分散環境における透過的データ利用の実現

植田和憲[†1] 神矢健一[†2]

学際的な研究プロジェクトを遂行するにあたり、他機関の研究者と協力して研究を進める場合がある。このようなとき、各研究者が保持する情報は物理的に分散しており、かつ、研究領域によってはその量が膨大となるのが普通である。情報ネットワーク技術が発達し、遠隔地の情報を活用できるようになってきているとはいえ、現状では膨大なデータを拠点間で転送するにはかなりの時間を要する。そこで、本研究グループでは、分散した拠点に散在するデータを透過的に扱うための枠組みを提案する。ここでの透過的とは、データが物理的に分散していることを意識せずに済むことを指す。

Transparency for Data Access on Distributed Network Environments

Kazunori Ueda [†1] and Ken-ichi Kamiya[†2]

In some occasions, interdisciplinary research project consists of researchers of several institutes in the world. In such case, data used by the researchers are distributed at the institutes. Although computer networking technologies have been progressed in recent years, it takes long time to transmit huge data such as 3D modeling data or high resolution movies. To solve this problem, we propose a new network environment that enables transparency of distributed database systems.

1. はじめに

高知工科大学総合研究所では、学術フロンティアプロジェクト “博物資源工学に基づく脳と知の共進化に関する実証的研究”（2006 年–2009 年）を推進した。このプロジェクトは、ネアンデルタール人・プロトクロマニョン人・クロマニョン人の三者の系統関係について、脳や全身骨格、運動機能、成長、言語能力等をコンピューターサイエンスの先端技術によって解明、復元することで人類の起源について検証し、旧人・新人交代劇の原因に迫ろうとするものである。具体的には、西アジア死海地溝帯における旧人ネアンデルタールと新人ホモ・サピエンスの交代劇の真相追求を行うものである。このプロジェクトにおける本学の役割とは、これまで培ってきた分析・解析技術を異分野の研究へ適用することで、工学のさらなる高度化および汎用化を目指し、さらに博物資源の学術価値の強化や考古学、人類学などへの貢献を果たすことである。具体的な技術としては、博物資源情報データ入出力に関するインタラクション技術開発、化石化した頭蓋骨の地層内変成過程の加速再現による古人類頭蓋骨の再生、光ファイバによるセンシング技術の応用、断片情報に基づく三次元可視化システムの開発、広域分散環境におけるデータ利用効率向上手法の提案、などである。

このような学際的な研究プロジェクトでは、各研究機関に存在する研究用データからなるデータ利用環境を構築することがある。これらのデータには、テキストファイルといった数 KB のデータや三次元画像や映像といった数 MB から数 GB ものデータなどが含まれる。この研究プロジェクトには、人類学や工学といったさまざまな分野の研究者が参加しており、各研究者が扱うデータの種類も多岐にわたっている。遺跡より発掘された旧人の骨の化石資料などを解析した CT スキャンや MRI の画像データなどは容量も大きく、研究者間の相互利用を難しくしている。このことからも分かるように、データ利用環境に含まれるデータは各拠点に存在するため、統合的に利用するためのシステムを構築する必要がある。

本研究の目的は、高知工科大学総合研究所において行われているプロジェクトの一環として構築される分散データ利用環境における大容量データの転送時間に関する問題を解決し、データ利用者に対し透過的なデータ利用環境を提供することである。ここで言う透過的な利用環境とは、集中データストレージに近い使用感が得られるような環境のことであり、本研究では、データの転送時間の短縮による実現を目指す。具体的には、研究内容やアプリケーション種別に基づいて今後利用すると予想されるデータをあらかじめ所属する拠点に転送しておき、利用時のデータ転送回数を減少させる。また、一度使用したデータを手元に保存しておくことで使用頻度の高いデータの重複転送を回避する。そして、これらを組み合わせることによってデータ利用者のデータ転送に伴う待機時間の短縮を実現する。

†1 高知工科大学
Kochi University of Technology
†2 株式会社ネオシステム
NeoSystem Co.,Ltd

2. 分散データアクセス技術

本研究では、分散データ利用環境における透過的なデータ利用の実現を目的としている。分散したデータを統合的に扱うものとして、分散データベースシステム、グリッドコンピューティング技術などが挙げられる。また、遠隔地に存在するデータ利用の効率化とデータストレージの問題を解決するものとしてコンテンツの先読み (prefetch) とキャッシュ (cache) とが挙げられる。以下、それぞれの詳細について述べる。

2.1 分散データベースの透過性

分散データベースとはネットワークを介して物理的に分散した複数の集中データベースを接続し、ユーザにはあたかもひとつの大きな集中データベースであるかのように利用できるようにしたものである。ユーザにデータベースの物理的分散を意識させず、集中データベースであるかのように利用できることを透過的であるといい、分散データベースシステムは透過性を実現するためにいくつかの機能を提供する。分散データベースシステムが提供する透過性の種類としては、位置に対する透過性、移動に対する透過性、分割に対する透過性などがある[3)]。

位置に対する透過性とは、データベースがネットワークにより接続されて物理的に分散していることを、ユーザに意識させないで利用可能とすることである。移動に対する透過性とは、運用の都合や性能向上の目的で分散されたデータや表の格納サイトが変更した後も、ユーザにそれらの移動を意識させることなく業務プログラムや操作手順を変更しなくても、移動したデータにアクセス可能とすることである。分割に対する透過性とは、一つとして表現されるデータが複数のデータベースサイトに分割して格納されていても、ユーザに意識させることなく利用可能とすることである。分散データベースが提供する透過性の種類には、さらに、重複に対する透過性、障害に対する透過性、データモデルに対する透過性などがある。

いずれの透過性も「ユーザがデータおよびデータベースの分散を意識しなくてもよい」ように利用できる環境を目指しており、分散データベースシステムにおける透過性を実現することは非常に重要である。

2.2 グリッドコンピューティング

グリッドコンピューティングとは、複数のコンピュータをネットワークを介して結合することで仮想的な一つの高性能コンピュータとみなし、その仮想コンピュータから必要なときに必要なだけ、処理能力や記憶容量を取り出して利用可能とするための次世代インフラである。グリッドコンピューティングにおける “The Grid” という名称の由来は「電力網」である。我々が電気を使用する際、その電気がどこで発電されどのような経路をたどって輸送されてきたのかを気にする人はなく、同様にコンピューティグパワーも、自由に簡単にいつでもどこでも必要なだけ使えるようにしたいという意味が込められている[5)]。

データグリッドは分散したストレージをまとめて使う利用形態である。膨大なデータを地理的に分散した場所に分割して保存し、それらを効率的に統合し、データに対する処理も効率化することを目的としている。仮想化する対象によって、データベースレベルの仮想化、ファイルレベルの仮想化、ブロックレベルの仮想化などに分類できる。データグリッドのように、分散したデータを統合的に利用するプロジェクトもいくつか進められてきた。

GEO Grid とは、Global Earth Observation Grid（地球観測グリッド）の意味で、グリッド技術を用いて、地球観測衛星データの大規模アーカイブ・高度処理を行い、さらに各種観測データベースや地理情報システム (GIS：Geographic Information System) データと融合し、ユーザが手軽に扱えることを目指したシステムである[4)]。大規模な衛星データに対応した高度な処理技術、協力機関とのセキュアな相互運用性、多様なユーザに対するセキュリティの維持を可能とするシステムを開発し、標準的な Web サービスのインターフェースを使用することで、ネットワーク上に分散する各種地球観測データと大規模な衛星データとの統合利用の実用化研究を行っている。

バイオグリッド・プロジェクトは、文部科学省における IT プログラム「スーパーコンピュータネットワークの構築」として平成 14 年度より 5 年間の研究プロジェクトとしてスタートしたグリッドプロジェクトである[2)]。スーパーコンピュータネットワーク上に分散された観測装置やデータベースを、統合的かつ安全に利用できるデータグリッド技術を開発するとともに、それぞれのデータベース間での真に有機的な連携利用、極めて高速な計算資源を必要とするデータ処理を橋渡しするプロセッシンググリッド技術を開発することを目的としている。

2.3 コンテンツの先読みとキャッシュ

分散データ利用環境におけるデータ取得時間を短縮することにより、ユーザにはデータが物理的に分散していることを意識させないようにする。これは透過的にデータを利用できる環境を実現したといえる。具体的には一度使用したデータを手元に保存、また近い将来に参照すると予想されるデータをあらかじめダウンロードしておくことによって、ユーザのデータ取得時間の短縮を図る。先読みとキャッシュという技術を併用することで透過的環境を実現できると考えられる。

先読みとは、ユーザがネットワークを使用していないうちに、近い将来アクセスが予想されるデータをあらかじめダウンロードしておく技術である。先読みをしておくことで、ユーザのデータへのアクセス時間を短縮することができる。先読みしたデータはキャッシュに保存しておく。ユーザがアクセスを要求したデータがすでに先読みされていれば、ネットワークを使ってダウンロードする時間が必要なく、すぐにそのデータにアクセスできる。本研究においては、参照順の傾向を指定しておき、その指示に従って「ユーザが使用中のデータに関連し、且つ近い将来にアクセスが予想されるデータ」をネットワークのアイドル時間を利用して先行ダウンロードしておくものである。

キャッシュとは、一度使用したデータを、データを使用する場所に近い高速な記憶装置に保存しておくことで、二回目以降のデータへのアクセス時間を短縮する技術、または装置のことである。キャッシュはさまざまな場所で使われており、典型的な例が Web で使用されているキャッシュ技術である。Web ブラウザは一度ダウンロードしたページのデータをハードディスクに保存し、二回目以降のサーバアクセスを高速化している。プロキシサーバでもキャッシュ技術が使用されており、一度アクセスしたページをサーバ内のハードディスクに保存し、二回目以降はインターネットにアクセスせずにページデータを渡すことができる。本研究においては、一度使用したデータを、ユーザが初めに接続するコンピュータ（キャッシュサーバ）に一時的に保存しておくことである。

また、CDN (Contents Delivery Network) 技術も負荷の集中回避とネットワーク資源の有効利用のために用いられている。この技術は、ユーザに広く要求される Web コンテンツなどを分散したサーバ上に複製しておくことで、単一のサーバへのリクエストの集中を防ぐことができる。ユーザからのリクエストはユーザの属するネットワーク上の位置によってネットワーク上の距離が近いコンテンツサーバへ振り分けられるようになっており、結果としてネットワーク全体のネットワーク資源を節約しつつ高速にコンテンツをダウンロードさせることが可能になる。

3. 透過的データ利用環境の実現

分散データ利用環境において考慮すべき点には以下のようなものがある。

- 利用するデータの種類と量
- データの利用傾向
- 利用者の拠点からデータの所在までのネットワーク上の距離

これまでに、分散したデータを統合的に利用するための技術が多く提案されているが、学際的なプロジェクトの推進のためのデータ利用環境として用いるためには必ずしも最適ではない。学際的な研究プロジェクトでは異分野の研究者が同一のデータに関連した研究を行うことがあるため、その組み合わせによって上記で述べた要件が異なる場合がある。そのため、さまざまなデータ利用形態に合わせて適当な効率化手法を採用する必要がある。そこで、本研究グループでは、分散データ利用環境において考慮すべき要件の内容の変化に対応できるシステムを提案する。

3.1 システム設計

提案システムでは、拠点間データ転送を管理することで透過的なデータ利用環境を提供する。具体的には、利用するデータの実体ができるだけ拠点内ストレージへ存在している状況となるように、拠点内ストレージにデータ転送管理機構を組み込む。拠点内ストレージには、利用可能なデータあるいはその参照先が格納されているとし、通常の集中ストレージを利用するときのような使用感を得られるようにする。実データをすべて拠点内ストレージに格納するのは総データ量が膨大である場合現実的ではないが、その参照先データのみであれば格納することができると思われる。利用するファイルシステムやアプリケーションにもよるが、参照データを実データであるかのように扱うことができれば、その区別をすることなく基本的には利用可能である。拠点内ストレージに組み込まれるデータ転送管理機構には、先読み機能とキャッシュ機能が実装され、ユーザの利用形態や利用状況に合わせて動作する。データ利用者は拠点ごとに存在する拠点内ストレージを介し仮想集中ストレージ内のすべてのデータへアクセスするが、その際に、参照情報のみで実際にネットワーク上での転送を伴った場合は該当データをキャッシュとして保存し、キャッシュの内容も加味してユーザが将来アクセスするであろうと予想されるファイル群をアイドル時間を利用して先行コピー（ダウンロード）する。図 1 に全体図を示す。

これまでに述べたように、データ利用者である研究者および研究内容によってデータへのアクセス傾向は異なると考えられる。つまり、データの利用形態は利用者・研究内容・アプリケーションによってさまざまであり、最適なキャッシュアルゴリズムや先読みアルゴリズムを一元的に決定することは不可能である。よって、アクセス傾向とそれに合わせた先読みアルゴリズム・キャッシュアルゴリズムをモジュールとして用意し、利用者やアプリケーションなどのそれぞれの利用形態に合わせてアルゴリズムを変更するアプローチを採用する。利用形態や傾向がある程度判明しているものについてはそれに合わせたアルゴリズムを設計することが可能であると考えられるが、そうでない場合は一般的な分散データアクセス技術で用いられているものを汎用アルゴリズムとして組み込んでおく。データ利用環境の利

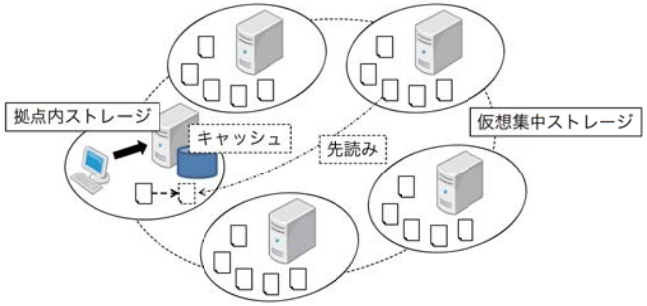

図 1　提案システム概要
Fig. 1　Proposal system design

用が進み、新たなアルゴリズムの採用が望まれれば、それに先立つデータ利用傾向を分析しそれに応じたアルゴリズムの設計及び実装を行い、モジュールとして追加すればよい。

3.2　プロトタイプシステムによる検証実験

キャッシュアルゴリズム・先読みアルゴリズムは利用形態に合わせたものを用意する必要がある。そこで、提案システムのプロトタイプとして、同じプロジェクト内の他の研究チームによって行われている研究で扱うデータを対象としたシステムを構築した。研究内容は、“化石化した頭蓋骨の地層内変成過程の加速再現による古人類頭蓋骨の再生” である。この研究は、地層内に埋没した頭蓋骨を対象として、温度や圧力の変化によって加速再現を行うものである。加速再現を試みた頭蓋骨のモデルの観測には CT スキャンを用いる。使用するアプリケーションは、CT スキャン画像を処理するものでデータ量が多く、ネットワークを介して利用するのには長い時間がかかるものである。このアプリケーションでは、CT スキャンの画像を順に読み込み、それらのデータを再構成して三次元モデルを作成することができる。図 2 にアプリケーションの使用中の画面を示す。

今回のプロトタイプの概要を図 3 に示す。データ利用者およびアクセス先である拠点内ストレージは同一ネットワークにあり、全実データが存在するストレージは別のネットワークにあるとみなすため、それぞれのストレージを接続するサーバ同士を峡帯域のネットワークで接続した。プロトタイプ上では、拠点内ストレージにはすべての利用可能なデータの参照先が格納されており、基本的にデータ検索などを行わずにデータが利用できるものとする。データ利用者は拠点内ストレージ上のデータをアクセスするようにアプリケーションを

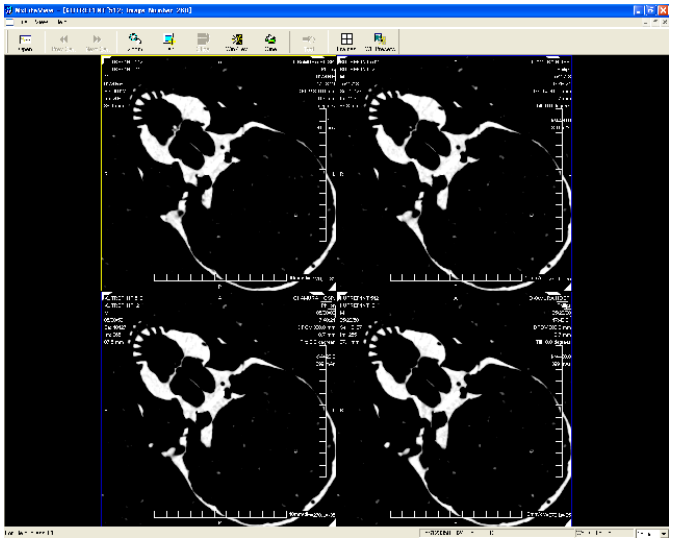

図 2　CT スキャン画像
Fig. 2　Graphics with CT scan

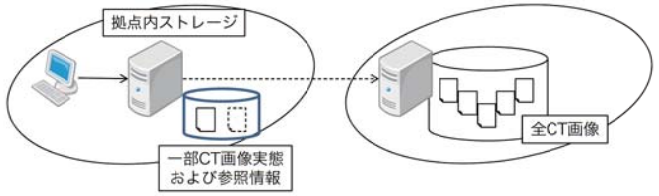

図 3　プロトタイプシステム
Fig. 3　Prototype system

利用するものとする。これは、データ位置に関する透過性についても考慮するためであるが、用いるアプリケーションにも特別な工夫を必要としないというメリットもある。

このプロトタイプシステム上において、先読みアルゴリズムとして文字列の類似度を考慮するアルゴリズムおよび連番のファイル名を想定したアルゴリズムを、キャッシュアルゴリズムとして最終アクセス時刻を考慮するものを用い、実際にデータが利用可能となるまでの

時間を測定した。

文字列の類似度判断にはレーベンシュタイン距離[1)]を基にする方法を用いた。レーベンシュタイン距離とは、2つの文字列がどの程度異なっているかを示す数値である。具体的には、二つの文字列において、文字の変更、削除、追加の操作を行い同一の文字列にすることができる最小の操作回数である。今回は参照中のファイル名とのレーベンシュタイン距離が一番小さなファイルを先読みさせるものとする。これにより、ユーザが参照中のファイルの名前を基に先読みデータ（ファイル）を決定し先読みするシステムとなる。さらに、ファイルの命名規則が連番であるような環境を想定し、ファイル名に含まれる数字に対して昇順または降順にアクセスされるファイルを予測するアルゴリズムも用いた。これは、実験で使用するアプリケーションがファイルを連番で作成するという特徴があることを把握したうえで用意したものである。アルゴリズムを複数用意したのは、ファイルアクセスの傾向が異なる場合にはそれに応じてアルゴリズムを追加するということを想定したものである。

先読みされた実データがキャッシュに蓄積されていくと、やがてキャッシュ容量が限界になる。新たに先読みされたデータによってキャッシュ容量の限界を超えた場合、アクセスされた時刻がもっとも古いキャッシュデータを削除していき、先読みしたデータを保存する。削除されたキャッシュデータは再び参照先を示すデータに置き換える。今回の実験では、ユーザがアクセスしたディレクトリとその日時をキャッシュリストとして記録しておき、キャッシュの個数が一定数を超えた場合、キャッシュリストの記録を基にアクセスした日時が最も古いキャッシュデータを消去していくこととした。

3.3 実験結果

提案システムを用い拠点内ストレージによってデータを利用することができるかどうかの検証実験を行った。また、今後の参考のため、プロトタイプを用いた場合、通常のファイル共有によって各データにアクセスした場合、データをローカルに配置した場合とでデータが利用可能になるまでの時間を計測した。

CT スキャンによって得られた頭蓋のスライスの画像から構成されるデータセット A（533KB × 909）を用いたとき、先読みとキャッシュをいずれも用いない場合約 150 秒、先読み（レーベンシュタイン距離によるもの）とキャッシュを併用する場合は約 49 秒、ローカルでは約 27 秒であった。また、CT スキャンの画像を基にして作成された 3 次元画像のデータで構成されるデータセット B（774KB × 40）の場合、先読みとキャッシュをいずれも用いない場合約 8 秒、、先読み（レーベンシュタイン距離によるもの）とキャッシュを併用する場合約 3 秒、ローカルでは約 2 秒であった。このとき、別で行った先読みに連番のファイルを昇順でアクセスするアルゴリズムを利用した場合でも先読みにレーベンシュタイン距離を利用するアルゴリズムを用いた場合とほとんど転送時間が同じであった。これは、レーベンシュタイン距離の小さいファイルへアクセスする場合でも連番のファイルへの昇順でアクセスする場合でも同一のファイルを先読みしたため差が出なかったものと思われる。

これらより、提案システムのプロトタイプとして想定する機能が動作しており、結果として差異はなかったが異なるアルゴリズムを同一のシステム上で利用することができることが確かめられた。しかし、依然として拠点内ストレージを介する方法では拠点内の転送速度の影響を受けるため、ローカルに近い利用環境の提供のために利用者のコンピュータへの機能実装や高速なネットワークインフラの整備が必要であることを再確認した。

4. おわりに

学際的な研究プロジェクトを推進するにあたり、各研究者の持つ資料などの研究データを相互に利用するためには、物理的距離あるいはネットワーク上での距離をできるだけ感じさせないデータ利用環境を提供することが重要である。本研究では、データへのアクセス高速化技術として、先読みとキャッシュを組み合わせることで効率的なデータ利用が可能なシステムを提案した。また、提案システムの動作を検証するために仮想的な分散データ利用システムのプロトタイプを構築し検証実験を行った。その結果、提案システムにおいて分散したデータに対する透過的なデータ利用環境を実現可能であることが分かった。さらに、採用する先読みアルゴリズムを交換してシステム上で利用できることも分かった。

今後の課題として、プロジェクトにおけるデータの利用形態を調査し、それぞれの場合に合わせた先読みアルゴリズム・キャッシュアルゴリズムを設計することが非常に重要である。さらに、データ利用者あるいは利用形態の違いの判別方法やシステムのローカルコンピュータへの適用方法の検討なども課題として考えられる。

謝　辞

この研究は、日本私立学校振興・共済事業団から私立大学等経常費補助金の特別補助によって一部援助を受けた。

参考文献

1) Levenshtein, V.I.: Binary Codes Capable of Correcting Insertions, Deletions and Reversals, *Cybernetics and Control Theory*, pp.707–710 (1966).

2) バイオグリッドプロジェクト：BioGrid. available at http://www.biogrid.jp/project/.
3) 疋田定幸：図解 分散型データベースシステム入門，株式会社オーム社，東京 (1989).
4) 独立行政法人 産業技術総合研究所 GEO Grid プロジェクト：GEO Grid. available at http://www.geogrid.org/jp/.
5) 日本アイ・ビー・エムシステムズ・エンジニアリング株式会社：グリッド・コンピューティングとは何か，ソフトバンク　パブリッシング株式会社，東京 (2004).