マルチメディア通信と分散処理 66－2
（１９９４．７．７）

分散環境における
部分データベースの複製配置法

高品 智一　宮西 洋太郎　渡辺 尚　水野 忠則

静岡大学

分散システムにおいて、回線やサイト等が故障してネットワークが分断された場合に、ある程度サービス品質が低下しても、そのネットワークのみで運転を継続することが望ましい場合がある。そこで、孤立したサイトで運転を継続するという観点からのシステム構成方式を確立することを目指した提案を行う。

ここでは、データベースの複製をとり、それをいくつかの部分データベースに分割し、データアクセスの地域的特性等の統計をもとにしてサイトへの配置を行う。それによって、孤立運転時における、ある程度までのサービス品質を維持しようと試みている。

Allocation methodologies of replicated partial databases in distributed environment

Tomokazu Takashina　Yohtaro Miyanishi
Takashi Watanabe　Tadanori Mizuno

Faculty of Engineering, Shizuoka University

3-5-1, Johoku, Hamamatsu, 432 Japan

In distributed system, the system had better run in a small domain, even if it is isolated because lines or sites's failure. This paper proposes to establish the system construction methodologies, which can perform restricted services in an isolated domain.

In this approach, database is replicated, it's DB divides into some partial databases, and they are allocated to each sites based on statistcal information of a local property for data access and so on. It makes an attempt to maintain restricted quality of service in an isolated domain.

1 はじめに

分散システムにおいては、位置透過性、アクセス透過性、複製透過性、障害透過性等により、分散システム全体の構成方法等をユーザ / 開発者は知る必要がない [1, 2]。通常、個々のサイトは自己以外のサイトと協調しながら正常な動作を行っている。また、自己と他とを接続する回線や他のサイト (特に中央のホスト) 等が故障してネットワークが分断された場合には、ある程度サービス品質が低下しても、そのネットワークのみで運転を継続することが望ましい場合もある。このため、データベースをはじめとする資源を有効に活用する研究がなされてきている [3, 4]。ここでは、孤立したサイトで運転を継続するという観点、即ち独立性の観点からのシステム構成方式を確立することを目指した提案を行う。

2 対象モデルと検討課題

2.1 対象データのモデル

対象となるデータは、データアイテム単位にアクセスされ、また1回のアクセスは地域的に分散した複数のサイトの中の1つを用いて行われるものとする。

また、センターには全体のデータを持つデータベース (以下 DB と略) が存在し、分散したサイトには、複製データを持つものとする。

2.2 対象データの特性

データの複製を行う際、対象データの特性を利用する。データ特性は、処理対象によって様々な特性を持つであろうが、一般的には地域的特性及び時間的特性を有すると考えられ、その統計的特性を利用することを提案する。

- 地域的特性 (局所性)
  地域的特性の表現方法として、データアイテムごとの、データアクセスを行ったサイトとの関連の強さ、即ち条件付き確率が考えられる。

- 時間的特性
  時間的特性の表現方法として、データアイテムごとのアクセス発生率、更新周期ごと (時、日、月等) のデータ、時系列履歴データ、バージョン管理等が考えられる。ここではアクセス発生率を考察の対象とする。

次に、上記のデータ特性を数式的に表現する。データアクセスが単位時間 $[Sec]$ 内に発生するとして、それがデータアイテム i についてであり、かつサイト j において発生する発生率 $\lambda(i,j)[1/Sec]$ をここでは一致発生率と称することにする。これは、

$$\lambda(i,j) = P(j|i) \cdot \lambda(i) \qquad (1)$$

であり、一致発生率は、データアイテム i の発生率 $\lambda(i)[1/Sec]$ とそのアクセスが j で行われるという条件付き確率 $P(j|i)$ との積で表される。地域的特性を $P(j|i)$ で、アイテムに対するアクセスの発生度合を $\lambda(i)$ で把握することにする。

2.3 データベースの水平分割、垂直分割

データベースは通常、一元管理されている。しかし、分散システムにおいて効率および耐故障性をあげるために、データベースの複製を持つことがある。[5] この複製は元のデータベースの完全な複製である場合もあるが、資源の許容量 (ディスク) や一貫性維持のためのコストから、データベースの一部だけの場合 (部分データベース) もあると考えられる。ここで、そのデータベースを分割する方法には大きく分けて2種類あると考えられる。つまり、水平分割と垂直分割の2種類である。次にリレーショナルデータモデルにおけるそれらの分割を示す。

- 水平分割
  タプルを単位としたデータの分割である。つまり、先に述べた地域的特性、時間的特性などにより、あるサイトで主に使うタプルが分かっているような場合に適用する。下図の例では、所属が '企画' であるタプルを集めることによってできる部分 DB を考える。この部分 DB を '企画' 課のサイトに配置することで、少ない資源消費で効率良く性能を向上させることができる。

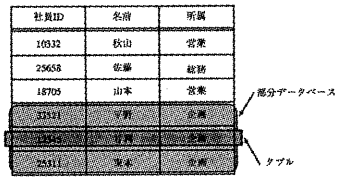

図 1: 水平分割の例

- 垂直分割
  属性を単位としたデータの分割である。これは属性によって使用される頻度に違いがあるので、使用頻度の高いものだけを部分DBとするものである。下図の例では、使用頻度の高い、'社員ID'と'名前'からなる部分DBをセンターのDBの複製としている。これにより複製にかかるコスト(資源)をおさえて、ある程度までのサービスを提供可能となる。

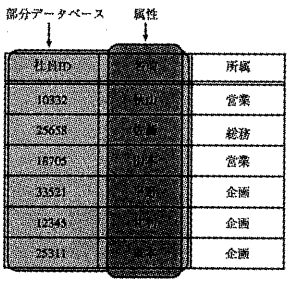

図 2: 垂直分割の例

また、これらの分割を組み合わせることによって、さらに性能を高めることができると考えられる。今回のモデルでは、それらのことを考慮して、データの利用度をアクセスの発生率 $\lambda(i)$ で表している。これによって、部分DBを作る際の指標を可能とした。

2.4 検討課題

以上のようなことを念頭において、検討すべき課題は、

- 複製データベースの分散配置方式
- 孤立運転時のサービス品質
- サービス品質と複製データベース維持負担とのトレードオフ

等についての定量的判断基準を確立していくことである。

3 複製データベースの分散配置方式

対象データの地域的特性に応じて、DBを分割し、サイトに配置するものとする。分割方式、配置方式にはいくつかの方式が考えられる。

- DB分割方式

1. 分割無し
   DBを分割せずに、DB全体を複製する。複製維持や資源にコストがかかるが、信頼性やサービス品質は向上する。
2. 重複を許容する分割
   DBを分割する際に、ある部分DBと別の部分DBとの間に、重複を許す。これにより、その重複した部分のサービス品質を高めることができる。しかし、その複製維持が複雑になりやすい。
3. 重複を許容しない分割
   部分DB間の重複を許さない。つまり、部分DBに分割するときに、DBの境界をはっきりと区別する。これにより、各々の部分DBは単体のDBのように扱うことができる。

- サイトへの配置方式

A 全てのサイトに配置
  複製されたDBを全てのサイトに配置する方式である。複製維持や資源にコストが非常にかかるが、信頼性やサービス品質は向上する。

B 特定のサイトに配置
  ある特定のサイト、例えば、各ネットワークごとで処理能力の高いサイトに配置する方式である。これは、比較的簡単に信頼性やサービス品質を向上できる方式であり、またサイトが特定されているため管理が容易である。

C アルゴリズムによるサイトへの配置
  本研究で主眼となっている方式で、地域的特性など統計をとり、それによって複製DBをサイトへ配置するものである。

これらの方式の組み合わせが考えられ、次のページの表のようになる。

これらの組み合わせからどれを選択するかは、データ特性、要求サービス品質、複製維持負担の総合的評価をすることが望ましいが、簡略のためデータ特性のみによる方式選択、特にここでは「重複を許容

		DB 分割		
		DB 全てを分散配置	重複の有る部分 DB を分散配置する	重複の無い部分 DB を分散配置する
サイト配置	全てのサイト	A1	A2(A1 と同じ)	A3(A1 と同じ)
	特定のサイト	B1	B2	B3
	判断アルゴリズムによるサイト	C1	C2	C3

表 1: DB の分割と配置の組み合わせ

しないアルゴリズムを使用したサイトへの配置方式 (C3)」について検討し、判断方法を提案する。

4 重複を許容しないアルゴリズムを使用したサイトへの配置方式 (C3) 選択方法

当該方式を選択する条件を考察する。サイト総数を S とすると、このアクセスはいずれかのサイトにおいてなされるので、

$$\sum_{j=1}^{S} P(j|i) = 1 \qquad (2)$$

これらの関係を図に示す。

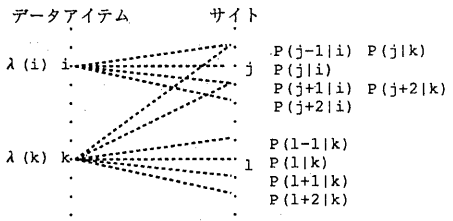

図 3: データアイテムとサイトとの関係

1. 地域的特性ありの判断
   地域的特性が全く無ければ、$P(j|i) = 1/S$ であるが、現実においては、厳密には $P(j|i) = 1/S$ となることはほとんどなく ($P(j|i)$ の推定値を求める意味でも)、多くの場合には多少とも地域的特性が存在する。従っていくつかの判断アルゴリズム案が考えられるがここでは、以下の案を提案する。

   (a) 各 i について、
   j について、$P(j|i)$ の標準偏差 σ_i を計算する。

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{S}\left(P(j|i) - \frac{1}{S}\right)^2}{S}} \qquad (3)$$

   (b) これを i について発生率で加重平均した標準偏差 σ を計算する。
   データアイテムの総数を I とすると、

$$\sigma = \frac{\sum_{i=1}^{I}(\sigma_i \cdot \lambda(i))}{\sum_{i=1}^{I}\lambda(i)} \qquad (4)$$

   (c) σ が S、I で定めた閾値 (threshold) を越える場合には地域的特性ありと判断する。

2. 複製データの配置サイトの決定

   (a) 複製データを持たせるか否かの判断
   アクセス発生率 $\lambda(i)$ が閾値を越えるデータアイテム i については複製を持つことにする。

   (b) 複製を配置するサイトの決定
   各 i について、$P(j|i)$ の大きい順に j を並べ直す。並べ直した j を $j_{1i}, j_{2i}, \cdots$ とする。すると大きい順に並べ直した $P(j|i)$ は、$P(j_{1i}), P(j_{2i}), \cdots$ となる。上位のものから、一定の閾値 P_0 を越えるまで加算する。

$$\lambda(i) \cdot \sum_{r=1}^{R_i} P(j_{ri}) \geq \lambda_m \cdot P_0 \qquad (5)$$

$$\lambda_m = \frac{1}{I} \cdot \sum_{i=1}^{I} \lambda(i) \qquad (6)$$

   ここで λ_m は平均発生率である。
   上記の式が成立する最初の r を R_i とし、このときのサイト $j_{1i}, j_{2i}, \cdots, j_{R_i i}$ に複製データを配置する。また、そのとき複製の

重みとしての値 $w(j_{ri})$ を $P(j_{ri})$ に比例させて定義する。

$$w(j_{ri}) = \frac{P(j_{ri})}{\sum_{r=1}^{R_i} P(j_{ri})} \quad (r = 1, 2, 3, \ldots, R_i) \tag{7}$$

重みの用途は後に述べる。(DB 更新の限界値等に使用)

データアイテム	複製配置サイト	サイト個数
1	$j_{11}\ j_{21}\ j_{31}\ j_{41}\ \cdots\ j_{R_1 1}$	R_1 個のサイト
2	$j_{12}\ j_{22}\ j_{32}\ j_{42}\ \cdots\ j_{R_2 2}$	R_2
3	$j_{13}\ j_{23}\ j_{33}\ j_{43}\ \cdots\ j_{R_3 3}$	R_3
4	$j_{14}\ j_{24}\ j_{34}\ j_{44}\ \cdots\ j_{R_4 4}$	R_4
・	・	・
・	・	・
i	$j_{1i}\ j_{2i}\ j_{3i}\ j_{4i}\ \cdots\ j_{R_i i}$	R_i
・	・	・
・	・	・
I	$j_{1I}\ j_{2I}\ j_{3I}\ j_{4I}\ \cdots\ j_{R_I I}$	R_I

表 2: データアイテムと複製配置サイト

3. 複製維持のための負担

複製を維持するための記憶容量、処理資源を評価する。

- 記憶容量の負担

  各々のサイト j に複製される複製データ容量を求める。上記の表において、j_{ri} が j に一致するデータアイテムのサイズの合計を求める。

  サイト j における複製データ量 c_j は、データアイテム i のデータ長を l_i として、

$$c_j = \sum_{i=1}^{I} \sum_{r=1}^{R_i} \delta_{jj_{ri}} \cdot l_i \tag{8}$$

  ただし、$\delta_{jj_{ri}}$ は $j = j_{ri}$ ならば1、そうでなければ0の値をとるものとする。c_j に単位長あたりの負担 (コスト) を乗じることによりサイト j における記憶容量負担を求めることができる。

- 処理能力の負担

  データアイテム i の1回のアクセスあたりの

  - CPU 使用 (複製作成、一貫性維持等) 時間 t_{c_i}
  - ディスク使用時間 t_{d_i}

  とする。一定期間 T 内において、サイト j における複製することによる処理時間 T_{p_j} は、一定期間 T 内のアクセス回数に1回あたりの使用時間を乗じて

$$T_{p_j} = \sum_{i=1}^{I} \sum_{r=1}^{R_i} \delta_{jj_{ri}} \cdot T \cdot \lambda(i,j) \cdot (t_{c_i} + t_{d_i}) \tag{9}$$

  によっておおよそ求められる。

5 孤立運転時のサービス品質の低下

ここでは1つの評価方法として、特定サイトが孤立運転時に、発生したアクセスが当該サイトに複製データを持つアイテムへのアクセスならばサービスを受けることができ、持たないならばサービスを受けることができない、という状況を評価する。

サイト j にアイテム i の複製を持っている確率 q_{ij} は、

$$q_{ij} = \sum_{r=1}^{R_i} \delta_{jj_{ri}} \tag{10}$$

であり、孤立運転時に特定サイト j においてデータアイテム i がサービスを受けられる確率 $\lambda'(i,j)$ は、複製が存在する場合には λ がそのまま維持され、複製が存在しない場合には λ は0となりアクセスが発生してもサービスは行われない。従って、

$$\lambda'(i,j) = \sum_{r=1}^{R_i} \delta_{jj_{ri}} \cdot \lambda(i,j) \tag{11}$$

ここで、$j = j_{ri}$の場合のみ$\delta_{jj_{ri}} = 1$である。これを全ての i について評価するため、サービス継続比率 SL として次の評価指数を提案する。

$$SL = \frac{\sum_{i=1}^{I} \lambda'(i,j)}{\sum_{i=1}^{I} \lambda(i,j)} \tag{12}$$

前述の一定閾値 P_0 を大きくすれば R_i が小さくなり、それだけ複製を持たせるサイトの数が少なくなり、その結果 $\lambda' = 0$ の i が多くなり、SL は小さくなる。

6 データの一貫性

今回提案した複製配置方式 (C3) では、複数のサイトに同一のデータアイテムを配置しているので、孤立サイトでデータをアクセス (特に更新) する場合には、他のサイトに配置されたデータとの間で一貫性が問題となってくる。

- 主コピー方式
  各データアイテム i ごとに $P(j|i)$ が最大の j を、この i の主コピーとする。一貫性を保持するため孤立運転時には、当該サイトを主コピーとするデータアイテムのみアクセス可能とする。つまり、次のような式になる。

$$\lambda'(i,j) = \delta_{jj_{1i}} \cdot \lambda(i,j) \qquad (13)$$

  また、孤立時には主コピー以外でのアクセスは不可となるのでサービスレベル SL はさらに低下する。協調運転回復時には他のサイトとの一貫性を回復させる。

- 上限設定方式
  孤立運転時には一時的に一貫性の保持を保留にする。つまり、楽観的並行性制御を使用する。前述の複製の重み $w(j_{ri})$ に応じた権限を付与するものとする。権限の具体的内容については様々であろうが、例えば預金口座のような場合には孤立運転直前の口座残高に $w(j_{ri})$ を乗じた額を引きだし限度額とする、といったことが考えられる。またこの例の場合に別のサイトでの預金の出し入れがあっても口座残高は伝わってこないので一貫性は犠牲になっている。

  この方式では、アクセス可否とは別のサービスレベルの低下が存在する。上記の例では預金引きだし額の制限によるサービスレベルの低下である。異質のサービスレベルをどのように総合評価するかは、今後の検討課題としたい。

  また、このような方式は在庫管理、座席予約等の共通資源を要求に対して割当てるような分野に適用できると考えられる。

7 おわりに

本論文では、DB の複製をとって部分 DB に分割し、それを配置する方式を提案した。また、複製維持のための負担や孤立運転時のサービス品質の低下などの評価方法を提案した。さらに、孤立運転時のデータの一貫性に関する考察を行った。これらによって、分散データベースを実装する際の評価が可能になった。今後の課題としては、上記の方法の妥当性の評価や、他の分割配置方式の検討等を行っていきたい。また、協調運転回復時には、サイト間の一貫性が必要となるが、変更したデータをできるだけ有効にする方法についても研究を推し進めていきたい。

参考文献

[1] 水野, 斉藤, 福岡, 落合 訳: 分散システム コンセプトとデザイン, 電気書院 (1991.6)

[2] Amjad Umar: DISTRIBUTED COMPUTING, Prentice-Hall (1993)

[3] 曽我, 谷林, 長田, 今井, 佐藤, 中川路, 水野: リソース指向分散環境 RODS の提案と実現, 情報処理, Vol.34, No.6, pp.1468-1477(1993.6)

[4] S.Ceri, B.Pernici, and G.Wiederhold: Distributed Database Design Methodlogies, IEEE, Vol.75, No.5, pp.533-546(1987.5)

[5] Susan.B.Davidson:Replicated Data and Partition Failures, Consistency in Partitioned Networks, pp.265-292(1985.9)

[6] 宮西, 高品, 渡辺, 水野: 分散システムにおける独立性の観点からの資源分散方法の提案, 情報処理学会第 48 回全国大会 7D-1, (1994)