関係の推移閉包の大きさの近似的推定法

宇野裕之† 茨木俊秀††

与えられた2項関係の推移閉包の大きさを知る方法としては，実際に推移閉包を計算してみるか，あるいはごく単純な近似的方法しか知られていない．そこで本論文では，対象とする関係中の各成分が，ある一般的な生起確率にしたがって，他の成分とは独立に定義域から選ばれているという前提のもとに，その確率分布と関係の大きさの情報だけから推移閉包の大きさを簡単に推定する近似法を提案する．さらに，得られた近似の精度を Zipf 分布を用いて実験によって調べる．

Approximate Evaluation to the Sizes of Transitive Closures of Relations

Yushi Uno† and Toshihide Ibaraki††

Given binary relations, only few very simple methods are known to evaluate the size of their transitive closures. In this paper, we assume that the components of relations are randomly chosen from the general probability distribution, and propose an approximate formula that estimates the sizes of the transitive closures to the original relations, based only on the sizes of the original relations and the given probability distribution. We also de monstrate the effectiveness fo our approximation by numerical experiments based on Zipf distribution.

1. はじめに

推移閉包の計算は，演算データベースにおいて再帰的な質問を処理する際に不可欠であり[1),2),6),8)]，推移閉包の大きさをあらかじめ推定しておくことは，必要な計算量の大ざっぱな評価を与えるという意味で，効率よい処理法を見出すための基本的な役割を果たす．しかし，推移閉包の大きさの推定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た[4),5),7)]．われわれは，文献10)において，もとになる関係の各成分が，一様分布にしたがって，他とは独立に確率的に生成されたという前提のもとに，推移閉包やその他の演算の計算量を簡便に評価する近似式を与えた．しかし，一様分布の仮定は一次近似として用いるという意味では妥当であっても，実際には関係をつくる定義域中に現れる値の生起確率が（たとえば日本における姓のように）事前にある程度知られていることも多く，その情報に基づくより精度の高い推定が望まれる．

そこで本論文では，最初に，演繹データベースの処理のさまざまな局面で用いることができるように，一般化された推移閉包を定義した後，各属性の定義域における値の生起確率が既知であるとして，推移閉包の大きさを近似的に評価する方法を提案する．そのために用いる方法は，文献10)の方法を拡張した形になっているが，一般分布を扱うために本質的に新しい工夫が各所で必要となる．得られた結果も文献10)の近似式の単なる拡張ではなく，一様分布の場合においても精度を高めたものになっている．

最後に，得られた近似式の精度を Zipf 分布[3),11)]を例として用いて調べ，定義域中の生起確率に極端なかたよりのある分布から一様分布に近い分布まで，広い範囲で有効であることを確かめる．

2. 関係および関係に関する仮定

2.1 関　　係

関係 (relation) R とは，定義域 (domain) $D_i(i=1,\cdots,m)$ の直積 (Cartesian product) $D_1\times\cdots\times D_m$ の部分集合である．定義域の数が m 個のとき m 項関係 (m-ary relation) といい，定義域の直積の要素を m

† 大阪府立大学総合科学部
College of Integrated Arts and Sciences, University of Osaka Prefecture
†† 京都大学工学部
Faculty of Engineering, Kyoto University

個組（m-tuple）（あるいは単に組（tuple））という．本文中では，通常の慣例にしたがい，各定義域 D_i の大きさ d_i は有限とする．

関係 R のすべての相異なる組を陽に表の形に並べたものを関係表（relational table）という．その各行は組であり，各列は各定義域 D_i の名前 $A_i(i=1,\cdots,m)$（これを属性（attribute）という）に対応する．各行の各成分は，その列の定義域に属するある値を保持している．一般に，関係 R の大きさ（size）あるいはその期待値を $|R|$ または r と記す．大きさがあらかじめ定まっている関係の大きさは確率変数ではなく，したがって期待値ではないが，関係代数の演算の結果できる関係の大きさは確率変数と考えることができる．本論文では簡単のためこれらを区別することをせず，ともに大きさ r と記す．

2.2 関係に関する仮定

本論文では，定義域 D_i の値 j の生起確率 p_{ij} はあらかじめわかっているとし，さらに関係 R の各成分が，この確率分布にしたがって互いに独立に生成されたと仮定して議論を進める．厳密には，1つの関係に同じ組は2つ以上存在しないことを前提としているので，重複する組を除いた結果，独立性は成立せず，値の分布も若干異なってくるが，近似的に独立性を仮定しても，大きな差異は生じないと考えられる．また，生起確率 p_{ij} は既知であるとするが，そうでない場合であっても，R の属性 A_i において値 j が生起している回数 n_{ij} を用いて，$p_{ij}\approx n_{ij}/r$ とするなどの簡便法をとれば，実用的には十分であろう．

この独立性の仮定に対し，現実に遭遇する関係では，属性間の独立性は仮定できないという議論もある．たとえば，親と子の関係を表す $par(X,Y)$ では，親の姓名と子の姓名は，（女性が結婚後，姓を変えるとしても）同じ姓である場合が多いという意味で，高い確率的相関性をもつ．しかし，推移閉包の例によく用いられる，空港 X から空港 Y へ飛行機の直行便があることを示す関係 $flight(X,Y)$ や，科目 X とその担当教官 Y の関係を表す $class(X,Y)$ などにおいて，それぞれの名前が生起する頻度には偏りがあるとしても（これは本論文では考慮される），空港名の間，あるいは科目名と担当教官名の間に，あらかじめ確率的な相関が存在するとは考え難い．したがって，本論文の結果は，独立性の仮定のもとでの結果であることを認識しつつ，それが近似的に許される場合に限定して利用すれば，有益な情報を提供すると考えられる．もちろん，理論的な立場からすれば，独立性を仮定しなければその解析はきわめて困難であり，意味のある結果を導けないという理由もある．

ところで，出発点となる関係 R の確率的性質を上のように仮定しても，推移閉包を計算するための各種の演算を適用すると，その都度生成される関係の属性における値の分布は変化する．したがって，次に施す演算の結果の評価にこの分布の変化を反映させなければならない．この点は，一様分布の場合[10]にはどのような演算を施してもその結果の関係における値の分布がまた一様であったのと大きく異なる．以下の解析では，この変化をうまく近似して評価することが主要なテーマとなる．

3. 推移閉包演算を構成する関係代数の基本演算

われわれは次の章で，推移閉包を Semi-Naive 法で実行計算する場合にもとづいてその大きさを評価するが，その方法を分解して考えると，いくつかの関係代数の基本的な演算から構成されている．そこでそれらの演算およびその他の基本的な演算[9]，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）について，それらによって生成される新しい関係の大きさについて，その期待値（の近似値）を評価する．本章での導出の考え方は，文献10）の一様分布の場合とほぼ同じであるので，詳しい説明は省略する．

3.1 直 積 ⊗

属性 $A_1,\cdots,A_m$ をもつ関係 R_1 と属性 $B_1,\cdots,B_n$ をもつ関係 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，属性 $A_1,\cdots,A_m$, $B_1,\cdots,B_n$ をもつ関係である．関係 R_1 と R_2 の大きさをそれぞれ r_1, r_2 とすると，

$$|R_1\otimes R_2|=r_1r_2 \tag{3.1}$$

である．

3.2 共通集合∩，集合和∪，集合差－

関係 R_1 と R_2 の大きさはそれぞれ r_1, r_2 であり，いずれも定義域 D_i の属性 A_i $(i=1,\cdots,m)$ をもつ．属性 A_i における値 j の生起確率はそれぞれ p_{ij}, q_{ij} とする．

● 共通集合∩

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．共通集合の大きさの期待値は

$$|R_1\cap R_2|\approx\sum_{j_1=1}^{d_1}\cdots\sum_{j_m=1}^{d_m}\{1-(1-p_{1j_1}\cdots p_{mj_m})^{r_1}\}\cdot\{1-(1-q_{1j_1}\cdots q_{mj_m})^{r_2}\}\qquad(3.2)$$

である．一般に，$p_{1j_1}\cdots p_{mj_m}$ や $q_{1j_1}\cdots q_{mj_m}$ は1に比べて十分小さくなるので，上式を線形近似して

$$|R_1\cap R_2|\approx r_1r_2\cdot\sum_{j_1=1}^{d_1}\cdots\sum_{j_m=1}^{d_m}p_{1j_1}\cdots p_{mj_m}q_{1j_1}\cdots q_{mj_m}$$

とすることができる．特に一様分布の場合は

$$|R_1\cap R_2|\approx r_1r_2\cdot\frac{1}{d_1\cdots d_m}$$

となる．

●集合和∪

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．共通集合の結果を利用して，次のようになる．

$$|R_1\cup R_2|=r_1+r_2-|R_1\cap R_2|.\qquad(3.3)$$

●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．

$$|R_1-R_2|=r_1-|R_1\cap R_2|.\qquad(3.4)$$

3.3 選択 σ，射影 π

●射影

射影 π_A は，1つの関係 R から特定の属性の集合 $A=\{A_{i_1},\cdots,A_{i_k}\}$ についてだけ抜き出して新しい関係をつくる演算である．このとき他の属性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．

$$|\pi_AR|\approx\sum_{j_1=1}^{d_{i1}}\cdots\sum_{j_k=1}^{d_{ik}}\{1-(1-p_{i_1j_1}\cdots p_{i_kj_k})^r\}.\qquad(3.5)$$

●選択

選択 σ_C とは関係 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をつくる演算である．いま条件 C を

$$C=\{A_{i_1}=j_1,\cdots,A_{i_k}=j_k\}$$

とすると，関係 R に対する選択後の関係の大きさの期待値は次式で与えられる．

$$|\sigma_CR|\approx p_{i_1j_1}\cdots p_{i_kj_k}\times r.\qquad(3.6)$$

3.4 自然結合 ⋈

属性 $A_1,\cdots,A_m$ からなる関係 R_1 と属性 $B_1,\cdots,B_n$ からなる関係 R_2 がある．R_1 の属性 A_i での値 j の生起確率を p_{ij}，R_2 の属性 B_i での値 j の生起確率を q_{ij} とする．R_1 と R_2 の等結合 (equijoin) とは，指定された属性の組 $A_{i_1},\cdots,A_{i_c}$ (定義域 $D_{i_1},\cdots,D_{i_c}$) と $B_{j_1},\cdots,B_{j_c}$ (定義域 $D_{j_1},\cdots,D_{j_c}$) に対し，$D_{i_1}=D_{j_1},\cdots,D_{i_c}=D_{j_c}$ が成立するとき，R_1 の組の属性 $A_{i_1},\cdots,A_{i_c}$ の成分と R_2 の組の属性 $B_{j_1},\cdots,B_{j_c}$ の対応する各成分が特にそれぞれ等しいときにそれらの組の連接をつくり，得られた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 $(m+n-c)$ 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．新しい関係の大きさの期待値は次のようになる．

$$|R_1\bowtie R_2|\approx r_1r_2\prod_{s=1}^{c}\sum_{k=1}^{d_{i_s}}p_{i_sk}\cdot q_{j_sk}.\qquad(3.7)$$

特別な場合として，共通の定義域 D をもつ属性 A_1，A_2 からなる2項関係 R を考え，属性 A_1 と A_2 の定義域 D における値 i の分布をそれぞれ p_i，q_i $(i=1,\cdots,d)$ とする．このとき A_2 と A_1 に関して自然結合 $R\bowtie R$ を行ってできる新しい関係の大きさの期待値は次式で与えられる．

$$|R\bowtie R|\approx r\times r\times\sum_{i=1}^{d}p_iq_i.\qquad(3.8)$$

4. 推移閉包

4.1 推移閉包の一般化とその定義

推移閉包は通常は2項関係に対して議論されるが，ここでは1つの関係 R に，複数の関係 $Q_1,\cdots,Q_u$ が結合され射影されることが繰り返される，一般化された推移閉包を定義する[10]．

ある m 項関係 R は属性集合 $A^{(0)}=\{A_1,\cdots,A_m\}$ からなり，この関係 R に対して u 個の関係 $Q_1,\cdots,Q_u$ がそれぞれ属性集合 $A^{(h)}(h=1,\cdots,u)$ を用いて順次自然結合され，$A^{(h)}$ に属する属性と共通の定義域をもつ属性が $A^{(0)}$ にもあり，それらどうしが自然結合されると仮定している．関係 Q_h と他の関係 Q_l $(l\neq h)$ の間では，共通の属性が存在しないか，あるいは存在していても，それらどうしが結合されることはない．このことを強調するために，これら一連の結合を

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

と表すことにする．次に上の結合の結果を，属性集合 $B=\{B_1,\cdots,B_m\}$ に対して射影する．ただし，各 B_i は R と $Q_1,\cdots,Q_u$ 内の属性から選ばれ，B_i の定義域 D_i は $A^{(0)}$ の属性 A_i の定義域に等しいとする．簡単のため，一般性を失うことなく，$B_1,\cdots,B_v$ は R から，残りの $B_{v+1},\cdots,B_m$ は $Q_1,\cdots,Q_u$ から選ばれていると

考える．そこで

$$\varphi(R)\equiv\pi_B(R\bowtie(Q_1,\cdots,Q_u)) \tag{4.1}$$

と定義すると，$\varphi(R)$ に対して同様の操作で

$$\varphi(\varphi(R))\equiv\pi_B(\varphi(R)\bowtie(Q_1,\cdots,Q_u))$$

をつくることができる．したがって，

$$R_1\equiv R$$
$$R_k\equiv\underbrace{\varphi\cdots\varphi}_{\varphi\text{ は }k-1\text{ 個}}(R),\quad k\geq 2 \tag{4.2}$$

と書くと，関係 R の一般化された推移閉包は

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

と定義される．

たとえば，もっとも典型的な同じ定義域 D をもつ2項関係 R の推移閉包 R^+ は，もとの関係 $R(A^{(0)}=\{A_1,A_2\})$ に対して，$Q=R(A^{(1)}=\{A_3,A_4\}$，ただし $A^{(1)}=A^{(0)}=\{A_1,A_2\})$ が，属性 A_2 と A_3 で自然結合される．さらにこれを属性集合 $B=\{A_1,A_4\}$ $(=\{A_1,A_2\})$ に対して射影すると，

$$\varphi(R)=\pi_B(R(A_1,A_2)\bowtie_{A_2=A_3}Q(A_3,A_4))$$

となり，以後 $\varphi(R)$ に対して同じ操作 φ を適用できる．

また，同世代問題を考えるときに現れる例[10]の場合には，関係 $R(X_1,Y_1)$ に対して $Q_1(X,X_1)$，$Q_2(Y,Y_1)$（すべての属性の定義域は共通の D）を，X_1 どうし，および Y_1 どうしで自然結合し，$B=\{X,Y\}$ に対して射影する

$$\varphi(R)=\pi_B(R\bowtie(Q_1,Q_2))$$

が用いられる．

4.2 推移閉包の大きさ

4.2.1 Semi-Naive 法

一般化された推移閉包 R^+ の大きさは，推移閉包を求める方法によらず同じである．実際に推移閉包を計算するためにはいろいろな方法が考案されているが，ここでは文献 10) と同様に Semi-Naive 法[6),9)]をとりあげ，それにもとづいて大きさの評価をすることを試みる．Semi-Naive 法は以下の計算を逐次行う．

$$\begin{cases}\Delta R_1:=R\\ R_1:=\Delta R_1\end{cases} \tag{4.3}$$

$$\begin{cases}R'_k:=\Delta R_{k-1}\bowtie(Q_1,\cdots,Q_u)\\ \delta R_k:=\pi_B R'_k\\ \Delta R_k:=\delta R_k-R_{k-1}\\ R_k:=R_{k-1}\cup\Delta R_k,\quad k=2,3,\cdots\end{cases} \tag{4.4}$$

$$R^+:=\bigcup_{k=1}^{\infty}\Delta R_k. \tag{4.5}$$

この方法の1回ごとの反復は，すべて前章で定義した関係代数の基本的な演算に分解することができる．具体的には，式(4.4)の上から順に自然結合，射影，集合差，集合和である．したがって，演算のたびに生成される関係の大きさと，関係の属性における値の分布を知れば，推移閉包の大きさの期待値が求められるはずである．

関係 R の各属性において値が一様分布していれば，各種演算を施しても一様分布の性質が保たれることは容易にわかるが，一般分布を想定している本論文の立場からは，中間関係における値の確率分布を反復のつど求めなければならない．しかし，互いに独立ではなくなるこれらの分布を忠実に再現するのは，実際上は不可能であり，もともと最初の関係 $\Delta R_1=R$ 内の値の分布が経験に基づいて近似的に与えられる場合が多いことを考慮すると，近似的な推定でも十分であろう．初期状態の $\Delta R_1=R_1=R$ から反復1回目で得られる $k=2$ では，各属性における値の確率分布はもとの分布に近いと考えられるが，反復とともにそれらの分布はなまり一様分布に近づいていく．これは，反復のたびに重複を削除するうえに，新しく付け加えられる組はそれまでに現れていないものだけであることから，各値の生起確率が均等になっていく傾向が強いからである．

そこで以下では，最初の分布をもとにした1回目の反復と，一様分布をもとにした2回目以降の反復とにわけて推移閉包の大きさの近似的評価を試みる．また簡単のため，以後の議論では各種期待値の近似値を単に期待値とよび，記号≈ではなく等号で表すことにする．

4.2.2 1回目の反復

まず，関係 R では，各属性 A_i $(i=1,\cdots,m)$ の値 j が確率 p_{ij} $(j=1,\cdots,d_i)$ で出現し，関係 Q_h $(h=1,\cdots,u)$ では属性 $A_i^{(h)}$ の値 j の出現確率は $t_{ij}^{(h)}$ であるとする．

式(4.4)の評価において，1回目の反復は，

$$\begin{cases}R'_2=R\bowtie(Q_1,\cdots,Q_u)\\ \delta R_2=\pi_B R'_2\\ \Delta R_2=\delta R_2-R\\ R_2=R\cup\Delta R_2\end{cases} \tag{3.6}$$

であり，R と Q_h での確率分布を用いて比較的実際に近く評価できる．

最初に関係 R'_2 の大きさの期待値を求める．各関係 Q_h（その大きさは q_h）の結合に用いられる属性 $A_i^{(h)}\in A^{(h)}$ のすべてが，R のそれぞれ対応する属性と結合

する確率 f_h は，

$$f_h=\prod_{A_i^{(h)}\in A^{(h)}}\sum_{j=1}^{d_i}p_{ij}\cdot t_{ij}^{(h)} \quad (h=1,\cdots,u)$$

となる．したがって，自然結合に関する式 (3.7) を反復して用いて，

$$|R_2'|=r\times f_1q_1\times\cdots\times f_uq_u$$

となる．つづいて関係 R_2' の属性集合 B に対する射影を考える．このとき，R_2' の任意の属性 A_i にはもとになった R や Q_h の対応する属性 A_i の成分としてすでに現れたものしか現れる可能性がないわけであるから，その定義域の大きさ d_i' を R や Q_h の属性 A_i および $A_i^{(h)}$ に現れる異なる値の種類の期待値

$$d_i'=\sum_{j=1}^{d}\{1-(1-p_{ij})^r\} \quad (i=1,\cdots,v)$$

$$d_i'=\sum_{j=1}^{d}\{1-(1-t_{ij}^{(h)})^{q_h}\} \quad (i=v+1,\cdots,m)$$

と評価する．これら d_i' 個の異なる値の分布については，厳密には実際に現れた値の集合それぞれを考慮して導出しなければならないが容易ではない．そこでここでは，初期分布が若干変形したものと考えて，最も生起確率の高い値から $\lceil d_i'\rceil$ 個（一般性を失うことなく添字の小さいものから $\lceil d_i'\rceil$ 個）が順に，初期分布を正規化した確率で出現すると仮定する．すなわち，

$$p_{ij}'=\frac{p_{ij}}{\sum_{j=1}^{\lceil d_i'\rceil}p_{ij}},\ t_{ij}'^{(h)}=\frac{t_{ij}^{(h)}}{\sum_{j=1}^{\lceil d_i'\rceil}t_{ij}^{(h)}} \quad (j=1,\cdots,\lceil d_i'\rceil)$$

で代替するわけである．結局，射影 δR_2 の大きさは，これらの属性の中で射影される属性 $A_i\in B$ に関する分布だけを用いて式(3.5)より次のように評価できる．

$$|\delta R_2|=\sum_{j_1=1}^{\lceil d_1'\rceil}\cdots\sum_{j_v=1}^{\lceil d_v'\rceil}\sum_{j_{v+1}=1}^{\lceil d_{v+1}'\rceil}\cdots\sum_{j_m=1}^{\lceil d_m'\rceil}\{1-(1-p_{1j_1}'\cdots p_{vj_v}'\cdot t_{v+1j_{v+1}}'^{(h)}\cdots t_{mj_m}'^{(h)})^{|R_2'|}\}\ (h\in\{1,\cdots,u\}).$$

つぎに，この δR_2 から R との重複を削除した関係 ΔR_2 の大きさを求めると，

$$\begin{aligned}|\Delta R_2|&=|\delta R_2|-\sum_{j_1=1}^{\lceil d_1'\rceil}\cdots\sum_{j_m=1}^{\lceil d_m'\rceil}\{1-(1-p_{1j_1}'\cdots p_{vj_v}'p_{v+1j_{v+1}}'\cdots p_{mj_m}')^r\}\{1-(1-p_{1j_1}'\cdots p_{vj_v}'t_{v+1j_{v+1}}'^{(h)}\cdots t_{mj_m}'^{(h)})^{|\delta R_2|}\}\\&\approx|\delta R_2|-r\cdot|\delta R_2|\cdot\sum_{j_1=1}^{\lceil d_1'\rceil}\cdots\sum_{j_m=1}^{\lceil d_m'\rceil}(p_{1j_1}'^2\cdots p_{vj_v}'^2\cdot p_{v+1j_{v+1}}'t_{v+1j_{v+1}}'\cdots p_{mj_m}'t_{mj_m}') \qquad (4.7)\end{aligned}$$

となる．最後に，

$$|R_2|=|R_1|+|\Delta R_2|$$

とすればよい．ただし，

$$|R_1|=|\Delta R_1|=|R|=r.$$

以上で1回目の反復が終了する．ここに現れた多くの式は近似式であるので，それを強調するために，今後これらの値を用いて求めることになる R_k', δR_k, ΔR_k, R_k の大きさを s_k', δs_k, Δs_k, s_k で表す．

4.2.3 2回目以降の反復

2回目以降の反復では，真の増分 ΔR_k の各属性における値の分布は，最初に与えられた関係 R での分布の性質を残しつつ，次第に一様分布に近づく．そこで，2回目の以後（つまり $k\geq 3$）の関係 R_k', δR_k, ΔR_k, R_k における値の分布をすべて一様分布と近似して処理する．ただし，関係 ΔR_2 の属性の定義域の大きさとしては，前節の d_i' がさらに

$$d_i''=\sum_{j=1}^{\lceil d_i'\rceil}\{1-(1-p_{ij}')^{\Delta s_2}\} \quad (i=1,\cdots,v)$$

$$d_i''=\sum_{j=1}^{\lceil d_i'\rceil}\{1-(1-t_{ij}'^{(h)})^{\Delta s_2}\} \quad (i=v+1,\cdots,m)$$

と制限されたと考える．

2回目以降の反復の中で $R_k'=\Delta R_{k-1}\bowtie(Q_1,\cdots,Q_u)$ の計算では，ΔR_{k-1} において，各属性 A_i で，異なる値の種類の数 $\lceil d_i''\rceil$ の一様分布をしていると近似しているので，関係 Q_h のすべての属性 $A_i^{(h)}\in A^{(h)}$ が対応する ΔR_{k-1} の属性と結合する確率 f_h' は次式のようになる．

$$f_h'=\prod_{A_i^{(h)}\in A^{(h)}}\sum_{j=1}^{\lceil d_i''\rceil}\frac{1}{\lceil d_i''\rceil}\cdot t_{ij}^{(h)} \quad (h=1,\cdots,u).$$

結局，2回目以降の反復は次のようにまとめられる．

$$\begin{cases}s_k'=\Delta s_{k-1}\times f_1'q_1\times\cdots\times f_u'q_u\\ \delta s_k=d_1''\cdots d_m''\left\{1-\left(1-\dfrac{1}{d_1''\cdots d_m''}\right)^{s_k'}\right\}\\ \Delta s_k=\delta s_k-\dfrac{\delta s_k\times s_{k-1}}{d_1''\cdots d_m''}\\ s_k=s_{k-1}+\Delta s_k\end{cases} \qquad (4.8)$$

ここで，

$$d^*=d_1''\cdots d_m'',$$

$$F=f_1'q_1\times\cdots\times f_u'q_u$$

とおき，$d_1''\cdots d_m''\gg 1$ であることを考慮して，式(4.8)の1式を代入した2式を

$$\begin{aligned}\delta s_k&\approx d^*\left\{1-\left(1-\frac{1}{d^*}\right)^{\Delta s_{k-1}\cdot F}\right\}\\&\approx\Delta s_{k-1}\cdot F-\frac{1}{2d^*}(\Delta s_{k-1}\cdot F)(\Delta s_{k-1}\cdot F-1)\\&\approx\frac{F^2}{2d^*}\left(\frac{2d^*}{F}-\Delta s_{k-1}\right)\cdot\Delta s_{k-1}\end{aligned}$$

のようにテイラー展開の2次の項まで近似し，その後

δs_k を消去すると，漸化式

$$\begin{cases} \Delta s_k \approx \dfrac{F^2}{2d^{*2}}\left(\dfrac{2d^*}{F}-\Delta s_{k-1}\right) \\ \qquad \times (d^*-s_{k-1})\Delta s_{k-1} \\ s_k = s_{k-1}+\Delta s_k \end{cases} \tag{4.9}$$

を得る．この漸化式を解くことによって

$$s^+ \approx \sum_{k=1}^{\infty} \Delta s_k = \Delta s_1 + \tilde{s} \tag{4.10}$$

$$\tilde{s} = \sum_{k=2}^{\infty} \Delta s_k$$

を求めることができる．

4.2.4 漸化式の近似解法

漸化式 (4.9) をこのままの形で厳密に解くことは容易ではないので，近似的に解くことを考える．式(4.9)の第1式をさらに変形すると

$$\frac{2d^{*2}}{F^2}\cdot\Delta s_k \approx \frac{2d^{*2}}{F}\cdot\Delta s_{k-1} - \frac{2d^*}{F}\cdot s_{k-1}\cdot\Delta s_{k-1} - d^*\cdot(\Delta s_{k-1})^2 + s_{k-1}\cdot\Delta s_{k-1}\cdot\Delta s_{k-1}$$

となるので，

$$C_1=\frac{2d^{*2}}{F^2},\quad C_2=\frac{2d^{*2}}{F},\quad C_3=\frac{2d^*}{F},\quad C_4=d^*$$

とおき，$k=3,4,\cdots$ として辺々加えると，

$$C_1(\tilde{s}-\Delta s_2) \approx C_2\cdot\tilde{s} - \frac{1}{2}C_3\left(\tilde{s}^2+\sum_{k=2}^{\infty}(\Delta s_k)^2\right) - C_4\sum_{k=2}^{\infty}(\Delta s_k)^2 + \sum_{k=2}^{\infty}(s_k\cdot\Delta s_k\cdot\Delta s_k)$$

となるが，

$$\sum_{k=2}^{\infty}(s_k\cdot\Delta s_k\cdot\Delta s_k) \approx \frac{1}{2}\left(\tilde{s}\sum_{k=2}^{\infty}(\Delta s_k)^2 + \sum_{k=2}^{\infty}(\Delta s_k)^3\right)$$

と近似したのち，さらに

$$\sum_{k=2}^{\infty}(\Delta s_k)^2 \approx \tilde{s}\cdot\Delta s_3,$$

$$\sum_{k=2}^{\infty}(\Delta s_k)^3 \approx \tilde{s}\cdot(\Delta s_3)^2$$

と近似することによって

$$C_1(\tilde{s}-\Delta s_2) = C_2\cdot\tilde{s} - \frac{1}{2}\cdot C_3(\tilde{s}^2+\tilde{s}\Delta s_3) - C_4\cdot\tilde{s}\cdot\Delta s_3 + \frac{1}{2}(\tilde{s}^2\cdot\Delta s_3 + \tilde{s}\cdot(\Delta s_3)^2)$$

と書き換えられる．ただし，式(4.7)の Δs_2 を用いて

$$\Delta s_3 = \frac{F^2}{2d^{*2}}\cdot\left(\frac{2d^*}{F}-\Delta s_2\right)\cdot(d^*-\Delta s_2)\cdot\Delta s_2.$$

そこでこれらを整理して $\tilde{s}$ について解くと

$$\begin{cases} \tilde{s} = \dfrac{-C+\sqrt{C^2+8C_1(C_3-\Delta s_3)\Delta s_2}}{2(C_4-\Delta s_3)}, \\ C = 2C_1-2C_2 \\ \qquad +C_3\cdot\Delta s_3+2C_4\cdot\Delta s_3-(\Delta s_3)^2 \end{cases} \tag{4.11}$$

となり，したがって推移閉包の大きさの期待値は

$$s^+ \approx \Delta s_1 + \tilde{s} \tag{4.12}$$

によって近似的に求められる．

5. 推移閉包の大きさの近似評価例

本章では，これまでの結果の適用例として，関係の成分が Zipf 分布にしたがって生成された場合を考察し，数値計算の結果を与える．

5.1 Zipf 分布

ここでは一般の分布の具体例として次の Zipf 分布[3),11)]をとりあげる．

$$p_j = \frac{1}{jH_n},\quad j=1,\cdots,n,$$

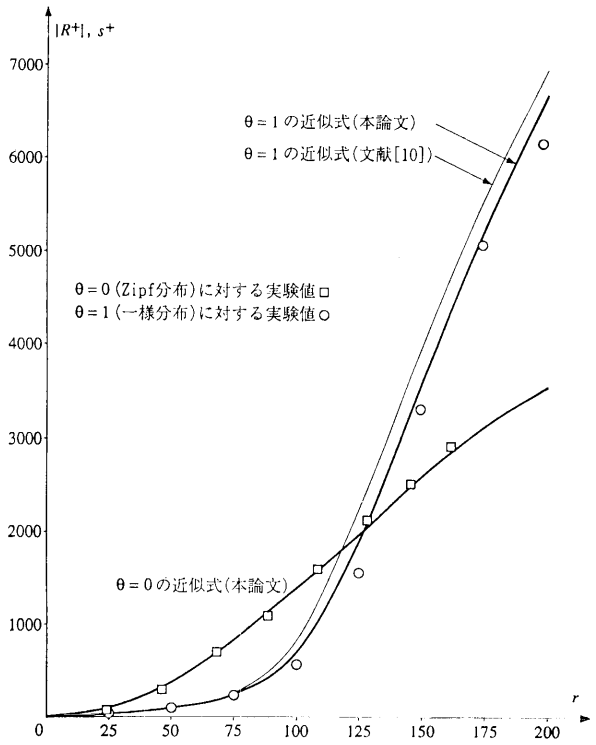

図 1 各 r に対する推移閉包の大きさ ($d=100$)
Fig. 1 Sizes of transitive closures as a function of r ($d=100$).

$$H_n=\sum_{j=1}^{n}\frac{1}{j},\qquad (5.1)$$

この分布は，文章中における単語や，全国民における名前の出現頻度のような具体例において観察されているという．Zipf 分布はしばしば次のように拡張される．

$$p_j=\frac{1}{j^{1-\theta}H_n^{(1-\theta)}},\quad j=1,\cdots,n,$$

$$H_n^{(1-\theta)}=\sum_{j=1}^{n}\frac{1}{j^{1-\theta}}.\qquad (5.2)$$

ここで，θ はかたよりの程度を表すパラメータであり，$\theta=0$ はもとの Zipf 分布を，$\theta=1$ は一様分布を表す．

5.2 Zipf 分布の場合の計算実験

実験では，簡単のため 2 項関係 R に対して同じ R を結合する推移閉包 R^+ を考える．すなわち，$B=\{A_1, A_2\}$, $R_1=R$, $R_2=\pi_B R\bowtie R$, $\cdots$ などである．関係 R の 2 つの属性 A_1 と A_2 は共通の定義域 D をもち，$R\bowtie R$ では前の R の属性 A_2 と後の R の属性 A_1 とを結合すると仮定する．また，R の 2 つの属性の値は，式(5.2)の Zipf 分布において，

(a) $\theta=0$ の純粋な Zipf 分布，

(b) $\theta=1$ の Zipf 分布，すなわち一様分布

の 2 通りを考えた．定義域の大きさ d と関係 R の大きさ r のいくつかの組合せのそれぞれについて

(1) 30 通りの関係をランダムに生成し，それらの推移閉包の大きさの平均値を求めた実験値 $|R^+|$,

(2) 近似式(4.11)と(4.12)による s^+ を求めたものの 2 種類の値を調べた．**図 1** に $d=100$ の場合を，**図 2** に $d=400$ の場合を示す．さらに $\theta=1$ の Zipf 分布（一様分布）に対しては，近似の精度を比較するために文献 10) で提案された近似による s^+ の結果もあわせて示した．

まず，$d=100$, 400 のどちらの場合も，$\theta=0$ と $\theta=1$ における推移閉包の大きさが大きく異なっている．このことは，推移閉包の大きさには定義域における値の生起確率の分布が大きく影響していることを示しており，本論文のように，確率分布を考慮した解析が必要であることがわかる．近似の精度については，かなり大胆な簡単化を行ったにもかかわらず，$\theta=0$, 1 のどちらの場合も，s^+ は比較的実験値と近い．さらに $\theta=1$ の一様分布の場合については，文献 10) の近似と比較して精度がかなり改善されている．その理由は，漸化式のテイラー展開の部分を線形近似から 2 次近似に改良した点にあると考えられる．

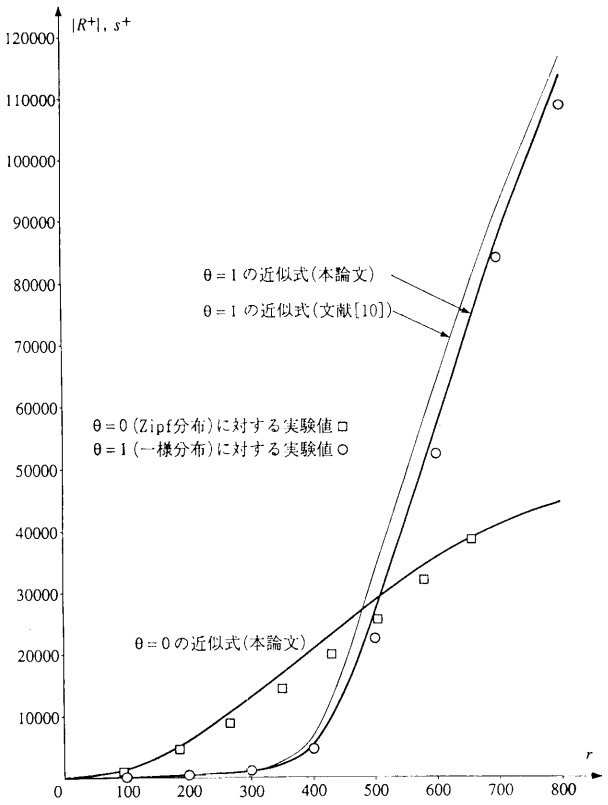

図 2 各 r に対する推移閉包の大きさ ($d=400$)
Fig. 2 Sizes of transitive closures as a function of r ($d=400$).

このように，$\theta=0$ と $\theta=1$ という両極端の Zipf 分布のどちらについても精度の高い近似値になっていることから，本論文の結果は，かなり広い範囲の一般分布に対して良好な近似を与えると予想される．

6. おわりに

本論文では，演繹データベースの処理などの応用例にも適用できるように推移閉包の定義を一般化したのち，各属性における値の確率分布と関係の大きさ r から，推移閉包の大きさを近似的に求める方法を提案した．このことは，関係データベースや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実用的に意義が大きいと考えられる．

なお，本論文の導出過程などから考えても，推移閉包の大きさをこのようなアプローチで推定するのは，精度の点で限界があると思われる．今後の研究課題として，近似値の正確さをさらに高めるための別のアプローチが必要であろう．

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東北大学経済学部の大西匡光助教授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, *Proceedings of the 13th VLDB Conference*, pp. 255-266 (1987).
2) Ioannidis, Y. E. and Ramakrishnan, R.: Efficient Transitive Closure Algorithms, *Proceedings of the 14th VLDB Conference*, pp. 382-394 (1988).
3) Knuth, D. E.: *The Art of Computer Programming, Vol. 3: Sorting and Searching*, Addison-Wesley (1973).
4) Lipton, R. J. and Naughton, J. F.: Estimating the Size of Generalized Transitive Closures, *Proceedings of the 15th VLDB Conference*, pp. 165-171 (1989).
5) Lipton, R. J. and Naughton, J. F.: Query Size Estimation by Adaptive Sampling, *Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on PODS*, pp. 40-46 (1990).
6) Lu. H.: New Strategies for Computing the Transitive Closures of a Database Relation, *Proceedings of the 13th VLDB Conference*, pp. 267-274 (1987).
7) Pittel, B.: On Distribution related to Transitive Closure of the Random Finite Mappings, *Annals of Probabilistics*, Vol. 11, No. 9, pp. 428-441 (1983).
8) Sippu, S. and Soisalon-Soninen, E.: A Generalized Transitive Closure for Relational Queries, *Proceedings of the 7th ACM SIGACT-SIGMOD-SIGART Symposium on PODS*, pp. 325-332 (1988).
9) Ullman, J. D.: *Principles of Database and Knowledge-Base Systems, Vol. 1*, Computer Science Press (1989).
10) 宇野，茨木：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，電子情報通信学会論文誌，Vol. J75-D-I, No. 9, pp. 855-863 (1992).
11) Wolf, J. L., Dias, D. M. and Yu, P. S.: An Effective Algorithm for Parallelizing Sort Merge Joins in the Presence of Data Skew, *Proceedings of the 2nd International Symposium on Database in Parallel and Distributed Systems*, pp. 103-115 (1990).

(平成5年7月19日受付)
(平成6年3月17日採録)

宇野 裕之（正会員）

1965年生．1987年京都大学工学部数理工学科卒業．1989年同大学院修士課程修了．1992年大阪府立大学総合科学部助手，現在に至る．関係データベースや演繹データベースの処理の最適化，オペレーションズ・リサーチ，組合せ最適化等の研究に従事．電子情報通信学会，人工知能学会，OR学会各会員．

茨木 俊秀（正会員）

1940年生．1963年京都大学工学部電気工学科卒業．1965年同大学院修士課程修了．1969年京都大学助手，1973年同助教授，1983年豊田技術科学大学教授，1985年京都大学工学部数理工学科教授，現在に至る．この間，Illinois大学，Waterloo大学，Simon Fraser大学，Rutgers大学等客員．工学博士．組合せ最適化，アルゴリズム，計算の複雑さ等の研究に従事．著書「アルゴリズムとデータ構造」（昭晃堂），「最適化プログラミング」（岩波書店），「最適化の手法」（共立出版），「Enumerative approaches to combinatorial optimization」（Baltzer），「Resource allocation problems: Algorithmic Approaches」（MIT Press），ほか．