

## 異種分散データベースへの質問処理

3 G-1

藤木 克至 伊藤 実

奈良先端科学技術大学院大学 情報科学研究科

## 1 まえがき

近年、生物学分野（特に、遺伝子、蛋白質）などの分野で、ネットワークを介したデータベースの検索サービスが行われている。しかし、同一分野においても、複数の異種データベースにわたってデータが分散しており、それらを統合的に利用するのが困難になっている[2]。また、ネットワークが輻輳しているとき、あるいは、外国のデータベースにアクセスするときには、通信に極端な時間がかかるため、必ずしもリアルタイムシステムが都合がよいとは限らない。本稿では、電子メールを介して検索を行うデータベースを想定し、複数の異種データベースに対する質問処理及び最適化の方式を考察する。

## 2 システム構成

図1は、異種分散データベースを利用するときのネットワークを示したものである。各データベースは、電子

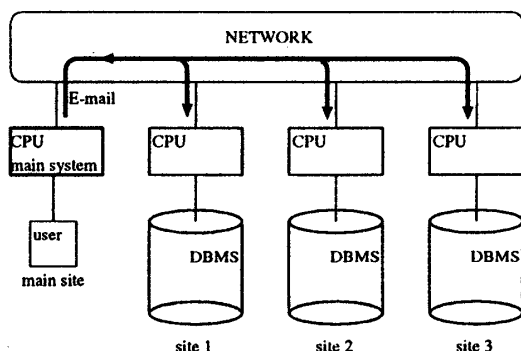


図1: ネットワーク

メールによる検索のみがサービスされ、外部利用者はそこで得られた結果を保存することができないと仮定する。また、各データベース間のデータのやりとりはなく、それぞれが独立したデータベースと仮定する。したがって、分散データベースで通常用いられる準結合 (semijoin) による質問処理の最適化は直接利用できない。すなわち、もし質問処理の中間結果が必要ならば、利用者サイト（メインサイト）に格納しなければならない。しかし、

On Processing Queries in a Multi-Database System  
Katsuji Fujiki and Minoru Ito  
Graduate School of Information Science, Nara Institute of Science and Technology

各データベースの関係をそのままメインサイトへ転送するのは、データ量の観点から現実的ではない。また、ネットワークの輻輳時や遠隔サイトとの通信を行う場合などには、レスポンスに問題が生じる。そこで、メインシステムにおいて、バッチ処理で質問結果を返すことを考える。利用者は、各データベースがどのサイトにあるのかを全く意識しないで、メインシステムに対して質問を行う。必要な関係のスキーマ情報のみを知っているだけでよい。本稿では、質問として連言質問を考える。メインシステムでの処理には、次にあげるデータをもとにした通信コスト（データ転送量）[1]を利用する。

- $r(R)$  関係.
- $R = (A, B, C)$  関係  $r$  のスキーマ.
- $n(R)$  関係  $r$  のタプルの数.
- $n_A = x(R)$  関係  $r$  の  $A = x$  であるようなタプルの数.
- $s(R)$  関係  $r$  の 1 タプルの長さ (bytes).

このとき、関係  $r$  の通信コストは、表の面積 ( $n(R)s(R)$  (bytes)) で表すことが出来る。メインシステムでは、こ

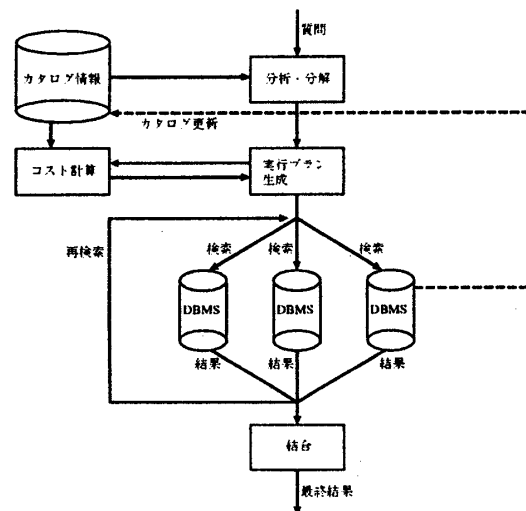


図2: 処理の流れ

の通信コストの低減をもとに、図2に示すようなバッチ処理を行う。メインシステムが保持するカタログ情報には、

1. 各データベースの質問言語
2. 各データベースのサイト情報
  - メインシステムからの距離
3. 各データベースのスキーマ情報
  - 各関係のドメイン
  - 各関係のキー
4. 各関係の統計情報

がある。メインシステムは、カタログ情報をもとに各データベースに検索を行い、その結果を利用者に返す。通信コスト低減のために、検索結果を待つて再検索を行う場合もある。

### 3 最適化の方針

$site1$ ,  $site2$  にそれぞれ存在する関係  $r(R)$ ,  $t(T)$  について考える。

|           |                                                                                            |         |   |   |  |  |  |           |                                                                                            |   |   |   |  |  |  |
|-----------|--------------------------------------------------------------------------------------------|---------|---|---|--|--|--|-----------|--------------------------------------------------------------------------------------------|---|---|---|--|--|--|
| 関係 $r(R)$ | <table><tr><td>A</td><td>B</td><td>C</td></tr><tr><td></td><td></td><td></td></tr></table> | A       | B | C |  |  |  | 関係 $t(T)$ | <table><tr><td>D</td><td>E</td><td>F</td></tr><tr><td></td><td></td><td></td></tr></table> | D | E | F |  |  |  |
| A         | B                                                                                          | C       |   |   |  |  |  |           |                                                                                            |   |   |   |  |  |  |
|           |                                                                                            |         |   |   |  |  |  |           |                                                                                            |   |   |   |  |  |  |
| D         | E                                                                                          | F       |   |   |  |  |  |           |                                                                                            |   |   |   |  |  |  |
|           |                                                                                            |         |   |   |  |  |  |           |                                                                                            |   |   |   |  |  |  |
| $site1$   |                                                                                            | $site2$ |   |   |  |  |  |           |                                                                                            |   |   |   |  |  |  |

スキーマはそれぞれ、

$$R = (A, B, C)$$

$$T = (D, E, F)$$

とし、 $A$  は関係  $r$  のキーとなっているものとする。ここで、次のような連言質問の処理を考える。

$$(A, B, F) : -r(A, B, C), s(D, E, F),$$

$$3 \leq A \leq 5, B = E$$

まず、 $site1$  の関係  $r$  に対して、

$$(A, B) : -r(A, B, C), 3 \leq A \leq 5$$

の検索を行い

|         |                                                                                                                                                                                             |   |   |   |       |   |       |   |       |
|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|---|---|-------|---|-------|---|-------|
| 関係 $r'$ | <table border="1"><tr><td>A</td><td>B</td></tr><tr><td>3</td><td><math>B_3</math></td></tr><tr><td>4</td><td><math>B_4</math></td></tr><tr><td>5</td><td><math>B_5</math></td></tr></table> | A | B | 3 | $B_3$ | 4 | $B_4$ | 5 | $B_5$ |
| A       | B                                                                                                                                                                                           |   |   |   |       |   |       |   |       |
| 3       | $B_3$                                                                                                                                                                                       |   |   |   |       |   |       |   |       |
| 4       | $B_4$                                                                                                                                                                                       |   |   |   |       |   |       |   |       |
| 5       | $B_5$                                                                                                                                                                                       |   |   |   |       |   |       |   |       |

を得たとする。しかし、 $site2$  では準結合をとることができないので、 $site2$  の関係  $t$  に対して、

$$(E, F) : -t(D, E, F)$$

の検索を行って得られる関係  $t'$  と関係  $r'$  とを、メインシステムにおいて準結合し、最終結果を求めることになる。しかし、この場合の通信コストは、

$$s(A, B) + s(E, F) \times n(T) \quad (1)$$

となり、関係  $t'$  のタプルの数によっては相当な通信コストがかかる。そこで、本システムでは、関係  $r'$  を得れば、その各タプルに対して

$$\text{for } 3 \leq x \leq 5$$

$$(E, F) : -t(D, E, F), E = x$$

end

を行って得られる3つの関係と関係  $r'$  とを結合して最終結果を得る。この場合の通信コストは、

$$s(B, C) + s(E, F) \times (n_{E=3}(T) + n_{E=4}(T) + n_{E=5}(T)) \quad (2)$$

となる。関係のサイズにも依存するが、(1) と (2) を比較すれば、通信コストは相当低減できることがわかる。

表1に、クライアント・サーバ(C/S)方式と本システムとの比較を示す。クライアント・サーバ方式では、複数の異種データベースからデータを検索する場合、複数のデータベースにクライアントとしてアクセスする必要がある。リアルタイムでデータの検索ができるという利点があるが、サーバ側の負荷が大きい場合、あるいはネットワークの輻輳時には、レスポンスが悪くなる。また、ネットワークにかかる通信コストも大きくなり得る。本システムでは、バッチ処理で質問結果を返すので、一旦質問を出すと、後は結果を待つのみで、システムに拘束されることもない。また、通信回数は増え得るが、通信コストは低減することができる。各データベースにおいても、電子メールで検索がなされるので、負荷が低減され、セキュリティの保全にもなっている。

表1: C/S方式と本システムの比較

|           | C/S方式 | 本システム |
|-----------|-------|-------|
| リアルタイム    | △     | ×     |
| レスポンス     | △     | ×     |
| 通信コスト     | △     | ○     |
| セキュリティの保全 | ×     | ○     |

### 4 あとがき

本システムでは、通信コストの低減において、カタログ情報内の統計情報が必要不可欠であり、スキーマの変更及びデータの登録・更新・削除に対応して、メンテナンスを行っていく必要があるが、そのメンテナンス時期及び間隔について検討する必要がある。また、本稿で示した以外の最適化方法を考慮中である。

### 参考文献

- [1] Hank Korth, and Abraham Silberschats, "DATABASE SYSTEM CONCEPTS," 2nd ed., McGraw-Hill, 1991.
- [2] Hongjun Lu, Beng-Chin Ooi, and Cheng-Hian Goh, "Multidatabase Query Optimization: Issues and Solutions", *Proc. Research Issues in Data Engineering: Interoperability in Multidatabase Systems*, pp.137-143, Apr. 1993.