

異種スキーマをもつデータベースへの統合的なアクセス手法

2F-1

西澤 格
東京大学工学部高須 淳宏 安達 淳
学術情報センター研究開発部

1 はじめに

本発表は、複数の情報源となるデータベースに対して統合的に情報検索を行う場合の情報検索システムの構成法と複数のデータベースの統合スキーマの構成法および発行された問い合わせの処理方法について述べる。

従来、複数の情報源に対して検索を行うには、それぞれの情報源に対して検索条件を考慮して検索式を書く必要があったが、本発表において提案する情報検索システム Robin では複数のデータベースに対して同時に問い合わせを生成することができる。

2 情報検索システム Robin

Robin は分散処理との親和性の高い情報検索システムを目標として設計されている。つまり、必要な情報が一つのデータベースに集中して蓄えられているのではなく、複数の独立したデータベースにわかれて蓄積されているという分散環境を前提としている。これらの複数のデータベースでは一般的にそのデータの構造・意味は異なり、連邦型データベースシステム¹⁾で用いられているようなスキーマ統合が必要となる。

連邦型データベースでは、粗に結合するデータベースを扱う際にはその統合のための自分の連邦スキーマと他のデータベースの輸出スキーマの間の写像の定義が難しいとされている。

Robin は複数のデータベースを統合的に扱うための情報検索のモデルとして、クライアント-サーバモデルを用い²⁾、情報検索を行おうとするユーザ側であるクライアントが各種の情報資源を用いることで問い合わせの評価を行うことによりサーバの差異を吸収する。このサーバの差異を吸収するために Robin はリソースオブジェクト (以下 RO) と呼ばれる情報資源を利用する。

具体的には Robin は RO とマネージャのセッションにより、システム上に仮想データベース (Virtual Database 以下 VD とする) を構成する。情報検索を行いたいユーザはこの仮想データベースに対して検索条

件を書くことになる。

以下この仮想データベースの構築方法と、その仮想データベースに対して発行された問い合わせを各要素データベースの問い合わせに分解する手法について説明する。

3 仮想データベース

Robin では個別のデータベース (以下これらのデータベースを要素データベースと呼ぶ) を統合的に取り扱うために、クライアント上に仮想データベースを構築する。この仮想データベースのスキーマは各要素データベースのスキーマ情報から構築される。

3.1 仮想データベース上の属性

要素データベース DB_i 上の属性 (項目) の集合を $U_i (i = 1, 2, \dots, N)$ とした時、仮想データベース上での属性の集合を統合スキーマを $U; U_i$ と定義する。ここで問題となるのは各要素データベース上の属性名の違いだが、これについては写像関係が RO によって与えられるものとする。なお、以下では要素データベース上の関係が 1 つであるとの前提をおいている。データベースは複数の関係から構成されるのが普通だが、本方式では普遍関係 (Universal Relation) の考え方³⁾によりあらかじめ仮想的に複数の関係の Join が行われていると考える。

3.2 仮想データベース上の関数従属性

一般にスキーマ上で関数従属性の集合 D が与えられた時、そのスキーマ上で成り立つ関数従属性は、 D^+ であり、成り立たない関数従属性は、 $U^D - D^+$ となる³⁾。つまり任意の関数従属性は、成り立つか成り立たないかがはっきりしている。いま要素データベース DB_i における関数従属性の集合を D_i と表現することにする。統合スキーマ上では、 U^D は以下の 3 つのクラスに分類することができる。

D_{valid} : 任意の要素データベースで成り立つ関数従属性の集合 $\{f \mid (\forall i)(f \in D_i^+)\}$

³⁾ここで D^+ は D によって論理的に含意される関数従属性の集合 (D の閉包)、 U^D は D の属性集合で可能な全ての関数従属性の集合を表すものとする。

D_{false} : いかなる要素データベースでも成り立たない関数従属性の集合 $\{f \mid (\exists i)(f \in (U^D - D_i^+))\}$

$D_{unknown}$: それ以外。 $U^D - D_{valid} - D_{false}$

このとき、仮想データベース VD 上の関数従属性を (1) 式のように定義する。この定義の意味は、いずれかの要素データベースで積極的に否定されない関数従属性は VD 上の関数従属性とみなすというものである。

$$D = \{f \mid (\forall i)(f \in D_i^+) \vee (f \in D_{unknown} \wedge (\exists i)(f \in D_i^+))\} \quad (1)$$

4 問い合わせの処理

4.1 発行される問い合わせ

まず各要素データベースに対して発行される問い合わせについての定式化を行う。なお本節では、問い合わせを安全な関係論理³⁾で限量子を含まないものに限定する。論理式で与えられた問い合わせ Q を論理和標準形 (選言標準形、(2) 式) に変換し、その中の項である Q_i について考える。ここで、検索の意味を考えた場合 OR で結合された問い合わせは個々に評価を行ってよいと考えられる。

$$Q = Q_1 \vee Q_2 \vee \dots \vee Q_n \quad (2)$$

Q_i を特定のデータベース DB に対する問い合わせに変換したものを $Q_{i(DB)}$ と表すものとするれば、

$$Q_{i(DB)} = \{q_{ij} \mid q_{ij} \in Q_i, \text{Item}(q_{ij}) \in \text{Items}(DB)\} \quad (3)$$

と書くことができる。但し、 $\text{Item}(q_{ij})$ はリテラル q_{ij} が参照する属性、 $\text{Items}(DB)$ は要素データベース DB 上に存在する属性の集合を表すものとする。

(3) 式の意味は、要素データベース DB 上に存在する属性に対する条件式のみを DB に発行するというものである。DB ではその中で定義されていない属性についての条件については評価を行うことは不可能であるのでこれは妥当であると思われる。

4.2 問い合わせ発行の判断

問い合わせの定式化の次に問題となるのがその生成された問い合わせを実際に発行するかどうかの判断である。ここでも 4.1 節で述べた理由から分解された論理式 Q_i について考えることにする。

VD に関して発行された問い合わせを実際に各データベースに問い合わせるかどうかの判断について考えた場合この結果は結局 '問い合わせる' あるいは '問い合わせない' の 2 値であり、論理式で扱うのに都合がよい。

この真偽値の定式化を行うため、要素データベース DB に問い合わせ Q_i を送るかどうかの真偽値を $J_{DB}(Q_i)$ で表すものとする。

この時、 $J_{DB}(Q_i)$ を以下の式に従って計算する。

$$J_{DB}(Q_i) = \bigwedge_{q_{ij} \in Q_i} (\{\text{Item}(q_{ij}) \in \text{Items}(DB)\} \vee (X \rightarrow \text{Item}(q_{ij}) \in D)) \quad (4)$$

ここで、

$$X = \{\text{Item}(q_{ik}) \mid \text{Item}(q_{ik}) \in \text{Items}(DB), q_{ik} \in Q_i, k \neq j\}$$

である。

4.3 判断式の解釈

(4) 式についてその意味を考える。まず第 1 項は要素データベースに、問い合わせの条件で参照される属性が全て存在するときのみみ検索を発行するものである。これは最も厳しい基準であり、この時には得られる解は求められるべき解の部分集合となり、論理的に正しいことが証明されている。しかしながら、情報検索システムとしては求められるべき解をその中に含むような集合 (解のスーパーセット) を得たいという要求があるため、これに答えるために関数従属性を考慮した第 2 項を判断式に追加した。この第 2 項は検索の意味の整合性を保ちながら解の集合を大きくするための項と解釈することができる。

5 おわりに

以上、分散環境との親和性の高い情報検索システム Robin におけるスキーマの統合手法と、問い合わせの処理方法について説明した。

データベースの統合的利用のためには不可欠なスキーマ統合に、関数従属性に注目したアプローチは有効であると考えられる。

参考文献

- [1] Amit P. Sheth, James A. Larson, "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases", ACM Computing Surveys, Vol.22, No.3, pp183-286.
- [2] 西澤 格, 安達 淳: "クライアント-サーバモデルによる情報検索システムの提案", 第 45 回情報処理学会全国大会 3R-6, (1992).
- [3] JEFFREY D. ULLMAN 著, 國井, 大保 訳, "データベースシステムの原理", (1985).