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

（創造的情報通信研究開発推進制度）

国立情報学研究所　小野 欽司
Kinji ONO

国立情報学研究所　北本 朝展
Asanobu KITAMOTO

国立情報学研究所　計 宇生
Yusheng JI

国立情報学研究所　フレデリック・アンドレス
Frederic ANDRES

1. 研究開発の概要

本研究は、衛星観測データを中心とした地球環境データの有効活用に必須となる大規模情報基盤における基幹要素技術、すなわちネットワーク技術およびデータベース技術に関する研究を進める。特にネットワーク技術としては、大容量の地球観測データをネットワーク上で効率的に配信・検索するための技術、またデータベース技術については、大容量の地球観測データから必要な情報を効率的に発見するための技術を中心に研究を進めることを目的とする。これらの技術は、地球環境データという文脈ではまだ課題も多く、特にデータサイズの巨大さ、分散したデータベースの連携、地球環境データのデータマイニング、などの観点に残された研究課題は多い。本研究ではこれら2本柱の研究テーマに関して、地球環境データ交換のためのネットワーク、地球環境データ定義・検索言語の開発、地球環境データへのアクセスパターンのモデル化、台風画像内容検索システムの構築、台風画像コレクションのデータマイニング、などに関する成果を得た。本報告はその成果の概要をまとめたものである。

2. 研究開発の内容

2.1 地球環境データのネットワーキング技術

地球環境データのネットワーキング技術に関する研究では、地球環境データの特性を考慮したネットワーキング技術に焦点を絞り、以下の項目について研究を進めた。

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
4. 多重解像度表現に基づく地球環境データのための画像符号化

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項では、国際的規模での衛星データ交換実証実験についてその経過を報告する。次に第2項では、上記実証実験のネットワーク構成の要となるメタサーバおよびデータ定義・検索言語のアイデア、およびメタサーバのためのデータ定義・検索言語について述べる。またこのテーマは、データ検索に関係するという意味では、後述の「分散型データベース技術」研究とも関連している。さらに第3項はデータ検索と並ぶメタサーバの機能であるデータ配信について、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を用いた高速化について考察する。最後に第4項では、大きなサイズの地球環境画像データを多重解像度表現することで、解像度に基づくアクセスパターンの偏りを活用しつつ、同時に段階的な画像伝送符号化を実現する方法を検討する。

2.2 地球環境データの分散型データベース技術

地球環境データの分散型データベース技術に関する研究でも、やはり地球環境データの特性を考慮し、以下の項目について研究を進めた。

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

これらも相互に関連するテーマである。まず第1項に関しては画像内容検索技術、すなわち地球環境画像データの画像解析に基づく画像データベース検索技術について研究を進める。一方第2項では、大量の地球環境データの中から、そこに埋もれた（統計的あるいは関係的な）性質を発見するための技術を研究する。データマイニングにおいて前者の内容検索技術はその1ステップに過ぎず、むしろ各種の手法を組み合わせ適用しながら有用な知識の発見に結びつけるというプロセス全体に重点を置く。これらの研究を進めるにあたって、本研究では一般的な地球環境データを対象とするのではなく、地球環境データの中でも気象学的および社会的に特に重要な気象現象である「台風」に対象を絞りこんだ。つまり、台風画像コレクションに対する内容検索技術およびデータマイニング技術に研究を集中する。このように対象を絞ることによって、一般論にとどまらない、より具体的な成果を得ることが可能となる。

3. 研究開発実績

3.1 地球環境データのネットワーキング技術

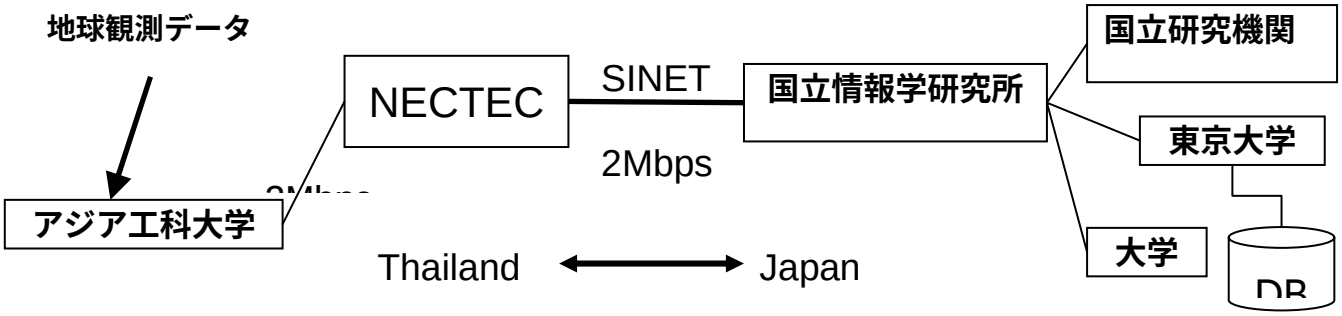

図 1　*SINET の国際回線を用いた大規模衛星データの準リアルタイム交換実証実験の構成図。*

3.1.1　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

最初に実証研究の報告として、国立情報学研究所とアジア工科大学（タイ王国）との国際共同研究に基づく、大規模衛星データの準リアルタイム交換実証実験について述べる。この実証実験では、アジア工科大学で受信するTERRA衛星MODISセンサーおよびNOAA衛星AVHRRセンサーからの受信データを、インターネット経由で準リアルタイムに交換するためのネットワーク基盤を確立する。日本およびタイで受信できる地球環境衛星データはそれぞれ観測範囲が異なるため、観測規模が国際的なものになると、これらの受信データを準リアルタイムで同期させるための国際的なネットワーク基盤が必要となる。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)のタイ国際回線を研究基盤として活用し、日本とタイとの間で地球環境データを準リアルタイムで同期させることを目指す。本実験のネットワーク構成図を図 1に示す。

実証実験の開始当初は、タイ国内の脆弱なネットワーク基盤がボトルネックとなっていたが、その後のネットワーク基盤の改善によって、現在はタイ国内でも十分高速なネットワークが稼動するようになった。その結果、国立情報学研究所とアジア工科大学との間で、おおよそ 130KBytes/sec すなわち 11.2GByte/day のビットレートで定常的に地球環境データを交換することを可能とした。このようなネットワーク基盤の構築によって、日本とタイの地球環境研究者がデータを準リアルタイムで共有しつつ、国際的な規模で連携しながら研究を進めていくことが可能となった。[2]

3.1.2　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

地球環境データベースに関しては、地球観測衛星データの受信局が世界各地に分散していることを考えると、複数のデータベースがネットワーク上に分散して存在するのが自然な構成である。このとき個々の受信局は、データの受信・蓄積・処理に責任を持ち各種データベースを提供する。しかしユーザの立場からは、分散データベースをいちいち検索するのは煩わしい。例えば、台風衛星画像とアメダス降水量データについて時刻をキーとして結合する場合に、これらが異種のデータベース（台風画像は画像データベース、降水量は関係データベース等）で提供されていれば、これらを結合して検索することは困難である。

そこで重要となるのがメタサーバの役割である。メタサーバはユーザからの検索要求に対し、個々の検索要求に応答可能なサーバに検索要求を分散し、その後複数の検索結果を結合してユーザに返す。つまりユーザから見れば、

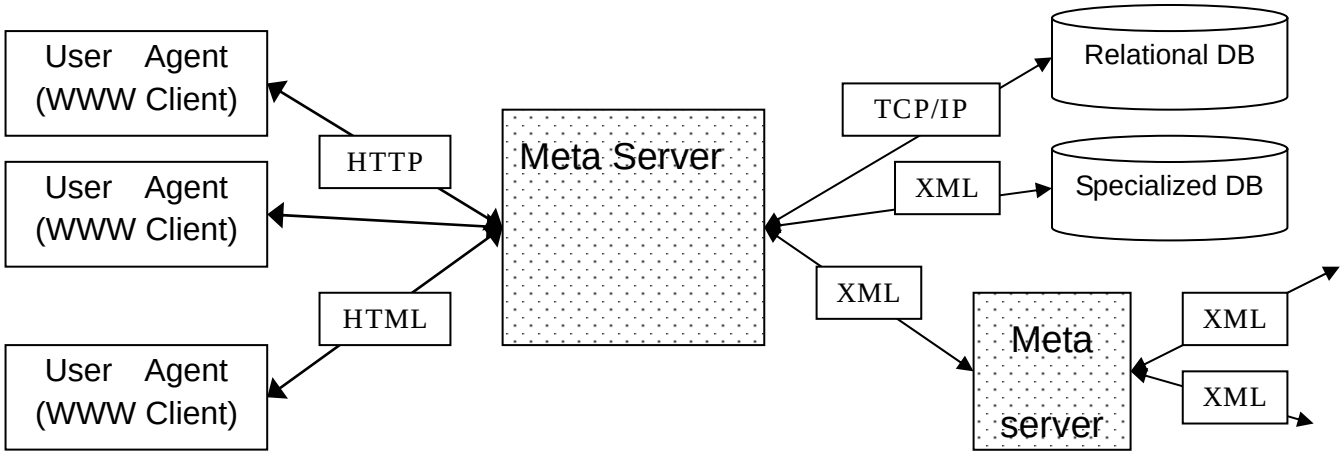

図 2　*XML に基づく分散衛星画像データベースシステムの構成図。*

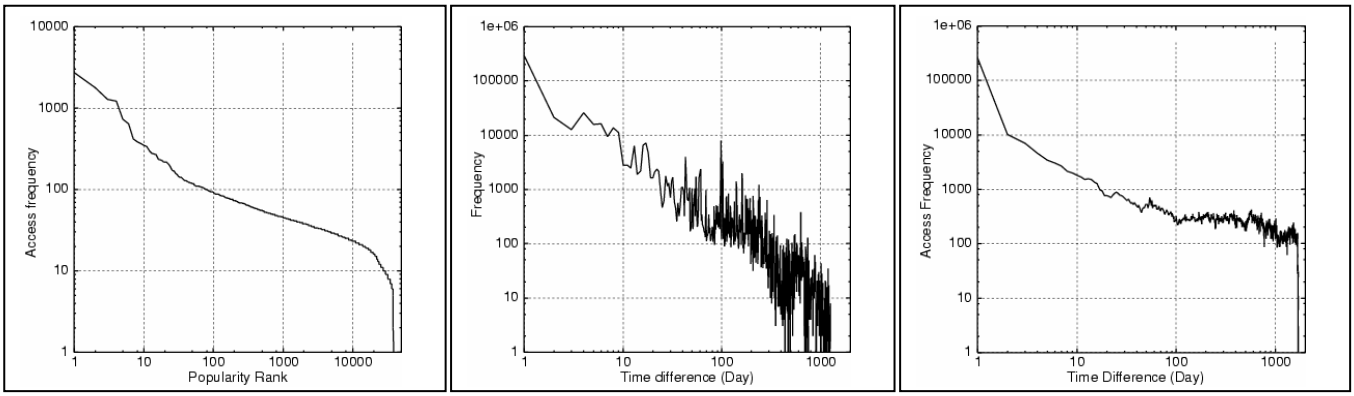

図 3　*衛星データに対するアクセスパターンのモデル化。左はデータに対するアクセス頻度の偏り、真中はデータに対するアクセス時刻間隔に対する偏り、右は観測時刻とアクセス時刻の間隔の偏り。*

データ検索が 1 箇所で完結するような、利便性の高いシステムを実現することができる。

このようなメタサーバを一般的な枠組で構築する場合には、個々のサーバが提供する情報およびインタフェースについて、サーバ間で情報を交換するためのプロトコルが必要になる。この種のプロトコルについては、メタデータ記述言語から Web Service の研究まで長い歴史と多様な試みがあるが、本研究では個々のサーバが提供する情報およびインタフェースは既知であると仮定し、その上でサーバ同士が検索要求･応答メッセージを交換するためのプロトコルについて検討する。ここでサーバとは、後述する画像内容検索技術を実装した画像検索サーバおよび関係データベースサーバを指す。そしてこれらのサーバとメタサーバとの間で検索要求･応答メッセージを交換するプロトコルを XML (extensible markup language)構文に基づき開発する。

本研究で XML 構文を用いるのは、1) XML の普及に伴いデータ互換性の面で有利になること、2) 多くの XML 関連ツールを活用するとフォーマット変換やシステム開発が容易になること、などが理由である。特に本研究では、以下の2種類のマークアップ言語およびプロトコルを定める。

1. データ定義言語 (Data Definition Language)
2. データ操作言語 (Data Manipulation Language)

まずデータ定義言語については、本研究では一般的で応用範囲の広い言語ではなく、地球環境データに対象を限定した言語を定義する。このように対象分野を限定して言語を定義するのは、地理データのマークアップ言語 GML などにも見られる一般的な傾向である。一方のデータ検索言語については、関係データベースにおける SQL (Structured Query Language) が代表的な存在であり、これは関係代数という理論的基盤を背景とすることも特徴である。しかし時系列データあるいは空間データ、マルチメディアデータなどに対する理論的基盤は現在でも確立されていない。そこで本研究では、SQL を XML 環境に拡張した XQuery あるいは XQueryX を参考にし、理論的な検討よりも実用性を重視した独自のデータ定義･操作言語を定める。

このような言語およびプロトコルを用いて、本研究のメタサーバは分散データベースと検索要求･応答メッセージを交換する。このようなメタサーバを含むシステムの構成図を図 2に示す。分散データベース間ではメッセージを以下のように交換する。

1. メタサーバは個々のサーバとの間に TCP コネクションを確立し、このコネクション上に XML 構文で記述した検索要求メッセージを送信する。
2. 個々のデータベースは同じく XML 構文で記述した検索応答メッセージをメタサーバに送信する。

現在のところこのプロトコルは単純な形式であるが、将来的には対話的なメタサーバ検索をサポートするようにプロトコルを拡張することを計画している。この考え方はネットワーク･エージェントの考え方に非常に近いものである。

*表 1*には、メタサーバから個々の検索サーバに送信するXML検索要求メッセージの例、および個々の検索サーバからメタサーバに送信するXML検索応答メッセージの例を示す。このようにXML構文を用いることで、複雑な検索要求および検索応答を統一的な形式で記述することができる。なお現在までにこのシステムは既に一部が稼動しておりWWW (World Wide Web)経由でアクセス可能である[1]。

3.1.3　キャッシング技術

次に効率的な分散データベース検索に不可欠となるデータ配信技術、具体的には図 2のメタサーバにおいて地球環境データをキャッシングし、ユーザへのデータ配信を効率化する方法について検討する。このとき重要となるのは、地球環境データへのアクセスパターンのモデル化という課題である。このときアクセスパターンがランダムでなければ、その偏りを活用したキャッシングポリシーを定め、デー

1 「デジタル台風」http://www.digital-typhoon.org/

タ配信を効率化できる。本研究では以下の4つの観点から地球環境データへのアクセスパターンをモデル化する。

表 1　XML を用いたデータ検索マークアップ言語。

Query specification

1. A single query example is chosen randomly from Typhoon 9903, and images that belong to this typhoon sequence are filtered out from subsequent tasks.
2. Images in the database are grouped by the name of typhoon sequences. Distance to the query example is calculated for each image, and images in each group are then sorted by distance in ascending order.
3. Fetch at most 2 similar images from each group. Those images are collected into the parent group, and again sorted by distance in ascending order. Finally top 5 images are fetched from the parent group, resulted in 5 most similar images in which at maximum 2 images are fetched from one typhoon sequence.
4. Return the list of similar images with the name of the typhoon sequence, the name of the image, and distance between the query example and each image, and the query example of this task.

XML encoding of a query

```
<?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>
<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>
<sort-by order="ascending">value</sort-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-each>
<task>
<let variable="value">
<function
target="example">distance</function>
<metric type="euclid" option="squared">
<min>0</min>
<max>30</max>
</metric>
</let>
</task>
<sort-by order="ascending">value</sort-by>
<fetch>
<from>0</from>
<size>2</size>
</fetch>
</for-each>
</query>
</body>
</envelope>
```

XML encoding of a result

```
<?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>
```

1. データに対するアクセス頻度の偏り
2. データに対するアクセス時刻間隔の偏り
3. データ受信時刻とアクセス時刻の間隔の偏り
4. データの解像度に対するアクセス頻度の偏り

これらを定量的にモデル化するため、本研究では東京大学生産技術研究所が提供する地球環境情報（特に衛星受信画像）サーバのアクセスログを実際に解析し、ここで収集した実データをもとにアクセスパターンを数学的にモデル化した。数学的モデルには最も単純なモデルであるジップ則 $P \sim \rho^{-\beta}$ を用い、その係数に基づきアクセスパターンの偏りを考察した結果、以下の知見を得た。

1. データに対するアクセス頻度の偏りは、2つのジップ則の混合、すなわち、頻度が大きい部分は $\beta = 0.8$ のジップ則、また頻度が中程度の部分は $\beta = 0.3$ のジップ則に近くなっている。その原因は、地球環境データへのアクセスパターンが、台風接近時のようにアクセスがバースト的に高まるパターンと、その他の定常的なアクセスが継続するパターンとの、2種類のアクセスパターンの混合であると考えればうまく説明できる。
2. データに対するアクセス時刻間隔の偏りは、$\beta = 1.3$ のジップ則に近い。とすると、個別のデータへのアクセス時刻間隔は単一のジップ則で近似できそうである。
3. 観測時刻とアクセス時刻の間隔に対する偏りは、主に 3 個の部分に分割して考察する。まず観測直後には、非常にアクセス頻度の高い期間が、1日から3日程度継続する。このように新鮮なデータへの需要が大きいのは直感的にも納得できるアクセスパターンである。このピークを過ぎると、時間の経過に伴って頻度は緩やかに減少する。この部分のジップ則は $\beta = 1.0$ に近い。最後に受信後 70 日から 100 日が経過すると、頻度はほぼ一定値となる。これはアクセスパターンがほぼランダムアクセスに近くなることを示唆している。このようなアクセスパターンは、受信時点からの経過時間でキャッシングポリシーを定めるための有用な指針となる。
4. データの解像度によるアクセスパターンのモデル化については、多重解像度表現を用いて衛星画像を表現し、解像度を指定してアクセスできる機能を実装し、その上でアクセスログを収集する必要があるが、本研究ではこの段階まで到達することができなかった。

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を複数のジップ則の混合を用いて近似することにし、その際のキャッシングポリシーについても基礎的な部分を考察した[1]。今後はこのようなキャッシングポリシーの有効性を実証することが重要な研究課題となる。

3.1.4 画像符号化技術

先述のように本研究では、データの解像度に応じたアクセスパターンの偏りをモデル化することで、キャッシングをより効率化したいと考える。ここでデータの解像度に着目する根拠は、衛星画像のように巨大な地球環境データに対しては、高解像度の原データよりも、データ内容の概略把握に便利な低解像度データへの需要の方が大きいと予想されるためである。また探索的な画像検索においては、ユーザ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て送信できる、段階的な符号化方式が有用であることも既に知られている。そこで本研究では、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て可逆な符号化を実現することを考える[1]。またこの符号化法を用いて各解像度のデータを別々に管理することで、解像度によるアクセスパターンの偏りを用いた効率的なキャッシングを実現することについても検討した。

このようなウェーブレット変換に基づく画像符号化は、本研究開始後に関連技術が進展したことで、JPEG (Joint Photographic Experts Group) 2000 という新規格に結実する見通しとなった。またこの符号化では可逆画像符号化および段階的符号化も統一的に実現されているため、現在となっては、独自の符号化を開発するよりは JPEG2000 を活用する方が有利となった。本研究では JPEG2000 のコーデックをこの目的に改良する段階までは到達できなかったが、今後 JPEG2000 がより世界に広まるにつれ、データ解像度に応じたキャッシング技術も広く使われるようになると考える。

3.2 地球環境データの分散型データベース技術

3.2.1 地球環境データベースの内容検索技術

以上に述べたネットワーク技術と両輪の役割を果たすのが、地球環境データベースの内容検索およびデータマイニング技術である。単純なテキストあるいは数値データを対象とする検索においては、データの意味は自明であることも多いが、検索対象が自然言語や画像などのようなあいまいさをもつデータになると、生のデータ自体よりもそれらを解析して得られる特徴量や意味などが重要な検索対象となることが多い。したがって、単純にデータそのものやメタデータをデータベースに蓄積するだけではなく、データが意味する内容も検索キーとして蓄積することが、画像データベースにおいて必須の技術になりつつある。

そこで本研究では、地球環境データベース、特に衛星画像データベースのモデルとして「画像内容素の階層モデル」を提案し、このモデルに基づく画像内容検索システムを実際に構築しながら研究を進めた。また衛星画像データに出現する「台風」に対象を絞り込んだ台風画像データベースシステムを重点的に研究した[4][5]。ここで台風を研究対象に選んだのは以下の理由による。

1. 気象学的にも社会的にも最重要の気象現象である。
2. パターンの変異度や観測データの充実度などの点で他の気象現象よりも研究対象として扱いやすい。

さて台風の雲パターンを表現するために、本研究では主に形状に基づくアプローチおよび全体論に基づくアプローチの二つを研究した。まず形状に基づくアプローチでは、台風の雲パターンを基本構成要素の集合に分解し、これらの空間的位置関係をグラフ構造で記述することにより、台風の雲パターンを数学的に表現する。一方全体論に基づくアプローチでは、台風の雲パターン全体を基本的な空間変動パターンの重ねあわせとして表現したり、台風の雲パターンを少数の基本パターンに代表したりさせることで、台風の雲パターンを数学的に表現する。

前者の形状に基づくアプローチは、気象学的に重要な現象を明示的に表現できるところに利点があるが、台風のような複雑な雲パターンに対して頑健な解析アルゴリズムを構成するのが難しいという欠点がある。本研究で提案するのは、楕円構成要素に基づく形状分解によって台風の雲パターンを基本構成要素の集合として表現し、その位置関係および数値属性を階層化属性つき関係グラフで表現するという方法である。またグラフ構造間の類似度を定義し、台風雲パターンをキーとする画像内容検索を実現した[3]。

一方、全体論的解析に基づく方法では、1）主成分分析を用いた台風雲パターンの固有表現、2）クラスタリングを用いた台風雲パターンの簡約表現、について研究を進めた。このアプローチは、雲パターンの形状を明示的に表現するのではなく、画素配列あるいは低レベルの特徴ベクトルの線形変換あるいは非線形変換を用いて雲パターンを表現するため、手法の頑健性の面で優れており、しかも台風雲パターンの全体的な構造を画像特徴に反映できる[4][5]。

さて全体論的解析に基づくアプローチの結果として、まず主成分分析に基づく台風表現を*図 5*に示す。この図から、台風雲パターンにおいて最大の分散を示すのは南北方向の構造であり、続いて台風のコア部分となる楕円形やらせん状のバンドパターンの部分にも大きな分散が存在することがわかった。これらのパターンは気象学的にも重要なパターンであり、その意味で妥当な結果といえる。そして*図 5*はいわば台風雲パターンの固有ベクトル表現（固有台風表現）であり、この結果に基づき固有台風表現を用いた台風画像検索システムを構築することができる。このシステムは現在 WWW で公開しており（URL については脚注1参照）、類似台風画像検索のスナップショットを*図 4*に示す。

図 4　「デジタル台風」での台風類似画像検索。

表 2　台風画像コレクションの概要。

	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 -- 180 度	東経 90 度 -- 170 度
台風シーズン	6 シーズン	5 シーズン
台風系列数	136 系列	62 系列
台風画像数	約 24500 件	約 9400 件
1系列あたり画像数	53 件 – 433 件	25 件 – 480 件

3.2.2　地球環境データベースのデータマイニング

上記の画像内容検索技術をさらに発展させ、本研究では地球環境データベースのデータマイニングにも研究を広げる。この研究テーマの目的は、大量の台風画像コレクションを対象とした機械学習を通して、2次元時系列パターンに隠された規則性を明らかにするという点にある。本研究で構築した台風画像コレクションの概要を表 2に示す。本研究で構築した台風画像コレクションは、北半球が約 24,500 件、南半球が約 9,400 件、南北両半球の合計で約 34,000 件に達する大規模な画像コレクションである。またデータマイニング研究のベンチマーキングにも使えるように前処理を工夫し、画像中心が常に台風中心と一致するような地図射影法を用いて台風画像を作成している。[4][5]

この大量の画像コレクションに対して、次にデータマイニングを適用する。最初に台風の典型的雲パターンを探ることを目的としたクラスタリングの結果を*図 5*に示す。これは空間的データマイニングの例であり、自己組織化マップを用いたクラスタリングによって、台風雲パターンの連続的な遷移を2次元平面上で眺めながら、台風雲パターンの多様性を把握することが可能となった。また時間的データマイニングの例として、台風の時間発展がカオス的であることに着目し、カオス時系列解析などを用いた台風ライフサイクルのモデル化などについても研究した。さらにグラフ理論的な手法で台風の状態遷移をモデル化する試みについても報告した[4]。

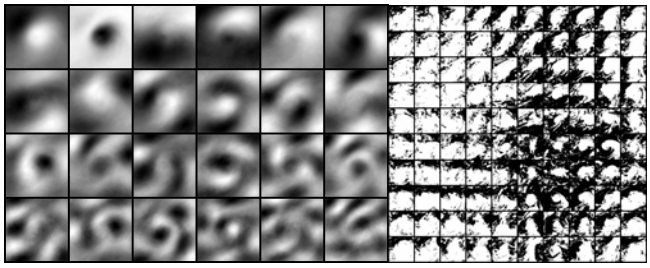

図 5 台風雲パターンの固有台風表現（左）および自己組織化マップによるクラスタリング（右）。

4.　結論

本研究は「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」と題し、地球環境データのネットワーキング技術および分散型画像データベースについて、幅広い観点から研究を進めてきた。研究課題の中には画像符号化技術の一部のように、技術の進展に伴って当初の計画が時代にそぐわない部分も生じた。しかし全体的に見れば、研究を開始した頃に叫ばれた大規模地球環境データベースの必要性は、現在もますます高まりつつあるのに対し、そのための基幹技術は未だ発展途上にとどまっている。本研究で提案した手法はその基幹技術の一角を占めうるものであり、今後の大規模地球環境データベースの構築には、これらの研究成果を有機的に結合していくことが重要な課題であると考えている。

謝　辞

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きました。深く感謝いたします。

文　献

[1] Kitamoto, A. "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

[2] Kitamoto, A. and Ono, K. “The Collection of Typhoon Image Data and the Establishment of Typhoon Information Databases Under International Research Collaboration between Japan and Thailand,” *NII Journal*, No. 2, pp. 15-26, 2001.

[3] 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

[4] Kitamoto, A., "Spatio-temporal Data Mining for Typhoon Image Collection", *Journal of Intelligent Information Systems*, Vol. 19, No. 1, 2002 (in press)

[5] Kitamoto, A., "IMET: Image Mining Environment for Typhoon Analysis and Prediction", *Multimedia Data Mining*, Djeraba, C. (編), Kluwer Academic Publishers, 2002 (in press)