ＡＤＴを含む問い合わせ処理の最適化方式の設計

7Ｈ－8　*矢島謙一 **大保信夫 **北川博之 **山口和紀 **藤原譲 **鈴木功
(*筑波大学工学研究科 **筑波大学電子情報工学系)

１．はじめに

近年、化学グラフ構造に対するデータベース化の要求が高まっているが、従来のデータベースシステムでは、複雑な構造を持つ化学グラフデータの支援は困難である。

この問題に対するアプローチとして、当グループでは、ＡＤＴ（抽象データ型）を用いて化学グラフ構造データの支援を可能とするシステムCHARMの研究開発を進めている[1]。化学グラフのようなＡＤＴを含むデータベースに対する問い合わせ処理を行なう場合、従来とは異なった最適化の基準が必要となる。例えば、従来コストが低いとみなされている選択演算についても、グラフの同型性判定といった莫大な計算コストを要する処理が必要な場合がある。

そこで、本稿では、ＡＤＴ属性に対する計算コストを考慮した場合の選択演算と接合演算の処理コストを定式化し、問い合わせの最適化に必要なこれらの演算の間の順序関係によるコストの依存性ついて述べる。

２．コストモデル

本節では、図１の２つのProcessor Tree（ＰＴ）について、それぞれ処理コストを定式化し、２つの演算の間の順序関係によるコストの依存性を考察する。

２．１．処理コストの前提条件

ここでは、各ＰＴに要するコストの定式化に必要な前提条件について述べる。

まず、２つのリレーションは次のスキーマに従うものとする。

R1(A,B) A：ＡＤＴ属性　B：接合属性
R2(B,C) B：接合属性 (Clustering Index)
　　C：基本データ型属性

つぎに、接合演算はLoop-Join方式で行うものとし、各演算のコストは [2]のコストモデルに従うものとする。また、各演算の結果得られるリレーションは、その都度２次記憶に戻されるものとする。

２．２．各ＰＴの処理コスト

２．１節で述べた前提条件に基づいて、図１の各ＰＴの処理コストを定式化する。

リレーション間のJoin Selectivity、及びＡＤＴの計算コストをそれぞれ、JS、f(n)（ｎ：タプル数）と表わすことにする。但し、

$JS = \|R1 \infty R2\| / (\|R1\| * \|R2\|)$

この時、(1) のＰＴのコストはＥ１式で与えられる。
（※ ‖Ｒ‖, |Ｒ| は それぞれ Ｒ のタプル数及びブロック数を表わす、また、$D(R.A) = \pi_A(R)$）

また、(2)のＰＴのコストはＥ２式で与えられる。

さらに、ＡＤＴコストを $f(n)=n*t(t>0)$ とすると(1)と(2)のＰＴのコスト差、(1)－(2)はＥ３式で与えられる。これからわかるように、(1)と(2)の差は JSとtを変数に持つ関数F(JS,t)として表わすことができる。そして、

F(JS,t)<0 → (1)のＰＴがコスト小
F(JS,t)>0 → (2)のＰＴがコスト小

となる。

２．３．２つのＰＴの処理コストの比較

２．２節の結果に基づいて、次の２つの場合について、それぞれ処理コストの比較を行なう。

(i) $JS \geqq 1/\|R2\|$

この場合、$\|R1 \infty R2\| \geqq \|R1\|$ となるため、接合コスト、選択コストともに(1)のＰＴが(2)に比べ小さいことは明らかである。

(ii) $JS < 1/\|R2\|$

この場合、$\|R1 \infty R2\| < \|R1\|$ となるため、JSと tの値に従って関数 F(JS,t)の正負が定まり、ＰＴの優劣が明らかになる。

図２に、関数F(JS,t)の例を示す。（但し、‖R1‖=1500 ‖R2‖=1000, |R1|=300, |R2|=200, ‖D(R1.A)‖=150, ‖D(R2.B)‖=500における $-200 \leqq F(JS,t) \leqq 0$ の領域を示す）

３．化合物データベースへの適用

本節では、第２節で述べた処理コストのモデルを、化合物データベースCHARMのリレーションに対して適用する。

３．１．リレーション間の関係

化合物データベース CHARMのリレーションの間には次の関係が存在する。

すなわち、接合演算が可能な２つのリレーションの接合属性は、互いに一方が他方のsupersetとなっている。したがって、図１のＰＴではR1とR2の間に、

$D(R1.B) \subseteq D(R2.B)$ あるいは $D(R1.B) \supseteq D(R2.B)$

という関係が成り立つ。そこで、上記の２つの場合について、それぞれ考察を行なうことにする。

(a) $D(R1.B) \subseteq D(R2.B)$

このとき、JSは $\|D(R2.B)\| \leqq \|R2\|$ から

$JS = 1/\|D(R2.B)\| \geqq 1/\|R2\|$

が成り立つ。これは、２．３節の(i)に該当するので、コストは(1)のＰＴの方が小さいことがわかる。

(b) $D(R1.B) \supseteq D(R2.B)$

このとき、JSは(a)と同様に

$JS = 1/\|D(R1.B)\| \geqq 1/\|R1\|$

が成り立つ。この場合、R2に対する条件は存在しないため、この条件だけではＰＴに対する優劣の判断はできない。

A Study on Query Optimization of ADT Database
K.Yajima, N.Ohbo, H.Kitagawa, K.Yamaguchi,
Y.Fujiwara, I.Suzuki
University Of Tsukuba

3．2．実際の適用例

ここでは、実際に化合物データベース CHAMRの一部のリレーションとそれに対する3つのPT（図3参照）を考え、それぞれについてコストの評価を行なう。

リレーションのスキーマは以下の通りである。

R1(C#,GRAPH,…)
R2(CG#,GRAPH,…)
R3(C#,CG#)

ここで、下線を持つ属性はリレーションのキーを表わし、タプルはこの値によってクラスタリングされている。さらに、これらのリレーションの間には以下の関係が存在する。

D(R2.CG#)⊃D(R3.CG#) ①　　D(R1.C#)⊇D(R3.C#) ②
‖R1‖＝‖R3‖　　③　　‖R2‖＞‖R3‖　　④
‖R2∞R3‖＝‖R3‖　⑤

まず、(2)と(3)のPTについて評価する。①、②より

D(R1.C#)⊇D((R2∞R3).C#)＝D(R3.C#)

これは、3．1節の(a)に該当するので、(2)のコストは(3)以下であることがわかる。

次に、(1)と(2)のPTについて評価する。①は、3．1節の(b)に該当するため、これだけでは(1)と(2)のコストの優劣は判断できない。そこで、他の条件について検討を行なうと④、⑤から2．3節の(ii)に該当することがわかる。したがって、コストの優劣はJSとtに依存した 関数 F(JS,t)の結果に従うことになる。この場合、⑤から

JS＝1／‖R2‖

となるため、(1)と(2)のコストの差は、F(1／‖R2‖,t)に一致する。

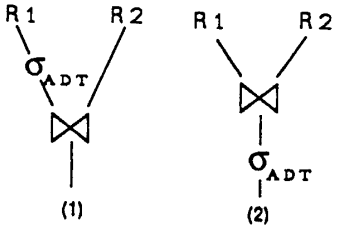

図1　2つのPT

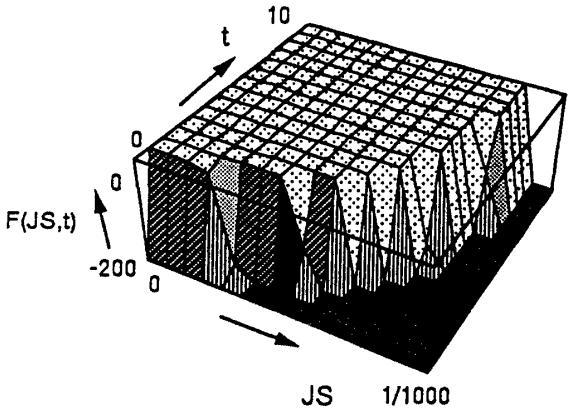

図2　関数F（JS、t）の例

4．おわりに

本稿では、ADT属性に対する選択の計算コストを考慮した場合の選択演算と接合演算の処理コストの定式化を行ない、それらの演算の間の順序関係によるコストの依存性について考察した。

その結果、ADT属性に対する選択演算を含む問い合わせに対し、最適化を可能とするための見通しが得られた。

今後は、ADTの計算コストについて詳細な解析を行ない、より現実に即した定式化を行なうとともに、既存の最適化ヒューリスティックスとの融合を図っていく予定である。

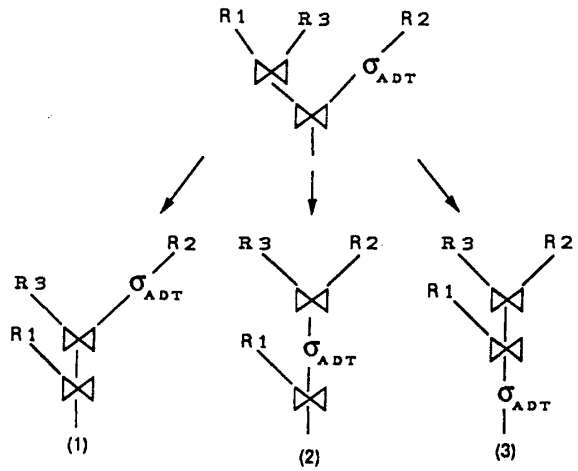

図3　3つのPT

参考文献

[1]山田et.al.,” 関数型モデルを用いた化学グラフデータベース CHARMのユーザインターフェース” 情報処理学会第39回全国大会予稿集 Oct.,1989
[2]J.D.Ullman.,"PRINCIPLES OF DATABASE AND KNOWLEDGE-BASE SYSTEMS" VOL2.,1989 Computer Science Press

$$(E1)\quad |R1| + f(\|R1\|) + \frac{2|R1|}{|D(R1.A)|} + \frac{|R1||R2| + \|R1\||R2|}{|D(R1.A)|}\cdot Js + \frac{|R1||R2|}{|D(R1.A)||D(R2.B)|}$$

$$(E2)\quad |R1| + f(\|R1\|\|R2\|JS) + \left(2 + \frac{1}{|D(R1.A)|}\right)(|R1||R2| + \|R1\||R2|)JS + \frac{|R1||R2|}{|D(R2.B)|}$$

$$(E3)\quad (\|R1\| - \|R1\|\|R2\|JS)t - 2(|R1||R2| + \|R1\||R2|)\cdot JS - \frac{|R1||R2|}{|D(R2.B)|} + \frac{2|R1|}{|D(R1.A)|} + \frac{|R1||R2|}{|D(R1.A)||D(R2.B)|}$$