アルゴリズム　23－8
（1991．9．20）

演繹データベースにおける質問処理に必要な計算コストの近似的評価法

宇野　裕之　　茨木　俊秀

京都大学　工学部

あらまし　演繹データベースに与えられた質問に対しその答えを導出する方法は一般に一通りではないので，あらかじめもっとも効率のよい導出方法を知り，それを実行することが望ましい．そのためには各導出方法が要する計算量を事前に評価することが必要となる．答えの導出は，関係表に対して，推移閉包を含む関係代数の諸演算を逐次適用することであると解釈できるので，ここでは，計算量をそれらの演算の結果生成される関係表の大きさの和としてとらえ，それぞれの大きさを評価する近似式を提案する．また，それらを実際の例に適用して各処理法による計算量の違いを評価する．

Approximate Evaluation of the Computing Cost for Answering Queries in Deductive Databases

Yushi Uno and Toshihide Ibaraki

Department of Applied Mathematics and Physics,
Faculty of Engineering, Kyoto University, Kyoto, Japan 606

Abstract　Given a query to a deductive database, methods to derive its answers are not unique in general. It is desirable, therefore, to know the most efficient one among them, before we carry out the actual derivation procedure. In order to do so, we have to evaluate beforehand the computing cost required by each method, which consists of successive executions of several kinds of operations of relational algebra including transitive closures. Therefore, the computing cost can be evaluated if we approximate formulas for them by considering the cost to be the sum of the sizes of the relational tables generated during execution of such operations. Finally, we apply the obtained formulas to some examples to demonstrate its effectiveness.

1 はじめに

演繹データベースの処理は関係代数の諸演算や推移閉包の演算を関係表に適用することで遂行される [9]．その中でも推移閉包は演繹データベースにおける再帰的な質問を処理する際に不可欠な演算で，その効率的な実行法に関する研究がなされている [1, 3, 6, 8]．ところで，一つの質問に対する答えの導出にはさまざまな方法が可能である場合が普通であって，最も効率のよい導出方法をあらかじめ求めて，それを実行することが望まれる．そのためには，各導出方法が要する計算量を事前に評価することが必要であり，これは導出アルゴリズムが各種演算の逐次実行と解釈できることを考えると，各演算の計算量を評価することに帰着する．ここでは，各演算の計算量を，その演算の結果得られる関係表の大きさとしてとらえる．関係代数の基本演算について，これらの評価を近似的に行うことは比較的容易であるが，推移閉包については，実際に関係のデータの一部をサンプリングして計算するなどごく限られた方法が知られているのみであって [4, 5, 7]，簡便な近似式は知られていない．

そこで本論文では，関係代数の諸演算を施して得られる関係表の大きさを評価する近似式を，データの確率的一様性および独立性の仮定の下で導き，つづいて 2 項関係の推移閉包を求める Semi-Naive 法 [2, 6] を，k項関係に対する一般化された推移閉包に拡張したのち，それによって得られる関係表の大きさを近似的に求める方法を提案する．最後に，演繹データベースに対する具体的な質問の例を用いて，ここで提案した方法を用いることによって，その質問を処理するさまざまな方法の計算量を近似的に評価できることを示す．

2 関係表

2.1 関係と関係表

関係 (relation) Rとは，定義域 (domain) D_i $(i = 1, \ldots, k)$ の直積 (Cartesian product) $D_1 \times \cdots \times D_k$の部分集合である．定義域の数が k個のとき k項関係 (k-ary relation) といい，定義域の直積の要素を k個組 (k-tuple) (あるいは単に組 (tuple)) という．本文中では，通常の慣例に従い，各定義域 D_iの大きさ d_iは有限とし，したがって 1 つの関係 Rに属する組の数も有限であるとする．

関係表 R とは，その関係に属するすべての相異なる組を陽に表の形にしたもので，その各行は組であり，各列は各定義域の名前 A_i $(i = 1, \ldots, k)$ (これを属性 (attribute) という) である．一般に，関係 Rの大きさ (size) の期待値を $|R|$ と記す．大きさ $|R|$ があらかじめ定まっている場合，その大きさは確率変数ではなく，したがって期待値とする必要はないが，記法を簡単にするため，本論文では特に区別しない．以下では $|R|$ を rと記すことが多い．

2.2 関係表に関する仮定

これから本文で扱う関係表 Rは以下の仮定にしたがって確率的に生成されているものと考える．すなわち，定義域が $D_1, \ldots, D_k$で大きさが rの関係表は，その表を構成する組の集合として $\binom{d_1 \cdots d_k}{r}$ 通りの組合せがあるが，これらはすべて等しい確率で生じる．また，1 つの関係に 2 つ以上の同じ組が現れることはない．以上の仮定の結果，各属性 A_i に任意の要素 $e \in D_i$の現れる確率は等しく $1/d_i$ である．

3 関係表の演算と演算結果の期待値

この章では，関係表に対する主要な演算に関して，それらの演算によって生成される新しい関係表の大きさの期待値，あるいは期待値の近似値を求める．ここでは関係表に対する演算 [9] として次のものをとりあげる．すなわち，直積 (Cartesian product)，共通集合 (intersection)，集合和 (union)，集合差 (set difference)，選択 (selection)，射影 (projection)，自然結合 (natural join) である．さらに次の章では推移閉包 (transitive closure) について考える．

3.1 直積 $\otimes$

定義域 $D_{i_1}, \ldots, D_{i_m}$からなる関係表 R_1と定義域 $D_{j_1}, \ldots, D_{j_n}$からなる関係表 R_2の直積 $R_1 \otimes R_2$は，R_1に属するすべての組と R_2に属するすべての組を接続してできるすべての組を要素とし，定義域 $D_{i_1}, \ldots, D_{i_m}, D_{j_1}, \ldots, D_{j_n}$ からなる関係である．関係 R_1と R_2の大きさをそれぞれ r_1，r_2とするとき，$R_1 \otimes R_2$の大きさは

$$|R_1 \otimes R_2| = r_1 r_2 \tag{3.1}$$

である．

3.2 共通集合 $\cap$，集合和 $\cup$，集合差 $-$

関係 R_1と R_2はいずれも定義域 $D_1, \ldots, D_m$をもち，確率的に独立に得られたものとする．

- 共通集合 $\cap$

R_1と R_2の両方に属する組を要素とする関係を求める演算である．R_1に属する任意の組が R_2にも存在する確率は $r_2/d_1 \cdots d_m$である．従って，共通集合の大きさの期待値は

$$|R_1 \cap R_2| = r_1 r_2 / d_1 \cdots d_m \tag{3.2}$$

である．残りの 2 つの演算についてはこの結果を用いることによってただちに導くことができる．

- 集合和 $\cup$

R_1と R_2のどちらかの関係に属する組を要素とする関係を求める演算である．その大きさの期待値は

$$|R_1 \cup R_2| = r_1 + r_2 - r_1 r_2 / d_1 \cdots d_m \tag{3.3}$$

である．

- 集合差 $-$

R_1の要素であって R_2の要素ではない組を要素とする関係を求める演算である．その大きさの期待値は

$$|R_1 - R_2| = r_1 - r_1 r_2 / d_1 \cdots d_m \tag{3.4}$$

である．

3.3 選択 σ，射影 π

これら2つの演算は，ある関係 Rに対するものである．Rの定義域は $D_1,\ldots,D_m$であるとする．

• 選択

選択σ_Cとは関係表 Rの中から，1つ以上の属性に関する条件の集合 Cをみたす組だけを選び出して新しい関係表をつくる演算である．いま条件 Cを

$$C=\{A_{i_1}=e_1,\ldots,A_{i_k}=e_k\}$$

(すなわち，属性 A_{i_j}の値が e_j $(j=1,\ldots,k)$ である行のみを選ぶ) とすると，選択後の関係の大きさの期待値は

$$|\,\sigma_C R\,|=r/d_{i_1}\cdots d_{i_k} \tag{3.5}$$

である．

• 射影

射影π_Aは，1つの関係表 Rから特定の属性の集合 $A=\{A_{i_1},\ldots,A_{i_k}\}$ についてだけ抜き出して新しい関係表をつくる演算である．このとき他の属性 (これらを $A_{i_{k+1}},\ldots,A_{i_m}$ と記す) を削除することによって生じる重複した組は，1つを残してすべて削除される．生成される関係の大きさの期待値は

$$\begin{aligned}|\,\pi_A R\,| &= d_{i_1}\cdots d_{i_k}\left\{1-\left(\frac{d_1\cdots d_m-d_{i_{k+1}}\cdots d_{i_m}}{d_1\cdots d_m}\times\right.\right.\\ &\qquad\left.\left.\cdots\times\frac{d_1\cdots d_m-d_{i_{k+1}}\cdots d_{i_m}-r}{d_1\cdots d_m-r}\right)\right\}\\ &\approx d_{i_1}\cdots d_{i_k}\left\{1-\left(1-\frac{1}{d_{i_1}\cdots d_{i_k}}\right)^r\right\}\end{aligned} \tag{3.6}$$

である．上式で，分数の積の部分は，全体で $d_1\cdots d_m$ とおりある組の中から1組ずつ取り出すときに，属性 $A_{i_1},\ldots,A_{i_k}$ の値がある特定の1組ではないことが連続して r回起こる確率を表している．{ } の中はその特定の組が大きさ rの関係表の中に少なくとも1つは現れる確率を，全体では属性 $A_{i_1},\ldots,A_{i_k}$ の部分に現れる異なる組の種類の数の期待値を表している．下式は上式を近似したものである．

3.4 自然結合 $\bowtie$

属性 $A_{i_1},\ldots,A_{i_m}$からなる関係 R_1と属性 $A_{j_1},\ldots,A_{j_n}$からなる関係 R_2がある．R_1と R_2の等結合 (equijoin) $\bowtie$とは，指定された属性の組 $A_{i_{s_1}},\ldots,A_{i_{s_p}}$ と $A_{j_{t_1}},\ldots,A_{j_{t_p}}$ (ただし $D_{i_{s_1}}=D_{j_{t_1}},\ \ldots,\ D_{i_{s_p}}=D_{j_{t_p}}$が成立する) に対し，$R_1$の組の属性 $A_{i_{s_1}},\ldots,A_{i_{s_p}}$の値と R_2の組の属性 $A_{j_{t_1}},\ldots,A_{j_{t_p}}$ の値がそれぞれ等しいときにその値を含む R_1と R_2の組の連接をつくり，それらの組の集合を新しい関係とする演算である．このとき属性の数は $m+n$ である．さらに，結合した属性どうしは互いにつねに同じ値を2重に保持していることになるので，そのうちの一方を省略して $(m+n-p)$ 個の属性を持つ関係にする演算が自然結合 (natural join) である．自然結合による関係の大きさは等結合のときと変わらない．等結合あるいは自然結合によってできる新しい関係の大きさの期待値は

$$|\,R_1\bowtie R_2\,|=r_1r_2/d_{i_{s_1}}\cdots d_{i_{s_p}} \tag{3.7}$$

となる．これは，R_1の任意の組の属性 $A_{i_{s_1}},\ldots,A_{i_{s_p}}$の値が R_2の任意の組の属性 $A_{j_{t_1}},\ldots,A_{j_{t_p}}$ の値と一致する確率が $1/d_{i_{s_1}}\cdots d_{i_{s_p}}$で，そのような R_1と R_2の組の組合せが全部で r_1r_2通りあるからである．

特別な場合として，R_1が属性 A_1,A_2をもち，R_2が属性 A_3,A_4 をもち，A_2と A_3の定義域が共通で Dであるとする．さらに定義域 Dの大きさを dとする．このとき A_2と A_3に関して自然結合 $R_1\bowtie R_2$を行ってできる新しい関係の大きさの期待値は

$$|\,R_1\bowtie R_2\,|=r_1r_2/d \tag{3.8}$$

である．

4 推移閉包

4.1 2項関係の推移閉包

推移閉包は一般には k項関係に対して考えらるが，ここではまず簡単のため，実際の場合によく現れる2項関係 Rについて述べる．ただし，Rの2つの属性の定義域は同じ Dであるとする．2項関係 Rの推移閉包 R^+は次のように定義できる．すなわち，ある整数 n (≥ 2) とある要素の列 $e_{i_1},\ldots,e_{i_n}$ があって，

(1) $e_{i_1}=e_1, e_{i_n}=e_2$,

(2) 任意の j $(1\leq j\leq n-1)$ に対して $(e_{i_j},e_{i_{j+1}})$ は Rの組，

であるとき，そのようなすべての2個組 (e_1,e_2) の集合を Rの推移閉包 R^+という．

関係 Rに対し，定義域 Dを節点集合 Vで表し，Rの各組を枝集合 Eの要素とすると，有向グラフ $G=(V,E)$ が得られる．有向グラフ $G=(V,E)$ において，節点 v_1から v_2へ長さ1以上の路 (path) で到達可能であるとき，そのようなすべての節点対 (v_1,v_2) の集合を枝集合 E^+としてできる新しいグラフ $G^+=(V,E^+)$ のことを G の推移閉包という．このとき，G^+の各枝は Rの推移閉包 R^+の各組を表す．

4.2 推移閉包を求める方法

推移閉包を求める効率的な方法は，グラフ理論にもとづいて数多く提案されている．最も直接的な方法は，G の枝集合をもとに，順に枝を接続するという操作をくりかえすことによって長さ $2,3,\ldots$ で到達できる節点の組を，それ以上組が増加しなくなるまでくりかえし求める方法であって，Warshall のアルゴリズムとして実現されている．

このアルゴリズムを関係表の立場から論ずると次のようになる．2項関係 Rの推移閉包 R^+の組 (e_1,e_2) で長さ kの系列 $e_1=e_{i_1},e_{i_2},\ldots,e_{i_{k+1}}=e_2$によって定まるものを求めるには，$R$の2番目の属性と1番目の属性について Rどうしの自然結合を行った後，結合に用いた2つの属性を削除するという操作を k回くりかえすことによって得られる．このとき得られた関係 R^kの属性の数は2で変わらない．すなわち，1回の結合ののち，結合に用いた属性を除いて属性の数を保存するこの演算を以後 $\bowtie$ と表すと，

$$R^k=\underbrace{R\bowtie R\bowtie\cdots\bowtie R}_{\bowtie\text{は}k-1\text{個}}$$

であり，R^+は

$$R^+=\bigcup_{k=1}^{\infty}R^k$$

によって求めることができる．この方法を Naive 法 [6] という．

ところが，上の計算では同じ組を何度も重複して求めていることになるので，各反復におけるむだを省くためには，反復のたびに新しく生成された組の増分の集合ΔR^kのみを求めればよい．この方法は Semi-Naive 法 [2, 6] とよばれ，次の手順で実現できる．

$$\begin{cases} \Delta R^1 := R \\ R^1 := \Delta R_1 \end{cases} \tag{4.1}$$

$$\begin{cases} \delta R^k := \Delta R^{k-1} \bowtie R \\ \Delta R^k := \delta R^k - R^{k-1} \\ R^k := R^{k-1} \cup \Delta R^k, \qquad k = 2, 3, \ldots \end{cases} \tag{4.2}$$

を実行し，最後に

$$R^+ := \bigcup_{k=1}^{\infty} \Delta R^k \tag{4.3}$$

とする．kの反復は実際には$\Delta R^k = \phi$となれば停止できるので，有限回でよい．

以後の議論では，この Semi-Naive 法で推移閉包 R^+の大きさを評価する手続きを考える．

4.3 推移閉包の一般化

2 項関係の推移閉包を一般化すると，1 つの k項関係 Rに対して u 個の関係 $Q_1, \ldots, Q_u$が結合されるという操作をひとまとめにして，これを逐次実行してできる推移閉包を考えることができる．ただし，1 回の操作の後には，先に定義した演算$\bowtie$と同じように，結合される前の属性がすべて保存されていなければならない．

いま関係 Rは属性として $A_1, \ldots, A_k$をもっており，それら全体の集合を $A^{(0)}$と書く．これに対して関係 $Q_1, \ldots, Q_u$がそれぞれ属性の集合 $A^{(i)}$ $(i = 1, \ldots, u)$ を用いて結合されるとする．すなわち属性集合 $A^{(i)}$に属する属性は関係 Q_iと Rに共通している．そこでこれら一連の結合を

$$R \bowtie (Q_1, \ldots, Q_u)$$

と表すことにする．次に上の結合の結果を，もとの関係 Rの属性集合 $A^{(0)}$に対して射影すると，その結果生成される関係は Rと同じ属性をもっているので，同様の操作を反復することができる．すなわち

$$f(R) \equiv \pi_{\{A^{(0)}\}}(R \bowtie (Q_1, \ldots, Q_u)) \tag{4.4}$$

と定義し，さらに

$$\begin{aligned} R^1 &\equiv R \\ R^k &\equiv \underbrace{f \cdots f}_{f\text{は}k-1\text{個}}(R), \quad k \geq 2 \end{aligned} \tag{4.5}$$

と書くことにすると，作用 f による関係 Rの一般化された推移閉包はやはり

$$R^+ = \bigcup_{k=1}^{\infty} R^k$$

と定義される．

実際にこの推移閉包 R^+を求めるには，前節で用いた Semi-Naive 法にもとづき，以下の計算を逐次行えばよい．すなわち式 (4.1), (4.2) と同様に，

$$\begin{cases} \Delta R^1 := R \\ R^1 := \Delta R^1 \end{cases} \tag{4.6}$$

$$\begin{cases} \delta R^k := f(\Delta R^{k-1}) \\ \Delta R^k := \delta R^k - R^{k-1} \\ R^k := R^{k-1} \cup \Delta R^k, \qquad k = 2, 3, \ldots \end{cases} \tag{4.7}$$

$$R^+ := \bigcup_{k=1}^{\infty} \Delta R^k \tag{4.8}$$

である．

4.4 推移閉包の大きさ

この節では，一般化された推移閉包 R^+の大きさを式 (4.6)〜式 (4.8) にもとづいて評価することを試みる．

4.4.1 作用 f による結果の大きさの評価

ここではまず，推移閉包を求める際の反復の 1 回分としての式 (4.4) の作用 f によって生成される関係の大きさの期待値を考える．

最初に自然結合 $R \bowtie (Q_1, \ldots, Q_u)$ によって得られる関係の大きさの期待値を求める．各関係 Q_iの結合に用いられる属性 $A_j^{(i)} \in A^{(i)}$ $(i = 1, \ldots, u)$ の定義域 $D_j^{(i)} \in D^{(i)}$の大きさを $d_j^{(i)}$，さらに

$$q_i \equiv |Q_i|, \quad d^{(i)} \equiv \prod_{A_j^{(i)} \in A^{(i)}} d_j^{(i)}$$

と定義すると，自然結合に関する式 (3.7) を反復して用いて

$$|R \bowtie (Q_1, \ldots, Q_u)| = r \times \frac{q_1}{d^{(1)}} \times \cdots \times \frac{q_u}{d^{(u)}} \tag{4.9}$$

となる．ここで右辺の分数の部分全体を関係 Rに対する自然結合の作用と考えて

$$p \equiv \frac{q_1}{d^{(1)}} \times \cdots \times \frac{q_u}{d^{(u)}} \tag{4.10}$$

とおき，式 (4.9) を

$$|R \bowtie (Q_1, \ldots, Q_u)| = r \times p$$

と記す．次に属性集合 $A^{(0)}$に対する射影を行うことになるが，このとき各 $A_j^{(0)} \in A^{(0)}$の定義域は $D_j^{(0)}$のままではなく，すでに関係 Rにそれぞれの属性の値として現れたものしか可能性がないわけであるから，自然結合の結果である $R \bowtie (Q_1, \ldots, Q_u)$ における属性 $A_j^{(0)}$の新しい定義域の大きさ $d_j'^{(0)}$は

$$d_j'^{(0)} \approx d_j^{(0)} \left\{ 1 - \left(1 - \frac{1}{d_j^{(0)}} \right)^r \right\} \tag{4.11}$$

と評価するのが妥当であると考える．この右辺は，定義域 $D_j^{(0)}$から r個の要素を等確率で独立に選んだときに得られる相異なる要素の数の期待値である．そこで

$$c \equiv \prod_{A_j^{(0)} \in A^{(0)}} d_j'^{(0)} \tag{4.12}$$

と定義してこれを用いると，最終的に $f(R)$ の大きさの期待値の近似値は，大きさ $r \times p$ をもつ関係 $R \bowtie (Q_1, \ldots, Q_u)$ について，新しい定義域の大きさ $d_j'^{(0)}$をもつ属性集合 $A^{(0)}$ に対して，射影の結果を表す式 (3.6) を用いて

$$|f(R)| \approx c\left\{1-\left(1-\frac{1}{c}\right)^{r\times p}\right\} \tag{4.13}$$

となる．この式が近似式である理由は，式(3.6)が近似式である上に，関係 $R \bowtie (Q_1,\ldots,Q_*)$ の大きさや定義域の大きさ $d_j'^{(0)}$ が確率変数であるにもかかわらず，その分布を考慮せずに期待値を用いているからである．このことを強調するため，式(4.13)を用いて求めることになる δR^k, ΔR^k, R^k の大きさの近似値を，以下ではそれぞれ δs^k, Δs^k, s^k で表すことにする．

4.4.2 漸化式

4.3節で述べたように，関係 R の作用 f による推移閉包 R^+ は式(4.6)~式(4.8)によって求められる．そこで，式(4.7)に対応して，それぞれの大きさの近似式

$$\begin{cases} \delta s^k = c\left\{1-\left(1-\frac{1}{c}\right)^{\Delta s^{k-1}\times p}\right\} & \text{(式(4.13)による)} \\ \Delta s^k = \delta s^k\left(1-\frac{s^{k-1}}{c}\right) & \text{(式(3.4)による)} \\ s^k = s^{k-1}+\Delta s^k \end{cases} \tag{4.14}$$

を得る．この漸化式の上の2式から δs^k を消去すると

$$\begin{cases} \Delta s^k = c\left\{1-\left(1-\frac{1}{c}\right)^{\Delta s^{k-1}\times p}\right\}\left(1-\frac{s^{k-1}}{c}\right) \\ s^k = s^{k-1}+\Delta s^k \end{cases} \tag{4.15}$$

となるので，この漸化式を解けば，

$$s^+ = \sum_{k=1}^{\infty} s^k \tag{4.16}$$

によって $|R^+|$ の期待値の近似値 s^+ を求めることができる．

4.4.3 漸化式の近似的な解法

漸化式(4.15)をこのままの形で厳密に解くことは容易ではない．そこでまず，実用上は $1/c \ll 1$ であることを考慮し漸化式(4.15)の上式を線形化して

$$\begin{aligned} \Delta s^k &= c\left\{1-\left(1-\frac{1}{c}\right)^{\Delta s^{k-1}\times p}\right\}\left(1-\frac{s^{k-1}}{c}\right) \\ &\approx c\left\{1-\left(1-\frac{1}{c}\times\Delta s^{k-1}\times p\right)\right\}\left(1-\frac{s^{k-1}}{c}\right) \\ &= p\left(1-\frac{s^{k-1}}{c}\right)\Delta s^{k-1} \end{aligned} \tag{4.17}$$

と近似する．この近似式は $(1-1/c)^{\Delta s^{k-1}\times p}$ のテイラー展開式の2次以下の項を考慮していないので，p が大きくなるにつれて Δs^k が実際の値より大きくなり，したがって式(4.16)による s^+ も，漸化式を厳密に解いた値より大きいほうへずれる傾向がある．つづいて，式(4.17)を

$$\begin{aligned} \frac{c}{p}\cdot\Delta s^k &\approx (c-s^{k-1})\cdot\Delta s^{k-1} \\ &= c\cdot\Delta s^{k-1}-\Delta s^{k-1}\cdot s^{k-1} \end{aligned}$$

と変形し，$k=2,3,\ldots$ として辺々加えると

$$\frac{c}{p}\left(s^+-\Delta s^1\right) \approx c\cdot s^+ - \frac{1}{2}\left(s^{+2}+\sum_{i=1}^{\infty}\{\Delta s^i\}^2\right) \tag{4.18}$$

となる．ここで右辺の最後の項 $\sum_{i=1}^{\infty}\{\Delta s^i\}^2$ を正確に求めることは困難であるので，これの近似として $\Delta s^1\cdot s^+$, $\Delta s^2\cdot s^+$, $\Delta s^3\cdot s^+$ などを考えたが，数値計算の結果

$$\sum_{i=0}^{\infty}\{\Delta s^i\}^2 \approx \Delta s^3\cdot s^+ \tag{4.19}$$

がもっともよく近似していると判断し(ただし，p が大きいと左辺が大きくなる傾向がある)，これを採用した．ただし，線形化した漸化式(4.17)を反復して用いて

$$\Delta s^3 \approx \frac{p^2}{c^3}\cdot\Delta s^1\cdot(c-p\cdot\Delta s^1)\cdot(c-\Delta s^1)^2 \tag{4.20}$$

である．この結果，式(4.18)より

$$p\cdot s^{+2}+\left(p\cdot\Delta s^3+2c-2cp\right)s^+-2c\cdot\Delta s^1\approx 0$$

が得られ，

$$\begin{aligned} s^+ \approx \frac{1}{2p}\Big\{&-\left(p\cdot\Delta s^3+2c-2cp\right) \\ &+\sqrt{\left(p\cdot\Delta s^3+2c-2cp\right)^2+8cp\cdot\Delta s^1}\Big\} \end{aligned} \tag{4.21}$$

となる．

式(4.21)の近似の精度は，式(4.17)と式(4.19)の誤差が p が大きいところで増大するため，その結果として，本来の値とは大きい方にずれるようである．しかし，後の2項関係に対する実験からわかるように，式(4.21)は $p \le 2$ の範囲なら十分実用的であると判断することができる．

4.4.4 2項関係の場合

4.4.3節で導出した式(4.21)の結果を，実際によく現れる2項関係 R (2つの属性の定義域は等しく D)の推移閉包に対して適用する．この場合1回の演算 $\bowtie$ による式(4.10)の p は

$$p = r/d$$

と考えられるので，これを式(4.13)に代入すると

$$|R\bowtie R| \approx c\left\{1-\left(1-\frac{1}{c}\right)^{r^2/d}\right\} \tag{4.22}$$

となる．この $p=r/d$ と

$$\Delta s^1 = r$$

を式(4.20)と式(4.21)に代入することによって

$$\Delta s^3 = \frac{r^3}{d^2}-\frac{2r^4}{cd^2}+\frac{r^5}{c^2d^2}-\frac{r^5}{cd^3}+\frac{2r^6}{c^2d^3}-\frac{r^7}{c^3d^3},$$

$$\begin{aligned} s^+ \approx \frac{1}{2}\Big\{&-\left(\Delta s^3+\frac{2cd}{r}-2c\right) \\ &+\sqrt{\left(\Delta s^3+\frac{2cd}{r}-2c\right)^2+8cd}\Big\} \end{aligned} \tag{4.23}$$

を得る．

4.5 数値実験

2項関係 Rの推移閉包について数値実験を行い，上記の近似式(4.22)と(4.23)の精度を評価した．

4.5.1 実験1

$|R\bowtie R|$の式(4.22)による近似:

$d=50,100,200$の場合について，rを種々の値に設定し，関係 Rを構成する組をr個ランダムに発生させた後，各Rについて$R\bowtie R$の大きさを得るという実験を行った．これを30回実行してそれらの平均を求めたものをプロットしたのが図1である．図中の実線は式(4.22)の値を，白丸$\circ$は実験結果を表している．

これより式(4.22)は結合$\bowtie$の比較的よい近似になっていることがいえる．

4.5.2 実験2

$|R^+|$の式(4.23)のs^+による近似:

$d=50\sim12800$のいくつかの場合について，4.5.1節と同様の実験を行った．この場合も，それぞれのd,rに対して大きさrの関係を30通りランダムに発生させて，各関係についてその推移閉包の大きさを計算してその平均を実験結果とした．

図2はその結果を表しており，図3は図2の中で$|R|$が1200以下，$|R^+|$が10000以下の部分を拡大したものである．実線は漸化式(4.15)を，$\Delta s^k<0.01$まで厳密に解いた値を，破線は式(4.23)の近似値を(大部分は厳密に解いた値とほぼ等しいので書かれていない)，白丸$\circ$は実験結果を表している．

これより，定義域の大きさdが十分大きく，それと比較して関係の大きさrが小さいような，現実によくおこりうる場合には，かなりよい近似になっていることがわかる．

5 質問に対する計算コストの評価例

この章では，これまでに提案した各種演算の計算コストを評価する方法を用いて，実際に簡単な演繹データベースとそれに対する質問が与えられたときに，必要な計算コストをあらかじめ評価することを試みる．ここで計算コストは，関係を各種演算で処理するたびに生じる新しい(中間)関係の大きさの総和と定義する．

適用例として，よく知られた"同世代関係(sg)"を表す演繹データベースに対して，ある人物aと同世代の人間をすべて求める質問をとりあげる．すなわち，

同世代関係0:

$$\begin{cases} sg(X,X)\leftarrow \\ sg(X,Y)\leftarrow par(X,X1),par(Y,Y1),sg(X1,Y1)\,, \end{cases} \tag{5.1}$$

質問:

$$query(Y)\leftarrow sg(a,Y)$$

である．ここで述語parはEDB(extensional database)述語と呼ばれ，陽に関係(表)の形で与えられるものである．一方，述語sgはIDB(intensional database)述語と呼ばれ，EDB述語に演算を施して求められる関係である．

以上の規則と質問に対して以下の3つの方法で答えることを考える．ただし，例をより具体的にするためにEDB述語である2項関係parの2つの属性の定義域Dは等しく，その大きさ$d=50$，関係parの大きさは$|par|=20$とする．

5.1 方法1

式(5.1)をそれと同値な次の定義式とみなし，答えを求める．

同世代関係1:

$$\begin{cases} sg(X,X)\leftarrow \\ sg(X,Y)\leftarrow par\otimes par(X,Y,X1,Y1),sg(X1,Y1)\,. \end{cases}$$

この方法では，親子関係$par(X,X1)$と$par(Y,Y1)$からその直積である4項関係$par\otimes par(X,Y,X1,Y1)$をつくり，その後その4項関係の属性の前2列と後ろ2列の自然結合を行い，その2列以外に射影することを繰り返すという推移閉包を求める．すなわち，

$$I_1(X,Y,X1,Y1)=par\otimes par(X,Y,X1,Y1)$$

に対して

$$f(I_1)=\pi_{\{X,Y,X2,Y2\}}(I_1(X,Y,X1,Y1)\bowtie par\otimes par(X1,Y1,X2,Y2))$$

によるI_1の推移閉包${I_1}^+$を求めることになる．つぎに，推移閉包${I_1}^+$の第3列と第4列の要素が等しいものだけを選択した後，第1列と第2列に関して射影することによって2項関係$sg(X,Y)$ができる．最後に同世代関係$sg(X,Y)$から第1列の要素がaであるものだけを選択する．

評　価

最初に$|par|=20,d=50$．さらに，関係parの各属性に現れる異なる値の数の期待値は

$$d'=50\left\{1-\left(1-\frac{1}{50}\right)^{20}\right\}=16.62\quad(\text{式}(4.11)\text{による}).$$

- $I_1(X,Y,X1,Y1)=par(X,X1)\otimes par(Y,Y1)$

  $|I_1|=20\times20=400$　(式(3.1)による)．

- $I_2(X,Y,X2,Y2)={I_1}^+(X,Y,X2,Y2)$

  これは上で求めた$I_1(X,Y,X1,Y1)$に対して

  $$f(I_1)=\pi_{\{X,Y,X2,Y2\}}(I_1(X,Y,X1,Y1)\bowtie par\otimes par(X1,Y1,X2,Y2))$$

  による推移閉包${I_1}^+(X,Y,X2,Y2)$を求める．この反復において，

  $$\Delta s^1=|I_1|=400,$$
  $$c=(16.62)^4\quad(\text{式}(4.12)\text{による}),$$
  $$p=\frac{1}{50}\times\frac{1}{50}\times400=0.16\quad(\text{式}(4.10)\text{による})$$

  より

  $$\Delta s^3=10.12\quad(\text{式}(4.20)\text{による}).$$

したがって

$|I_2| = 475.90$ (式(4.21)による).

- $I_3(X,Y,X2,Y2) = \sigma_{\{X2=Y2\}} I_2(X,Y,X2,Y2)$

$|I_3| = 475.90/16.62 = 28.63$ (式(3.5)による).

- $I_4(X,Y) = \pi_{\{X,Y\}} I_3(X,Y,X2,Y2)$

$$|I_4| = 16.62^2 \left\{ 1 - \left(1 - \frac{1}{16.62^2} \right)^{28.63} \right\} = 27.25$$

(式(3.6)による).

- $I_5(Y) = \sigma_{\{X=a\}} I_4(X,Y)$

$|I_5| = 27.25/16.62 = 1.64$ (式(3.5)による).

したがって，方法1によって生成される中間関係の大きさの和(つまり計算コスト)は

$$|I| = |I_1| + |I_2| + |I_3| + |I_4| + |I_5| = 932.52.$$

5.2 方法2

式(5.1)をそれと同値なつぎの定義式とみなして答えを求める方法である.

同世代関係2:

$$\begin{cases} sg(X1,Y1) \leftarrow par(X1,Z), par(Y1,Z) \\ sg(X,Y) \leftarrow par(X,X1), sg(X1,Y1), par(Y,Y1). \end{cases}$$

この方法では最初に親子関係 $par(X1,Z)$ と $par(Y1,Z)$ の第2列目どうしの自然結合を行い関係 $par'(X1,Y1,Z)$ をつくり，その第1列と第2列に対してと射影を行って関係 $sg(X1,Y1)$ を求める．次に，上で求めた関係 $sg(X1,Y1)$ に対して関係 $par(X,X1)$, $par(Y,Y1)$ の共通する属性どうしの結合を行ったのち属性 X と Y に対する射影を行う，という操作を繰り返して推移閉包を求める．すなわち，

$$I_1(X1,Y1) = sg(X1,Y1)$$

に対して

$$f(I_1) = \pi_{\{X,Y\}}(I_1(X1,Y1) \bowtie (par(X,X1), par(Y,Y1)))$$

である．最終的に，求められた関係 $sg(X,Y)$ の第1列の要素が a であるものについてだけ選択する.

評 価

最初に $|par| = 20, d = 50, d' = 16.62$.

- $I_1(X1,Y1) = \pi_{\{X1,Y1\}}(par(X1,Z) \bowtie par(Y1,Z))$

この式において

$|par| = 20$,
$c = 16.62^2$ (式(4.12)による).

したがって

$$|I_1| = 16.62^2 \left\{ 1 - \left(1 - \frac{1}{16.62^2} \right)^{\frac{20\times 20}{16.62}} \right\} = 23.09$$

(式(3.6)による).

- $I_2(X,Y) = {I_1}^+(X,Y)$

これは上で求めた $I_1(X1,Y1)$ に対して

$$f(I_1) = \pi_{\{X,Y\}}(I_1(X1,Y1) \bowtie (par(X,X1), par(Y,Y1)))$$

による推移閉包 ${I_1}^+(X,Y)$ を求めるものである．この反復において

$\Delta s^1 = 23.09$ (上の I_1 の結果より),
$c = 16.62^2$ (式(4.12)による),
$p = \frac{1}{50} \times \frac{1}{50} \times 20 \times 20 = 0.16$ (式(4.10)による)

より

$\Delta s^3 = 0.49$ (式(4.20)による).

したがって

$|I_2| = 27.23$ (式(4.21)による).

- $I_3(Y) = \sigma_{\{X=a\}} I_2(X,Y)$

$|I_3| = 27.23/16.62 = 1.64$ (式(3.5)による).

したがって，全体の計算コストは

$$|I| = |I_1| + |I_2| + |I_3| = 51.95.$$

5.3 方法3

同世代関係3:

$$\begin{cases} m\text{-}sg(a) \leftarrow \\ m\text{-}sg(Y) \leftarrow m\text{-}sg(X), par(X,Y), \end{cases}$$

$$\begin{cases} sg(X1,Y1) \leftarrow m\text{-}sg(X1), par(X1,Z), par(Y1,Z) \\ sg(X,Y) \leftarrow m\text{-}sg(X), par(X,X1), sg(X1,Y1), par(Y,Y1). \end{cases}$$

この方法はMagic Set法と呼ばれ，上の2式であらかじめ a および a の先祖だけを関係 $m\text{-}sg(Y)$ として求め，次に $m\text{-}sg(Y)$ の要素だけに範囲を限定してそれらと同世代であるものだけを求めるものである．最後に $sg(X,Y)$ の第1列の要素が a であるものについて選択するのは前の方法と同じである.

評 価

まず，前の方法と同様に，$|par| = 20, d = 50, d' = 16.62$.

- $I_1(Y) = m\text{-}sg^+_{(J_1)}(Y)$

これは $m\text{-}sg(X) = m\text{-}sg(a)$ に対して

$$f_1(m\text{-}sg(a)) = \pi_{\{Y\}}(m\text{-}sg(X) \bowtie par(X,Y))$$

による推移閉包 $m\text{-}sg^+_{(J_1)}(Y)$ を求めるものであるが，その結果は $par(X,Z)$ に対して

$$f_1'(par) = \pi_{\{X,Y\}}(par(X,Z) \bowtie par(Z,Y))$$

による推移閉包 $par^+_{(J_1')}(X,Y)$ を求め，さらに

$$I_1(Y) = \sigma_{\{X=a\}} par^+_{(J_1')}(X,Y)$$

という選択を行った結果と同じである．すなわち，式(4.23)において

$$r = 20, d = 50, c = 16.62^2, \Delta s^3 = 2.67$$

とおいて $|par^{+}_{(I_1')}| = 31.99$ を得て，したがって式(3.5)より

$$|I_1| = 31.99/16.62 = 1.93\,.$$

- $I_2(X, X1) = I_1(X) \bowtie par(X, X1)$

この計算は，関係 $I_1(X)$ と $par(X, X1)$ の属性 X を用いての結合を行う．したがって，ここでは関係 I_1 の第1列の X と関係 par の第1列の X の任意の要素が結合される確率を知る必要がある．しかし，この時点では関係 I_1 は確率的にランダムに構成されているとはいえないので，その確率は $1/d$ や $1/d'$ ではない．ここでは，関係 $I_2(X, X1)$ が，親子関係 $par(X, X1)$ に対して1番目の属性 X の値が関係 $m\text{-}sg(X)$ の要素である組だけについて選択を行ったものであることに注目して，関係 I_2 の大きさを評価することにする．すなわち

$$|I_2| = 20 \times \frac{1}{16.62} + 20 \times \left(1.93 - \frac{1.93 \times 1}{50}\right) \times \frac{1}{50} = 1.96\,.$$

この式の右辺の和の左半分は関係 $par(X, X1)$ において $X = a$ である組の個数を，右半分は X が a 以外の a の先祖である組の個数の期待値を表している．

- $I_3(X1, Y1) = \pi_{\{X1,Y1\}}(I_2(X1, Z) \bowtie par(Y1, Z))$

この式において $|I_2| = 1.96$ で，いま関係 $I_2(X1, Z)$ の属性 $X1$ に現れる異なる値の数は，a と a の先祖の数であるから，式(3.3)より

$$1 + 1.93 - \frac{1.93 \times 1}{50} = 2.89$$

である．したがって，

$$c = 2.89 \times 16.62 = 47.97$$

より

$$|I_3| = 47.97\left\{1 - \left(1 - \frac{1}{47.97}\right)^{\frac{1.96 \times 20}{16.62}}\right\} = 2.32\,.$$

- $I_4(X, Y) = I^{+}_{3(f_2)}(X, Y)$

これは $I_3(X1, Y1)$ として求めた $sg(X1, Y1)$ に対して

$$f_2(sg(X1, Y1)) = \pi_{\{X,Y\}}(I_3(X1, Y1) \bowtie (I_2(X, X1), par(Y, Y1)))$$

による推移閉包 $I^{+}_{3(f_2)}(X, Y)$ を求めるものである．ここで

$$\Delta s^1 = |I_3| = 2.32,$$
$$c = 47.97$$

である．次に1回の結合による式(4.10)の p を求めるために，対応する各属性が結合される確率を考える．はじめに関係 I_3 の属性 $X1$ が関係 I_2 の属性 $X1$ と結合される確率を考えると，関係 I_3 の属性 $X1$ の要素は，a と a 以外の a の先祖が，関係 I_2 と同じく 1.20 : 1.75 の割合で出現すると考えられる．これらが関係 I_2 の属性 $X1$ と結合される確率は，a の場合は a の先祖として a が現れる確率であるが，これは関係 par のランダム性より 1/50 と考えられる．一方，a 以外の a の先祖の場合は，関係 I_2 の属性 $X1$ に a の先祖が等確率で現れると考え，$1/|I_1|$ とする．したがって全体としての確率は

$$\frac{1.20}{1.95} \times \frac{1}{50} + \frac{1.75}{1.95} \times \frac{1}{1.93} = 0.21$$

である．一方，関係 I_3 の属性 $Y1$ と関係 par の属性 $Y1$ が結合される確率は関係 par のランダム性より 1/50 である．以上のことから

$$p = 0.21 \times \frac{1}{50} \times 1.96 \times 20 = 0.17$$

と推定される．そこで

$$\Delta s^3 = 0.058 \quad (\text{式}(4.20)\text{による})$$

と求めることができるので

$$|I_4| = 2.77\,.$$

となる．

- $I_5(Y) = \sigma_{\{X=a\}} I_4(X, Y)$

最後に関係 $I_4(X, Y)$ の中で $X = a$ である組だけを選択する．ここで属性 X には a と a 以外の a の先祖が 1.20 : 1.75 の割合で存在しているはずなので

$$|I_5| = 2.77 \times \frac{1.20}{1.96} = 1.70\,.$$

したがって，全体としての計算コストは

$$|I| = |I_1| + |I_2| + |I_3| + |I_4| + |I_5| = 8.08\,.$$

となる．

このように，同世代問題の答えを導く3とおりの方法に対してそれぞれの計算コスト $|I|$ は

方法1： $|I| = 932.52$,
方法2： $|I| = 51.95$,
方法3： $|I| = 8.08$

となり，それらの差は歴然としている．すなわち，一般に認められているようにMagic Set法の効率が良いことが確かめられる．なお，最終的に求められる答えの大きさは，3方法それぞれについて

$$|I_5| = 1.64, |I_3| = 1.64, |I_5| = 1.70$$

と推定され，いくつかの演算の複合であるにもかかわらずほぼ近い値となっている．このことからも，本論文の近似式がかなり精度の良いものであることが予想される．

5.4 数値実験：同世代問題の方法1による近似

$d = |D| = 50, 100, 200$ に対して $r = |par|$ のいくつかの組合せについて，それぞれ30通りランダムにデータを発生させ，5.1節の方法1にもとづいて実行処理し，各 $|I_j|$ の平均値を求めた．表1の(a)から(c)にその結果を示す．

これらの実験の結果，$|I_2|$ と $|I_3|$ の値に多少の誤差はあるものの，それらも含めて全体によい近似になっているといえよう．

6 おわりに

関係に適用される典型的な演算に対し，それらを適用した結果の関係の大きさを知るために本文で提案した近似式は，期待値の比較的よい評価値をあたえることがわかった．新しい関係の大きさを，このように定義域ともとの関係の大きさだけから簡単に推定できることは，演繹データベースの処理

にともなう計算コストをあらかじめ評価できることを意味し，効率よい処理法を見出す上で意味が大きいと考えられる．実際，同世代問題の例では，質問に対する答えを導く方法によってそれらの計算コストが大きく異なることが簡単に評価でき，一般に認められているように，Magic Set 法が有効であることも確かめられた．

なお，現実の関係では，定義域の各要素が一様に分布している場合は少なく，その意味で本文での関係表に対する仮定は一般的ではない．したがって，今後の課題として，定義域中の各値の生起確率が偏っている場合の計算コストの近似評価式について考えていきたい．また，各 r や d に対して，演算の結果得られる関係の大きさの期待値ばかりでなく分散も求めることができればよいが，これはかなり困難である．

謝　辞

本論文を書くにあたって，貴重な助言やアイデアをいただいた本研究室の大西匡光助手に感謝の意を表します．また本研究は一部文部省科学研究費によるものである．

参考文献

[1] Agrawal, R. and Jagadish, H. V.: "Direct Algorithms for Computing the Transitive Closure of Database Relations," Proceeding of the 13th VLDB Conference, 1987.

[2] Bayer, R.: "Database Technology for Expert Systems," International GI-Kongress, 85, Wissenbasiete Systeme, Informatik Fachberichte 112, 1985.

[3] Ioannidis, Y. E. and Ramakrishnan, R.: "Efficient Transitive Closure Algorithms," Proceeding of the 14th VLDB Conference, 1988.

[4] Lipton, R. J. and Naughton, J. F.: "Estimating the Size of Generalized Transitive Closures," Proceeding of the 15th VLDB Conference, 1989.

[5] Lipton, R. J. and Naughton, J. F.: "Query Size Estimation by Adaptive Sampling," Proceeding of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on PODS, 1990.

[6] Lu, H.: "New Strategies for Computing the Transitive Closures of a Database Relation," Proceeding of the 13th VLDB Conference, 1987.

[7] Pittel, B.: "On Distribution related to Transitive Closure of the Random Finite Mappings," Annals of Probabilistics, 11, 1983.

[8] Sippu, S. and Soisalon-Soininen, E.: "A Generalized Trasitive Closure for Relational Queries," Proceeding of the 7th ACM SIGACT-SIGMOD-SIGART Symposium on PODS, 1988.

[9] Ullman, J. D.: *Principles of Database and Knowledge-Base Systems, Vol. 1*, Computer Science Press, 1989.

r	区別	$\|I_2\|$	$\|I_3\|$	$\|I_4\|$	$\|I_5\|$	$\|I\|$
20	理論	475.90	28.63	27.25	1.64	533.42
	実験	471.40	39.40	24.05	1.50	536.35
30	理論	1404.00	61.78	58.28	2.56	1526.62
	実験	1365.80	103.60	49.95	2.10	1521.45
40	理論	4410.81	159.15	143.83	5.19	4718.98
	実験	5334.30	317.60	93.35	5.00	5750.25

(a) $d = 50$ の場合.

r	区別	$\|I_2\|$	$\|I_3\|$	$\|I_4\|$	$\|I_5\|$	$\|I\|$
40	理論	1904.47	57.53	56.07	1.69	2019.76
	実験	1952.50	96.80	53.65	1.85	2104.80
50	理論	3332.54	84.37	82.15	2.08	3501.14
	実験	3437.85	146.05	70.50	2.20	3656.60
60	理論	5622.71	124.16	120.51	2.66	5870.04
	実験	5779.00	230.60	93.65	2.60	6105.85
70	理論	9600.21	190.04	183.17	3.63	9977.05
	実験	10410.30	421.55	147.30	3.65	10982.80

(b) $d = 100$ の場合.

r	区別	$\|I_2\|$	$\|I_3\|$	$\|I_4\|$	$\|I_5\|$	$\|I\|$
40	理論	1666.63	45.87	45.10	1.24	1758.84
	実験	1739.25	65.45	45.25	1.35	1851.30
60	理論	3955.94	76.15	75.10	1.45	4108.64
	実験	4008.10	111.50	71.00	1.45	4192.05
80	理論	7618.75	115.31	113.82	1.72	7849.60
	実験	7760.05	185.75	101.25	1.45	8048.50
100	理論	13332.53	169.10	166.83	2.12	13670.58
	実験	13697.80	321.00	153.35	1.90	14174.65

(c) $d = 200$ の場合.

表 1: 同世代問題の答えを方法 1 で求める際に生成される中間関係の大きさの評価例

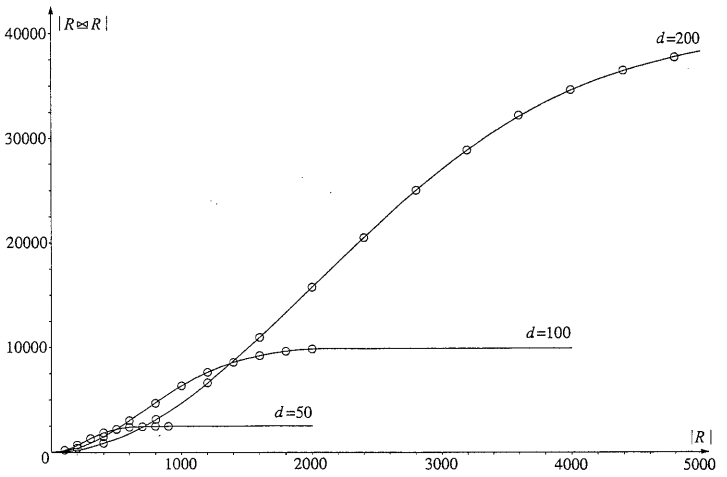

図 1 : $|R \bowtie R|$ の式 (4.22) による近似

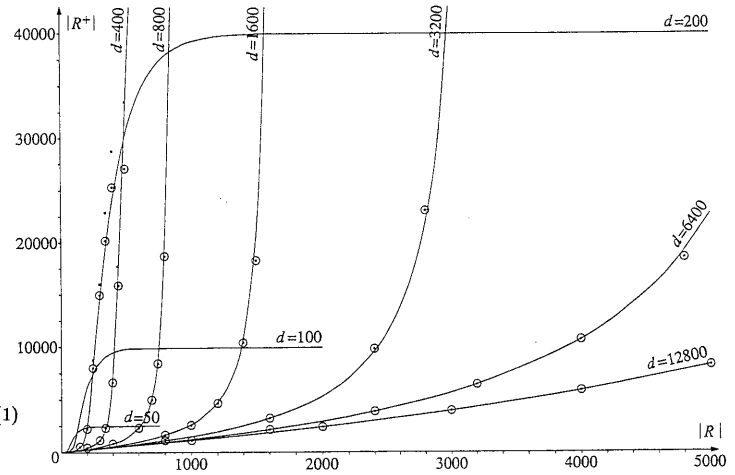

図 2 : $|R^+|$ の式 (4.23) の s^+ による近似 (1)

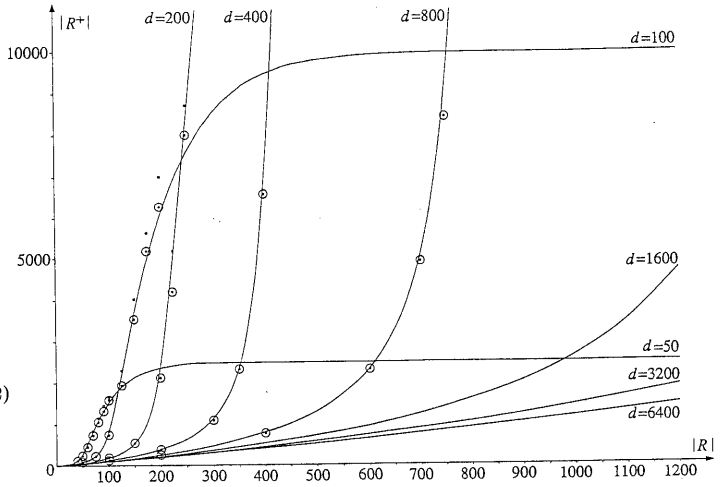

図 3 : $|R^+|$ の式 (4.23) の s^+ による近似 (2)