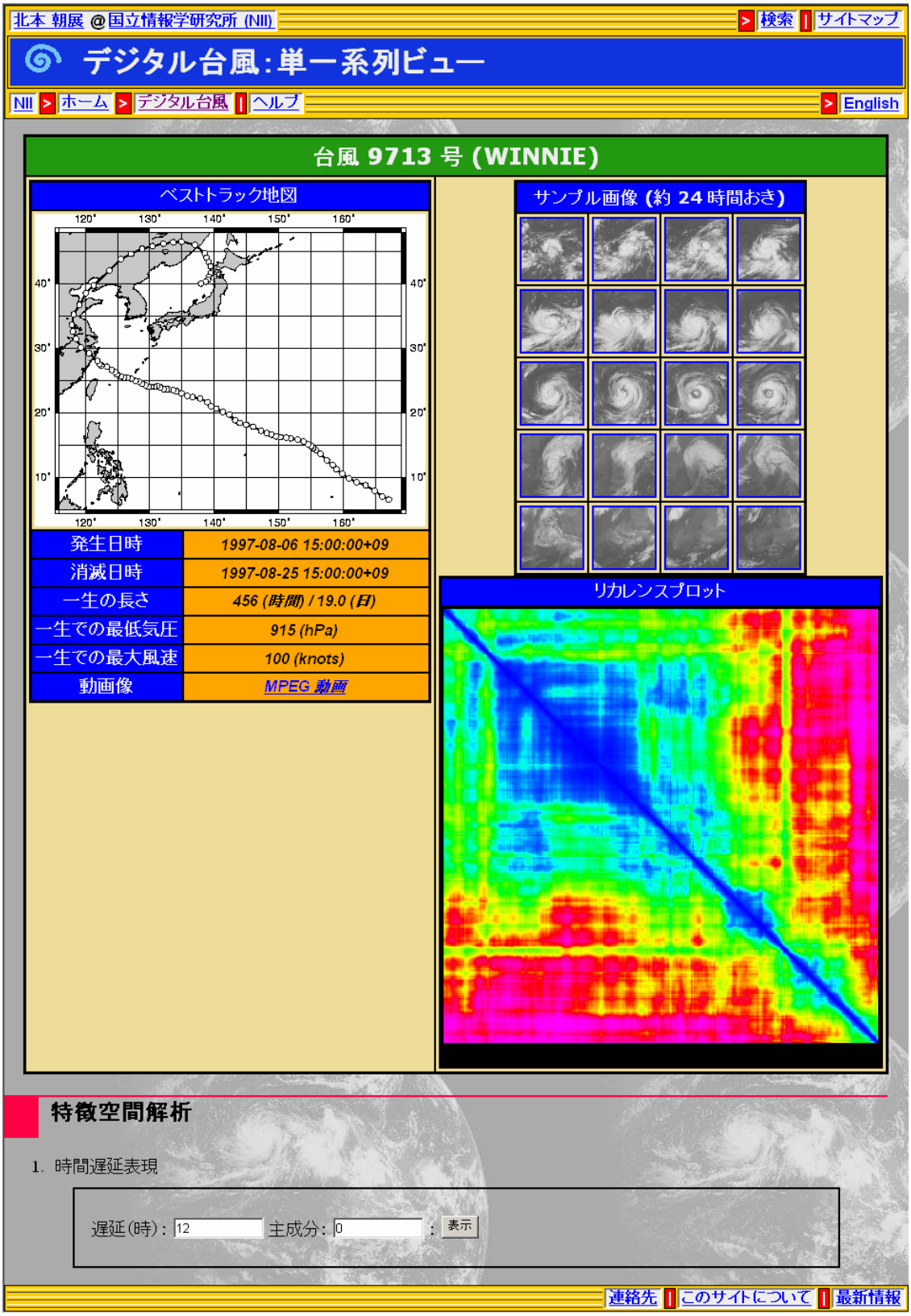


図 39　単一系列ビュー．台風系列の情報をまとめて表示する．

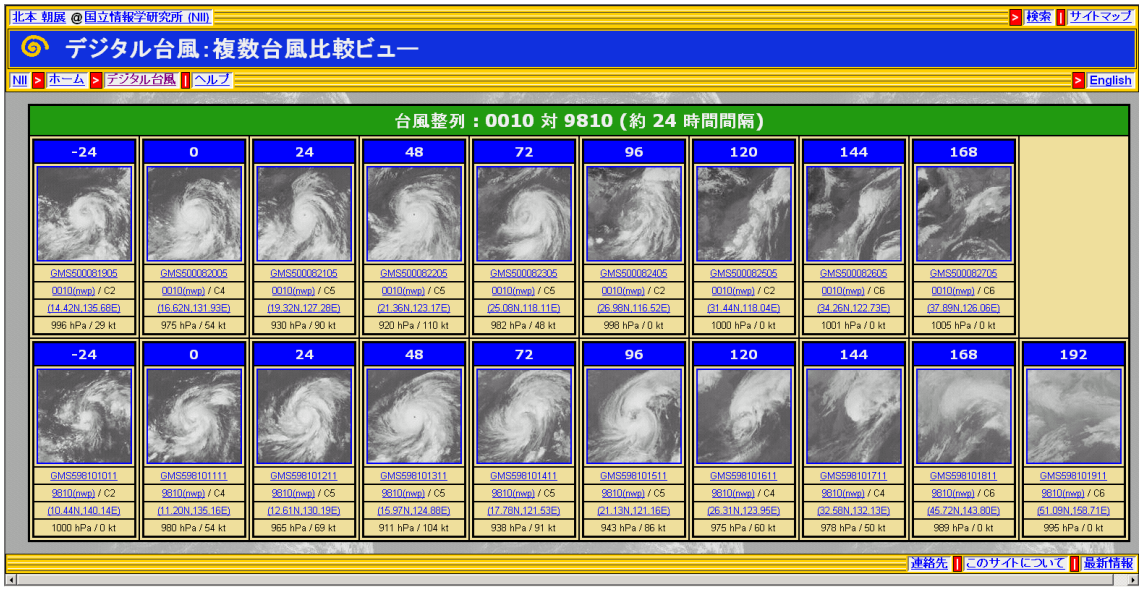


図 40　複数台風比較ビュー．二つの台風を最も類似した画像のペアで整列し，その前後の雲パターンの時間変化を比較している．

8.4.3　グルーピング・順序付きリスト

先述のようにデジタル台風で用いるデータ操作言語GRQLにおいては，グルーピングを基本的な操作としている．またこの言語ではグルーピングという概念を拡張し，クラスタリングも一種のグルーピング操作に含めている．ここでクラスタリングの中心座標を決定するアルゴリズムは種々のアルゴリズムに依存するものの，クラスタの中心座標さえ決定すれば，それぞれのデータを最も近い中心座標をもつクラスタにグルーピングする機能は共通である．またデータをグループに分割してしまえば，グループに対して適用可能なデータ操作（例えば並べ替えなど）もクラスタリングアルゴリズムに依存しない．ゆえにグループという概念を導入することで，単なるクラスタリングよりも高度な操作を組み合わせて適用することができる．以下では拡張したグルーピング操作の例をいくつか示す．

射影行列

射影行列は高次元の特徴空間を低次元の特徴空間に射影するための線形演算子である．主に高次元の特徴空間を2次元の特徴空間に射影し，2次元空間中でデータの分布を調べるために用いる．線形演算子であるため特徴空間の分布を捉える能力には限界があるが，この行列の選び方によって平面状に実際に出現する分布は異なることになる．例えばこの行列を固有ベクトルの行列とすると，低次元の空間は第一主成分および第二主成分によって張られる空間となるため，これらの主成分が示唆する意味などを把握するにはこのような射影行列を使うべきである．あるいは極端な場合は，この行列をランダムに選んでも，射影された平面は高次元特徴空間の2次元断面を示しており，その断面も何かしらの雲パターン分布を示すことになる．

このような平面上での雲パターンの分布を可視化するためには，平面上の位置に応じて代表的な雲の

パターンを並べて表示すればよい. そのような2次元平面グリッドの可視化は, 拡張したグルーピング操作を活用して以下のようなアルゴリズムで実現することができる.

1. すべてのデータを2次元平面に射影し, この2次元平面空間上でデータが存在する領域を計算する.

2. データが存在する領域を N 等分し, 2次元平面を $N \times N$ 個のグリッドに分割する. これらのグリッドの中心をグリッドの代表点とする.

3. 2次元平面上の代表点の座標を, 射影行列のムーアペンローズ逆行列を用いて, 高次元空間に射影しなおす.

4. 高次元空間でこれらの代表点に基づくクラスタリングをおこない, すべてのデータを最も近いクラスタに所属させるグルーピング操作を適用する.

5. グループ内で代表点への距離を用いてデータを並べ替え, 最も代表点に近いものからグループの代表画像とする.

6. これらの代表画像を2次元平面上に並べて表示する.

この方法を用いて特徴空間を可視化したビューの例として, 固有ベクトル(主成分)を用いた場合の例を図 41に, またランダムに生成した行列によるランダム射影を用いた場合の例を図 42に示す. いずれの場合も, 特徴空間中の雲パターンの分布を表現している. 特に図 41では, 横軸が第一主成分, 縦軸が第二主成分という統計的意味をもっており, それぞれの主成分がどのようなパターンの変動を表現しているかがわかるようになっている.

8.4.4 メタデータによる検索

メタデータによる検索は, 最も伝統的な検索方法である. メタデータとは「データを説明するデータ」と言われるが, 一般にデータそのものではなくデータに関する情報を記述する付加的な要素という位置づけで用いられる. 例えば本研究でもメタデータによる検索として, 台風の名前や気象衛星の観測日時, 地理的位置などを問合せとして検索する方法を提供している. つまり, 台風画像に対してあらかじめ人間が与えたメタデータ(台風の名前など), あるいは台風画像に対して自動的に付与されるメタデータ(観測日時など)などのメタデータを用いて検索する機能である. その一例として図 43には地理的位置というメタデータを問合せに用いるためのグラフィカルユーザインタフェースを示す. 利用者は地図の上で探したい地点をクリックすると, それが第4章で述べたデータ操作言語GRQLをエンコードしたXML検索要求メッセージとなり, その検索結果はクリック地点からの距離順に並べ替えられた単一観測のリストとして得られる. これは類似画像検索の結果として得られるリストと同じであり, このリストの任意の観測を新たに基準とすることにより, さらにデータベースのナビゲーションを進めていくことができる.

メタデータによる検索方法は, 格別新しいものではなく, 特に使いやすいわけでもないが, 日時や地域など馴染み深い情報をキーとして検索することができるだけに, 例えば特定地域の台風を調べたり, 特定期間の台風を調べたり, といったデータマイニングには効果を発揮する.

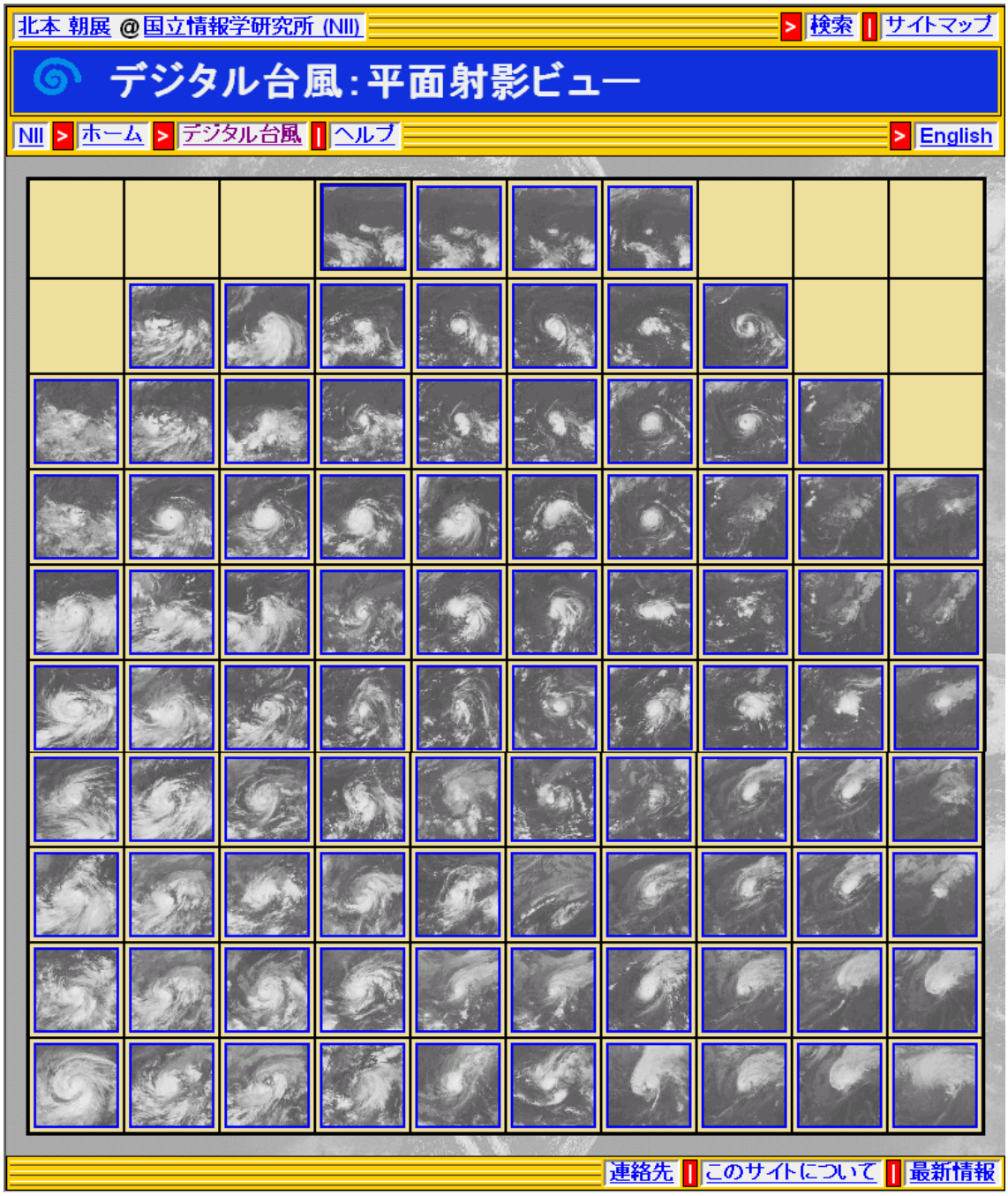


図 41　主成分による平面射影ビュー．2次元平面グリッドを用いたグルーピング法を用いている．

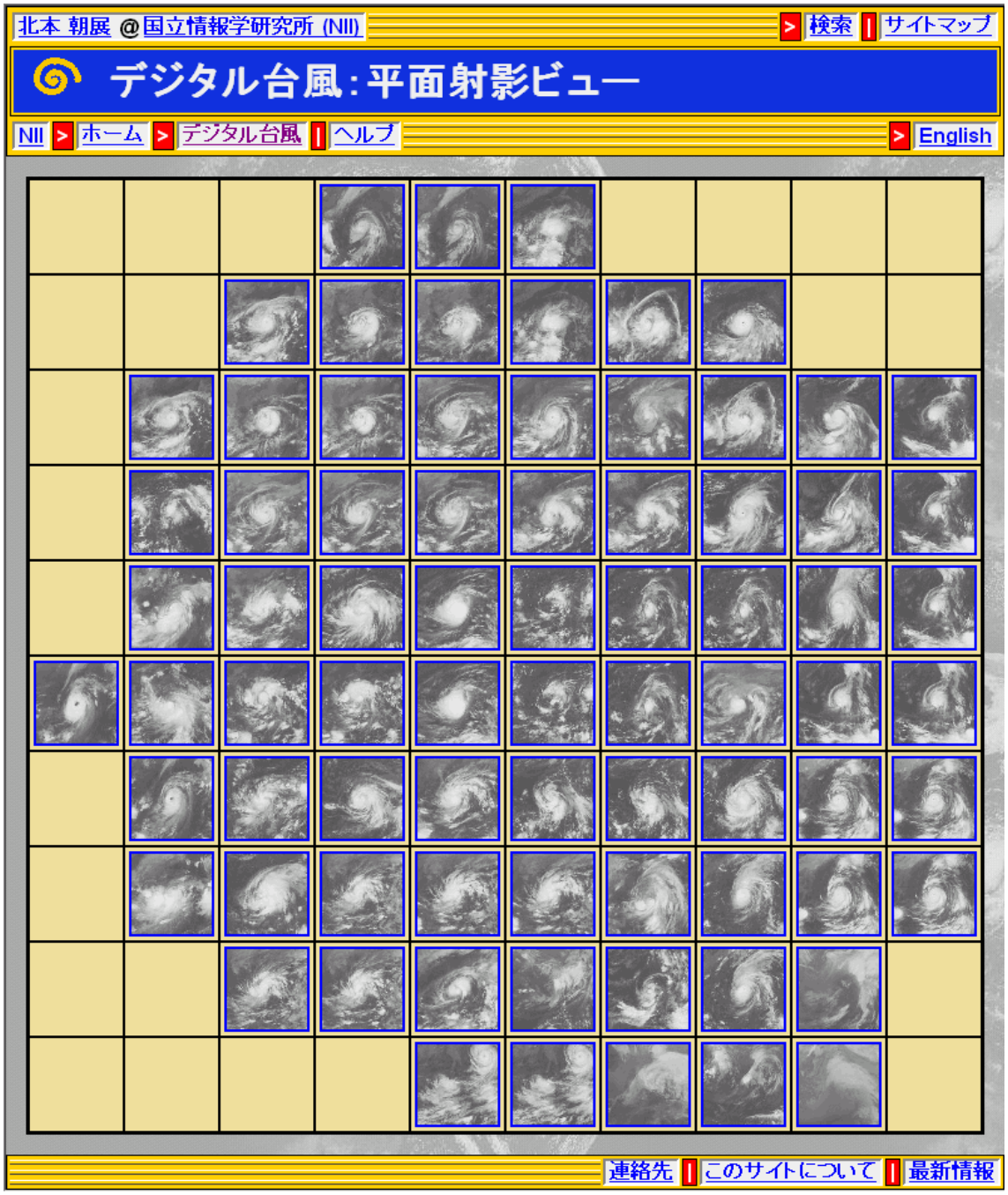


図 42　ランダム射影による平面射影ビュー．2次元平面グリッドを用いたグルーピング法を用いている．

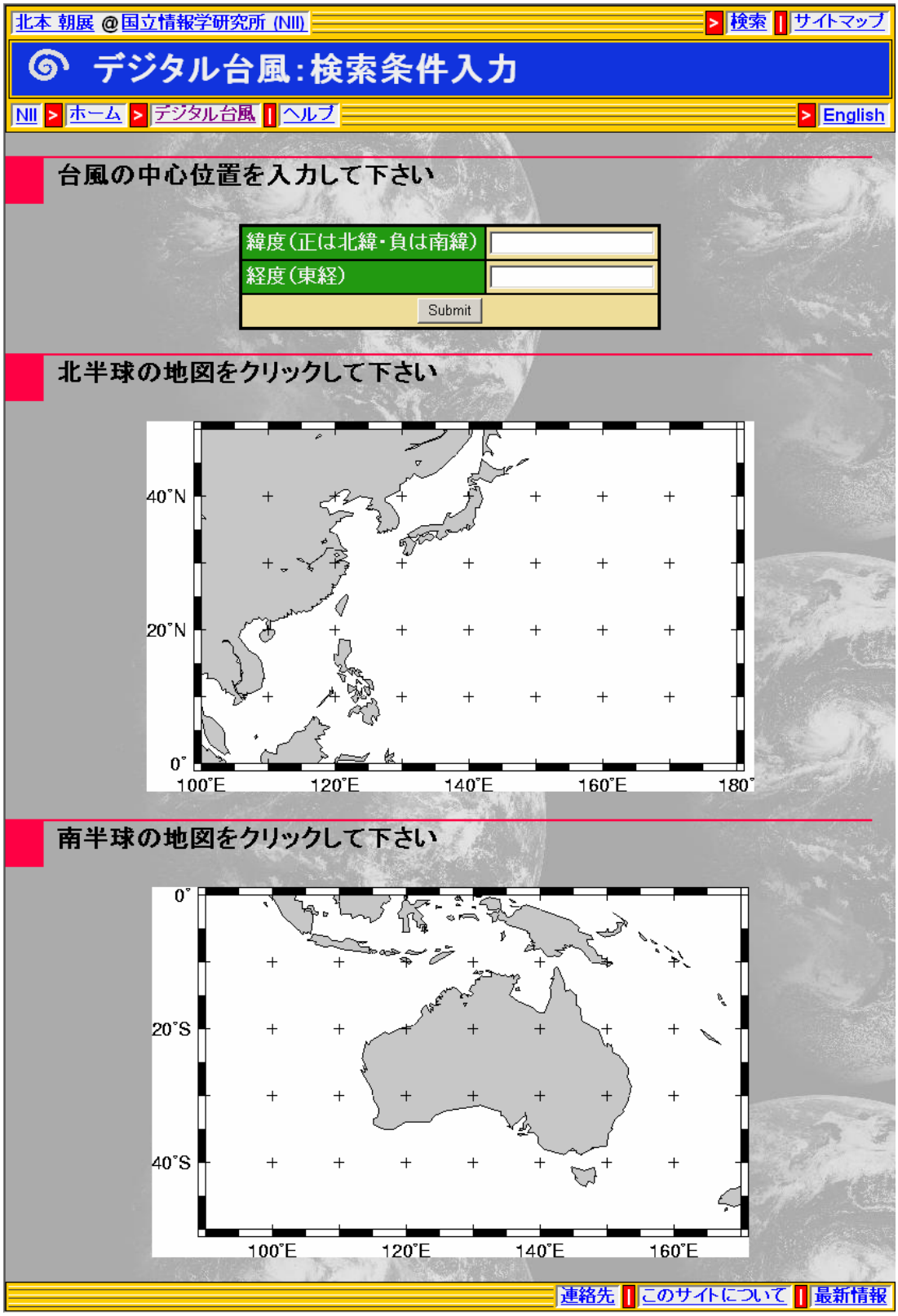


図 43　地理的位置というメタデータを用いた検索のためのインタフェース.

第9章 終章

9.1 研究成果の総括

本研究プロジェクト「ネットワークに基づく分散型地球環境データベースの構築」に関する研究は5年間にわたって進められ，以上に述べた研究成果を得た．これらはいずれも地球環境データベースに不可欠の要素技術であり，分散型地球環境データベースの発展に貢献しうるものである．

地球環境データに関する最大の問題点は，データの死蔵問題であると考えている．つまりせっかく観測した貴重なデータが，磁気テープの山の中に埋もれつつ，誰からも活用されずにひっそりと消えつつあるという問題である．これは確かに非常に無駄なことであるが，現在の地球環境データの利便性の悪さを考えれば，致し方ないという側面が否めないのも事実である．利便性の悪さを具体的に言えば，データを入手する際の利便性の悪さ，データを解析する際の利便性の悪さ，データを検索する際の利便性の悪さ，などを挙げることができるだろう．

このうち，データを解析する際の利便性の悪さについては，ある程度はやむを得ない部分もある．というのは，データ解析は結局のところ利用者の目的に左右される部分が大きいからである．しかし現状ではそのような程度以上に，データ解析に必要な手間は大きいのが現状である．またそれよりも早急に解決する必要があるのは，データを入手する際の利便性の悪さであろう．これには，1)データがオンラインでアクセスできない，2)リモートサイトからデータをダウンロードできない，という2種類の問題点がある．

これらの問題点のうち，1)については東京大学生産技術研究所が研究を進めており，2)は今回我々の研究グループがタイとの共同研究でまさに取り組んだテーマである．国立情報学研究所がたまたまタイとの間に高速回線を運営していたため，これらの問題点を解消することができたが，そうでない場合にはネットワークの帯域を増強することが結局のところ最も根本的な解決策であろう．つまり分散型地球環境ネットワークにとっては，このような広帯域のネットワークを確保することが大前提であり，このことは大容量のデータ共有を必要とする研究者に特に当てはまる．

しかし，より広い人々にとって重要なのは，データを検索する際の利便性の悪さ，つまりデータの検索機能にあると考える．つまり「地球環境データベース」にはそもそもどんなデータがあり，それらはどんな関連を持っているのか，といったことは，データベースが高度な機能として提供する必要があるだろう．本研究の大部分は，このテーマに関連する研究に費やしている．そしてこの部分が将来にわたって，地球環境データの利便性向上に決定的な役割を果たすとも考えている．

このような問題意識に基づき，本研究ではさまざまな観点からデータベース技術を研究してきた．ここで特に，地球環境データベースという文脈を常に意識しながら研究を進めてきたのが，本研究の特色と言

えるのではないかと考えている. このように対象を決めることによって, 技術的課題をより明確に浮き彫りにすることができ, そこに面白い問題も発見できるようになるためである. もちろん究極的には, わざわざ「地球環境」データベースを構築する必要はなく, 一般のデータベースを「地球環境データのために」使えばいいという状況になるのかもしれない. しかし現状でこのようなアプローチを取ると, 地球環境データに特有の具体的な性質には目をつぶり, 一般のデータとしての抽象的性質のみを拾い上げて地球環境データを構築することになる. このようなアプローチは技術的には実現しやすいものの, 地球環境データに特有のセマンティクス(意味論)を無視しては, 結局のところ地球環境データ利用者のニーズに答えられないものができてしまう.

つまり今の段階ではまだ, 一般解をめざすのではなく, 個別解をめざすべきではないかと考える. このような過度の一般化, つまり技術を無理やり一般化することで逆に何にも使えない技術になること, は特に画像データベースの研究などでよく見られる現象であり, 一般的な技術であると主張する技術が実際のところは必ずしもより広い分野で実用的となっているわけではない. ゆえに, 本研究の階層的なモデルのように, 抽象的な計算論を考えるメタなモデルと, 具体的な対象をしっかり捉えたプラグインモデルの双方を研究する必要があると考えるのである.

本研究はこうした階層モデルの下部の階層から上部の階層までを総合的に研究することを目指す研究であった. 実際のところこのようなスタイルは, 地球環境データベースに対する取り組みの中では, やや風変わりな研究であったかもしれない. というのも, 地球環境データベースに関するほとんどの研究は, ある特定の問題領域のある特定の事象を検索するために用いる地球環境データベースであったり, 一方で既存の技術の組み合わせを地球環境データに適用したり, といった試みが多くを占めていたからである. その意味で本研究のように, 地球環境データベースの性質を上の階層から下の階層まで総合的に調べ, そこに使われるべきアルゴリズムを考えていくという方法は, これまでにあまりなかったアプローチかもしれない. しかし, このようなアプローチのおかげで, 今まで気づかなかった微妙な問題点に気づくことができたなどの利点を考えれば, このアプローチは成功だったと考えている.

最後に, 具体的な研究対象として台風に出会ったのは, 幸運なことだったかもしれない. これによって, 地球環境データベースの一般論では見えにくかった, モデルやアルゴリズムの様々な問題点が見えてきたこと, また台風という対象を突き詰めていくことで, 情報学的に興味深い様々な問題を発見できたこと, などの点で幸運だったと考える. 情報学者の任務とは, 具体的な対象物を情報の観点から抽象化することにあり, 最初から存在する抽象的な情報概念を具体的なものに無理やり当てはめる, というのでは順序が逆ではないだろうか. その意味でも, 今後しばらくは, 台風という具体的な対象物にこだわってみる計画である. また, このような経験をうまく抽象化できたとき, そこに新しい地球環境データベースを生み出すことができるのではないかと期待している.

9.2 今後の展望

最後に本研究プロジェクトの今後の展開について説明する. 本研究の主要な研究成果である「デジタル台風」については今後も継続的にプロジェクトを進めていく計画である. 台風データは毎年台風が発生するたびに増えていくため, データコレクションとしては毎年充実していくことになる.

このプロジェクトの発展の方向として, 最初に地球環境データを地理的・時間的座標で結合する, 統合地球環境情報基盤の構築という課題がある. これはいわゆるデジタルアース計画の台風データ版であり, 種々の台風関連メタデータをいったん共通基盤の上で束ねることにより, 複合的な観点から情報を関連付けナビゲートできるようになるというアイデアである. ただしこれはどうしてもシステム寄りの発想に

近くなり，デジタル台風でどのような情報を提供するのか，というデザインまではなかなか話が及ばない傾向がある．

もう一つの方向性は， 台風データを「体感」するための可視化，可聴化技術の研究である．台風データがひまわり画像の2次元平面上をちょろちょろと動き回っているだけでは，台風の怖さや迫力はなかなか伝わってこない．このような情報を立体的な映像や音などの様々なメディアを用いて表現することで，台風のさまざまな性質を具象化しようというのが研究のアイデアである．ここで注意すべきことは，この研究の目的は「台風らしい」映像をつくることではなく，実在の台風そのものを再現しようという試みである点である．したがって，この技術の基盤となるのはコンピュータグラフィックス技術ではなく，あくまで実在の台風を解析する技術，すなわち我々が本研究で培ってきたさまざまな地球環境データ処理のための要素技術である．このような確固とした台風解析技術があって初めて，台風をデジタル地理空間に再現し，情報環境の中で人間が動き回りながら新たな発見に出会えるような，そんな情報環境を実現することができるのではないかと考えている．

参考文献

[1] 北本 朝展, 高木 幹雄, "類似画像検索システム構築のフレームワークとしての階層モデル", 電子情報通信学会技術報告, Vol. PRMU97-58, pp. 25-32, 1997年7月

[2] 北本 朝展, "パイプライン型遺伝的アルゴリズムによる類似画像検索パラメータの対話的な最適化", 情報処理学会 第55回全国大会, Vol. 2, pp. 457-458, 1997年9月

[3] 北本 朝展, 高木 幹雄, "パイプライン型遺伝的アルゴリズムを用いた対話的な画像散策", ワークショップ「インタラクティブ進化的計算論」, pp. 31-36, 1998年3月

[4] 北本 朝展, 高木 幹雄, "ミクセル密度を含む混合密度推定を用いたミクセルの面積占有率推定", 電子情報通信学会論文誌, Vol. J81-D-II, No. 6, pp. 1160-1172, 1998年6月

[5] 北本 朝展, 高木 幹雄, "ミクセルの内部構造を反映する確率モデルを用いた画像分類法", 画像の認識・理解シンポジウム (MIRU'98), Vol. I, pp. 87-92, 1998年7月

[6] Asanobu KITAMOTO and Mikio TAKAGI, "Image Classification Using a Stochastic Model that Reflects the Internal Structure of Mixels", Advances in Pattern Recognition (SPR'98) : Lecture Notes in Computer Science 1451, pp. 630-639, 1998年8月

[7] 北本 朝展, 高木 幹雄, "待ち行列型遺伝的アルゴリズムを用いた対話的な画像散策法", 人工知能学会誌, Vol. 13, No. 5, pp. 728-738, 1998年9月

[8] 北本 朝展, 高木 幹雄, "画像の空間的量子化を考慮したしきい値選定法", 電子情報通信学会1998年ソサイエティ大会, Vol. D-12, pp. 36, 1998年10月

[9] 北本 朝展, 高木 幹雄, "待ち行列型遺伝的アルゴリズムの特徴と応用", 情報処理学会 第57回全国大会, Vol. 2, pp. 376-377, 1998年10月

[10] 北本 朝展, 高木 幹雄, "ミクセルの内部構造を反映する面積占有率密度を用いた画像分類法", 電子情報通信学会論文誌, Vol. J81-D-II, No. 11, pp. 2582-2597, 1998年11月

[11] 北本 朝展, 高木 幹雄, "進化的計算論に基づく対話的な画像散策法", 第4回知能情報メディアシンポジウム, pp. 173-180, 1998年12月

[12] Asanobu KITAMOTO, "Toward Content-Based Satellite Image Database Systems over the Network", Proceedings of the 5th International Workshop on Academic Information Networks and Systems (WAINS), pp. 31-38, 1998年12月

[13] Asanobu KITAMOTO and Mikio TAKAGI, "Image Classification Using Probabilistic Models that Reflect the Internal Structure of Mixels", Pattern Analysis and Applications, Vol. 2, No. 1, pp. 31-43, 1999年4月

[14] 北本 朝展, "ミクセルの影響を考慮した最大ゆう度しきい値選定法", 電子情報通信学会技術報告, Vol. PRMU99-166, pp. 7-14, 1999年12月

[15] Asanobu KITAMOTO, "The Development of Typhoon Image Database with Content-Based Search", Proceedings of the 1st International Symposium on Advanced Informatics (AdInfo), pp. 163-170, 2000年3月

[16] Asanobu KITAMOTO, "Multiresolution Cache Management for Distributed Satellite Image Database Using NACSIS-Thai International Link", Proceedings of the 6th International Workshop on Academic Information Networks and Systems (WAINS), pp. 243-250, 2000年3月

[17] Asanobu KITAMOTO and Mikio TAKAGI, "Area Proportion Distribution -- Relationship with the Internal Structure of Mixels and its Application to Image Classification", Systems and Computers in Japan, Vol. 31, No. 5, pp. 57-76, 2000年5月

[18] Asanobu KITAMOTO, "The Moments of the Mixel Distribution and Its Application to Statistical Image Classification", Advances in Pattern Recognition (SPR'00), Lecture Notes in Computer

Science 1876, pp. 521-531, 2000年8月
[19] 北本 朝展, "台風雲パターンの衛星時系列画像を対象とした楕円形状分解手法", 電子情報通信学会 2000年ソサイエティ大会, Vol. D-12-2, pp. 189, 2000年9月
[20] 北本 朝展, "「デジタル台風」 --- 人工知能的アプローチに基づく台風解析", 情報処理学会技術報告, Vol. CVIM123-8, pp. 59-66, 2000年9月
[21] 北本 朝展, "台風雲パターンの衛星画像解析に基づく台風データベースの構築", 情報処理学会 第61回全国大会, Vol. 2-3V-5, pp. 217-218, 2000年10月
[22] Asanobu KITAMOTO, "Enhancing the Quality of Typhoon Image Database Using NOAA AVHRR Data Received in Thailand", Proceedings of the 7th International Workshop on Academic Information Networks and Systems (WAINS), pp. 1-11, 2000年12月
[23] 北本 朝展, 小野 欽司, "台風画像コレクションの構築および台風解析への応用", NII Journal, No. 1, pp. 7-22, 2000年12月
[24] 北本 朝展, "Holistic Analysisを用いた台風雲パターンの解析", 電子情報通信学会技術報告, Vol. PRMU2000-240, pp. 129-136, 2001年3月
[25] 北本 朝展, 小野 欽司, "日本とタイの国際共同研究に基づく台風データの収集および台風画像データベースの構築", NII Journal, No. 2, pp. 15-26, 2001年3月
[26] Asanobu KITAMOTO, "Data Mining for Typhoon Image Collection", Proceedings of the 2nd International Workshop on Multimedia Data Mining, pp. 68-77, 2001年8月
[27] Asanobu KITAMOTO, "FCA: The Fractional Component Analysis", 第4回情報論的学習理論ワークショップ, pp. 297-302, 2001年8月
[28] Asanobu KITAMOTO, "Analysis and Prediction of the Typhoon from an Informatics Perspective", Proceedings of the 8th International Workshop on Academic Information Networks and Systems (WAINS), pp. 43-52, 2001年10月
[29] 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年1月
[30] Asanobu KITAMOTO, "Spatio-temporal Data Mining for Typhoon Image Collection", Journal of Intelligent Information Systems, Vol. 19, No. 1, pp. 25-41, 2002年7月
[31] Asanobu KITAMOTO, "IMET: Image Mining Environment for Typhoon Analysis and Prediction", Multimedia Data Mining, Djeraba, C. (編), pp. (in press), Kluwer Academic Publishers, 2002年12月

[32] 三浦 孝夫, "データモデルとデータベース", 第1巻・第2巻, サイエンス社, 1997年
[33] 北川 博之, "データベースシステム", 昭晃堂, 1996年
[34] Extensible Markup Language (XML), http://www.w3.org/XML/
[35] Earth Science Markup Language (ESML), http://esml.itsc.uah.edu/
[36] Geography Markup Language (GML) 2.0, http://opengis.net/gml/01-029/GML2.html
[37] G-XML, http://gisclh.dpc.or.jp/gxml/
[38] MPEG-7, http://mpeg.telecomitalialab.com/
[39] XML Schema, http://www.w3.org/XML/Schema
[40] Resource Description Framework (RDF), http://www.w3.org/RDF/
[41] Semantic Web, http://www.w3.org/2001/sw/
[42] OMF, http://zowie.metnet.navy.mil/~spawar/JMV-TNG/XML/OMF.html
[43] 西尾 章次郎 監修, "実践SQL教科書", アスキー出版局, 1996年
[44] XML Query, http://www.w3.org/XML/Query
[45] Multimedia Retrieval Markup Language (MRML), http://www.mrml.net/
[46] The Extensible Stylesheet Language (XSL), http://www.w3.org/Style/XSL/

[47] J. Shim, P. Scheuermann, and R. Vingralek. “Proxy cache algorithms: Design, implementation, and performance”. IEEE Transactions on Knowledge and Data Engineering, 11(4):549-562, 1999.
[48] C. Aggarwal, J.Wolf, and P.Yu. Caching on the World Wide Web. IEEE Transactions on Knowledge and Data Engineering, 11(1):94-107, 1999.
[49] Satellite Imagery Archive at Institute of Industrial Science, University of Tokyo, http://www.tkl.iis.u-tokyo.ac.jp/SIAIIS/
[50] L. Breslau, P. Cao, L. Fan, G. Phillips, S. Shenker, “Web Caching and Zipf-like Distributions: Evidence and Implications”, Proceedings of the IEEE INFOCOM’99, pp. 126-134, 1999.
[51] Y. Yasuda, M. Takagi, S. Kato, and T. Awano. Step by step image transmission and display from gross to fine information using hierarchical coding. The Transactions of the Institute of Electronics and Communication Engineers, J63-B(4):379-386, 1980.
[52] A. Said and W. Pearlman. An image multiresolution representation for lossless and lossy compression. IEEE Transactions on Image Processing, 5(9):1303-1310, 1996.
[53] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1998.
[54] JPEG 2000, http://www.jpeg.org/JPEG2000.htm
[55] 加藤俊一, 栗田多喜夫. 画像の内容検索 --- 電子美術館への応用 ---.情報処理, Vol.33, No.5, pp.466-477, 1992.
[56] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. Query by Image and Video Content: The QBIC System. Computer, Vol.28, No.9, pp.23-32, 1995
[57] Tanenbaum, A.S. コンピュータネットワーク, 丸善, 第2版, 1992.
[58] デビッド・マー. ビジョン. 産業図書, 1987.
[59] D.A. Quattrochi and M.F. Goodchild, Scale in Remote Sensing and GIS, Lewis Publishers, 1997.
[60] M. Girolami, Advances in Independent Component Analysis, Springer, 2000.
[61] A. Hyvarinen, J. Karhunen and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
[62] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, vol. 401, pp. 788-791, 1999.
[63] R.A. Redner and H.F. Walker, Mixture Densities, Maximum Likelihood and the EM algorithm, SIAM Review, Vol. 26, No. 2, pp. 195-239, 1984
[64] G.J. McLachlan and T. Krishnan, The EM Algorithm and Extensions, John Wiley & Sons, 1997
[65] T. Hastie. R. Tibshirani and J. Friedman, The Elements of Statistical Learning, Springer, 2001
[66] M. Kass. A. Witkin. D. Terzopoulos, Snakes : Active Coutour Models, Proc. of Int. Conf. on Computer Vision, pp. 259-268, 1987.
[67] A. Blake and M. Isard, Active Contours, Springer, 1998
[68] P.J. Brockwell and R.A. Davis, 入門 時系列解析と予測, CAP出版, 2000
[69] 片山 徹, 新版 応用カルマンフィルタ, 朝倉書店, 2000
[70] 鈴木 和史, 元木 敏博, 台風―解析と予報―, 気象研究ノート, Vol. 197, 日本気象学会, 2000
[71] D.E. Goldberg, Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley, 1989
[72] H. Takagi, Interactive Evolutionary Computation: Fusion of the Capacities of EC Optimization and Human Evaluation, Proceedings of the IEEE, Vol. 89, No. 9, pp. 1275-1296, 2001
[73] K. Sparck Jones and P. Willett Eds., Readings in Information Retrieval, Morgan Kaufmann Publishers, 1997
[74] Y. Rui, T.S. Huang, M. Ortega and S. Mehrotra, Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 8, No. 5, pp. 644-655, 1998

[75] 山岬正紀. 台風. 東京堂出版, 1982
[76] Sidney, V.D.B. Galaxy Morphology and Classification. Cambridge University Press, 1998.
[77] R.M. Zehr. Tropical cyclone research using large infrared image datasets. In 24th Conference on Hurricanes and Tropical Meteorology, pp. 486-487. American Meteorological Society, 2000
[78] W.R. Moninger, J. Bullas, B. de Lorenzis, E. Ellison, J. Flueck, J.C. McLeod, C. Lusk, P.D. Lampru, R.S. Phillips,W.F. Robers, R. Shaw, T.R. Stewart, J.Weaver, K.C. Young, and S.M. Zubrick. Shootout-89, a comparative evaluation of knowledge-based systems that forecast severe weather. Bulletin American Meteorological Society, Vol. 72, No. 9, pp. 1339-1354, 1991
[79] L.E. Carr, R.L. Elsberry, and J.E. Peak. Beta test of the systematic approach expert system prototype as a tropical cyclone track forecasting aid. Weather and Forecasting, Vol. 16, pp. 355-368, 2001.
[80] L. Zhou, C. Kambhamettu, and D.B. Goldgof. Fluid structure and motion analysis from multi-spectrum 2D cloud image sequences. In Proc. of Conference on Computer Vision and Pattern Recognition. IEEE, 2000.
[81] R.S.T. Lee and J.N.K. Liu. An automatic satellite interpretation of tropical cyclone patterns using elastic graph dynamic link model. Pattern Recognition and Artificial Intelligence, Vol. 13, No. 8, pp. 1251-1270, 1999.
[82] F. Dvorak. Tropical cyclone intensity analysis using satellite data. NOAA Technical Report NESDIS, Vol. 11, pp. 1-47, 1984.
[83] 木本昌秀. 天気予報とカオス. 合原一幸(編), 応用カオス, 第4-5 章, pp. 313-325. サイエンス社, 1994.
[84] M. Turk and A. Pentland. Eigenfaces for recognition. J. of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991.
[85] D.S. Wilks. Statistical Methods in the Atmospheric Sciences. Academic Press, 1995.
[86] H. Kantz and T. Schreiber. Nonlinear Time Series Analysis. Cambridge University Press, 1997.
[87] E.N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. Journal of the Atmospheric Sciences, Vol. 26, pp. 636-646, 1969.
[88] Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Croz, J. Du, Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. LAPACK Users' Guide. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.
[89] Kohonen, T., Self-Organizing Maps. Springer, second edition, 1997
[90] J.W. Sammon, Jr., A Nonlinear Mapping for Data Structure Analysis, IEEE Transactions on Computers, Vol. C-18, No. 5, pp. 401-409, 1969
[91] 立平 良三, 気象予報による意思決定, 東京堂出版, 1999
[92] 合原 一幸, 池口 徹, 山田 泰司, 小室 元政, カオス時系列解析の基礎と応用, 産業図書, 2000
[93] Digital Earth, http://www.digitalearth.gov/

謝辞

本研究では種々の地球観測衛星データを活用した．その中でも中心的な存在は気象衛星「ひまわり」画像である．この気象衛星ひまわり画像は，東京大学生産技術研究所のグループが受信およびアーカイブに多大な努力を払って運営しているものである．貴重なデータへのアクセスを許可くださる，東京大学生産技術研究所の安岡善文教授および喜連川優教授，根本利弘助手に深く感謝したい．ここで，本研究のように台風衛星画像を網羅的に収集するというタイプの研究には，すべてのデータがオンラインで入手できる環境が不可欠であるという事実を強調しておきたい．これがオンラインになっていなければ，人手や時間の関係から，網羅的な研究をは事実上不可能である．したがって，東京大学生産技術研究所で整備・運用されている大規模地球環境アーカイブシステムは，台風画像コレクションの構築を可能とした影の立役者であると言っても過言ではない．さらに，気象データを集約する気象庁でさえ，過去の大部分のひまわり衛星画像はテープにオフラインで保存されており，これだけの規模のひまわり衛星画像をオンラインでダウンロードできる環境は全く実現できていないことを考えれば，このアーカイブシステムがいかに多くの研究者の研究支援に役立っているかということを，ここでぜひ強調しておきたい．

さらに，東南アジアで受信する地球観測衛星データは，アジア工科大学・アジアリモートセンシング研究センターの提供を受けた．本多潔ディレクターおよび岩男弘毅シニアリサーチアソシエートに深く感謝したい．

最後に通信・放送機構には，これまで5年間の長きにわたり研究を支援して頂いた．最大の感謝を表したい．